UNIVERSITY OF CALIFORNIA
SANTA CRUZ

DARK MATTER, BLACK HOLES, AND NEW PHYSICS

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

PHYSICS

by

Benjamin V. Lehmann

September 2022

The Dissertation of Benjamin V. Lehmann
is approved:

Professor Stefano Profumo, Chair

Professor Wolfgang Altmannshofer

Professor Michael Dine

Peter Biehl
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Dark matter, black holes, and new physics

by

Benjamin V. Lehmann

The unknown nature of dark matter already represents one of the greatest gaps in our understanding of the universe. But the study of dark matter is now encountering a crisis: collider searches and direct detection experiments are quickly ruling out the strongly-motivated WIMP models that have guided theoretical progress for decades. Departing from the WIMP paradigm opens vast regions of parameter space across the scales, from ultralight bosons to objects at the Planck scale and beyond, and effectively probing this space of possibilities calls for new tools. Fortuitously, several such tools are available to us in the form of new quantum sensors, new astrophysical observables, and the new science of gravitational wave astronomy. In this thesis, I show how these methods can be combined to probe well-motivated dark matter candidates across an enormous range of masses.

This thesis is dedicated to the many curious and capable people who have been denied a place in the scientific community by no fault of their own, whether by financial pressures, systemic barriers, abusive mentors, or any other circumstances beyond their control.

Acknowledgments

I have been extremely fortunate to have had the support of many mentors, colleagues, friends, and family members over the course of this work.

I am especially grateful to my advisor, **Stefano Profumo**, for his steadfast coaching through the entire process. The best academic advice I have ever seen is to "find a mentor who believes in you, and believe them." Stefano has been that mentor to me almost from the moment we met. He is an outstanding model of life in and beyond physics. Being his student made graduate school a delightful and fruitful experience.

I also thank my other committee members, **Wolfgang Altmannshofer** and **Michael Dine**, each mentors to me in their own manner of approaching science. When I wanted to understand the broader variety of perspectives in particle physics, they each took me more seriously than I deserved, and helped me grow tremendously as a citizen of the field.

Each of my committee members is part of the SCIPP theory group, and I am grateful to the entire group for creating a welcoming and stimulating environment that hatched most of the work described in this thesis. I especially thank my direct collaborators in both published and unpublished work, including **Adam Coogan, Francesco D'Eramo, Will DeRocco, Jeff Dror, Stefania Gori, Logan Morrison, Hiren Patel,** and **Nolan Smyth**.

The theory group has also been a welcoming place for undergraduate researchers, and I have benefitted immeasurably from mentoring these junior students. I am grateful to former undergraduates **Anikeya Aditya, Paul Andreini, Sam En-**

glish, Mason Hargrave, Olivia Ross, Tom Schwemberger, Ava Webber, Jackson Yant, and **Jianhong Zuo**, each of whom trusted me to be part of their introduction to research. Each of them has shown me how amazingly capable they are, and I look forward to watching their careers bloom. In the language of my Jewish heritage, I shep nachas from each of them.

Beyond the theory group at SCIPP, I have had the privilege of working with and learning from outstanding scientists at UC Santa Cruz and around the world. I especially thank collaborators **Karl Berggren, Christian Boyd, Ilya Charaev, Max Gaspari, Christian Johnson, Eric David Kramer, Sae Woo Nam, Tao Xu,** and **To Chin Yu**. I also thank **Piero Madau** for valuable consultations.

I have reserved a few names from this list for special mention: in addition to my advisor and committee members, several wonderful people beyond the theory group have taken me under their wings as surrogate advisors, putting forth their time and effort to support my professional future. I am deeply thankful to my mentors **Yonit Hochberg, Yoni Kahn, Noah Kurinsky,** and **Steven Ritz**. My introduction to the field would not have been the same without them.

I am also grateful to many people at UC Santa Cruz who have enriched my experience in other ways. Among the departmental staff, I especially thank **Vicki Johnson, Ben Miller, Cathy Murphy,** and **Amy Radovan** for outstanding support and guidance. I have benefitted greatly from participating in many departmental initiatives, including the Diversity & Climate Committee and the Project for Inmate Education, and I thank all of the people who were involved in making those possible.

Through the entire process, I have been sustained by my relationships with

friends and family. I am grateful to my entire graduate cohort for creating a strong and healthy culture among the students, overflowing with mutual support and devoid of competition. In this regard, special thanks are due to **Daniel Davies, Dominic Pasquali,** and **John Tamanas** for much good and honest conversation about both physics and life. Beyond my department, I am grateful for the unwavering support of many friends and family members—parents, siblings, aunts, uncles, and cousins—who validated my belief that this path was right for me.

Sadly, on the subject of family, I must conclude these acknowledgments on a painful note. Throughout my time in the program, I have received invaluable support from **Natalie Telis**, my wife and partner of the last fourteen years. She has been a constant source of reassurance, courage, and meaning in the moments when all else has seemed fleeting. I wish I could write these thanks in the joyful tone that her support deserves. Unfortunately, at the time of this writing, our marital status is in grave doubt.

I considered omitting any mention of our personal difficulties from these acknowledgments, but chose to include it for two reasons. First, anybody interested enough in my personal journey to be reading these acknowledgments deserves an authentic account that normalizes the struggle with personal suffering. Second, it is likely that my own anxious overwork in creating this thesis has played a role in fraying our relationship. I say now for myself and as a warning to others: not one thing in this thesis, nor its overall completion, nor the professional future it has enabled, has been worth the loss of my beloved partner. This *can* happen to you: I thought our foundation was unbreakable, and I was wrong. It is bitter to know that however the future plays out, I will count my decisions in this time among the biggest regrets of my entire life.

Introduction

Expectations, crisis, and possibility

The identity of dark matter (DM) is perhaps the widest single gap in our understanding of the physical universe today.

Here I mean "widest" in a very specific sense: this is a problem that touches on an enormous breadth of systems, processes, timescales, and indeed subfields of physics. Virtually every corner of modern cosmology and particle physics has some relationship to DM. The existence of DM is perhaps the most concrete motivation for new degrees of freedom beyond the Standard Model of particle physics, a clear and present problem that is no matter of fine-tuning. A great variety of models that resolve problems in the Standard Model naturally provide a DM candidate. DM underlies the formation of galaxies, and literally creates our place in the Universe. Uncovering the identity of DM would surely give us new ingredients to understand the very early history of the cosmos—and the possibilities are endless.

However, despite this breadth, the community was for many years all but certain of the particle identity of DM: DM was sure to be a weakly interacting massive particle, or WIMP. There are two key hints for DM near the weak scale, at $\mathcal{O}(100\,\mathrm{GeV})$. First, the hierarchy problem of the Standard Model strongly suggests that there ought to be new degrees of freedom near the weak scale, a hypothesis that was powerful mo-

tivation for the construction of the Large Hadron Collider (LHC). Second, a weak-scale mass and annihilation cross section automatically give rise to DM production in the thermal bath of the early universe in roughly the observed amount. This "WIMP miracle" meant that models of beyond-Standard-Model particle physics with new degrees of freedom at the weak scale could very easily include a DM candidate.

No discussion of WIMPs would be complete without mention of supersymmetry (SUSY). For decades, SUSY was by far the favorite paradigm to resolve the SM hierarchy problem, and it did so only with the introduction of numerous new degrees of freedom—superpartners—including species with weak interactions. Phenomenologically viable SUSY models also included a simple mechanism to prevent DM from decaying. To prevent baryon- and lepton-number violation, many such models imposed R-parity, a $\mathbb{Z}_2$ symmetry acting on a field ψ with spin s_ψ as $\psi \to (-1)^{3(B-L)+2s_\psi}$, where B and L are baryon and lepton number. This phase is $+1$ for all Standard Model fields and -1 for all superpartner fields, so in a model with R-parity, the lightest supersymmetric particle (LSP) is absolutely stable. The LSP of any R-parity-conserving SUSY model is a potential DM candidate.

Beyond its phenomenological appeal at the weak scale, SUSY has received a great deal of attention for its formal properties. On the one hand, there is the matter of simple aesthetics: due to the Coleman–Mandula theorem, SUSY is the unique symmetry involving nontrivial combinations of spacetime symmetry transformations and internal symmetry transformations. Numerous conceptually-difficult computations are dramatically simplified in supersymmetric theories, and there is even a connection to quantum gravity: if string theory provides the correct description of nature, the exis-

tence of fermionic degrees of freedom *requires* the presence of (broken) SUSY. It seems that if nature were kind, or at least as clever as we are, then SUSY would surely be realized as the simplest explanation of numerous puzzles in particle physics.

These coincidences set great expectations for the discovery of SUSY at the weak scale. Thus, despite the breadth of the possibilities for the nature of DM, the lion's share of experimental effort has gone to searching for WIMPs. Such searches have mainly taken three forms: collider searches test the production of WIMPs from Standard Model species; direct detection searches probe the scattering of astrophysical WIMPs with Standard Model particles; and indirect detection searches look for signs of WIMP annihilation into Standard Model final states. However, despite substantial improvements to experimental sensitivity, none of these searches has found definitive evidence of WIMPs, and tightening constraints have started to put pressure on the entire WIMP paradigm [1–5].

It is hard to overstate the significance of this null result. While the experimental picture has only become clear in the last few years, it has already had an enormous if premature effect on the DM community, and on the particle physics community more broadly. If SUSY and the WIMP paradigm indeed turn out to be red herrings, the current slough of results will be remembered alongside the Michelson–Morley experiment. SUSY was once synonymous with the future of particle physics, and "LSP" was once used interchangeably with "dark matter". The notion that nature is uglier than we realized—or perhaps more clever than we are—has sent phenomenologists scrambling to understand the space of possibilities for DM. Cosmological DM may or may not be a WIMP, but identifying this elusive species will absolutely provide a stepping stone to

the next paradigm of particle physics.

That is the context of this thesis: the field is in a moment of crisis, with decades-long hopes recently dashed by inconvenient experimental truths. The task before us is to identify new possibilities and new routes to discovery.

Before we strategize to that end, let us take a step back to the most basic requirements for dark matter candidates. What do we already know about DM? Arguably, we know quite a bit from astronomy, even without characterizing DM microphysics. We know that:

1. DM is *dark*. Its interactions with Standard Model species are highly constrained by experimental data.

2. DM is *cold*. We know that DM collapses into the structures that host galaxies in the late Universe, and such structures would not form if DM was relativistic.

3. DM is *matter*. That is, when modeled as a perfect fluid on cosmological scales, DM has the equation of state of matter at all epochs that are presently accessible to observations.

4. DM is *abundant*. DM accounts for roughly one quarter of the energy density of the Universe today, and roughly 80% the matter density. Some mechanism must have produced it in that amount.

5. DM is *stable*. Cosmological observations from multiple epochs imply that at most a few percent of DM is allowed to decay over the lifetime of the universe.

6. DM is *ghostly*. By this I mean that DM particles largely pass by one another

without colliding, i.e., the self-interaction cross section is constrained.

7. DM is *structural*. The relationship between galaxies and DM halos indicates that DM collapses into structures at least as small as dwarf galaxies. Among other things, this implies that the mass of the DM species is at least $\mathcal{O}(\mathrm{keV})$ if it is fermionic, and at least $\mathcal{O}(10^{-19}\,\mathrm{eV})$ if it is bosonic.

8. DM is *fluid-like*. On galactic scales, DM is indistinguishable from a noninteracting fluid, so the particle mass must at least be much lower than the masses of DM halos.

9. DM is *discrete*. That is, DM behaves in every observable respect like a particle species, and is *not* readily compatible with a modification to gravity. This is still disputed by very smart, respectable people, but for the moment, let me say only that theirs is a minority viewpoint for good reasons.

SUSY WIMPs met all of the above requirements. But that is not necessarily so miraculous. Numerous other paradigms automatically accomplish most or all of these feats, in what would have been titled miracles if they were attached to a framework as compelling as SUSY. Indeed, there are viable models of DM populating the entire mass range we have laid out above: from $10^{-19}\,\mathrm{eV}$ at the low end (or $1\,\mathrm{keV}$ for fermions) up to many thousands of solar masses, it is possible to write down a viable model that satisfies all of the constraints above. It is worth pointing out that two familiar species, neutrinos and black holes, each meet most of the requirements above, and indeed *do* constitute some fraction of DM, and in each case, close relatives have been proposed as DM candidates.

How should we proceed given such a vast array of options? There is certainly good reason to deeply study single well-motivated paradigms, as has been done to date: the axion is perhaps the rising star of DM candidates, for good reason, and even the SUSY WIMP is not at the end of its journey. However, this moment demands a level of humility. Our most promising theoretically-motivated speculation has yielded little fruit so far. Despite this, discovering the identity of the DM species would surely provide a clear path forward to new physics. Thus, in parallel with the study of individual "lampposts", there is ample motivation to broaden the search for DM. This is the ideal time to construct new probes that test wide categories of DM models, agnostic to any specific Lagrangian.

A broad search across the enormous DM parameter space can only succeed by leveraging new tools, and that is the unifying principle of this thesis. Several new tools are available to the DM community by no merit of our own—parallel developments in other fields now provide us with new astronomical observables, new quantum sensing technology, and new gravitational wave observatories, to name three key opportunities. The open question is how to leverage these new tools to discover the identity of DM.

In this thesis, I describe a number of explorations in the DM parameter space taking advantage of new tools to extract generic insights. The work is presented in three parts, with somewhat artificial delineations. In Part I, I will begin by exploring the prospects for ultraheavy DM in the form of black holes, and I will describe new observables for this scenario. In Part II, I will leap from black holes to gravitational waves, and show how gravitational wave astronomy can probe not only ultraheavy DM, but also ultralight DM. I will promptly segue to the use of other cosmological observables

to probe "light" DM between 1 meV and 1 GeV. Critically, this discussion will expose complementarity between cosmology and direct detection experiments for light DM. In Part III, I will discuss new concepts and prospects for direct detection at low masses, taking advantage of new quantum sensor technology to probe key new parameter space. Taken together, these studies provide a new set of opportunities to explore DM models across the entire viable mass range.

I wish to be direct now about two things: first, the arc of this story is necessarily artificial. Given the context of DM phenomenology today, even the most organized research program must ultimately be based on opportunism. I am extremely gratified that the work I will discuss in this thesis has indeed probed such a wide range of model space—that was always my hope, but to call it intentional would understate the role of chance and surprise in the scientific process. This chapters of this story have been cherry-picked to fit the overall message, and the exclusion of a few papers from discussion here should not be taken as a slight against the importance of these other directions.

Second, after this introduction, I will switch from the first-person singular to the first-person plural. This is no accident. All of the work discussed here was performed together with an outstanding set of collaborators. While I am proud to have been a primary contributor to each of the works represented in this thesis, it would be disingenuous to claim that they are the fruits of my efforts alone. I am extremely grateful to all those who have been part of this work, whether as a supervisor, as a colleague, or as a more junior student. I hope that all three categories will take some pleasure in the story as rendered here.

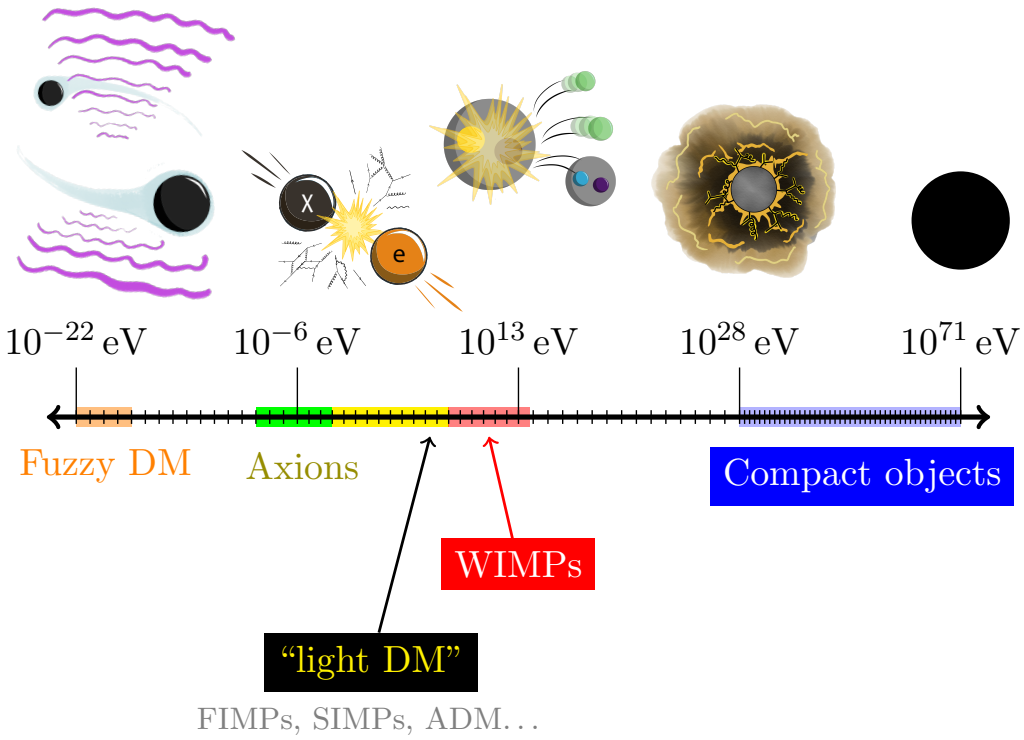

Figure 0.1: Viable mass range for dark matter. Several important classes of candidates are highlighted, but giving a full accounting of proposed candidates and their favored mass ranges seems environmentally unfriendly, as it would require a number of pages comparable to the length of this thesis.

Part I

Primordial black holes

and dark matter

Invitation

We begin at the upper end of the mass range in Fig. 0.1 by considering a dark matter particle that is hardly a particle at all. Black holes *are* DM: as components of a fluid, they are non-relativistic, redshift as matter, and interact primarily via gravity. The trickier question is whether such objects constitute more than a very small fraction of cosmological DM. This is indeed possible: every scale on which we observe DM is so large that sparse compact objects would be dynamically indistinguishable from a fluid even for black hole masses as large as $10\,\mathrm{M}_\odot$. Testing the notion of black holes as DM relies on a host of specialized observables. Our goal in this part, as in the rest of this thesis, will be to connect the the phenomenology of such black holes to the capabilities of upcoming and existing tools.

As with any DM model, the most exciting aspect of black holes as a dark matter candidate is what such a discovery would tell us about physics beyond the Standard Model. If anything more than a small fraction of DM is composed of black holes, the implications would be staggering: such black holes would have to arise from some mechanism besides stellar collapse. To understand what sort of mechanism would be viable, consider the problems with stellar collapse as a candidate:

1. Stars form from baryons, but the baryon density is smaller than the dark matter density by a factor of $\mathcal{O}(10^{-1})$, as measured before stars form.

2. Stars themselves account for only a $\mathcal{O}(10^{-2})$ fraction of the baryon density, so if every star ever born formed a black hole, these black holes would account for at most a $\mathcal{O}(10^{-3})$ fraction of dark matter. Moreover, in reality, only a $\mathcal{O}(10^{-3})$ fraction of stars produce black holes, so stellar evolution is not an efficient mechanism for black hole production.

3. Stars form fairly late in the scheme of cosmic history. Star formation peaked at redshifts $2 \lesssim z \lesssim 3$, whereas the dark matter density is already well characterized by cosmological probes at the epoch of recombination ($z \approx 1100$).

From these three challenges, we learn that any mechanism capable of producing enough black holes to account for dark matter must begin with a sufficiently large reservoir of energy density, must efficiently convert that energy density into black holes, and must produce black holes at extremely early times. Any such mechanism relies on ingredients that lie beyond the Standard Models of particle physics and cosmology.

Perhaps the most exciting mechanism is also the simplest: the formation of *primordial black holes* (PBHs) by the collapse of density perturbations in the early universe. If the power spectrum of density perturbations after inflation has a large amplitude at very small scales, which are not well constrained observationally, such perturbations would collapse directly into black holes with no need for the intermediate steps of stellar evolution. This was first envisioned over fifty years ago by Zel'dovich & Novikov [6]. Since then, many inflation models have been proposed that provide

the requisite shape to the power spectrum, predicting a population of primordial black holes. Here the word *primordial* is surely appropriate, referring to black holes produced directly by the physics of the inflationary universe. However, as numerous formation mechanisms have been proposed [6–9], the term has come to apply to black holes formed by any non-stellar mechanism at early times.

Can PBHs constitute the entirety of DM? This was a widely acknowledged possibility for decades following the initial proposal, with a healthy competition between WIMPs and MACHOs (MAssive Compact Halo Objects). Interest in MACHOs declined with the advent of powerful limits from microlensing (e.g. Ref. [10]). But the parameter space of PBH dark matter is quite broad: in principle, such black holes could be extremely massive, or microscopically small, or indeed any combination of masses across this spectrum. This parameter space has never been fully closed by even the most aggressive interpretation of observational data. Interest in PBHs surged after the detection of the first gravitational wave signatures from black hole binaries: Refs. [11, 12] immediately pointed out that these observations are consistent with PBH mergers, given the status of independent constraints on the PBH population.

The current enthusiasm for PBHs is doubtless driven in part by waning optimism for WIMPs, and in part by the availability of new tools to study black holes. In light of these developments, the level of tuning in typical PBH DM models seems less garish. Still, PBHs are undeniably an attractive DM candidate: they automatically satisfy all of the requirements for DM, with their properties determined by gravity, and their formation requires no new fields apart from the inflaton. In a sense, PBHs are the result of a macroscopic analogue of gravitational particle production.

Indeed, the distinction between black holes and elementary particles is somewhat arbitrary. It is tempting to think of black holes as composite objects, with an internal structure composed of more fundamental constituents. But the lesson of no-hair theorems is precisely the opposite: to our perspective as external observers, black holes have essentially no internal structure. With few exceptions, stable black hole solutions are fully characterized by mass, spin, and U(1) charges. What, then, is the difference between a black hole and any other particle? The key feature is the presence of a horizon, which is natural for objects with mass above the Planck scale. It is amusing to compare a particle's Schwarzschild radius, $2GM/c^2$, with its reduced Compton wavelength, $\hbar/(Mc)$. The former exceeds the latter—i.e., the particle can be localized on a length scale smaller than its Schwarzschild radius—only when $M > M_{\rm Pl}/\sqrt{2}$. This heuristic suggests that any elementary particle with a mass above the Planck scale ought to exhibit a horizon, and thus behave exactly like a black hole, further blurring the line between black holes and particles. For our purposes, the key implication is that PBHs are the most natural extension of "particle" DM to macroscopic scales.

We will shortly quantify the actual status of bounds on the total abundance of PBHs. For now, it is sufficient to point to three mass ranges in which bounds are absent or less robust:

1. 10^2–$10^5\,{\rm M}_\odot$: constraints in this range are based on dynamical disruption of wide binary systems and CMB distortion due to energy injection from black hole accretion. The dynamical constraints depend on the modeling of particular systems, and their validity has been disputed. The accretion constraints are based on spherically-symmetric, quasi-static accretion flows that are known to be unrealis-

tic at larger masses. This situation was part of the motivation for the conclusions of Refs. [11, 12]. Refinement of these accretion bounds is one goal of my ongoing work.

2. 10^{-16}–$10^{-12}\,M_\odot$: previously-claimed bounds from microlensing in this regime have been completely erased by two effects. First, the size of the black hole horizon becomes comparable to the wavelength of the light being observed [13, 14]. Second, the angular sizes of the light sources being lensed are typically larger than originally assumed [15]. All of dark matter can be composed of PBHs in this mass range alone.

3. $M_{\rm Pl}$ ($10^{-38}\,M_\odot$) and below: constraints on very low-mass PBHs are based on the physics of black hole evaporation. However, at the Planck scale, quantum gravity effects cannot be neglected, and semiclassical predictions for the evaporation rate become unreliable. If evaporation slows down or stops in this regime, as some models suggest, then remnants of evaporating black holes could be stable enough to constitute cosmological dark matter. We will explore one potential observable of this scenario shortly.

There is thus plenty of room for PBHs as a DM candidate. However, it is important to note that the discovery of a population of PBHs would carry enormous implications even if these objects account for a negligible fraction of DM. PBHs have long been studied as a potential signature of BSM particle physics [16–18]. They have been invoked as candidates for the origin of the baryon-antibaryon asymmetry [19–24], the production of particle DM [19, 21, 24, 25], the source of high-energy photon and

cosmic-ray positron emission [26, 27], and the constituent *per se* of cosmological DM [28–33].

More poetically, PBHs are messengers from the early universe. The conditions of the early universe before the epoch of big bang nucleosynthesis (BBN) are effectively screened from us today by thermal equilibrium. The attractor nature of thermal equilibrium prevents us from accessing information about the state of the Universe at higher temperatures. But a population of PBHs is never in equilibrium. As stable, non-annihilating "particles" with masses well above the temperature of the surrounding thermal bath, the properties of the PBH population would retain sensitivity to the conditions of the universe at the epoch of their formation, typically far earlier than BBN. If observed today, PBHs would serve as a probe of the high-scale physics which sets the conditions of their formation [16].

Given the remarkable potential of PBHs as a stepping stone to new physics, and given the new tools available to study them, this is a crucial moment in which to understand the prospects for and routes to their discovery. In the following chapters, we explore the possibilities in three steps. First, in Chapter 1, we develop the mathematical technology to determine the overall bounds on the PBH abundance, and discuss the implications for the status of PBHs as a DM candidate. Next, in Chapter 2, we explore the capture of PBHs in stellar systems and associated observables. Finally, in Chapter 3, we return to the possibility of microscopic PBH remnants, and demonstrate an opportunity to detect these objects in terrestrial laboratories—and many of the conclusions from this discussion will apply equally well to ultraheavy particle DM.

One obvious question will not be treated in these three chapters: can we

identify a population of PBHs using gravitational wave observatories such as LIGO? We will defer this question to Part II, where we will discuss other opportunities associated with cosmology and gravitational waves.

Chapter 1

Bounding the abundance of primordial black hole dark matter

1.1 Introduction

First, we consider the question of how to bound the total abundance of PBHs.

Depending on the formation mechanism, PBHs may exist today with masses as small as $10^{-16} M_{\odot}$, or as large as those of supermassive black holes. Thus, constraining the total density contained in PBHs requires the combination of constraints that span this vast range of mass scales. Such observables include microlensing surveys [10, 13, 34, 35], CMB data [36], and the statistics of wide binaries [37]. In general, constraints from these observables have been computed under the assumption that all PBHs have the same mass. The corresponding mass functions, comprising a single Dirac delta, are said to be *monochromatic*. However, as realistic production mechanisms necessarily result in an extended (non-monochromatic) mass function, it is essential to correctly combine

constraints across all masses.

This problem has recently been studied by several authors [38–40]. In general, the constraints depend non-trivially on the functional form of the mass function, and statements about the implications of constraints for properties of the PBH population can be difficult to generalize. In particular, the total fraction $f_{\rm PBH}$ of dark matter that may be accounted for by PBHs varies with the form of the mass function, so $f_{\rm PBH} = 1$ is ruled out for some forms of the mass function, and allowed for others. This has led to confusion regarding the observational viability of the PBH dark matter scenario, and while prior work has established procedures for comparing specific extended mass functions with observables, general statements regarding the allowed total fraction of dark matter in PBHs are lacking.

Depending on the set of constraints considered, observational data may or may not already rule out $f_{\rm PBH} = 1$ for monochromatic mass functions. Since the many constraints span a wide mass range, and since several do not overlap significantly, some authors have argued that broadening the mass function might relax constraints on PBHs [41, 42], possibly allowing for $f_{\rm PBH} = 1$ even if that possibility were excluded by constraints for monochromatic mass functions. However, [39, 40] have evaluated the constraints *numerically* for several forms of extended mass functions, and found that extended mass functions are typically subject to stronger constraints than monochromatic mass functions.

These findings motivate the question we now pose: what is the theoretical maximum density of PBHs permitted by constraints for a fully general mass function? Our goal is ultimately to clarify the observational status of PBH dark matter, and

to understand the circumstances under which extending the mass function can relax constraints. We also seek a procedure which is flexible and simple enough to allow us to compare results for different sets of constraints, and to elucidate the dependence of the maximal density on the form of the constraints themselves. To that end, we derive the form of the mass function which optimizes the density subject to all observational constraints combined. This allows us to obtain a general bound on the density of PBHs with minimal numerical computation, independently of the true form of the PBH mass function. Note that we do not propose a new prescription for the evaluation of constraints for a given extended mass function. Rather, we maximize the PBH density subject to constraints as evaluated using existing methods from the literature. The maximal-density mass functions we derive then provide insights into the overall impact of each individual observable.

This chapter is organized as follows. In Section 1.2, we establish conventions and notations, and review the application of constraints from the monochromatic case to extended mass functions. In Section 1.3, we present a pedagogical derivation of our main results regarding the maximum density of PBHs, and we apply them to current data. We consider the impact of gravitational wave constraints separately in Section 1.4. We discuss these results in Section 1.5 and conclude in Section 1.6. Finally, in Appendix A, we validate our analytical results with direct numerical techniques.

1.2 Interpreting constraints for extended mass functions

1.2.1 The interpretation problem

Applying observational constraints to generic extended mass functions is non-trivial. It is not sufficient to check that the mass function does not intersect constraint curves, as experiments are typically sensitive to the *integral* of the mass function in each of a set of mass bins. Thus, most constraints are only trivial to interpret for monochromatic mass functions, i.e., mass functions of the form

$$\psi_{\mathrm{mono}}\left(M_0, f_0; M\right) \equiv f_0\, \delta\left(M - M_0\right) \tag{1.1}$$

whose integrals are non-zero in only one bin. In this case, an observational constraint curve $f_{\mathrm{max}}(M)$ imposes the requirement that $f_0 < f_{\mathrm{max}}(M_0)$. Transforming such constraints to the parameter space of a more general extended mass function involves summing contributions to observables from all mass bins. Multiple prescriptions for this procedure have been used in the literature.

The earliest systematic treatment of constraints for extended mass functions is due to [29]. They divide the mass range into N bins $I_1, \ldots, I_N$, approximating the constraint functions as step functions on these bins. Within each bin, only the strongest constraint function $f_{\mathrm{max}}(M)$ is considered. A mass function ψ is excluded if

$$\int_{I_k} \mathrm{d}M\, \psi(M) > \max_{M \in I_k} f_{\mathrm{max}}(M) \tag{1.2}$$

for any k. This prescription is used by [38] to numerically transform observational constraints to the parameter space of a lognormal mass function. Their findings suggest that broadening the mass function does not generally relax constraints. However, as [38] treat the problem computationally, it is difficult to determine the relevance to their

results of any particular constraint, or of the lognormal form of the trial mass functions. This provided partial motivation for the analysis of [39], who obtain similar numerical results for several additional constraints and forms of the mass function. Further, [39] derive a more rigorous prescription for transforming observational constraints to general extended mass functions. We review their derivation in Section 1.2.2.

Similar questions motivate the recent analysis of [40]. Rather than develop a prescription for translating constraints for monochromatic mass functions to suit a given extended mass function, the authors develop a prescription for converting the extended mass function into a set of monochromatic mass functions, each accounting for the contribution of the PBH population to one observable. The extended mass function is then subject to each constraint as it applies to the corresponding monochromatic mass function. This approach is used to constrain the parameter spaces of lognormal and power law mass functions, with results similar to those of [38] and [39].

Our methods bear some similarities to [40], in that we also find it sufficient to work with sets of monochromatic mass functions. However, the monochromatic mass functions we consider have a different interpretation, as discussed in Section 1.3.3. Our goal is not to place constraints on any specific extended mass function, but rather to place bounds on PBH dark matter while allowing complete freedom in the mass function. Thus, our formalism is structured around the maximization problem, and we use our results to study both the current status of the PBH dark matter paradigm and the potential impact of future observables.

1.2.2 Constraint prescription

In this chapter, we seek a general result for the maximum allowed fraction of dark matter in PBHs, independent of the form of the mass function, and in a form that elucidates the relevance of each observable. As such, it is necessary that we adopt a prescription for constraining a given mass function that allows for multiple simultaneous constraining observables, a requirement most naturally satisfied by that of [39]. Their prescription is thus the basis for our analytical work. We numerically confirm that similar results are obtained under the prescriptions of [40] and [29] (see Appendix A.2).

We follow [39] to convert constraints for monochromatic mass functions to constraints for extended mass functions. We denote the mass function by ψ and adopt their normalization and conventions, such that

$$\psi \propto M\frac{\mathrm{d}n}{\mathrm{d}M}, \qquad \int \mathrm{d}M\,\psi(M) = \frac{\Omega_{\mathrm{PBH}}}{\Omega_{\mathrm{DM}}} \equiv f_{\mathrm{PBH}} \tag{1.3}$$

where n is the number density of PBHs at fixed mass. Most observables that can constrain primordial black holes are determined by the properties of single black holes, with no need to consider relationships between them. In such a case, an observable quantity A receives a linear combination of contributions from each mass bin, and the contribution from black holes of mass M is proportional to $\psi(M)$. As such, the observable can be written as a functional of ψ in the form

$$A[\psi] = A_0 + \int \mathrm{d}M\,\psi(M)K_1(M). \tag{1.4}$$

We note in passing that there are some observables for which relationships between black holes are significant. For example, gravitational wave observations of

mergers are dependent on the properties of pairs of black holes, and so one must combine contributions from pairs of mass bins. In the simplest case, where the contributions scale linearly with number in each mass bin, such an observable can clearly be written in the form

$$A[\psi] = A_0 + \int \mathrm{d}M\, \psi(M) K_1(M) + \int \mathrm{d}M\, \mathrm{d}M'\, \psi(M)\psi(M') K_2(M, M') \tag{1.5}$$

and one can always express a generic observable by including higher-order terms of this form. Note that higher-order terms also account for non-linear dependence of A on ψ at fixed mass. For example, an observable which scales as $\psi(M)^2$ can be expressed exactly at second order by setting $K_2(M, M') \propto \delta(M - M')$.

We study the potential impact of gravitational wave observations in Section 1.4. All of the other constraints that we consider in this chapter are of the simplest kind, and we will find Eq. (1.4) sufficient. In this case, it is straightforward to relate constraints for a monochromatic mass function to constraints for a generic mass function, and we briefly review the argument given in [39]. Let $\psi_{\mathrm{mono}}(M_0; M) \equiv f_{\mathrm{max}}(M_0)\, \delta(M - M_0)$, where $f_{\mathrm{max}}(M_0)$ is the largest coefficient allowed by constraints for a mass function of this form. If we take $\psi(M) = \psi_{\mathrm{mono}}(M_0; M)$ in Eq. (1.5), we obtain

$$K_1(M_0) = \frac{A[\psi_{\mathrm{mono}}] - A_0}{f_{\mathrm{max}}(M_0)} \tag{1.6}$$

Suppose that the difference $A[\psi] - A_0$ is observable with the desired significance when $A[\psi]$ crosses a threshold value A_{obs}. Then $A[\psi_{\mathrm{mono}}] = A_{\mathrm{obs}}$ by definition of f_{max}, so Eq. (1.6) gives $K_1(M)$ independent of ψ. Substituting for $K_1(M)$ in Eq. (1.5) while leaving ψ generic gives the condition

$$\mathcal{C}[\psi] \equiv \int \mathrm{d}M \frac{\psi(M)}{f_{\mathrm{max}}(M)} \leq 1. \tag{1.7}$$

This expresses the constraint on a mass function $\psi(M)$ when the constraint for a monochromatic mass function is $\int \mathrm{d}M\, \psi_{\mathrm{mono}}(M_0; M) \leq f_{\mathrm{max}}(M_0)$.

1.3 The optimal mass function

1.3.1 Single-constraint case

For pedagogical purposes, we first consider the case of a single constraining observable. For such situations, when all observables can be expressed in the form of Eq. (1.4), the constraint on the mass function has the form $\mathcal{C}[\psi] \leq 1$, with $\mathcal{C}[\psi]$ as defined in Eq. (1.7). The problem is then to maximize $\int \mathrm{d}M\, \psi(M)$ subject to this constraint. The optimal mass function saturates the constraint, so it suffices to require $\mathcal{C}[\psi] = 1$.

Naively, this problem looks as though it can be solved using the method of Lagrange multipliers, by finding stationary points of the functional

$$\mathcal{S}[\psi, \lambda] = \int \mathrm{d}M\, \left(\psi(M) - \lambda \frac{\psi(M)}{f_{\mathrm{max}}(M)} \right). \tag{1.8}$$

However, the Euler-Lagrange equation in ψ admits no non-trivial solutions. This is because $\int \mathrm{d}M\, \psi(M)$ can be made arbitrarily large, even subject to $\mathcal{C}[\psi] = 1$, unless $\psi(M) > 0$ is imposed. Positivity can be imposed by setting $\psi = \phi^* \phi$ and performing an unconstrained optimization in ϕ, but the corresponding Euler-Lagrange equation leads to the condition that ϕ is, at every point, either zero or non-analytic.

The variational approach does not generalize to the case of multiple constraints, so we do not pursue it any further. Rather, we observe that since $\mathcal{C}[\psi]$ is linear, we have $\mathcal{C}\big[\mathcal{C}[\psi]^{-1}\psi\big] = 1$. Thus, we can impose $\mathcal{C}[\psi] = 1$ by rescaling $\psi \to \mathcal{C}[\psi]^{-1}\psi$, and then

the problem is to maximize the functional

$$\mathcal{M}[\psi] \equiv \int \mathrm{d}M \left(\mathcal{C}[\psi]^{-1}\psi(M)\right) = \frac{\int \mathrm{d}M\, \psi(M)}{\int \mathrm{d}M\, \frac{\psi(M)}{f_{\mathrm{max}}(M)}} \tag{1.9}$$

subject only to positivity. We call $\mathcal{M}[\psi]$ the *normalized mass* of ψ.

It is now simple to show that $\mathcal{M}[\psi]$ is maximized by taking ψ to be a monochromatic mass function. Let $M_{\mathrm{max}} \equiv \mathrm{argmax}\, f_{\mathrm{max}}(M)$ and $f_{\mathrm{mono}} \equiv f_{\mathrm{max}}(M_{\mathrm{max}})$, and define

$$\psi_0(M) \equiv f_{\mathrm{mono}}\ \delta(M - M_{\mathrm{max}}) \tag{1.10}$$

so that $\psi_0(M)$ is the monochromatic mass function which maximizes the PBH density, and f_{mono} is the maximum PBH density allowed for a monochromatic mass function. Choose any mass function $\psi \equiv \psi_0 + \delta\psi$. Since ψ_0 vanishes everywhere except for M_{max}, positivity of ψ requires that $\delta\psi(M) \geq 0$ for all $M \neq M_{\mathrm{max}}$. Then we have

$$\mathcal{M}[\psi] = \frac{\int \mathrm{d}M \big[\psi_0(M) + \delta\psi(M)\big]}{\int \mathrm{d}M \big[\psi_0(M)/f_{\mathrm{max}}(M) + \delta\psi(M)/f_{\mathrm{max}}(M)\big]}. \tag{1.11}$$

Since ψ_0 saturates the constraint of Eq. (1.7), we must have $\int \mathrm{d}M\, [\psi_0(M)/f_{\mathrm{max}}(M)] = 1$ and $\int \mathrm{d}M\, \psi_0(M) = f_{\mathrm{mono}}$, so we write

$$\mathcal{M}[\psi] = \frac{f_{\mathrm{mono}} + \int \mathrm{d}M\, \delta\psi(M)}{1 + \int \mathrm{d}M\, \delta\psi(M)/f_{\mathrm{max}}(M)} \tag{1.12}$$

but $f_{\mathrm{max}}(M) \leq f_{\mathrm{mono}}$ by definition, so we have

$$\mathcal{M}[\psi] = \frac{f_{\mathrm{mono}} + \int \mathrm{d}M\, \delta\psi(M)}{1 + \int \mathrm{d}M\, \delta\psi(M)/f_{\mathrm{max}}(M)} \leq \frac{f_{\mathrm{mono}} + \int \mathrm{d}M\, \delta\psi(M)}{1 + \int \mathrm{d}M\, \delta\psi(M)/f_{\mathrm{mono}}} = f_{\mathrm{mono}}. \tag{1.13}$$

Thus we have shown that $\mathcal{M}[\psi] \leq f_{\mathrm{mono}} \equiv \mathcal{M}[\psi_0]$, so no functional form allows a higher total PBH density than does the Dirac delta. In particular, for fixed PBH density, we conclude that an extended mass function is always more strongly constrained than the

optimal monochromatic mass function. While this will not hold for the case of multiple constraints, it remains an excellent approximation if the constraints are weakest by far in a mass range where a single observable dominates.

1.3.2 Combining constraints

Realistically, the single-constraint case is too simplistic. In general, a mass function is ruled out on the basis of a χ^2 test statistic. If PBHs are constrained by multiple observables A_j, then the test statistic is found by adding the individual χ^2 statistics in quadrature. That is,

$$\chi^2[\psi] = \sum_{j=1}^{N} \chi_j^2 = \sum_{j=1}^{N} \left(\frac{A_j[\psi] - A_{\mathrm{obs},j}}{\sigma_j} \right)^2 . \tag{1.14}$$

To fail to reject ψ at some significance level requires that $\chi^2[\psi] \leq \gamma^2$ for some threshold value γ^2, i.e.,

$$\sum_{j=1}^{N} \left(\int \mathrm{d}M \, \psi(M) \frac{K_{1,j}(M)}{\gamma \sigma_j} \right)^2 \leq 1. \tag{1.15}$$

If we set $N = 1$, this reduces to

$$\int \mathrm{d}M \, \psi(M) \frac{K_{1,1}(M)}{\gamma \sigma_1} \leq 1 \tag{1.16}$$

so matching with Eq. (1.7) gives $K_{1,j}(M)/(\gamma\sigma_j) = 1/f_{\mathrm{max},j}(M)$, where $f_{\mathrm{max},j}(M)$ is the analogue of $f_{\mathrm{max}}(M)$ for the jth constraint alone. For general N, [39] show that the constraint takes the form

$$\sum_{j=1}^{N} \left(\int \mathrm{d}M \, \frac{\psi(M)}{f_{\mathrm{max},j}(M)} \right)^2 \leq 1. \tag{1.17}$$

Since the individual constraints are added in quadrature, the argument applied to the single-constraint case does not extend to the case of multiple constraints, and indeed,

there are cases in which the density is not maximized by a monochromatic mass function. However, we will show that the maximizer is in general a linear combination of N monochromatic mass functions.

1.3.3 The general problem

For the case of several constraining observables, one has N constraint functions denoted by $f_{\max,1}, \ldots, f_{\max,N}$. For brevity, we define $g_j(M) \equiv 1/f_{\max,j}(M)$, and by analogy with Eq. (1.7), we define

$$\mathcal{C}_j[\psi] \equiv \int \mathrm{d}M\, \psi(M)\, g_j(M). \tag{1.18}$$

Then the problem is to find ψ to maximize

$$\mathcal{M}[\psi] \equiv \frac{\int \mathrm{d}M\, \psi(M)}{\left(\sum_{j=1}^{N} \mathcal{C}_j[\psi]^2\right)^{1/2}} = \frac{\int \mathrm{d}M\, \psi(M)}{\|\boldsymbol{\mathcal{C}}[\psi]\|} \tag{1.19}$$

where $\boldsymbol{\mathcal{C}}[\psi]$ denotes the vector with components $\mathcal{C}_j[\psi]$. We define $f_{\max,\mathrm{all}} = \max \mathcal{M}[\psi]$.

Since rescaling ψ does not change $\mathcal{M}[\psi]$, we can always set $\int \mathrm{d}M\, \psi(M) = 1$, and then the problem is equivalent to minimizing $\|\boldsymbol{\mathcal{C}}[\psi]\|$ subject to this constraint. For convenience, we cast the integral in discrete form, writing

$$\|\boldsymbol{\mathcal{C}}[\psi_Q]\|^2 = \sum_{j=1}^{N} \left(\sum_{k=1}^{Q} a_k g_j(M_k) \right)^2 = \left\| \sum_{k=1}^{Q} a_k \mathbf{g}(M_k) \right\|^2 \tag{1.20}$$

where Q is not restricted to be finite. Thus, the problem is to minimize the norm of a sum of $a_k \mathbf{g}(M_k)$ for some $\{M_k\}_{k=1,\ldots,Q}$, subject to our normalization condition, which now takes the form $\sum_{k=1}^{Q} a_k = 1$. Geometrically, this is the same as minimizing the norm over the convex hull of the $\mathbf{g}(M)$, i.e., to compute

$$\min \{\|\mathbf{x}\| \mid \mathbf{x} \in \mathrm{conv}\, \{\mathbf{g}(M) \mid M \in U\}\} \tag{1.21}$$

where U is the mass range under consideration. We henceforth denote conv $\{\mathbf{g}(M) \mid M \in U\}$ by conv($\mathbf{g}$). Since the minimizer is the projection of the origin onto a convex set, it is unique in the sense that any optimal mass function ψ must have the same $\boldsymbol{\mathcal{C}}[\psi]$. This does not require that the minimizing mass function is itself unique.

Such a geometric formulation simplifies the interpretation of the problem. In particular, the result for the case of a single constraint is now immediate: the convex hull is 1-dimensional, so the point with minimum norm is simply the minimum value of $g(M)$. The corresponding mass function is monochromatic, with a peak at argmin $g(M)$. It is also clear that the monochromatic mass function is not generally the minimizer of the norm in the case of multiple constraints: we have no guarantee that $\|\mathbf{g}(M)\|$ attains the minimum of the norm on conv($\mathbf{g}$) for any single M.

Still, minimizing the norm over the convex hull of a discretization of $\mathbf{g}(M)$ is a simple computational problem, and it is easy to validate the result. We find an optimal mass function in three steps:

1. Choose a discretization of $\mathbf{g}(M)$ of the form $G = \{\mathbf{g}(M_1), \ldots, \mathbf{g}(M_R)\}$. We choose the M_k using adaptive sampling to capture features of the constraint functions as precisely as possible. The convex hull of G is now a polytope A.

2. Find the point $p_{\min} \in A$ with minimum norm. We implement the algorithm of [43], which requires only the extreme points of A as inputs. To avoid computing the convex hull in a high-dimensional space, we supply all of the points of G, of which the extreme points of A form a subset. The algorithm determines the facet S of A which contains $p_{\min}$, and gives the barycentric coordinates of $p_{\min}$ in S as a vector

$\mathbf{w}$.

3. Define a mass function

$$\psi_{\text{opt}}(M) = \sum_{k=1}^{|\mathbf{w}|} w_k \delta\left(M - M_k\right) \tag{1.22}$$

where $\mathbf{g}(M_k)$ is the kth point of S. Note that $\mathbf{g}(M_k) \in G$ for each M_k since $S \subset A$.

Observe that $\boldsymbol{\mathcal{C}}[\psi_{\text{opt}}] = \sum_{k=1}^{|\mathbf{w}|} w_k \mathbf{g}\left(M_k\right) \equiv p_{\min}$. Thus, ψ_{opt} is a mass function which attains the maximum total dark matter fraction. In particular, for any mass function ψ, we have $\mathcal{M}[\psi] \leq \mathcal{M}[\psi_{\text{opt}}] = \|p_{\min}\|^{-1}$, so $f_{\text{max,all}} = \|p_{\min}\|^{-1}$ is an upper bound on the fraction of dark matter in PBHs irrespective of the functional form of the mass function. We will refer to ψ_{opt} as the *semi-analytical optimum* mass function.

We can now explain geometrically why the maximizing mass function is a linear combination of no more than N monochromatic mass functions. Observe that for any $\mathbf{g}(M)$, the minimum of the norm must lie on the boundary of the convex hull $\text{conv}(\mathbf{g})$, and since $\mathbf{g}(M_k) \in \mathbb{R}^N$, this boundary has dimension at most $N-1$. One can construct an arbitrarily refined triangulation of this boundary formed from $(N-1)$-simplices, each with N points of G as vertices. The minimizer of the norm is a linear combination of these vertices, each of which is one of the original $\mathbf{g}(M_k)$, corresponding to a monochromatic mass function. We emphasize that at no step do we impose that the optimal mass function is a discrete linear combination of a finite number of monochromatic mass functions. This is a consequence of the fact that the optimum corresponds to a point in an $(N-1)$-simplex, meaning that this mass function lies in a space which is spanned by at most N monochromatic mass functions.

Our method is deceptively similar to the procedure of [40], in that we also

work with sets of monochromatic mass functions. However, the monochromatic mass functions considered in that work are used only to study the consequences of a given extended mass function for a given observable. The sum of the effective monochromatic mass functions corresponding to each observable does not generally give a single mass function with equivalent consequences for all observables combined. This is appropriate for the purposes of [40] because they investigate which constraints are most effective for mass functions of a fixed functional form.

In this chapter, the mass functions we derive maximize the density of PBHs with respect to all constraints simultaneously. Since the constraints are statistical in nature, the combination of multiple independent constraints at a single mass is stronger than any one of them individually. We follow [39] in treating constraints simultaneously, and our resulting semi-analytical optima are indeed sums of monochromatic mass functions. This approach is necessary for our purposes because we investigate which constraints and mass ranges are most significant for overall constraints on PBH dark matter, irrespective of the functional form of the mass function.

1.3.4 Results

We perform the maximization explicitly for several sets of constraints. Set **A** includes robust constraints from evaporation [28]; GRB lensing [44]; microlensing from HSC [13], Kepler [34], EROS [35], and MACHO [10]; and CMB limits from Planck [36]. Set **B** includes dynamical constraints from Segue I [45], Eridanus II [46], and non-disruption of wide binaries [37]. Set **C** includes a constraint from white dwarf explosions [47], a constraint from neutron star capture [48] and a recently claimed constraint

from SNe lensing in the LIGO window [49]. The constraints from evaporation and from Planck in **A** have been estimated differently in the literature, with important consequences for our analysis. Set **A** itself contains relatively non-restrictive estimates of these constraints. We incorporate more stringent versions (see Section 1.5) of these constraints in a set $\bar{\mathbf{A}}$, which is otherwise identical to **A**.

We determine optimal mass functions for sets **A**, $\bar{\mathbf{A}}$, and all of their combinations with sets **B** and **C**. The results are summarized in Table 1.1 and illustrated in Figs. 1.1 and 1.2. We do not include cosmological constraints on the total matter density, so these values of $f_{\rm max,all}$ may exceed 1. In particular, note that all combinations containing **A** have $f_{\rm max,all} > 1$, while all combinations containing $\bar{\mathbf{A}}$ and **C** have $f_{\rm max,all} < 1$. The set $\bar{\mathbf{A}}$ on its own has marginal status if only monochromatic mass functions are considered, but clearly $f_{\rm max,all} > 1$ in this case. With the constraints we consider in this chapter, $f_{\rm PBH} = 1$ is always allowed when using the less stringent set **A**, regardless of additional constraints.

1.4 Prospects for gravitational wave constraints

Gravitational wave observables are the major exception to the rule that measured quantities are linear in the PBH mass function in each mass bin. There are several methods by which gravitational waves might constrain the primordial black hole population. In principle, the simplest constraint arises from present-day measurements of the black hole binary (BHB) merger rate, but this is weak for two reasons: first, it is difficult to distinguish primordial black holes from astrophysical black holes, for which

	f_{mono}	$f_{\text{max,all}}$	$f_{\text{max,GW}}$	$\sigma[\psi]/M_\odot$	$\langle M/M_\odot \rangle$
A	27.17	27.25	2.580	2.259	31.09
AB	1.372	1.965	5.139	0.162	0.009
AC	1.371	1.443	0.566	7.294	1.807
ABC	1.371	1.402	2.936	0.220	0.015
$\bar{\textbf{A}}$	0.991	1.502	2.171	4.827	1.492
$\bar{\textbf{A}}$**B**	0.991	1.437	11.07	0.221	0.017
$\bar{\textbf{A}}$**C**	0.330	0.484	0.364	7.963	5.430
$\bar{\textbf{A}}$**BC**	0.330	0.405	0.982	0.741	0.182

Table 1.1: Optimal mass function properties for each of several sets of constraints. The column f_{mono} gives the maximum DM fraction allowed for a monochromatic mass function, and the column $f_{\text{max,all}}$ gives the maximum DM fraction across all functional forms. The column $f_{\text{max,GW}}$ gives the maximum DM fraction obtained by scaling the semi-analytical optimum while remaining consistent with gravitational wave constraints (see Section 1.4). Also given here are the mean PBH mass and the standard deviation for the semi-analytical optimum mass function.

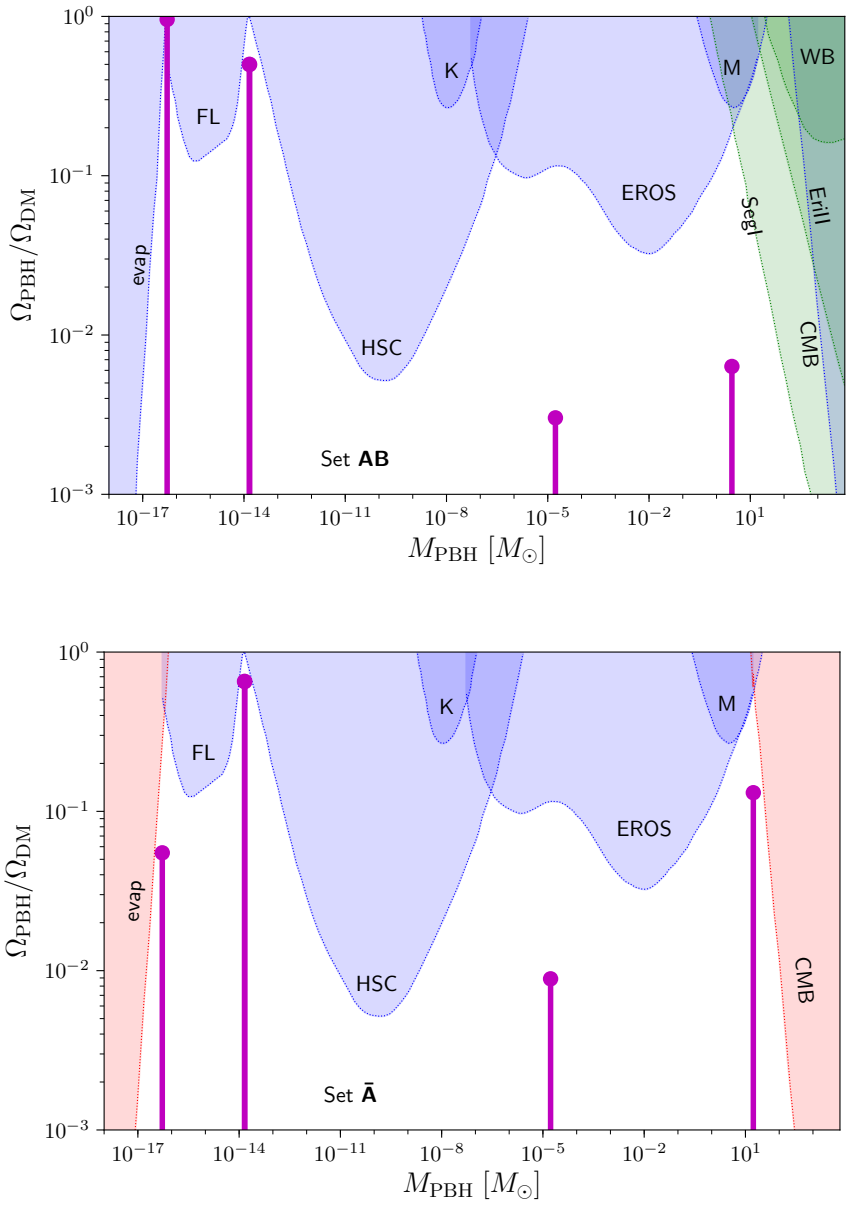

Figure 1.1: The semi-analytical optimum mass function for two sets of constraints. Constraint functions for monochromatic mass functions are shown in blue (**A**), red ($\bar{\mathbf{A}}$), and green (**B**). Vertical lines denote the locations of Dirac deltas in the semi-analytical optimum mass function, with height indicating the weight given to each one. The labeled constraints are from BH evaporation (`evap`, [28]), GRB femtolensing observations (`FL`, [44]), Hyper Suprime-Cam (`HSC`, [13]), Kepler (`K`, [34]), EROS-II (`EROS`, [35]), MACHO (`M`, [10]), Segue I dynamics (`SegI`, [45]), Eridanus II dynamics (`EriII`, [46]), wide binary dynamics (`WB`, [37]), and CMB observables (`CMB`, [36, 39]).

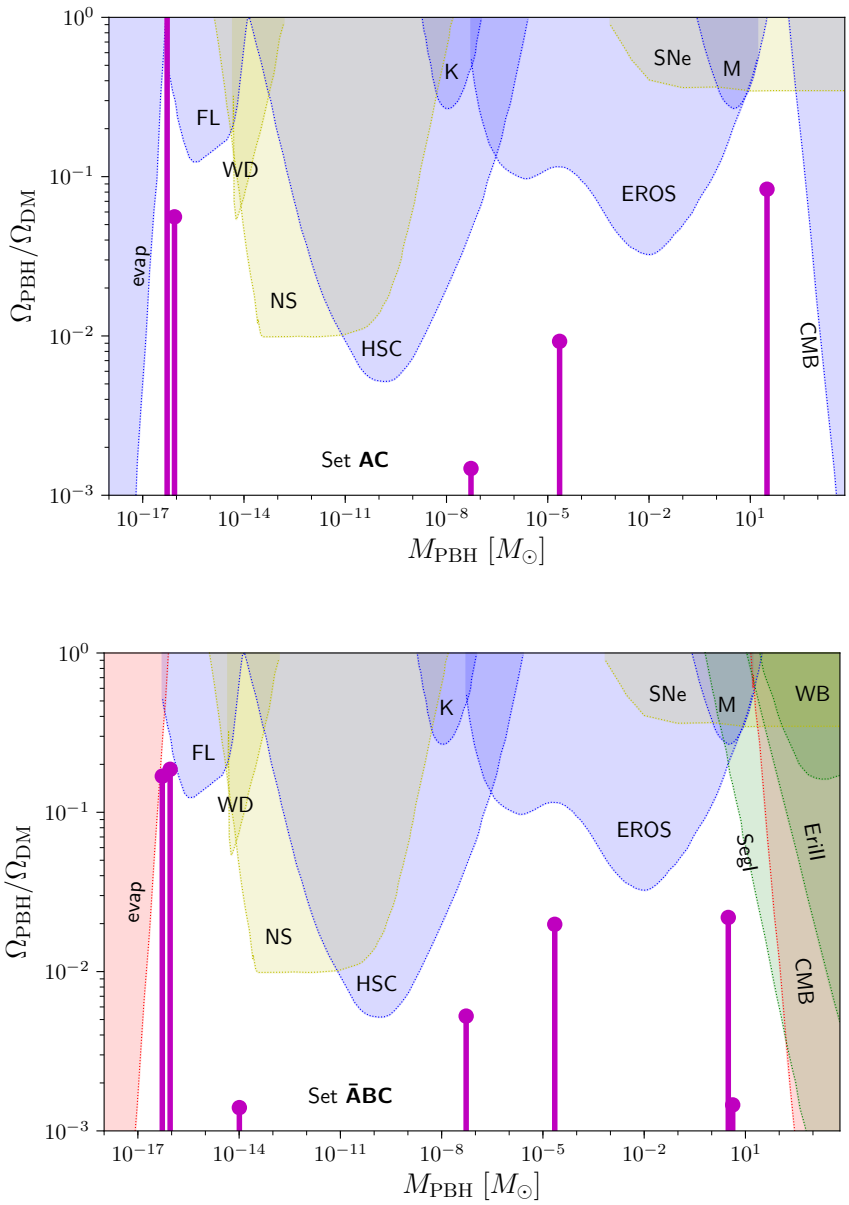

Figure 1.2: The semi-analytical optimum mass function for two sets of constraints. Constraint functions for monochromatic mass functions are shown in blue (**A**), red ($\bar{\mathbf{A}}$), green (**B**), and yellow (**C**). Vertical lines denote the locations of Dirac deltas in the semi-analytical optimum mass function, with height indicating the weight given to each one. The labeled constraints are from BH evaporation (`evap`, [28]), GRB femtolensing observations (`FL`, [44]), white dwarf explosions (`WD`, [47]), Hyper Suprime-Cam (`HSC`, [13]), Kepler (`K`, [34]), EROS-II (`EROS`, [35]), supernova lensing (`SNe`, [49]), MACHO (`M`, [10]), Segue I dynamics (`SegI`, [45]), Eridanus II dynamics (`EriII`, [46]), wide binary dynamics (`WB`, [37]), and CMB observables (`CMB`, [36, 39]).

a variety of additional physical mechanisms may affect the merger rate; and second, the observed merger rate is sufficiently uncertain as to be compatible with a wide range of PBH dark matter models [11, 50].

An alternative method is to search for the stochastic gravitational wave background from primordial density fluctuations associated with inflationary production mechanisms [51]. However, such constraints are only effective within the context of this class of formation models, and within such a limited scope, our level of generality is excessive. Here it is sufficient to consider the mass functions that can be reasonably produced by such formation mechanisms, and constraints for such mass functions have been treated elsewhere in the literature.

A third technique is to search for the stochastic gravitational wave background due to BHB mergers throughout cosmic history [52]. While such an approach may ultimately produce strong constraints, there remains a great deal of uncertainty in modeling the merger rate, particularly for extended mass functions. This problem has only recently been treated in the literature [53, 54], and the resulting constraints may not be robust. Still, it is useful to estimate these constraints, even imprecisely, in order to determine their relevance in the case of our semi-analytical optimum mass functions.

We now consider constraints from the non-detection of a stochastic background of gravitational waves from BHBs throughout cosmic history. This background is qualitatively different from all other observables considered in this chapter, since it has complicated non-linear dependence on the mass function. This means that determining constraints on a general mass function is non-trivial. In particular, one needs to include the higher order terms in the expansion of Eq. (1.5), and the analogue of Eq. (1.6) is

then

$$\sum_{n=1}^{\infty} K_n(M_0, \ldots, M_0) = \frac{A[\psi_{\rm mono}] - A_0}{f_{\rm max}(M_0)} \tag{1.23}$$

which does not constrain off-diagonal values of the kernels K_n. Thus, gravitational wave constraints on the parameter space of monochromatic mass functions are insufficient to determine constraints on an extended mass function, even when the functional form is specified. This reflects the fact that gravitational wave constraints on extended mass functions are inherently model-dependent, in that one must determine the contribution to the background from binaries whose partners have unequal masses.

The results of [53] provide a simple method for estimating the stochastic gravitational wave background given a particular mass function, which we now review. The observable characteristic strain amplitude is given by

$$h_c^2(\nu_{\rm GW}) = \frac{4A_1}{3\pi^{1/3}\left(\log 10\right)^2}\left(\frac{GM_\odot}{c^2}\right)^{5/3}\left(\frac{\nu_{\rm GW}}{c}\right)^{-4/3}\int \frac{{\rm d}m_1}{m_1}\frac{{\rm d}m_2}{m_2}\,\tau_{\rm merge}\left(m_1, m_2\right)\mathcal{M}_c^{5/3} \tag{1.24}$$

where $A_1 \simeq 0.7642 H_0^{-1}$ is a cosmology-dependent constant, $\nu_{\rm GW}$ is the gravitational wave frequency, $\tau_{\rm merge}$ is the mass-dependent binary merger rate per unit volume, and $\mathcal{M}_c \equiv (m_1 m_2)^{3/5}(m_1 + m_2)^{-1/5}$ is the chirp mass. The merger rate is determined by consideration of the capture rate for formation of PBH binaries, given in the Newtonian approximation by

$$\tau_{\rm capture}(m_1, m_2) = 2\pi n_{\rm PBH}(m_1) v_{\rm PBH}\left(\frac{85\pi}{6\sqrt{2}}\right)^{2/7}\frac{G^2(m_1+m_2)^{10/7}(m_1 m_2)^{2/7}}{(v_{\rm rel}/c)^{18/7}c^4} \tag{1.25}$$

where $n_{\rm PBH}$ is the local number density of PBHs, $v_{\rm PBH}$ is the characteristic velocity of a single black hole, and $v_{\rm rel} \equiv \sqrt{2} v_{\rm PBH}$ is the characteristic relative velocity of two black holes. The local number density of PBHs of mass m is parametrized as $n_{\rm PBH} =$

$\delta_{\rm PBH}\rho_{\rm PBH}(m)/m$, where $\rho_{\rm PBH}(m)$ is the cosmological average density of PBHs of mass m and $\delta_{\rm PBH}$ is the local density contrast of PBHs. In particular, in our notation, this number density is given by

$$n_{\rm PBH}(m) = \delta_{\rm PBH}\frac{\Omega_M \rho_c \psi(m)}{m} \tag{1.26}$$

Estimates for $\delta_{\rm PBH}$ range from 10^6 to 10^{10}. In order to estimate conservative constraints, we henceforth take the relatively low value $\delta_{\rm PBH} = 10^7$.

In terms of the capture rate, the merger rate per unit volume is $\tau_{\rm merge}(m_1, m_2) = (f_{\rm PBH}/\delta_{\rm PBH})\tau_{\rm capture}(m_1, m_2)n_{\rm PBH}(m_2)$. Thus the strain amplitude is given by

$$h_c^2[\psi](\nu_{\rm GW}) = \left(\frac{\nu_{\rm GW}}{c}\right)^{-4/3} C \int {\rm d}m_1\, {\rm d}m_2\, \frac{\psi(m_1)}{m_1}\frac{\psi(m_2)}{m_2}\,(m_1 m_2)^{2/7}\,(m_1+m_2)^{23/21} \tag{1.27}$$

where C is a cosmology-dependent factor given by

$$C = \frac{2\pi^{20/21}170^{2/7}c^{1/7}A_1}{3^{9/7}(\log 10)^2}\left(\frac{GM_\odot}{c^2}\right)^{5/3}\frac{\rho_M^2\delta_{\rm PBH}f_{\rm PBH}}{(v_{\rm PBH}/c)^{11/7}}. \tag{1.28}$$

In particular, the dependence of $h_c^2(\nu_{\rm GW})$ on ψ admits a simple expansion of the form of Eq. (1.5). If $f_{\rm PBH}$ is fixed independently of ψ, then the only non-vanishing term has the form $\int {\rm d}m_1\, {\rm d}m_2\, \psi(m_1)\psi(m_2)K_2(m_1, m_2)$, with

$$K_2(m_1, m_2) = C\left(\frac{\nu_{\rm GW}}{c}\right)^{-4/3}\frac{(m_1+m_2)^{23/21}}{(m_1 m_2)^{5/7}}. \tag{1.29}$$

Note that K_2 varies with $\nu_{\rm GW}$, reflecting the fact that measurements of $h_c(\nu_{\rm GW})$ at each frequency $\nu_{\rm GW}$ are independent. Thus, if the instrument used has sufficiently high frequency resolution, a large number of independent constraining observables can be measured, meaning that the maximizing mass function need not resemble a linear

combination of a small number of monochromatic mass functions. In this case, direct numerical methods are necessary to determine the maximum density of PBHs.

It would be inappropriate to perform a full numerical optimization within our framework, since there are considerable theoretical uncertainties in the determination of merger rates. However, we can estimate the potential impact of gravitational wave constraints by checking compatibility of our optimal mass functions with existing gravitational wave observations. This serves to indicate the potential for future modeling work to constrain the PBH population: if the semi-analytical optimum for a given set of constraints is already consistent with gravitational wave observations as well, then we can predict that the detailed inclusion of this additional constraint will have a minimal impact on $f_{\mathrm{max,all}}$.

Current aLIGO bounds on $h_c(\nu_{\mathrm{GW}})$ are strongest at $\nu_{\mathrm{GW}} \simeq 100\,\mathrm{Hz}$, and we represent the current limit by $h_c(100\,\mathrm{Hz}) \lesssim 10^{-22}$. Given a functional form for the mass function ψ, we can compute the maximum f_{PBH} for which $h_c[\psi](100\,\mathrm{Hz})$ satisfies this bound. We denote this maximum by $f_{\mathrm{max,GW}}$, and the values of $f_{\mathrm{max,GW}}$ for our semi-analytical optima are shown in Table 1.1. Of the sets of constraints we consider, only sets **AC** and $\bar{\mathbf{A}}\mathbf{C}$ have $f_{\mathrm{max,GW}}$ significantly less than one. This is to be expected, since the maximizing mass function in both of these cases has a large variance: [53] show that the gravitational wave background is strongly enhanced as the variance is increased. In the case of set $\bar{\mathbf{A}}\mathbf{C}$, we have $f_{\mathrm{max,GW}} \simeq f_{\mathrm{max,all}} < 1$, meaning that the overall maximum is minimally impacted by gravitational wave constraints. In particular, $f_{\mathrm{PBH}} = 1$ is ruled out regardless.

Only set **AC** has $f_{\mathrm{max,GW}} < 1 < f_{\mathrm{max,all}}$, which, in general, is difficult to

interpret: in principle, there may be a different form for the mass function which relaxes gravitational wave constraints, retains $f_{\rm PBH} \geq 1$, and remains consistent with the other constraints we consider in this chapter. A simple way to check this is to consider the maximizing monochromatic mass function, for which gravitational wave constraints should be relaxed compared with the high-variance semi-analytical optimum. Indeed, there is a monochromatic mass function which satisfies all non-gravitational constraints in set **AC** with $f_{\rm PBH} = 1.371$, and we compute $f_{\rm max,GW} > 10$ for this mass function. Thus, while it may not be possible to attain $f_{\rm max,all}$ in this case without violating gravitational wave constraints, $f_{\rm PBH} = 1$ clearly remains allowed. As such, the addition of gravitational wave constraints does not change the overall status of the PBH dark matter paradigm for any of the constraint sets we consider. This reflects both the current status of observations and the large uncertainties in modeling the background. However, future experiments are expected to improve limits on $h_c(\nu_{\rm GW})$ by 2–4 orders of magnitude, which would be sufficient to rule out all of the mass functions represented in Table 1.1 even under fairly conservative assumptions.

1.5 Discussion

With the maximization procedure introduced in Section 1.3.3, it is simple to determine the maximum PBH density consistent with constraints. We stress that this is a bound that applies for mass functions of all forms. Thus, given a set of observational constraints, we can determine a model-independent bound on the density of PBHs.

Our results quantify, for the first time, the risks of using monochromatic mass

functions to assess the overall status of the PBH dark matter paradigm. So long as one window in the constraint functions is much less constrained than all others, the difference between $f_{\rm max,all}$ and $f_{\rm mono}$ is generally very small. Set **A** is a clear example of such a case, and the correction is of order 0.1%. On the other hand, if PBHs are constrained to a similar extent in multiple windows, the correction can be large. The most dramatic example is provided by set $\bar{\mathbf{A}}$, for which $f_{\rm max,all}$ is larger than $f_{\rm mono}$ by $\sim 50\%$. We conclude that, at worst, the bound on the total PBH density is related to the monochromatic bound by an $\mathcal{O}(1)$ factor.

The optimal mass functions themselves (Figs. 1.1 and 1.2) do not correspond to any well-motivated production scenario that we are aware of, and we certainly do not claim that the maximal density can be attained by producing PBHs monochromatically at a discrete collection of masses spanning 15 orders of magnitude. Instead, the panels of Figs. 1.1 and 1.2 should be interpreted as a tool to relate monochromatic constraint functions to their impact on the allowed total density of PBHs. In particular, an immediate and non-trivial conclusion that can be drawn from the figures is that the addition of any new constraint which does not overlap the peaks of the optimal mass function will not reduce $f_{\rm max,all}$.

Further, the functional form of the optimal mass function clarifies the dependence of constraints on the variance of the mass function. In the single-constraint case, we showed that an extended mass function never outperforms the optimal monochromatic mass function. Indeed, in this case, increasing the variance of a narrow mass function will only relax constraints if $f_{\rm max}$ is concave-up in the mass range of interest, i.e., if the monochromatic mass function under consideration is not the optimal

one. When multiple constraints are considered, the relationship between the variance of the mass function and the allowed density is less obvious. Our semi-analytical optimum mass functions all exhibit some non-zero spread, and they definitively allow higher PBH densities than any zero-variance (i.e., monochromatic) mass function. However, extending a monochromatic mass function only slightly, without overlapping additional points of the semi-analytical optimum mass function, is not useful for relaxing constraints. In this respect, our findings are consistent with those of [39, 40].

The most substantial differences in $f_{\mathrm{max,all}}$ arise from differences between **A** and $\bar{\mathbf{A}}$. Set $\bar{\mathbf{A}}$ contains more stringent forms of constraints from CMB anisotropy and PBH evaporation. The CMB constraint is strongly dependent on modeling poorly-understood accretion processes. Both versions of the constraint used in this chapter are drawn from [36]: the version in set **A** is obtained by considering only collisional ionization of the accreted gas, while the version in set $\bar{\mathbf{A}}$ is obtained by including photoionization as well. The evaporation constraint is sensitive to uncertainties in the spectrum of extragalactic background radiation. We adopt the extreme cases considered by [39], with the relaxed form contained in set **A** and the more stringent form in set $\bar{\mathbf{A}}$.

1.5.1 Relative impact of constraints

The values of $f_{\mathrm{max,all}}$ in Table 1.1 demonstrate that the present observational status of PBH dark matter is strongly dependent on the constraints adopted. However, to rule out $f_{\mathrm{PBH}} = 1$, it is necessary to both take the more stringent constraints $\bar{\mathbf{A}}$ in place of **A**, and to include at least one of the constraints from set **C**: supernova microlensing [49], neutron star capture [48], and white dwarf explosions [47].

The supernova microlensing constraint is the most recent of those we consider, and its robustness is the subject of ongoing discussion in the literature [see e.g. 50]. We note that this constraint is dominant in the LIGO window only when dynamical constraints from set **B** are neglected, so the addition of this constraint alone to set **AB** or $\bar{\mathbf{A}}\mathbf{B}$ will have a small impact on $f_{\mathrm{max,all}}$. The constraint from neutron star capture is also subject to astrophysical uncertainties, since it is dependent on the dark matter density in the cores of galactic clusters [48]. We consider the relatively restrictive constraint obtained by taking $\rho_{\mathrm{DM}} = 10^4\,\mathrm{GeVcm}^{-3}$. The strength of the constraint scales linearly with ρ_{DM}, and more conservative estimates take ρ_{DM} smaller by an order of magnitude or more. However, this constraint is most effective in a window shared with constraints from white dwarf explosions, so even if one of the two is subject to substantial uncertainties, the effect of set **C** on $f_{\mathrm{max,all}}$ remains large.

The form of the optimal mass function allows us to rapidly identify the potential impacts of prospective constraints from future observations. For instance, constraints on intermediate-mass black holes with $M \gtrsim 10^2\,M_\odot$ are already strong enough that our semi-analytical optimal mass functions are negligibly small throughout this region. Thus, the identification of additional dynamical systems that might tighten constraints in this region will not affect the overall bound on the PBH density at a level greater than one part in 10^4. On the other hand, GRB femtolensing limits lie in a mass range where some of the semi-analytical optima have a large peak, and strengthening these constraints will have an immediate impact on the overall bound. In particular, upcoming Fermi GRB observations are expected to substantially strengthen constraints in this window, improving by a factor of five after 10 years of operation [55]. These

results may ultimately rule out the PBH dark matter paradigm, with the exception of non-evaporating Planck-mass relics.

Constraints from gravitational wave observations are a special case, as they do not admit the linear interpretation that we take for the impact of extended mass functions on other observables. Further, the strain amplitude at each frequency is sensitive to PBHs in a wide range of masses. Thus, it is not trivial to predict the effect of future gravitational wave constraints on our overall bound without direct numerical optimization. However, constraints from LISA and DECIGO [see e.g. 56] will eventually be capable of ruling out all of our semi-analytical optima, potentially lowering the upper bounds we set in this chapter.

1.6 Conclusions

We have found the form of the mass function which maximizes the PBH density subject to observational constraints, and we have used this to calculate an upper bound on the fraction of dark matter in PBHs. Depending on the constraints adopted, we find $f_{\mathrm{max,all}}$ as large as 27.25 (set **A**) or as small as 0.405 (set $\bar{\mathbf{A}}\mathbf{BC}$). The scenario in which all dark matter is composed of PBHs is ruled out by stringent limits from evaporation and Planck if combined with the constraints from white dwarf explosions, neutron star capture and SNe lensing (set **C**). However, if relaxed constraints from evaporation and Planck are adopted, PBH dark matter is not ruled out by the addition of any other constraints we consider in this chapter. Estimated gravitational wave constraints do not affect these conclusions at the sensitivity of current instruments.

Our method provides a fast and robust technique to determine the total allowed density of PBHs given a set of constraints ($f_{\rm max,all}$), independent of the form of the PBH mass function. The optimal mass function itself allows an easy test of the impact of additional constraints on $f_{\rm max,all}$. While the optimal mass function is not exactly monochromatic, it is very nearly so for realistic constraints. The optimal mass function corresponding to each set of constraints we consider is approximately monochromatic, with additional components scaling the total allowed fraction by no more than an $\mathcal{O}(1)$ factor. Our results explain the findings of [39, 40] that extended mass functions are generally more strongly constrained than monochromatic mass functions, and confirm that the monochromatic maximum density $f_{\rm mono}$ is a good approximation of the allowed density across all mass functions.

Chapter 2

Primordial black holes in extrasolar systems

2.1 Introduction

In this chapter, we turn to a different aspect of PBH phenomenology: the prospect of PBH capture in binary systems.

The future of cosmology and particle physics rests heavily on new astrophysical probes. A growing cast of observational programs offers numerous opportunities to test new physics, even with tools that were designed for entirely different purposes. In particular, new instruments and observational methods have led to surging interest in extrasolar planetary systems within the astronomy community [57–62], with potential implications for beyond–Standard-Model (BSM) particle physics. Several recent proposals demonstrate that exoplanets and other small bodies can sensitively probe BSM scenarios, including e.g. dark matter (DM) interactions [63] and new long-range forces

[64]. In this chapter, we study the prospects for using these systems to detect primordial black holes, i.e., black holes that formed at early times from mechanisms besides stellar collapse.

Many PBH searches are designed to target rare but distinctive signatures. In particular, a *single object* can be identified as a PBH if it lies outside the mass range achievable by stellar collapse, providing clear evidence of new physics and defining a clear direction for subsequent DM searches. Amid this context, it is critical to understand the various astrophysical environments in which one might expect to find PBHs.

If PBHs do make up a significant fraction of cosmological DM, then they should be scattered throughout our galactic DM halo, with a comparable phase space distribution. For particle DM, the phase space distribution is generally sufficient to determine any observable at any time. However, for PBHs, many observables of interest are discrete events that are rare on the timescale of observations. For instance, lensing events [10, 13, 65, 66] or low-mass PBH mergers [9, 67–69] would each occur infrequently during the corresponding observations. While the time-averaged rate of such events is determined by the DM phase space distribution, the events themselves are stochastic.

In this chapter, we consider a scenario which translates this stochasticity from the *timing* of rare events to the distribution of systems in which they occur: we consider the capture of PBHs into bound orbits in an extrasolar stellar system. If PBHs are indeed captured in this manner, and if such captured orbits are long-lived, then some fraction of stellar systems should stably host PBHs at any given time. In the limit that captured objects are permanently bound, a stellar system would only need one encounter with a PBH over its entire history in order to host such an object today. In

particular, this means that the rapidly advancing observational techniques probing the dynamics of such systems may also provide a new probe of the PBH population.

Capture requires the incoming PBH to lose mechanical energy in order to become bound to a stellar system. There are several different physical mechanisms that can lead to such a loss of energy, and these mechanisms can be classified by the sink that absorbs energy from the PBH:

1. In an encounter with a single other body, the PBH can be rapidly accelerated, causing it to lose energy to gravitational radiation.

2. In an encounter with a few-body system, the PBH can lose mechanical energy to one object and become bound to another object.

3. In passing through a many-body system, the PBH can dissipate energy and effectively heat the system.

These energy sinks are each associated with unique phenomenology. In the first case, gravitational waves from such close encounters are potentially detectable directly [70, 71]. In the second case, if the energy transfer is purely mechanical, then the process is time-reversible, and the PBH may be ejected once captured. Finally, in the case of a dissipative process, the deposited energy itself may have observable consequences for the host system.

Making precise predictions for the population of captured objects is inherently challenging. In the case of few-body encounters, such processes are governed by a relatively simple set of parameters, but can nonetheless exhibit complicated chaotic dynamics. By contrast, the dynamics of many-body processes are comparatively simple,

but the parameters of these systems are subject to significant astrophysical uncertainties. In this chapter, we establish order-of-magnitude predictions for the abundance of PBHs captured by each of the above mechanisms across a wide variety of systems.

Throughout this chapter, we will assume that PBHs make up all of the DM, and we will assume that their velocities are described by a truncated Maxwell–Boltzmann distribution. We focus mainly on two classes of objects: first, black holes at masses near Earth mass $M_\oplus$, which may account for excess microlensing observations by OGLE [66], and second, microscopic black holes from $10^{-16}\,\mathrm{M}_\odot$ to $10^{-12}\,\mathrm{M}_\odot$, where current constraints are ineffective. In particular, at the lower end of this mass range, active evaporation may make it possible to detect these objects.

This work is organized as follows. We devote one section to each sink of energy that can lead to captures: in Section 2.2, we study the capture of PBHs due to gravitational radiation; in Section 2.3, we evaluate the abundance of PBHs captured by few-body interactions; and in Section 2.4, we study captures that take place via dissipative dynamics. We discuss our findings and implications for observables in Section 2.5.

2.2 Gravitational radiation in two-body encounters

In this section, we first review basic principles that apply to all captures, and we then estimate the rate and lifetime of captures due to gravitational wave (GW) emission.

2.2.1 Generalities of capture

A capture takes place when a free PBH becomes bound to some stellar system, i.e., when its total mechanical energy changes sign from positive to negative. This requires that the object give up energy to the surroundings. Thus, to evaluate the rate of captures, we can evaluate the cross section for an incoming object to lose an amount of energy commensurate with its initial mechanical energy. For an object that originates far from the stellar system, as a free object should, the initial mechanical energy is simply the initial kinetic energy. Thus, given an initial velocity v_∞, the capture cross section for a PBH of mass $M_{\rm PBH}$ is just the cross section for the object to lose an amount of energy greater than its initial kinetic energy:

$$\sigma_{\rm cap}(v_\infty) = \int_{\frac{1}{2}M_{\rm PBH}v_\infty^2}^{\infty} \mathrm{d}E_{\rm loss} \frac{\mathrm{d}\sigma_{\rm cap}}{\mathrm{d}E_{\rm loss}}. \tag{2.1}$$

We will assume that PBHs have a velocity distribution like that of halo DM, with probability density function (pdf) given by

$$f_\infty(v) \propto v^2 \exp\left(-\frac{v^2}{v_0^2}\right) \Theta\left(v_{\rm esc} - v\right). \tag{2.2}$$

We take $v_0 = 220\,\mathrm{km/s}$ and $v_{\rm esc} = 550\,\mathrm{km/s}$. This pdf is an approximate description of the equilibrium distribution of DM particles—in our case, PBHs—throughout the halo. Near a point mass like a star, the velocity distribution is modified by the local gravitational potential. Thus, $f_\infty(v)$ should be treated as the distribution of particles far from the stellar system. In particular, the existence of a low-velocity tail of the distribution does not imply that low-velocity objects are "born" captured. However, per Eq. (2.1), the capture cross section is typically largest for the smallest values of v_∞, and in some cases, the low-velocity tail dominates the capture rate.

Each energy loss mechanism leads to some differential cross section $\mathrm{d}\sigma_{\mathrm{cap}}/\mathrm{d}E_{\mathrm{loss}}$, and thus to some total cross section $\sigma_{\mathrm{cap}}(v_\infty)$. Once these quantities are calculated, the total capture rate is

$$R_{\mathrm{cap}} = n_\infty \langle \sigma_{\mathrm{cap}} v_\infty \rangle = n_\infty \int \mathrm{d}v \, f_\infty(v) \sigma_{\mathrm{cap}}(v) v, \tag{2.3}$$

where n_∞ is the number density of objects far from the system and angle brackets denote the average over velocities. Some systems can also lose captured objects, particularly by ejection in few-body systems with conservative dynamics. In this case, the rate for a particular object to be ejected is independent of the number of objects captured in the system, so we represent this rate by a single quantity R_{ej}. Thus, in a system with N objects captured, the rate for any one object to be lost is NR_{ej}. On a sufficiently long timescale, capture and loss are in equilibrium, meaning that the expected number of captured objects in a system is $\langle N \rangle = R_{\mathrm{cap}}/R_{\mathrm{ej}}$. Assuming that $R_{\mathrm{ej}} \ll R_{\mathrm{cap}}$, equilibrium is attained on a timescale of order $t_{\mathrm{eq}} \simeq \langle N \rangle / R_{\mathrm{cap}} = 1/R_{\mathrm{ej}}$.

2.2.2 Capture cross section from gravitational radiation

We now consider gravitational wave emission as the physical mechanism for energy loss, and evaluate the expected number of objects that are captured by this route.

A PBH that undergoes a close encounter is rapidly accelerated, losing a significant amount of energy to gravitational radiation in the process. If enough energy is lost, the PBH can become bound as a result of the encounter. Such a capture can be much more stable than a capture produced by few-body dynamics. In particular,

this process can take place in a *two*-body encounter, or if in a few-body system, it can take place far from the orbital trajectory of the any third body, minimizing the rate of subsequent close encounters that could lead to ejection.

To compute the energy lost to gravitational waves, we follow Ref. [71]. We consider a close encounter between the PBH and a stellar or planetary body S. The energy loss in the encounter is given by

$$\Delta E_{\rm GW} = \frac{8}{15}\frac{M_{\rm PBH}^2 M_S^2}{(M_{\rm PBH}+M_S)^3}\frac{p(e)}{(e-1)^{7/2}} \times v_\infty^7, \tag{2.4}$$

where v_∞ is the relative speed at infinity, e is the eccentricity of the inbound orbital trajectory, and $p(e)$ is given by

$$p(e) = (e+1)^{-7/2}\left[\arccos\left(-\frac{1}{e}\right)\left(24+73e^2+\frac{37}{4}e^4+\frac{\sqrt{e^2-1}}{12}\left(602+673e^2\right)\right)\right]. \tag{2.5}$$

The eccentricity can be written in terms of v_∞ and the impact parameter b as

$$e = \sqrt{1+\frac{b^2 v_\infty^4}{(GM_S)^2}}. \tag{2.6}$$

Here we are interested in cases in which the PBH is captured, i.e., in which $\Delta E_{\rm GW}$ exceeds the kinetic energy of the PBH at infinity. We are especially interested in the possibility that the PBH is captured without passing through object S, so that it does not become captured *within* object S and settle to the center. Here there is a very strong dependence on the impact parameter of the encounter, and thus on the radius of object S. For example, if object S is a Jupiter-like planet, then the energy loss will be extremely small for all impact parameters that avoid collisions with the planet. On the other hand, if object S is a compact object like a neutron star, then

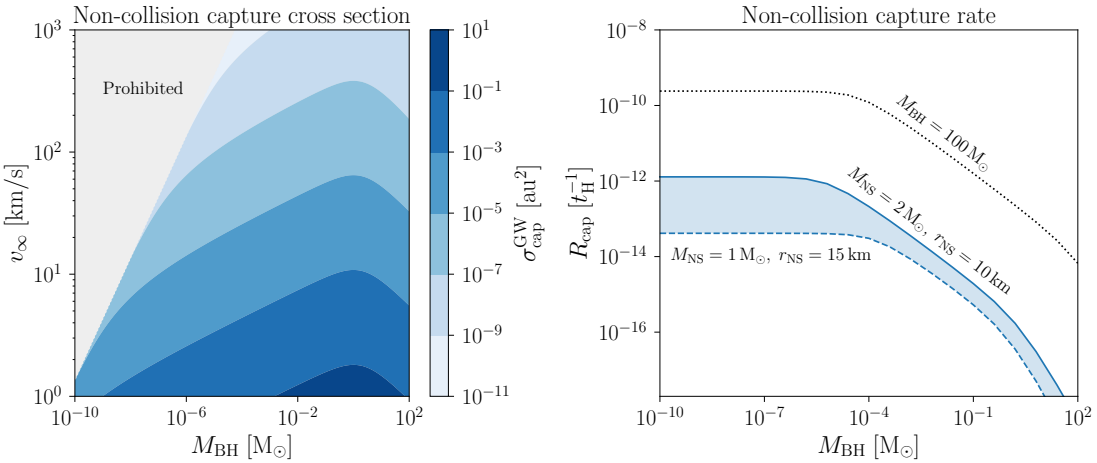

Figure 2.1: *Left:* Cross section for capture by gravitational wave emission without collision. A neutron star of mass $M_{\rm NS} = 2\,{\rm M}_\odot$ and radius $R_{\rm NS} = 10\,{\rm km}$ is assumed. In the gray region, capture without collision is not possible. *Right:* integrated capture rate as a function of PBH mass in units of the Hubble rate. The shaded area shows the region between two neutron star configurations: one with mass $M_{\rm NS} = 2\,{\rm M}_\odot$ and radius $R_{\rm NS} = 10\,{\rm km}$, and one with mass $M_{\rm NS} = 1\,{\rm M}_\odot$ and radius $R_{\rm NS} = 15\,{\rm km}$. These correspond roughly to the most and least compact neutron stars expected to form [72]. The dotted black curve corresponds to encounters with a $100\,{\rm M}_\odot$ black hole. Substantially less compact objects such as main-sequence stars cannot capture BHs of any size at realistic velocities by GW emission.

small impact parameters without collisions are indeed realizable, as discussed in detail by Ref. [71]. Such captures are stable on fairly long timescales. In particular, for PBH masses $M_{\rm PBH} \lesssim 10^{-14}\,{\rm M}_\odot$, these captures survive longer than a Hubble time.

Captures without collision are possible for a bounded range of impact parameters. The capture condition $\Delta E_{\rm GW} > \frac{1}{2} M_{\rm PBH} v_\infty^2$ gives a critical impact parameter, $b_{\rm max}$, below which an encounter will lead to capture. While solving for $b_{\rm max}$ is in general quite complicated, it is a good approximation to set $e = 1$ in Eq. (2.5), corresponding

to a free object with minimal kinetic energy. This gives

$$e = 1 + \left(\frac{bv_\infty^2}{GM_S}\right)^2 + \mathcal{O}\left[\left(\frac{bv_\infty^2}{GM_S}\right)^4\right] \simeq 1 + 10^{-9}\left(\frac{b}{100\,\mathrm{km}}\right)^2\left(\frac{v_\infty}{220\,\mathrm{km/s}}\right)^4\left(\frac{M_S}{1\,\mathrm{M}_\odot}\right)^{-2}. \tag{2.7}$$

Taking $e = 1$ gives a 7th-order polynomial equation in $b_{\max}$, which is readily solved semi-numerically to find the maximal impact parameter for capture. On the other hand, to avoid a collision, there is a minimum impact parameter: the point of closest approach in a Kepler orbit, $r_{\min}$, is related to the impact parameter by

$$r_{\min} = \frac{GM_S}{v_\infty^2}\left[\left(1 + \frac{b^2 v_\infty^4}{(GM_S)^2}\right)^{1/2} - 1\right], \tag{2.8}$$

so the minimum impact parameter to avoid a collision is found by setting $r_{\min} = R_S$ in Eq. (2.8), where R_S is the radius of the object S. That is, we take

$$b_{\min} = \sqrt{R_S^2 + \frac{2GM_S R_S}{v_\infty^2}}. \tag{2.9}$$

The cross section for capture by gravitational wave emission without collision is then given by $\sigma_{\mathrm{cap}}^{\mathrm{GW}} = \pi(b_{\max} - b_{\min})^2\Theta(b_{\max} - b_{\min})$. This cross section is shown as a function of M_{PBH} and v_∞ in Fig. 2.1. As expected, $\sigma_{\mathrm{cap}}^{\mathrm{GW}}$ is larger for small v_∞, but it also increases moderately for larger PBH massses M_{PBH} due to the non-linear dependence on M_{PBH} in the energy emitted in GWs.

For our cases of interest, i.e., microscopic PBHs and Earth-mass PBHs, capture by GW emission is exceedingly rare. As shown in Fig. 2.1, the capture rate is extremely small both at small PBH mass, due to the inefficiency of GW emission, and at $M_{\mathrm{PBH}} \sim M_\oplus$, due to the very small number density of such objects. Given the sharp dependence on the lowest velocities, it is possible that a cold feature in the phase space distribution

of the halo could substantially enhance the capture rate, but typical capture rates are well below the Hubble rate (inverse Hubble time). Indeed, the capture rate is below the rate at which captured objects sink to the center of the NS by further GW emission.

2.3 Few-body interactions

It is this sort of temporary capture which concerns us in this chapter. We are motivated by an apparently simple question: what are the properties of the population of captured objects in a given binary system? The resolution of this question is relevant to the study of free-floating exoplanets and their bound counterparts [73–77], for example, but is also significant for less familiar objects. In particular, it is important for assessing the population of captured dark matter particles, or for characterising the demographics of compact objects that might be temporarily captured in observable binary systems, including the capture of interstellar objects in the solar system [78–84].

The capture of unbound objects into bound orbits by binary systems has been studied by many authors in widely varying contexts. Three-body capture and ejection were studied systematically by Ref. [85], who obtained approximate forms for the rates of these processes in cases where detailed balance can be applied. Subsequently, the theory of capture and ejection was extended by several authors to study comets in the solar system [78, 79, 86, 87], interstellar panspermia [81, 83], and the population of captured dark matter particles in the vicinity of Earth [88–93]. A comprehensive account of results and astronomical applications is given by Ref. [94].

Many of these studies are based on the results of detailed numerical simula-

tions, which make it possible to study the properties of the captured population both immediately after capture and at late times. However, the results of such simulations are specific to the solar system. In scenarios involving extrasolar binary systems, it is important to have a simple description of capture processes that holds for a wide range of systems and interloper velocities. For such purposes, it is desirable to have a flexible semi-analytical framework for describing the population of captured objects—not only the capture and ejection rates, but also the distributions of orbital parameters of captured objects. Moreover, it is valuable to describe the dependence of each of these on the parameters of both the binary and the third body prior to capture. Finally, it is useful to obtain a simple geometric description of the types of encounters that lead to captures, and to understand the behavior of these captured trajectories at late times.

In this extended section, we develop such a formalism. We focus on captures resulting from a close encounter between a test particle and the smaller body of a binary, and we demonstrate that the form of the capture cross section in this case lends itself well to predictions of orbital parameters and ejection time-scales. In particular, within certain approximations, we show that the set of impact parameters leading to captures forms a disc whose parameters can be written in closed form. We use this result to derive analytical approximations for the capture rate, the orbital parameters of captured objects, and the ejection time-scale. Our results generalize those of Ref. [79] and provide an analytical interpretation of the sorts of trajectories studied therein. We further extend the results to give a simple prescription for the ejection rate of captured objects as a function of their parameters upon capture.

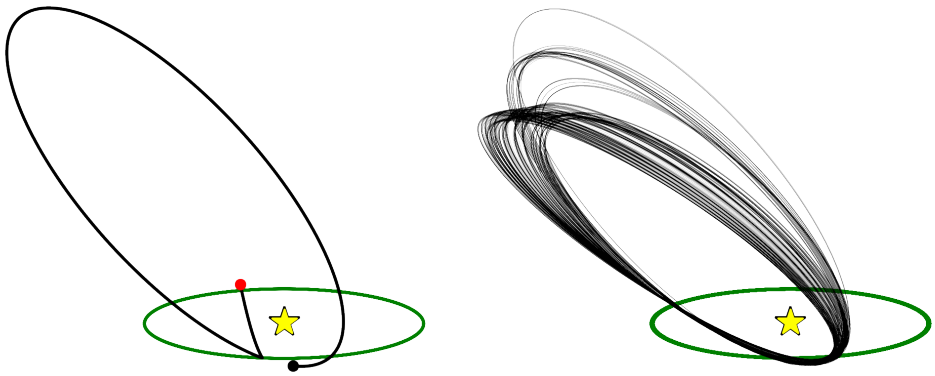

Figure 2.2: Left: numerical simulation of a single three-body capture of a test particle by the sun–Jupiter system. The simulation begins at the red dot. Right: long-term evolution of the captured object, showing successive changes in the orbital parameters.

2.3.1 The capture cross section

Our goal is to identify the sorts of close encounters in which the incoming object is slowed enough to enter a bound orbit. In this subsection, we describe the set of impact parameters leading to captures, and connect this with both the capture cross section and the distributions of orbital parameters. We first establish our notation and approximations, which largely follow the presentation of Ref. [79]. The notation is summarized in Fig. 2.3.

We assume that the binary system is composed of two objects A and B with masses M_A and M_B, and we take $M_A \gg M_B$. We use $\mu_X \equiv GM_X$ to denote the standard gravitational parameter for any object X, where G is Newton's constant, and we denote the distance between any two objects X and Y by r_{XY}. While our formalism can be naturally extended to accommodate eccentric binaries, we take the orbit to be circular ($e = 0$) in this chapter, so that r_{AB} is constant. Unprimed quantities are measured in the frame of A and primed quantities are measured in the frame of B. We

assume that a test particle C is incident from infinity with velocity $\mathbf{v}_\infty$, has a close encounter with object B, and thereafter becomes bound to object A. We write $\mathbf{v}_1$ and $\mathbf{v}_2$ to denote the velocity of C just before and just after the close encounter.

The state of the binary is described by a single phase λ_1, and we define $\lambda_1 = 0$ to be the phase such that the A–B axis is parallel to the projection of $\mathbf{v}_\infty$ in the plane of the orbit. We will assume that the time-scale of the close encounter is much smaller than the orbital time-scale of the AB system so that λ_1 does not change significantly during the close encounter, i.e., we work in the impulse approximation. In general, $\mathbf{v}_\infty$ is inclined with respect to the orbital plane by an angle β_1, and $\mathbf{v}_2$ is inclined by an angle β_2. Additionally, we will speak of the impact parameter for the close encounter as a vector $\mathbf{b}$ in the frame of object B, spanning from B to the point of closest approach of C if the latter were to continue travelling undeflected with velocity $\mathbf{v}_1'$ (see Fig. 2.3, inset). We define $\mathbf{b}$ in the plane orthogonal to $\mathbf{v}_1'$, endowing this plane with polar coordinates (b, ϕ). We will fix the axis $\phi = 0$ shortly, and we will also return to the subtlety of frame-dependence in the definition of $\mathbf{b}$. First, however, we quantify the meaning of a close encounter.

For our purposes, a close encounter takes place when C passes close enough to B so that tidal acceleration by A can be neglected. Then the encounter can be treated purely as a two-body problem in the frame of object B, greatly simplifying the analysis. This translates to the condition

$$\frac{\mu_A}{(r_{AB} - r_{BC})^2} - \frac{\mu_A}{r_{AB}^2} < \epsilon \frac{\mu_B}{r_{BC}^2} \quad \text{for some } \epsilon \ll 1. \tag{2.10}$$

Note that r_{BC} is not a fixed parameter of the encounter, but rather evolves throughout

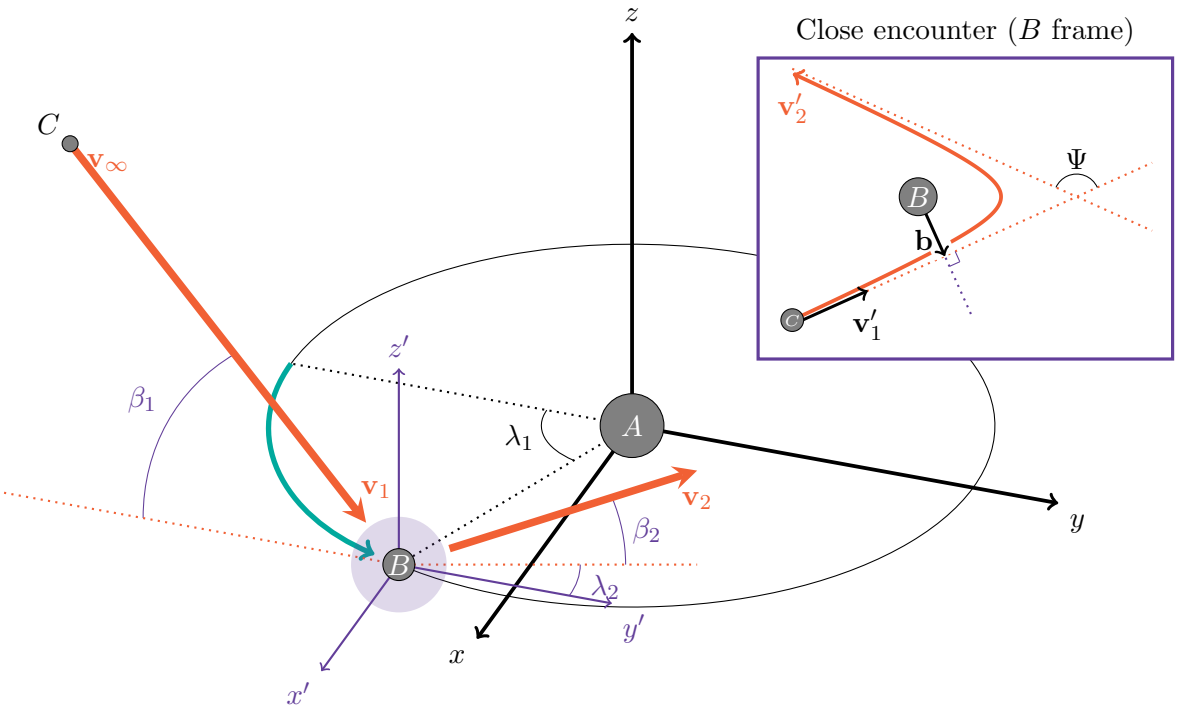

Figure 2.3: Configuration and notation assumed in Section 2.3.1. *Centre:* 3d illustration of the encounter on the scale of the AB system. Dotted lines lie in the plane of the AB system. Notation largely follows Ref. [79]. Object C, with velocity $\mathbf{v}_\infty$ at infinity, has a close encounter with object B in the shaded region with initial velocity $\mathbf{v}_1$ and exits the encounter with velocity $\mathbf{v}_2$. Note that on the scale of the system as drawn, the trajectory of C should be curved throughout due to acceleration by A, a feature we omit for simplicity. *Inset:* 2d illustration of the close encounter in the frame of object B. Dotted lines lie in the plane of the two-body scattering process. The inset is intended only to illustrate the notation, and is not drawn to scale with respect to the centre image.

the scattering process. The condition above determines which values of r_{BC} are small enough to indicate a close encounter. To leading order in ϵ, this condition can be written in the form

$$r_{BC} \lesssim r_{\text{close}}(\epsilon) \equiv r_{AB} \left(\frac{M_B \epsilon}{M_A} \right)^{1/3}. \tag{2.11}$$

Note that $r_{\text{close}}(\epsilon)$ is smaller than the Hill radius for $\epsilon \ll 1$, and for a fixed choice of ϵ, the value of $r_{\text{close}}(\epsilon)$ defines what we mean by a close encounter. Later, when computing the capture cross section numerically, we will take $\epsilon = 0.1$ and neglect trajectories for which $\min r_{BC} > r_{\text{close}}(\epsilon)$. This leads to a conservative result for the capture cross section, but has the opposite effect on the ejection cross section, as we will discuss later. Since $M_A \gg M_B$, we will assume that $r_{\text{close}} \ll r_{AB}$.

Having made this definition of a close encounter, we can compute v_1 as a function of v_∞. Our approach assumes that the close encounter can be treated as an isolated two-body problem, which is only appropriate if the gravitational potential of object B is small at r_{close}. Otherwise, the acceleration of C is dominated by the potential of A for a significant part of the encounter, and by the time the two-body treatment is applicable, C is already well within the potential of B. In the case that this effect can be neglected, it is sufficient to account for acceleration of C by A during infall from infinity to r_{close}, which gives

$$v_1 = \sqrt{v_\infty^2 + 2\mu_A / r_{AB}}. \tag{2.12}$$

On the other hand, if the potential of B is not small at r_{close}, then C has now been non-negligibly accelerated by B prior to the close encounter, but v_1 must still be fixed where the close encounter begins. Thus, in general, we will include this additional prior

acceleration, and we take

$$v_1 = \sqrt{v_\infty^2 + 2\mu_A/r_{AB} + 2\mu_B/r_{\text{close}}(\epsilon)}. \tag{2.13}$$

For the sun–Jupiter system, this additional acceleration contributes only a fraction of a percent to v_1, but in other realistic systems, the effect can be significantly larger. Note that this expression fixes only the speed v_1 in terms of v_∞, and does not specify the vectorial relation between $\mathbf{v}_\infty$ and $\mathbf{v}_1$. We will return to the implications of directionality shortly.

Now, presuming a close encounter, we determine the conditions leading to capture of C. Under the stated assumptions, the relative velocity of B and C evolves as in the two-body problem from $\mathbf{v}_1'$ to some $\mathbf{v}_2'$. Object C is bound after the close encounter if its speed is sufficiently low in the A frame, i.e., if $v_2 < v_{\text{esc}}$, where $v_{\text{esc}} = \sqrt{2\mu_A/r_{AB}}$ is the escape velocity of object A at the location of the close encounter. The key feature of the two-body encounter for our purposes is that the speed of recession is equal to the speed of approach, i.e., $v_1' = v_2'$. This makes the outcome of the encounter very simple to describe analytically: the trajectory of C is simply deflected by an angle Ψ about the axis parallel to $\mathbf{b} \times \mathbf{v}_1'$. The angle Ψ is related to the impact parameter b via

$$\cos\Psi = \frac{b^2 v_1'^4 - \mu_B^2}{b^2 v_1'^4 + \mu_B^2}. \tag{2.14}$$

We can now compute $\mathbf{v}_2$ in terms of $\mathbf{b}$ algebraically. To be concrete, we first rotate the coordinate system so that $\mathbf{v}_1' \propto \hat{\mathbf{z}}$ and $\mathbf{b} \propto \hat{\mathbf{x}}$ via a rotation R_1. Then the deflection of $\mathbf{v}_1'$ into $\mathbf{v}_2'$ is computed by performing a rotation by Ψ in the xz-plane. This procedure allows us to define ϕ unambiguously: the impact parameter $\mathbf{b}$ lies in the plane orthogonal

to $\mathbf{v}_1'$, so in the rotated coordinate system, it takes the form $\mathbf{b} = (b_1, b_2, 0)$. We define $\phi = 0$ such that $\mathbf{b}$ lies in the xy-plane in the *original* coordinate system. That is, we require that $(R_1^{-1}\mathbf{b}) \cdot \hat{\mathbf{z}} = 0$. If we further choose that the x-component is positive, we can solve for b_1 and b_2 uniquely:

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix}_{\phi=0} = \frac{b \operatorname{sign}(v_{1y}')}{\sqrt{v_{1x}'^2 + v_{1y}'^2}} \begin{pmatrix} v_{1y}' \\ -v_{1x}' \end{pmatrix}. \tag{2.15}$$

Now $\mathbf{b}$ can be obtained for arbitrary ϕ by rotation of Eq. (2.15).

To implement the deflection by Ψ, observe that in the new coordinate system, the rotation axis $\hat{\mathbf{r}}$ is related to $\mathbf{b}$ by a $\pi/2$ rotation. It is convenient to change coordinates with a rotation R_2 in the xy-plane to align $\mathbf{b}$ with the x-axis and $\hat{\mathbf{r}}$ with the y-axis. In the coordinate system produced by the rotation R_2R_1, the deflection corresponds to a simple rotation by Ψ in the xz-plane, which we denote by R_Ψ. It follows that the deflection $\mathcal{R}\colon \mathbf{v}_1' \mapsto \mathbf{v}_2'$ is implemented by the matrix $\mathcal{R} = R_1^{-1}R_2^{-1}R_\Psi R_2 R_1$. Using Eq. (2.14) to write $\cos\Psi$ in terms of b, and using $\mathbf{v}_2 = \mathbf{v}_2' + \mathbf{v}_B$, we can now write $\mathbf{v}_2'$ in terms of b. For brevity, we define

$$a_{xy} \equiv \sqrt{a_x^2 + a_y^2}, \qquad q \equiv \sqrt{1 + v_{1z}'^2 \Big/ v_{1xy}'^2}, \tag{2.16}$$

for any vector $\mathbf{a}$. Then $\mathbf{v}_2$ is given by

$$\mathbf{v}_2 = \mathbf{v}_B + \left(\frac{b^2 v_1'^4 - \mu_B^2}{b^2 v_1'^4 + \mu_B^2}\right)\mathbf{v}_1' + \frac{2\operatorname{sign}(v_{1y}')\mu_B v_1' b}{b^2 v_1'^4 + \mu_B^2} \begin{pmatrix} q\left(v_{1x}' v_{1z}' \sin\phi - v_1' v_{1y}' \cos\phi\right) \\ q\left(v_{1y}' v_{1z}' \sin\phi + v_1' v_{1x}' \cos\phi\right) \\ -v_1' v_{1xy}' \sin\phi \end{pmatrix}. \tag{2.17}$$

Neglecting collisions with B, the condition for capture of C can now be expressed succinctly as $v_2 < v_{\text{esc}}$. Conveniently, it can be shown with some algebraic

effort that saturation of this condition produces the equation of a circle in the plane of impact parameters. To this end, observe that the boundary relation $v_2 - v_{\text{esc}} = 0$ can be factored in the form

$$v_2 - v_{\text{esc}} = -F(\mathbf{v}_1, \mathbf{b}) \operatorname{sign}(v'_{1y}) v'_{1xy} v'^4_1 \left(\frac{r_{AB} v_1^2 - 2\mu_B}{b^2 v'^4_1 + \mu_B^2} \right), \tag{2.18}$$

for some factor $F(\mathbf{v}_1, \mathbf{b})$. The remainder of the right-hand side depends on $\mathbf{b}$ only through the factor $b^2 v'^4_1 + \mu_B^2$, which is positive-definite. Thus, the right-hand side apart from $F(\mathbf{v}_1, \mathbf{b})$ is non-zero almost everywhere, so our original condition can be rewritten in the form $F(\mathbf{v}_1, \mathbf{b}) = 0$. Carrying out the factorization explicitly, F has the form

$$F(\mathbf{v}_1, \mathbf{b}) = b^2 + \left(\frac{g_5}{g_4} - \frac{g_2}{g_1} \right) + b \left(\frac{g_3}{g_4} \cos\phi + \frac{g_1}{g_4} \sin\phi \right), \tag{2.19}$$

where the coefficients g_i are given in Table 2.1. In fact, the relation $F(\mathbf{v}_1, \mathbf{b}) = 0$ is simply the equation of a circle in the plane orthogonal to $\mathbf{v}'_1$, with radius R and centre $\mathbf{b}_c$ given by

$$R(\mathbf{v}_1) = \sqrt{\frac{g_2}{g_1} + \frac{g_1^2 + g_3^2}{4g_4^2} - \frac{g_5}{g_4}}, \qquad \mathbf{b}_c(\mathbf{v}_1) = -\frac{1}{2g_4} \begin{pmatrix} g_3 \\ \\ g_1 \end{pmatrix}. \tag{2.20}$$

This allows us to make an extremely simple estimate of the capture cross section: we have simply

$$\sigma_{\text{cap}}(\mathbf{v}_1) \simeq \pi \min \left[R(\mathbf{v}_1),\ r_{\text{close}}(\epsilon) \right]^2. \tag{2.21}$$

When $R(\mathbf{v}_1) < r_{\text{close}}(\epsilon)$, this takes the form

$$\sigma_{\text{cap}}(\mathbf{v}_1) \simeq \frac{\pi \mu_B^2 \left[\left(v'^2_1 - v_B^2 \right)^2 - v_{\text{esc}}^4 \right]}{\left(v_1^2 - v_{\text{esc}}^2 \right)^2 v'^4_1}. \tag{2.22}$$

This simple expression gives the capture cross section as a function of the incoming object's direction with respect to the axis of the binary—again, assuming a circular

$$
\begin{aligned}
g_1 &= 4\mu_B(\mathbf{v}_B \cdot \mathbf{v}_1')v_1'^6 v_{1z}' \\
g_2 &= 4\mu_B^3(\mathbf{v}_B \cdot \mathbf{v}_1')v_1'^2 v_{1z}' \\
g_3 &= 4\mu_B v_1'^7 \left(\mathbf{v}_1' \times \mathbf{v}_B\right)_z \\
g_4 &= \mathrm{sign}\left(v_{1y}'\right) v_1'^8 v_{1xy}' \left(v_1^2 - v_{\mathrm{esc}}^2\right) \\
g_5 &= 2\,\mathrm{sign}\left(v_{1y}'\right) \mu_B^2 v_1'^4 v_{1xy}' \left(v_1'^2 + v_B^2 - v_{\mathrm{esc}}^2\right)
\end{aligned}
$$

Table 2.1: Coefficients g_i appearing in Eqs. (2.19) and (2.20). Here $\mathbf{v}_1 = \mathbf{v}_1' + \mathbf{v}_B$ and v_{1xy}' is the magnitude of the projection of $\mathbf{v}_1'$ onto the xy plane.

binary and working within the impulse approximation. When computing rates, the cross section should be multiplied by a factor of v_1/v_∞ to account for gravitational focusing. Since v_1 and v_1' scale with v_∞, the cross section vanishes rapidly for $v_\infty \gg v_{\mathrm{esc}}$. On the other hand, as $v_\infty \to 0$, the velocity v_1 is nearly equal to v_{esc}, up to the small correction due to the potential of object B (see Eq. (2.13)). Thus, the cross section becomes very large, and is eventually subject to the cutoff in Eq. (2.21).

Equation (2.22) only holds for parameters such that $R(\mathbf{v}_1)$ is real in Eq. (2.20), which is a non-trivial constraint. In particular, there is a maximum change in velocity that can be imparted to object C during the encounter: the speed of approach is equal to the speed of recession in the frame of object B, so the maximum impulse corresponds to the case in which the direction of object C is exactly reversed in the frame of B (i.e., $\cos\Psi = -1$). In this case, $|\Delta v| = 2v_B$ in the frame of A. This means that there is a maximum velocity $v_{\mathrm{max}} = v_{\mathrm{esc}} + 2v_B$ such that objects with $v_1 > v_{\mathrm{max}}$ cannot be captured regardless of impact parameter. Such velocities correspond to non-real values of $R(\mathbf{v}_1)$, and for these velocities, the capture cross section is exactly zero.

We may now average over the binary phase λ_1 and arrival angle β_1 to obtain the directionally averaged cross section $\overline{\sigma_{\rm cap}}$. Note that we use an overbar to indicate the directional average, reserving $\langle\cdot\rangle$ for the average over speeds. This requires care, however: not all arrival directions are kinematically allowed for fixed v_1 and $v_{\rm esc}$, and it is difficult to analytically integrate only over parameters for which the expression of Eq. (2.22) is positive-definite. Explicitly, the directional average should be computed by an integral of the form

$$\overline{\sigma_{\rm cap}}(\mathbf{v}_1) = \int \frac{\mathrm{d}\lambda_1}{2\pi}\, \mathrm{d}\cos\beta_1\, \sigma_{\rm cap}(\mathbf{v}_1)\, \chi(\mathbf{v}_1, \lambda_1, \beta_1), \tag{2.23}$$

where χ is an indicator function equal to one when the arguments are kinematically allowed and zero otherwise. This average is readily carried out numerically, but χ is difficult to represent in closed form. However, for simplistic estimates, we can obtain an order-of-magnitude calculation of $\overline{\sigma_{\rm cap}}$ by integrating over all arrival directions, including non-physical directions. We denote this quantity by $\widetilde{\sigma_{\rm cap}}$, and it takes the form

$$\widetilde{\sigma_{\rm cap}} \equiv \pi \left(\frac{\mu_B}{v_1^2 - v_{\rm esc}^2}\right)^2 \left[-1 - \left(\frac{v_{\rm esc}^2 - v_B^2}{v_1^2 - v_B^2}\right)^2 + \frac{v_{\rm esc}^2 + v_B^2}{v_1 v_B} \operatorname{arctanh}\left(\frac{2 v_1 v_B}{v_1^2 + v_B^2}\right)\right]. \tag{2.24}$$

This is by no means a precise calculation, but the result is nonetheless quite useful, particularly for exhibiting the parametric dependence of the capture cross section on the binary configuration. The approximation breaks down most severely when v_∞ is so small that $v_1 \sim v_B$, but it is quite effective for larger values of v_∞. For the sun–Jupiter system, we find $\widetilde{\sigma_{\rm cap}} = 7.9 A_J$ for $v_\infty = 20\,\mathrm{km\,s^{-1}}$, where A_J is the cross-sectional area of Jupiter. Full numerical integration over kinematically allowed angles gives $\overline{\sigma_{\rm cap}} = 9.9 A_J$. To illustrate the applicability of this approximation, we compare the approximate and numerical results for several configurations in Table 2.2.

r_{AB}	M_B	v_∞ [km s^{-1}]	$\widetilde{\sigma_{\text{cap}}}$ [A_J]	$\overline{\sigma_{\text{cap}}}(\mathbf{v}_1)$ [A_J]
r_{SE}	M_E	46.28	2.78×10^{-6}	3.50×10^{-6}
r_{SJ}	M_J	20.23	7.133	9.074
r_{SN}	M_N	8.436	0.732	0.924
r_{SE}	M_J	46.15	0.263	0.335
r_{SJ}	M_N	20.28	2.19×10^{-2}	2.77×10^{-2}
r_{SN}	M_E	8.439	2.51×10^{-3}	3.16×10^{-3}
r_{SE}	M_N	46.26	8.10×10^{-4}	1.02×10^{-3}
r_{SJ}	M_E	20.29	7.52×10^{-5}	9.47×10^{-5}
r_{SN}	M_J	8.417	238	303

Table 2.2: Approximate and numerically averaged cross-sections for several configurations of object B. In each case, the velocity v_∞ of the incoming object is fixed such that $v_1 = \frac{1}{2}v_{\max}$, where $v_{\max}$ is the maximum velocity with non-zero capture cross section (see text). This is chosen only as a representative velocity for typical captures. The subscripts E, J, and N refer to Earth, Jupiter, and Neptune, respectively. For $X \in \{E, J, N\}$, M_X denotes the mass of the planet X, and r_{SX} denotes the distance between the Sun and the planet. The approximate cross section $\widetilde{\sigma_{\text{cap}}}$ slightly underestimates $\overline{\sigma_{\text{cap}}}$ by a consistent factor across configurations with widely varying parameters.

We can further directly obtain the differential cross section for a fixed specific energy transfer $\Delta\mathcal{E} \equiv \Delta E_C/M_C$. Since the potential energy is nearly the same immediately before and after the close encounter, we have $\Delta\mathcal{E} \approx \frac{1}{2}(v_2^2 - v_1^2)$, and thus we need only substitute $v_2(\mathcal{E}_2)$ for $v_{\rm esc}$ in Eq. (2.22). This gives the total cross section to final states with specific energy below $\mathcal{E}_2$. Differentiating the resulting expression with respect to $\mathcal{E}_2$, and writing $\mathcal{U}(\epsilon) = -\mu_A/r_{AB} - \mu_B/r_{\rm close}(\epsilon)$, we find

$$\frac{\mathrm{d}\sigma_{\rm cap}(\mathbf{v}_1)}{\mathrm{d}\mathcal{E}_2} = \frac{16\pi\mu_B^2\left[\mathcal{E}_2 + \mathcal{U}(\epsilon)\right]}{v_1'^4\left(v_1^2 - 2\left[\mathcal{E}_2 + \mathcal{U}(\epsilon)\right]\right)^3} \times \left[v_B^2\left(2v_1'^2 + \mathbf{v}_B\cdot\mathbf{v}_1'\right) + \left(v_1'^2 - 2\left[\mathcal{E}_2 + \mathcal{U}(\epsilon)\right]\right)\left(\mathbf{v}_B\cdot\mathbf{v}_1'\right)\right], \quad (2.25)$$

as long as $R(\mathbf{v}_1, \mathcal{E}_2) < r_{\rm close}(\epsilon)$. Otherwise, while the desired specific energy transfer may not be kinematically prohibited, it cannot be attained by a two-body encounter with the specified value of ϵ. We can approximate the directional average of this expression by starting instead with Eq. (2.24), which yields

$$\frac{\mathrm{d}\widetilde{\sigma_{\rm cap}}(v_1)}{\mathrm{d}\mathcal{E}_2} = \frac{\pi\mu_B^2}{(\mathcal{E}_2 - \mathcal{E}_1)^3}\left[1 + \frac{\mathcal{E}_2 - \mathcal{E}_1}{2[\mathcal{E}_1 - \mathcal{U}(\epsilon)] - v_B^2} - \frac{\mathcal{E}_1 + \mathcal{E}_2 - 2\mathcal{U}(\epsilon) + v_B^2}{2v_B\sqrt{2[\mathcal{E}_1 - \mathcal{U}(\epsilon)]}}\,\mathrm{arctanh}\left(\frac{2v_B\sqrt{2[\mathcal{E}_1 - \mathcal{U}(\epsilon)]}}{2[\mathcal{E}_1 - \mathcal{U}(\epsilon)] + v_B^2}\right)\right]. \quad (2.26)$$

Our computations thus far neglect the possibility of collisions with object B. In principle, it is possible that collisions also contribute to captures for compact objects such as light black holes. However, the relevant physics is quite different: energy is lost dissipatively by deformation of object B. For most cases of interest, the capture cross section is much larger than the collision cross section, but it is a simple matter to compute and subtract the latter if desired. The eccentricity e_1' and semimajor axis a_1'

of the two-body hyperbolic orbit in the frame of object B are given by

$$e_1' = \sqrt{1 + \frac{b^2 v_1'^4}{\mu_A^2}}, \qquad a_1' = -\frac{b}{\sqrt{e_1'^2 - 1}}. \tag{2.27}$$

Then the pericentre is given by $r_{\min} = a_1'(1 - e_1')$, or

$$r_{\min} = \frac{\sqrt{\mu_A^2 + b^2 v_1'^4} - \mu_A}{v_1'^2}. \tag{2.28}$$

Requiring $r_{\min} > r_B$, we obtain the condition

$$b > b_{\min} \equiv \frac{1}{v_1'} \sqrt{2\mu_B r_B + (r_B v_1')^2}. \tag{2.29}$$

The set of impact parameters leading to collisions is, of course, also a circle. We can now write the cross section for captures without including collisions by simply subtracting the area of intersection of the two circles from our prior result. This is given by

$$\sigma_{\text{int}} = -\left[\frac{1}{2}\left(-b_c + R + b_{\min}\right)\left(b_c + R - b_{\min}\right)\left(b_c - R + b_{\min}\right)\left(b_c + R + b_{\min}\right)\right]^{1/2} + R^2 \arccos\left(\frac{b_c^2 + R^2 - b_{\min}^2}{2 b_c R}\right) + r^2 \arccos\left(\frac{b_c^2 + b_{\min}^2 - R^2}{2 b_c b_{\min}}\right). \tag{2.30}$$

In general, σ_{int} can be subtracted from σ_{cap} to exclude collisions from the cross section. For our present purposes, we neglect the possibility of collisions altogether, so we do not carry out this subtraction in our subsequent results.

We can now use the capture cross section in Eq. (2.24) to estimate the capture rate of test particles with velocity v_∞ far from the binary system. First, however, it is necessary to convert $\sigma_{\text{cap}}(v_1)$ to the cross section $\sigma_{\text{cap}}(\mathbf{v}_\infty)$ pertinent to the rate calculation. The relationship between $v_1 \equiv \|\mathbf{v}_1\|$ and $v_\infty \equiv \|\mathbf{v}_\infty\|$ is specified by Eq. (2.13). But the arrival direction of object C at object B is also influenced by acceleration due

to object A, so the relationship between $\mathbf{v}_1$ and $\mathbf{v}_\infty$ has a non-trivial angular dependence. However, we expect this effect to have only a small impact on the directionally averaged cross-section: any modifications to λ_1 must disappear from the time-averaged cross section by azimuthal symmetry, so the sole effect of such deflection is to change the distribution of inclination angles β_1 of incoming objects. We are already treating this distribution crudely by integrating over non-physical arrival angles in Eq. (2.24), so we neglect this additional deflection, assuming that $\mathbf{v}_1 \propto \mathbf{v}_\infty$.

With this assumption, we can write $\sigma_{\rm cap}(\mathbf{v}_\infty) = \sigma_{\rm cap}(\mathbf{v}_1(\mathbf{v}_\infty))$. Now, given a distribution function $f(\mathbf{v}_\infty)$ for the velocity at infinity, the capture rate can be estimated as $n\,\langle\sigma_{\rm cap} v\rangle$, where n is the number density of objects and the velocity-averaged cross section is given by

$$\langle\sigma_{\rm cap} v\rangle = \int \mathrm{d}^3\mathbf{v}_\infty\, f(\mathbf{v}_\infty)\, \sigma_{\rm cap}(\mathbf{v}_\infty) v_1(v_\infty). \tag{2.31}$$

Note the appearance of v_1 in place of v_∞, accounting for the gravitational focusing factor v_1/v_∞.

This formalism also lends itself well to describing the orbital parameters of captured objects. Since we have obtained $\mathbf{v}_2$ explicitly as a function of the impact parameter, we can readily compute the specific orbital energy $\mathcal{E}$ and specific angular momentum $\mathcal{L}$ of the captured object as

$$\mathcal{E}_2 = \frac{1}{2}\mathbf{v}_2^2 + \mathcal{U}(\epsilon), \qquad \mathcal{L}_2 = \|\mathbf{r}_{AB} \times \mathbf{v}_2\|\,, \tag{2.32}$$

whereupon the eccentricity e and semimajor axis a of the captured object's orbit take the form

$$e = \sqrt{1 + \frac{2\mathcal{E}_2\mathcal{L}_2^2}{\mu_A^2}}, \qquad a = -\frac{\mu_A}{2\mathcal{E}_2}. \tag{2.33}$$

The resulting expressions are algebraically complicated but are nonetheless tractable, and in closed form. Obtaining the full distributions of orbital parameters is analytically challenging, but readily performed semi-analytically: uniformly sampled points in the (b, ϕ) plane can now be converted to orbital parameters. In particular, we can evaluate $\bar{e}$ and $\bar{a}$ by numerically integrating over initial configurations which produce captures, i.e., over the circle described by Eq. (2.20).

For an analytical estimate, we can translate Eq. (2.26) to an approximate differential cross section with respect to a, using

$$\frac{\mathrm{d}\widetilde{\sigma_{\mathrm{cap}}}(\mathbf{v}_1)}{\mathrm{d}a} = \frac{\mu_A}{2a^2}\frac{\mathrm{d}\widetilde{\sigma_{\mathrm{cap}}}(\mathbf{v}_1)}{\mathrm{d}\mathcal{E}_2}, \tag{2.34}$$

and thus obtain a probability distribution for a as a function of $\mathbf{v}_1$. The binary is assumed to be circular, with fixed separation r_{AB}, and the captured orbit must cross the trajectory of object B, so we impose a lower cutoff $a > r_{AB}$. The resulting distribution is sharply peaked at small a, but does not have a well-defined mean. For comparison with numerical results, it suffices to evaluate $\bar{a}$ considering only captured orbits with $a < a_{\mathrm{max}}$. We denote this approximate mean by $\tilde{a}$. For instance, for the sun–Jupiter system with $v_\infty = 20\,\mathrm{km\,s^{-1}}$, taking $a_{\mathrm{max}} = 120\,\mathrm{au}$ gives $\tilde{a} = 15.5\,\mathrm{au}$. This result is comparable to that described in fig. 5 of Ref. [79], although note that the latter gives an approximate result computed only for a fixed value of β_1. Alternatively, one can compute the median value of a, which is analytically challenging but readily performed numerically. For the aforementioned Solar system configuration, we estimate the median semimajor axis of captured objects at 13.7 au. The distribution of Eq. (2.34) is also in excellent agreement with numerical experiments, as we shall see in Section 2.3.3.

Estimating the eccentricity after capture is substantially more complicated, since the specific angular momentum is independent of the specific energy after capture. There is no obvious geometric structure to the final angular momentum, in contrast to the circular regions we have identified for the final energy, and in general, the average over arrival angles must be performed numerically. However, we can exploit the semimajor axis distribution to make a simplistic estimate, as follows. Generally $\bar{a} > r_{AB}$, but the orbit of object C after capture must cross the orbit of object B. Thus, given a value of a, there is a minimal eccentricity $e_{\min}(a)$ needed to ensure that the perihelion of C lies within the orbit of B, i.e., $a(1-e) < r_{AB}$. Saturating this condition gives the lowest possible eccentricity for a capture with a given value of the semimajor axis. In general, highly eccentric captures are possible at the extremes of the parameter space. Thus, for a first estimate of the orbital parameter distribution, we assume that eccentricity is uniformly distributed on $(e_{\min}(a), 1)$ for fixed a. That is, we take

$$\frac{\mathrm{d}^2\widetilde{\sigma_{\mathrm{cap}}}(\mathbf{v}_1)}{\mathrm{d}a\,\mathrm{d}e} = \frac{\mathrm{d}\widetilde{\sigma_{\mathrm{cap}}}(\mathbf{v}_1)}{\mathrm{d}a}\frac{\Theta(1-e)\,\Theta(e-e_{\min})}{1-e_{\min}}, \tag{2.35}$$

where Θ is the Heaviside function. While crude, this is in reasonably good agreement with eccentricities extracted from numerical experiments, as we shall demonstrate in Section 2.3.3. We define a typical eccentricity $\tilde{e}(a)$ as the mean of the corresponding uniform distribution at fixed a, i.e., $\tilde{e}(a) = \frac{1}{2}(1 + e_{\min})$.

We now pause to compare our results to those of Ref. [79] more generally. Figure 4 of that reference shows impact parameters leading to capture for several values of the orbital phase λ_1, similar to our Fig. 2.4. While the shape and position of each capture region is generally comparable to the circular region of Eq. (2.20), there is clear

distortion away from a circular shape. This is presumably due to one or both of two effects. One is our neglect of angular deflection between $\mathbf{v}_\infty$ and $\mathbf{v}_1$, but another is the definition of the impact parameter—and while the consequences for the capture rate are ultimately insignificant at the order-of-magnitude level, it is nonetheless important to understand the distinction between the two definitions.

Our formalism relies on the premise that the close encounter between objects B and C can be treated as a two-body encounter. Thus, working in the frame of object B, there is a natural definition of the impact parameter, which we temporarily denote by $\mathbf{b}'$: it is simply the vector of closest approach between B and the ray $\mathbf{x}'_C\big|_{t=0} + \mathbf{v}'_1 t$ over all t. This is equivalent to the vector of closest approach between B and C in the absence of any interaction. The vector $\mathbf{b}'$ is orthogonal to $\mathbf{v}'_1$, but notice that it is *not* orthogonal to $\mathbf{v}_1$, the initial velocity in the frame of object A. The impact parameter in the frame of A has a different meaning. Indeed, in general, the magnitude of the impact parameter, as defined via the closest approach of the initial velocity ray to the second object, is only invariant between frames in which the initial velocities of B and C are parallel. The frame of B is of course such a frame, but the frame of A is generally not.

This means that any statements involving the impact parameter require us to specify the choice of frame. For our purposes, there are two relevant statements with such a dependence. One statement is the relationship of Eq. (2.14) between the impact parameter and the deflection angle Ψ. This is formulated in the two-body problem, where the impact parameter is specified in a frame where the velocities are parallel. Thus, for calculation of the deflection angle, we must use the impact parameter $\mathbf{b}'$, as calculated in the frame of B, and not its equivalent in frame A. The other statement

concerns the relationship of the impact parameter to the cross section. Ultimately, the set of impact parameters that result in capture forms a region in the plane orthogonal to velocity whose area is the capture cross section. While the total cross section is the same between the frames of A and B, the impact parameters are not, and thus, the shape of the capture region must transform in a complicated way to compensate.

We have checked that defining the impact parameter in the frame of A produces regions in the impact parameter plane that more closely resemble the non-circular shapes of Ref. [79]. In Section 2.3.3, we numerically validate our analytical prescription, and show that the capture regions are indeed circular under our stated assumptions and conventions.

2.3.2 Estimating the ejection rate

In two-body dynamics, a pair of gravitationally bound objects remain bound forever. This is not the case in a three-body system for exactly the same reason that capture of the third body is possible: since the system is time-reversal invariant, the same process can take place in the opposite direction. A close approach between two bodies in a three-body bound system can transfer energy between them and lead to ejection of one of the two bodies from the system.

Unfortunately, estimating the rate of ejection from first principles is very challenging. As Ref. [85] explains, the complicated dynamics of the three-body system mean that the orbital configurations are constantly changing in an unpredictable fashion. The most reliable estimates of ejection time-scales come from direct numerical simulation of such systems, and even these are difficult to execute reliably over the potentially long

time-scales involved. However, short of such a calculation, it is nonetheless useful to have an order-of-magnitude estimate of the lifetime of bound orbits under particular conditions. In the present context, our interest lies in estimating the statistics of the population of captured particles across a variety of systems *without* expensive simulations, so it is useful to at least understand the basic dependence of the ejection rate on binary parameters.

In practice, ejection time-scales are often estimated using simplified Monte Carlo algorithms based on Öpik theory [95–98] instead of full numerical simulations, an approach known as the Öpik–Arnold method [99]. In our framework, since we can estimate the relevant cross-sections analytically, we can perform a semi-analytical analogue of the Öpik theory estimate without any actual simulation. Since this approach is fundamentally rooted in the same approach as Öpik–Arnold codes, we first review the typical algorithmic method.

The Öpik–Arnold estimate of the ejection rate relies on the assumption that the ejection process is driven by close encounters. The problem can then be decomposed into two parts: (1) determining the rate of close encounters, and (2) determining the outcome of each close encounter as it affects the orbital parameters of the captured object. Ref. [95] estimates the time-scale between close encounters as a function of the orbital parameters of both objects, providing a solution to the first part of the problem. The second part can be approached iteratively via a Monte Carlo algorithm, randomly choosing an impact parameter for each close encounter and determining the new set of orbital parameters. While the algorithmic estimate is not in perfect agreement with numerical integration, it is capable of giving an inexpensive order-of-magnitude estimate

of the ejection time-scale (see Ref. [87] for an extensive discussion).

However, despite the simplicity of the Öpik–Arnold algorithm, it is inherently stochastic and iterative. This makes it difficult to produce straightforward analytical estimates of the ejection time-scale without a computational implementation. Thus, the primary advantage of the algorithmic approach is that it is much faster and simpler to implement than full numerical integration. For our purposes, however, we would like to have an order-of-magnitude estimate of the ejection rate that can be written in closed form, or at least evaluated semi-analytically. Our explicit algebraic results derived in the previous subsection make such a simplistic estimate possible, under the following assumptions:

1. ejection of object C is driven by close encounters, and

2. close encounters take place mainly with object B.

Note that since the initial orbital parameters of object C are determined during a close encounter with object B, its initial orbit includes the point of the close encounter. It follows that the orbit of object C crosses the orbit of object B, at least initially, justifying our second assumption.

There are now two strategies one could use to estimate the ejection time-scale. The first is to follow essentially the same strategy as the Öpik–Arnold algorithm, but to use semi-analytical averages rather than iterative Monte Carlo computations. In particular, in the limit that there is a large number N of close encounters prior to ejection, the specific energy transfer $\Delta\mathcal{E}$ can be treated differentially, writing $\mathrm{d}\mathcal{E}_C/\mathrm{d}N = \langle\Delta\mathcal{E}\rangle|_{\mathcal{E}_C}$. In principle, using the differential cross section in Eq. (2.26), one can explicitly evaluate

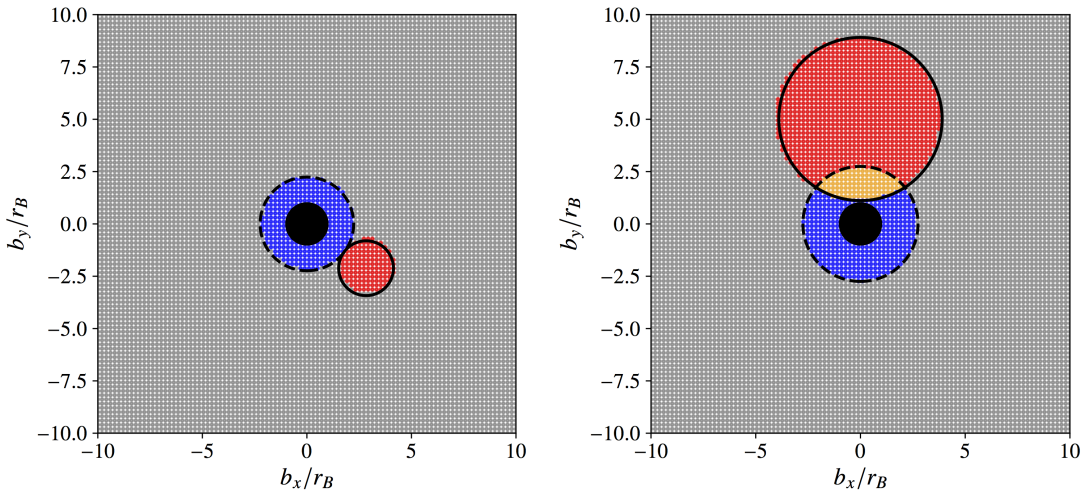

Figure 2.4: Outcomes of close encounters with Jupiter as determined by numerical integration, with $v_\infty = 20\,\mathrm{km/s}$, $\beta_1 = \pi/3$, and two values of λ_1: in the left panel, $\lambda_1 = 0$, and in the right panel, $\lambda_1 = \pi/2$. Each point represents an independent simulation with a different impact parameter. Points are shown in the plane of the impact parameter orthogonal to the velocity $\mathbf{v}_1'$, i.e., from the perspective of object C in the frame of object B. The angular coordinate is fixed by the prescription in Section 2.3.1. Red points indicate capture of object C, and gray points indicate that object C was unbound after departing from the close-encounter region. Orange points indicate that object C would have been captured if Jupiter were replaced by a point mass, but instead suffered a collision. Blue points indicate collisions that would not yield captures even if Jupiter were compressed to a point. The solid circle shows the analytical prediction of the capture region in Eq. (2.20), and the dashed circle shows the prediction of the collision region in Eq. (2.29). Compare with fig. 4 of Ref. [79]. Note that in the bottom panel, the red points are shifted very slightly to the left of the analytical prediction. This shift is in the direction of the sun and signals the presence of tidal forces.

$\langle \Delta \mathcal{E} \rangle|_{\mathcal{E}_C}$, integrate this separable differential equation, and then solve $\mathcal{E}_C(N_{\rm ej}) = 0$ to determine the number $N_{\rm ej}$ of close encounters required to produce an ejection event. Once $\mathcal{E}_C(N)$ is obtained in closed form, one can approximate the time-scale between close encounters as a function of $\mathcal{E}_C$, and integrate on $N \in (0, N_{\rm ej})$ to finally estimate the ejection time-scale.

While certainly possible numerically, this process is algebraically formidable, and thus offers no great advantage over the Öpik–Arnold treatment for an order-of-magnitude estimate. We therefore choose radical acceptance of our limitations, and propose an alternative method for an even simpler estimate of the ejection time-scale. While the orbital parameters of object C certainly change significantly over the life-time of the bound configuration, we make the following assumptions in addition to the previous two:

3. most close encounters do not substantially change the *ejection cross section* in subsequent orbits, and

4. most encounters at distance r_1 do not substantially change the time between subsequent close encounters at distances $r \ll r_1$.

We caution that these assumptions are almost certainly flawed in most cases, but they may nonetheless suffice for a very simplistic parametric estimate.

The value of these approximations, on the other hand, is significant: taken together, they imply that we may ignore all close encounters except those which lead directly to ejection. Given the ejection cross section, we can then use the same Öpik formalism to estimate the rate of such close encounters, and thus produce an estimate

of the ejection rate. In principle, neglecting distant encounters is not all that different from what is typically done in Öpik–Arnold codes, which themselves neglect encounters falling beyond the influence radius of object B: implementations of the algorithm often include an enhancement factor alongside the cross section of the sphere of influence to account for the aggregate effects of such distant encounters. We do the same to a somewhat greater extent, as we will detail shortly.

Now all that remains is to compute the ejection cross section σ_{ej}. Fortunately, this much is easy in our formalism. The ejection cross section is simply the cross section for a close encounter with object B in which the energy exchange is large enough that object C becomes unbound, but apart from the amount of energy to be transferred, this is identical to the capture cross section, and we can thus use the same technology to compute the ejection cross section. In particular, Eq. (2.24) holds in identical form, with v_1 replaced by v_2, the speed of object C immediately after the close encounter leading to capture.

To implement this calculation, we follow the Öpik-theory estimate of the close encounter time-scale as presented by Ref. [87]. With non-canonical units restored, the close encounter rate is given by

$$\frac{\mathrm{d}N}{\mathrm{d}t} = \left(\frac{v_B\sqrt{r_{AB}}}{2\pi}\right)\frac{KW\tau^2}{\pi S W_x r_{AB}^2 a^{3/2}}. \tag{2.36}$$

Here τ is the length associated with the encounter cross section, i.e., $\sigma = \pi\tau^2$; a is the semimajor axis of object C; W is the approach speed, analogous to v_1 in the capture case; W_x is the component of $\mathbf{W}$ parallel to $\mathbf{r}_{AB}$; and K is the enhancement factor to the cross section mentioned previously, whose value we will address shortly. We determine

W and W_x following Ref. [87],[1] and we likewise set $S = \max(\sin i,\ \tau/r_{AB})$, where i is the orbital inclination of object C.

We assume that the orbital parameters of object C change rapidly enough on the time-scales of ejection that we may average over i. The average can be performed explicitly in terms of elliptic integrals, and since we may safely assume that $\sigma \ll r_{AB}^2$, the result simplifies to

$$\widetilde{R_{\rm ej}} \simeq \frac{K v_B^2 \widetilde{\sigma_{\rm ej}}}{2\pi^{5/2} r_{AB}^{3/2} a^{3/2} \sqrt{W_x}} \times \left\{ 2\sqrt{\xi} - \kappa_- \arctan\left(\frac{\sqrt{\xi}}{\kappa_-}\right) - \kappa_+ \arctan\left(\frac{\sqrt{\xi}}{\kappa_+}\right) + i\left[\kappa_- \operatorname{arctanh}\left(\frac{\kappa_+}{\kappa_-}\right) + \kappa_+ \operatorname{arctanh}\left(1 + \frac{\eta\widetilde{\sigma_{\rm ej}}}{2\pi r_{AB}^2 \kappa_+^2}\right) - 2i\kappa_+\right]\right\}, \quad (2.37)$$

where for brevity we define

$$\eta = \sqrt{a(1-e^2)/r_{AB}}, \quad \xi = 3 - r_{AB}/a, \quad \kappa_\pm = \sqrt{-\xi \pm 2\eta}. \quad (2.38)$$

The ejection cross section can be written explicitly as

$$\widetilde{\sigma_{\rm ej}} = \pi \left(\frac{2M_B r_{AB}}{5M_A}\right)^2 \left[-1 - \left(\frac{v_B^2 r_{AB} - 2\mu_A}{2v_B^2 r_{AB} + \mu_A}\right)^2 - \frac{v_B^2 r_{AB} + 2\mu_A}{v_B\sqrt{\mu_A r_{AB}/2}} \arctan\left(\frac{2v_B\sqrt{2\mu_A r_{AB}}}{\mu_A - 2v_B^2 r_{AB}}\right)\right]. \quad (2.39)$$

Taken together, Eqs. (2.37) to (2.39) allow for an analytical estimate of the ejection rate. We can certainly average the ejection rate over a and e values using the joint distribution of Eq. (2.35). However, by simply substituting $\tilde{a}$ and $\tilde{e}(\tilde{a})$ for a and e, we obtain a crude but closed-form estimate for the typical lifetime of a captured orbit in a given binary system.

This estimate should be understood as an estimate of the mean of some distribution of lifetimes of captured orbits. The shape of this distribution reflects our

[1] Note that Ref. [87] denote our W and W_x by U and U_x. We use W to avoid confusion with $\mathcal{U}(\epsilon)$.

assumption that close encounters can be treated as a Poisson process: if this were exactly true, the distribution of lifetimes T would be exponential, with the probability distribution $f(T) = R_{\rm ej} \exp(-R_{\rm ej} T)$. This is potentially complicated by the effects of other close encounters: in principle, as in the Öpik–Arnold approach, the trajectory of a typical capture is influenced by several other close encounters before the one which leads directly to ejection. If ejection is modelled as the cumulative outcome of some N close encounters, each of which takes place with a comparable time-scale T_1, then the lifetime is distributed as a sum of N exponentially distributed random variables, i.e., according to the Erlang distribution $E(N, T_1^{-1})$. Thus, the shape of the lifetime distribution is a key test of our simplistic ejection model: an exponential distribution is compatible with our assumptions, while a more general Erlang distribution signals the non-trivial involvement of multiple close encounters. In Section 2.3.3, we will see that the distribution of lifetimes in numerical experiments is well-fit by an exponential distribution, justifying the assumptions of this subsection.

With a complete estimate in hand, we can now compare to numerical benchmarks to estimate an appropriate value for K. We will carry this out in detail in Section 2.3.3, but for the moment, we note that $K \sim 25$ is appropriate for order-of-magnitude estimates. As expected, this is somewhat larger than the value $K \sim 10$ preferred by Öpik–Arnold codes to account for encounters lying beyond the influence radius.

Having developed a set of analytical approximations for the rates of capture and ejection, we now turn to the properties of the equilibrium population: in the limit of long times, what is the expected number of captured objects bound to object A? In

equilibrium, the ejection rate balances the capture rate. Now, if the captured objects do not interact among themselves, then the capture rate is independent of the number of captured objects, while the ejection rate is proportional thereto. Thus,

$$\overline{N} = R_{\rm cap}/R_{\rm ej}. \tag{2.40}$$

We can thus estimate $\overline{N}$ by $\widetilde{N} \equiv \widetilde{R_{\rm cap}}/\widetilde{R_{\rm ej}}$ for fixed v_∞. If the population of free objects interacting with the binary has a distribution $f(v_\infty)$, then we can average over the population and write

$$\langle \widetilde{N} \rangle = n_\infty \int \mathrm{d}v_\infty \, f(v_\infty) \frac{\widetilde{\sigma_{\rm cap}}(v_\infty) v_1(v_\infty)}{\widetilde{R_{\rm ej}}(v_\infty)}, \tag{2.41}$$

where n_∞ is the number density far from the binary. In general, this integral must be performed numerically. Nonetheless, this procedure allows for a rapid order-of-magnitude estimate of the equilibrium number of captured objects.

To demonstrate, we apply this method to the capture of particle dark matter with no non-gravitational interactions. This scenario has been studied extensively for the case of the solar system [88–90, 92], so we likewise make an estimate for the sun–Jupiter system. We can make a simple semi-analytical estimate using an isotropic Boltzmann distribution for $f(v_\infty)$, i.e., neglecting the dark matter wind. Such a distribution has the form $f(v_\infty) \propto v_\infty^2 \, \mathrm{e}^{-v_\infty^2/v_0^2}$, so that $f(v_\infty) \sim v_\infty^2/v_0^3$ at low velocities, with an exponential cutoff for $v_\infty \gtrsim v_0$. Note that v_0 for the local dark matter distribution is much larger than the orbital speed of Jupiter, so the low-velocity tail dominates the capture rate. We can numerically evaluate Eq. (2.41), taking $\widetilde{\sigma_{\rm cap}}$ from Eq. (2.24), $v_1(v_\infty)$ from Eq. (2.13), and $\widetilde{R_{\rm ej}}$ from Eq. (2.37). Taking an rms velocity of $220\,\mathrm{km\,s^{-1}}$ for the dark matter particles, we find $\langle \widetilde{N} \rangle \simeq (0.1\,\mathrm{au}^3) n_\infty$. Compared to the number den-

sity n_∞ in the spherical volume within Jupiter's orbit, this corresponds to an $\mathcal{O}(10^{-4})$ enhancement. This is reasonably consistent with detailed simulations by Ref. [92], who finds that the density enhancement at Earth is sub-per cent.

2.3.3 Comparison with numerical integration

In the previous subsection, we obtained analytical results for the capture cross section, and semi-analytical results for the distribution of orbital parameters. These results are only reliable within the context of the stated approximations, and it is thus important to compare them with numerical results to be assured of their validity in the regimes of interest. We will begin our numerical analyses with the sun–Jupiter system, since this system has been extensively studied by prior authors, and thus serves as a well-understood benchmark.

We numerically integrate the equations of motion using the MERCURIUS integrator [100] via the publicly-available REIN:2011UW code [101, 102]. In each simulation, we configure the three bodies A, B, and C according to fixed values of λ_1, β_1, and v_1. We set the initial position of object C in the frame of object B, offset by a vector of length $r_{\rm close}(\epsilon = 0.1)$ in the direction of $-\mathbf{v}_1'$ and by an orthogonal vector $\mathbf{b}$. In the following, we shall describe $\mathbf{b}$ as a 2d vector in the plane orthogonal to $\mathbf{v}_1'$. We always fix v_∞ and derive v_1' from Eq. (2.13) to avoid unphysical speeds.

We begin with the dynamics of captures. Our first goal is to confirm our statements regarding the *shape* of the capture region in the plane of the impact parameter $\mathbf{b}$. To that end, we configure simulations with varying impact parameter $\mathbf{b}$, and for each such configuration, we test whether C becomes bound to the sun before

leaving the close-encounter region. We diagnose a capture trajectory as one for which object C is initially free, i.e., $\mathcal{E}_1 > 0$, and for which object C becomes bound to object A at some later time, i.e., $\frac{1}{2}v_C^2 - \mu_A/r_{AC} < 0$ in the frame of object A. Figure 2.4 shows the results of our numerical simulations for the same parameters used in fig. 4 of Ref. [79], demonstrating excellent agreement with our analytical predictions. Note that the impact parameter used in Fig. 2.4 is defined as in Section 2.3.1. As a benchmark, the capture cross section for objects with $v_\infty = 20\,\mathrm{km\,s^{-1}}$ and inclination $\beta_1 = \pi/3$ is $4.8A_J$, where A_J is the cross-sectional area of Jupiter. This agrees with the result of Ref. [79], who finds this cross section to be "roughly five times the area of Jupiter."

We compare analytical predictions of the orbital parameter distributions to numerical results in Fig. 2.5. The analytical semimajor axis distribution is in good agreement with numerical results. Our estimate of the eccentricity distribution is very crude, based only on heuristic arguments, but it nonetheless traces the essential behavior of the numerical results. We stress that these orbital parameters are not time-invariant, but evolve after the capture. This is a key difference between two-body and three-body dynamics. Figure 2.5 shows the orbital parameters only immediately after capture.

Finally, we test our prediction of the ejection time-scale against numerical integration. For the sun–Jupiter system with $v_\infty = 20\,\mathrm{km/s}$, our prescription estimates the typical ejection time-scale at $1/\widetilde{R_{\mathrm{ej}}} = (3.0 \times 10^7\,\mathrm{yr})/K$. We determine the mean ejection time-scale numerically by integrating an ensemble of initial conditions, randomly sampled with isotropic arrival directions and with impact parameters sampled uniformly in the plane orthogonal to $\mathbf{v}_1'$. As in Fig. 2.4, we include impact parameters that lie outside the capture region according to our analytical prediction, but we discard

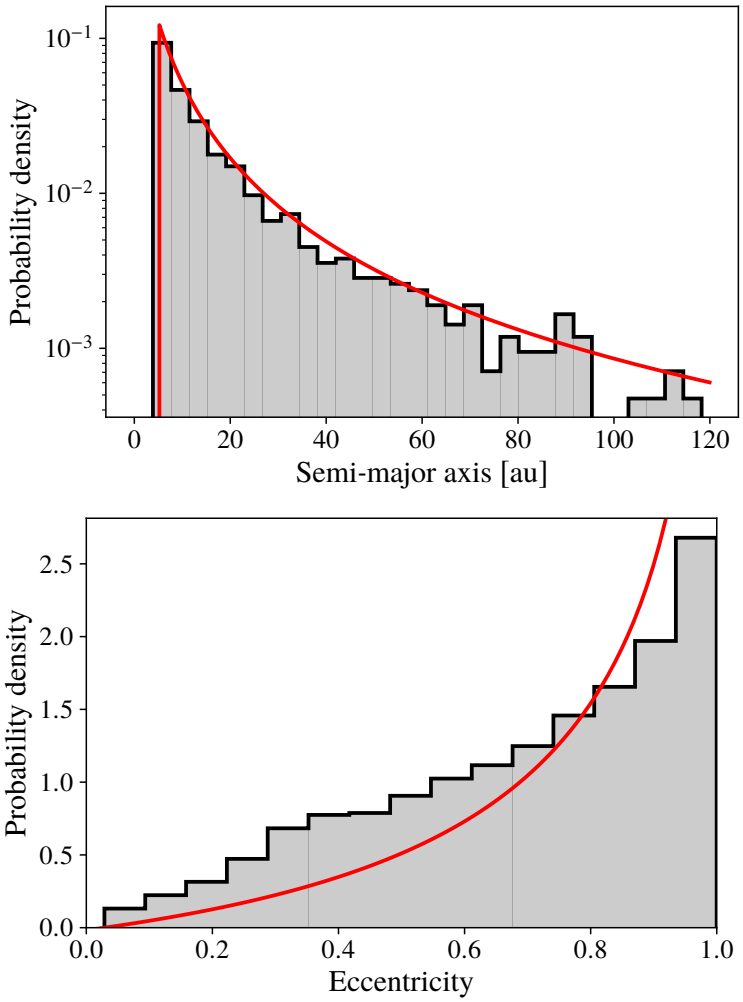

Figure 2.5: Top: distribution of semimajor axes immediately after capture by the sun–Jupiter system for $v_\infty = 20\,\mathrm{km/s}$. The histogram shows the distribution extracted from an ensemble of simulations (see text for details). The red line shows the prediction of Eq. (2.34). Bottom: distribution of eccentricities. The solid red line shows the prediction of Eq. (2.35), marginalizing over a.

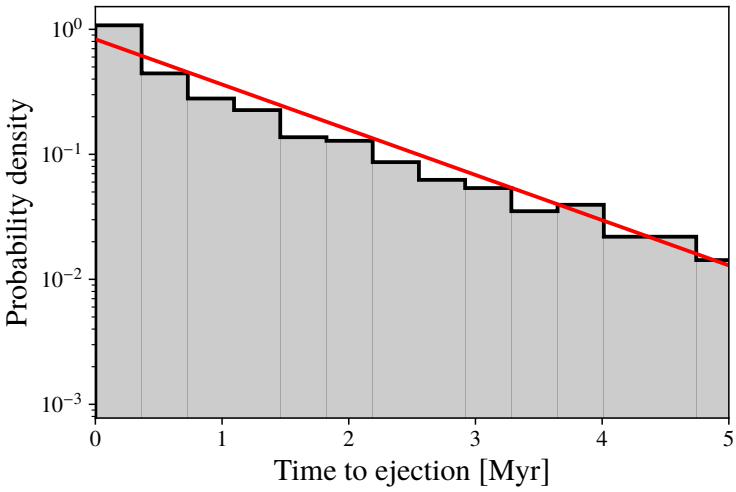

Figure 2.6: Distribution of capture lifetimes (i.e., time from capture to ejection) in an ensemble of 2500 simulations in the sun–Jupiter system with $v_\infty = 20\,\mathrm{km/s}$. The red curve shows an exponential distribution with the estimated ejection rate of Eq. (2.37) ($K = 25$).

all configurations which do not result in capture of object C. We integrate forward in time until object C is ejected. This ensemble of simulations gives the mean ejection time-scale as $\bar{t}_{\mathrm{ej}} \simeq 1.2 \times 10^6\,\mathrm{yr}$, suggesting $K \sim 25$, as noted in Section 2.3.2. A very small number of initial conditions lead to long-lived captures that are not ejected within the running time of our simulations, and the impact of these points in our subsequent analysis is negligible.

It is certainly encouraging that our analytical estimate can reproduce numerical results with a value of K only an $\mathcal{O}(1)$ factor larger than that used in Öpik–Arnold codes. A larger value of K is expected, of course—our analytical estimate neglects contributions from a larger set of close encounters than are neglected in the Öpik–Arnold approach. Nonetheless, a dramatically larger value of K would signal the failure of our method to account for most of the dynamics relevant to ejection. Moreover, we verify in Fig. 2.6

that our estimated ejection rate, interpreted as the rate of an exponential distribution, produces a good fit to the entire distribution of lifetimes extracted from simulations. As discussed in Section 2.3.2, if the dynamics of ejection were not dominated by a single close encounter, we would expect a more general Erlang distribution rather than the simple exponential distribution seen here.

However, our main goal is to produce an estimate of the ejection time-scale that remains valid across a wide variety of systems. Thus, the real test of our result is the extent to which a fixed value of K can be used to obtain an order-of-magnitude estimate of the ejection rate not only in the sun–Jupiter system, but in binaries with different mass ratios and semimajor axes. Indeed, even in the sun–Jupiter system, a single value of K must be sufficient to predict the ejection rate for objects captured with many values of v_∞.

We thus vary these parameters and compare the outcomes of numerical simulations with the analytical prediction, with the results shown in Fig. 2.7. Some of the behavior in these results is easy to understand: in particular, the M_B dependence can be estimated by the impact on the ejection cross section. Naïvely, increasing the mass M_B of the companion increases the ejection cross section as $\sigma_{\rm ej} \sim M_B^2$, comparably decreasing the ejection time-scale. On the other hand, the dependence of the analytical estimate on r_{AB} and v_∞ is much more complicated. Note that even the M_B dependence is not as straightforward as our heuristic argument would suggest, because the orbital parameter distribution of captured objects also has non-trivial M_B dependence. Thus, even for this case, we must rely on the numerical results to benchmark the analytical calculation. Figure 2.7 shows that Eq. (2.37) provides an excellent order-of-magnitude

estimate of the ejection time-scale, generally lying within a factor of 2 of the numerical mean.

Finally, we note that for some parameter values, the lifetime distribution is sensitive to the approximations that we make in deriving the orbital parameter distributions. In particular, for small values of v_∞, our formalism can fail to accurately predict the distribution of semimajor axes after capture, resulting in disagreement between the analytical result and simulation outputs (see Fig. 2.7, bottom panel). This is to be expected due to tidal forces. Our approach assumes that the capture is driven by a close encounter, i.e., $\min r_{BC} \lesssim r_{\rm close}(\epsilon)$ (see Section 2.3.1). But for small values of v_∞, the capture cross section becomes large, and in particular, it is possible that $\sqrt{\sigma_{\rm cap}} \gtrsim r_{\rm close}(\epsilon)$. In this case, the close-encounter condition is not satisfied for all impact parameters leading to capture, and our estimate of the orbital parameter distributions breaks down. A similar condition is produced by taking small values of r_{AB}, which causes $r_{\rm close}(\epsilon)$ to shrink.

2.3.4 Compact object capture in different systems

In the previous subsections, we have outlined a simplified calculation of the capture and ejection rates, and in particular, we have arrived at a relatively simple estimate of the equilibrium population of captured objects. We now consider the classes of systems which are most and least effective at capturing and retaining PBHs.

The equilibrium number of captured objects is shown in Fig. 2.8 as a function of M_A and R_{AB}, for two fixed values of M_B. In each panel, it is assumed that all of the DM is in the form of PBHs with mass $10^{-13}\,{\rm M}_\odot$. As long as the PBH mass is

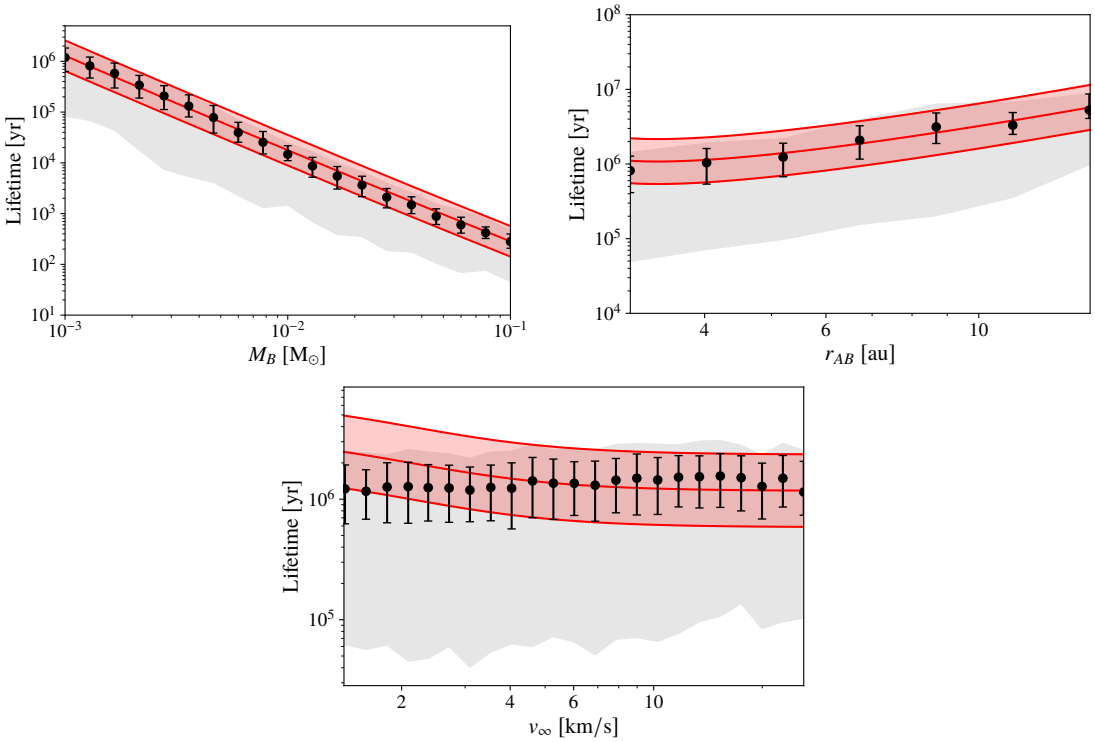

Figure 2.7: Mean ejection timescale for captured objects as predicted by Eq. (2.37) (red) and in an ensemble of numerical simulations (black). Error bars show $\pm 1\sigma$ bootstrap confidence intervals. The gray regions show $\pm 1\sigma$ quantiles for the lifetime distribution at each point. In each panel, one parameter is varied with respect to the base configuration, consisting of the sun–Jupiter system with $v_\infty = 20\,\mathrm{km\,s^{-1}}$. The top-left panel varies the companion mass M_B, the top-right panel varies the binary radius r_{AB}, and the bottom panel varies the initial speed v_∞ of object C prior to capture. The analytical prediction is shown for three values of K: 50, 25, and 12.5 from bottom to top. Each black point shows the mean time to ejection after capture in an ensemble of simulations with randomized initial configurations. Note that at very small values of v_∞, and potentially r_{AB}, our prediction becomes unreliable (see text).

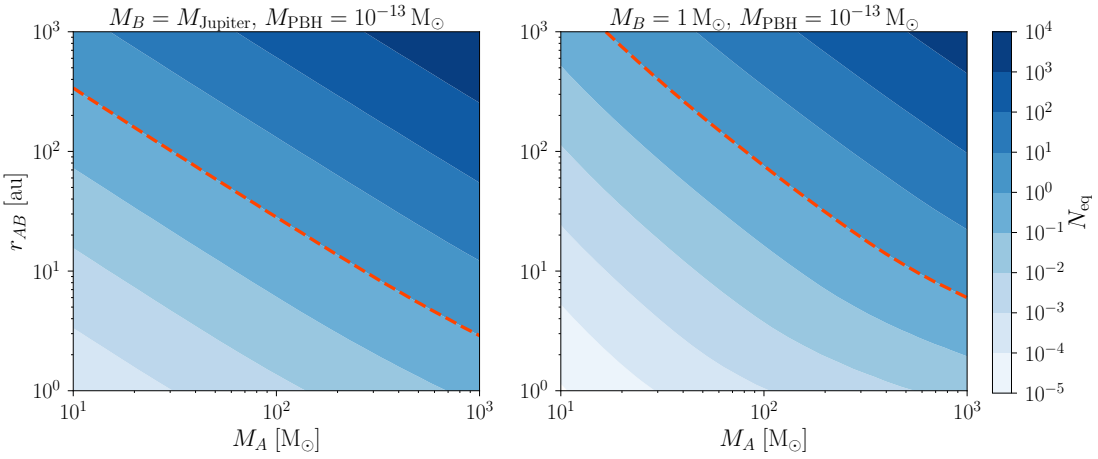

Figure 2.8: Equilibrium number of captured PBHs as estimated in Eq. (2.41), assuming all DM is in the form of $10^{-13}\,{\rm M}_\odot$ PBHs, shown as a function of the mass of the heavier object in the binary (M_A) and the binary separation (R_{AB}). Note that the PBH number density is inversely proportional to $M_{\rm PBH}$, so both panels can be recast to other PBH masses by multiplication by $10^{-13}\,{\rm M}_\odot/M_{\rm PBH}$. The left and right panels fix $M_B = M_{\rm Jupiter}$ and $M_B = 1\,{\rm M}_\odot$, respectively. In each panel, the dashed red curve indicates $N_{\rm eq} = 1$. Note that $N_{\rm eq} \propto M_{\rm PBH}^{-1}$ in this regime. The nearly power-law structure of the equilibrium number and the weak dependence on M_B can both be extracted from analytical arguments (see text).

well below the masses of the objects in the binary, the equilibrium number captured scales with their ambient number density, and thus, fixing the DM density, this means that $N_{\rm eq} \propto M_{\rm PBH}^{-1}$. The equilibrium number captured increases nearly as a power law with the orbital separation of the binary and with the mass of the heavier object in the system, but is only weakly dependent on the mass of the lighter object. We will explain this behavior shortly. For the moment, we note that with all of DM in the form of $10^{-13}\,{\rm M}_\odot$ PBHs, massive wide binary systems ($M_A \gtrsim 10^2\,{\rm M}_\odot$, $R_{AB} \gtrsim 100\,{\rm au}$) would *typically* host a large number of captured objects.

To understand the features of Fig. 2.8, we consider a simpler version of the capture cross section. The capture cross section of Eq. (2.24) reflects the average over incoming directions. At the order-of-magnitude level, it is analytically simpler to choose a particular orbital phase and inclination angle and evaluate the capture cross section for varying binary parameters. We choose the single configuration that maximizes the energy loss, and thus the capture cross section, for high-velocity compact objects. To identify this configuration, consider the kinematics of three-body captures. When the PBH has a close encounter with object B, they can be treated as a two-body system, and in particular the speed of recession of the PBH is equal to the speed of approach in the frame of object B. Thus, the maximum energy loss takes place when the direction of the PBH is reversed by the encounter in the frame of object A. This is only possible when the PBH velocity before the encounter, $\mathbf{v}_1$, is parallel to $\mathbf{v}_B$, and the velocity after the encounter, $\mathbf{v}_2$, is antiparallel to $\mathbf{v}_B$. Taking this directional configuration for $\mathbf{v}_1$ and $\mathbf{v}_B$ corresponds to fixing $(\lambda_1, \beta_1) = (\pi/2, 0)$. This provides a useful reference point for comparison between different systems. The resulting capture cross section

takes a relatively simple form:

$$\sigma_{\rm cap}(v_\infty) = \pi \left(\frac{M_B}{M_A}\right)^2 R_{AB}^2 \times \frac{1 + 2\zeta^2 \left[1 - \left(2 + 2/\zeta^2\right)^{1/2}\right]}{\left[1 - (2 + 2/\zeta^2)^{1/2}\right]^4}\,, \qquad \zeta = \frac{v_{\rm esc}}{v_\infty}. \tag{2.42}$$

This expression can be simplified further in the regime relevant for captures: $\sigma_{\rm cap}$ is maximized at $\zeta \approx 0.37$, and drops sharply for higher values of ζ, so the capture rate is dominated by objects with $\zeta \ll 1$. In this limit, the cross section simplifies to $\sigma_{\rm cap} \simeq \frac{\pi}{4}(M_B/M_A)^2 R_{AB}^2 \zeta^4$. Further, the DM velocity distribution in Eq. (2.41) can be considerably simplified for realistic systems. The capture rate is dominated by the peak in the cross section at $\zeta \approx 0.37$, corresponding to $\mathcal{O}(1)$ values of $v_\infty/v_{\rm esc}$. In turn, typical values of $v_{\rm esc}$ are on the order of 10 km/s, far below $v_0 \approx 220$ km/s. Thus, captures should be dominated by the low-velocity tail of the PBH velocity distribution, which has the form

$$f(v_\infty) \simeq \frac{4 v_\infty^2}{\sqrt{\pi} v_0^3} \qquad (v_\infty \ll v_0). \tag{2.43}$$

Finally, we fix the semimajor axis and eccentricity of the captured PBH's orbit to representative values $a = 3R_{AB}$ and $e = 1 - R_{AB}/(2a) = 5/6$. Together with Eqs. (2.41) and (2.43), this enables a rapid estimate of the equilibrium population of captured objects in a wide variety of systems. Taking $M_B \ll M_A$, the result is

$$N_{\rm eq} \simeq 0.9 \left(\frac{0.65 + \log_{10}\frac{M_A}{M_B}}{3.7}\right)^{-1} \left(\frac{M_A}{1\,{\rm M}_\odot}\,\frac{R_{AB}}{5\,{\rm au}}\right)^{3/2} \left(\frac{v_0}{220\,{\rm km/s}}\right)^{-3} \left(\frac{n_\infty \times 10^{-16}\,{\rm M}_\odot}{0.3\,{\rm GeV/cm^3}}\right), \tag{2.44}$$

where the base values for the parameters are chosen to be roughly representative of the capture of objects of mass $10^{-16}\,{\rm M}_\odot$ by the Sun–Jupiter system, assuming they account for the entirety of the local DM density. Strictly speaking, this is an estimate of an

upper bound on the capture rate, since the angles λ_1 and β_1 are chosen in the most favorable configuration possible. Nonetheless, this serves as an informative estimate of the capture rate at the order of magnitude level and applies to a wide range of systems. Indeed, this result is a reasonably good match to the numerical results in Fig. 2.8, overestimating the number of captured objects by about an order of magnitude.

PBHs at masses below $\sim 10^{-16}\,\mathrm{M}_\odot$ are strongly constrained by evaporation, so this optimistic estimate indicates that capture in the Sun–Jupiter system is only possible for PBH DM in a narrow mass range. Nonetheless, this estimate suggests that if a substantial fraction of the DM is composed of PBHs with significant evaporation luminosity, then it is possible that a bright point source could be found captured within the solar system. Recently, Ref. [103] studied the potential implications of discovering such a low-mass PBH nearby: since such an object would be actively evaporating, the relationship between the object's mass and evaporation rate would enable a direct count of the number of dark-sector degrees of freedom. Our calculation suggests that if a population of such objects were maintained for a sufficiently long time, then there would be good prospects to find one close enough to be studied in this manner. However, since such a measurement relies on the rapid evaporation of the observed PBH, such a population would not be stable on cosmological timescales.

In the limit of asymmetric masses $M_B \ll M_A$, the equilibrium number of captured objects is only very weakly dependent on the mass of the lighter object in the binary. This is to be expected: in such a case, the cross sections for capture and ejection both scale with M_B^2. On the other hand, systems with larger M_A and R_{AB} are much more efficient at capturing and retaining light PBHs. At the upper ranges

of Fig. 2.8, a wide binary with a $100\,\mathrm{M}_\odot$ central object and an orbital separation of $10^3\,\mathrm{au}$ has $\langle \widetilde{N} \rangle \sim 10^3$ for all of DM in the form of PBHs with mass $10^{-13}\,\mathrm{M}_\odot$. Thus, such a system has an $\mathcal{O}(1)$ probability of hosting at least one PBH in a bound orbit if the PBH mass is below $10^{-10}\,\mathrm{M}_\odot$. The capture rate of Earth-mass objects is very low in all realistic binary systems, so such captured objects cannot account for OGLE microlensing events. If instead DM is in the form of light PBHs with mass between $10^{-16}\,\mathrm{M}_\odot$ and $10^{-14}\,\mathrm{M}_\odot$, as is allowed by current constraints, then such objects should be commonly bound in systems only slightly heavier and wider than the Solar system.

2.4 Dissipative dynamics

The treatment of the previous section is limited to capture by three-body interactions. We now turn our attention to capture by many-body interactions, which are qualitatively distinct due to dissipation: such captures are not time-reversible. When PBHs are captured around single objects, ejection is impossible. Even in multi-component systems, dissipative captures are much less prone to ejection than their few-body counterparts. Thus, even though the rates of dissipative captures are naively much smaller, it is important to evaluate their contribution to the population of bound objects.

2.4.1 Gas drag and dynamical friction

As an unbound object such as a planet passes through a gaseous environment, its kinetic energy is dissipated via interactions with many gas particles, potentially resulting in a capture [104]. A similar mechanism may lead to captures of certain types

of compact objects. However, only a particular class of compact objects are subject to the usual physics of gas drag. As usually treated, gas drag presumes that the object efficiently displaces gas in its path, but this is not the case for dark compact objects such as PBHs.

A black hole will still accrete gas particles in its path, which will reduce the object's specific kinetic energy. However, this effect is suppressed by the very small geometric cross section of the black hole. Including gravitational focusing, this cross section is

$$\sigma_{\rm PBH} = \pi \left(\frac{2GM_{\rm PBH}}{c^2} \right)^2 \left(1 + \frac{c^2}{v_{\rm rel}^2} \right), \tag{2.45}$$

where we have used the Schwarzschild radius $r_{\rm PBH} = 2GM_{\rm PBH}/c^2$. Now suppose that a black hole with initial velocity v_∞ transits through a spherical gas cloud of density ρ and radius R, accreting a mass $\Delta M_{\rm PBH} \approx 2\rho R \sigma_{\rm PBH}$, and suppose that the accreted gas particles are slow compared to the accreting PBH. At the end of the transit, the potential energy is reduced by $\Delta U_{\rm PBH} \simeq -GM_{\rm cloud}\,\Delta M/R$ due to the accreted mass. Capture requires that the total mechanical energy becomes negative, and since the accreted mass leaves the kinetic energy constant, the potential energy must decrease by at least $T_\infty = \frac{1}{2} M_{\rm PBH} v_\infty^2$. Taking $v_{\rm rel} \ll c$, and neglecting changes in the PBH velocity due to accretion, we have

$$\frac{\Delta U_{\rm PBH}}{T_\infty} \simeq \frac{32\pi^{3/2} G^3 \rho^2 R^2 M_{\rm PBH}}{v_\infty^2 c^2 \left[3G\rho \left(v_\infty^2 + 4\pi G\rho R^2 \right) \right]^{1/2}} \operatorname{arctanh} \left[2R \left(\frac{\pi G\rho}{3(v_\infty^2 + 4\pi G\rho R^2)} \right)^{1/2} \right]. \tag{2.46}$$

For massive clouds with $4\pi G\rho R^2 \gg v_\infty^2$, this simplifies to

$$\frac{\Delta U_{\rm PBH}}{T_\infty} \simeq 10^{-13} \times \left(\frac{M_{\rm PBH}}{M_\oplus} \right) \left(\frac{R}{10^5\,{\rm au}} \right) \left(\frac{\rho}{10^{-10}\,{\rm g/cm^3}} \right) \left(\frac{v_\infty}{220\,{\rm km/s}} \right)^{-2}. \tag{2.47}$$

Thus, even for densities and radii well in excess of those of typical gas clouds, simple accretion is not an efficient energy loss mechanism for black holes.

However, black holes are still subject to dynamical friction, i.e., energy loss due to gravitational interactions with the gas, and we now estimate the rate of captures by this mechanism. First, consider a perturber of mass $M_{\rm PBH}$ moving with velocity v in the rest frame of a uniform gaseous medium with density ρ. Dynamical friction in this scenario has beeen treated by Ref. [105], and previously applied to planet formation by Ref. [106]. The frictional force on the perturber depends first on whether the relative velocity is subsonic or supersonic. Recall that for an ideal gas with adiabatic index γ and molecular mass $m_{\rm mol}$, the sound speed is given by $c_s^2 = \gamma k_{\rm B} T / m_{\rm mol}$. In terms of the Mach number $\mathcal{M} \equiv v/c_s$, the frictional force is given by

$$F_{\rm DF} = -\frac{2\pi\rho G^2 M_{\rm PBH}^2}{\mathcal{M}^2 c_s^2} \begin{cases} \log\left(\dfrac{1+\mathcal{M}}{1-\mathcal{M}}\right) - \mathcal{M} & \mathcal{M} < 1 \\ \log\left(1 - 1/\mathcal{M}^2\right) + 2\log\left(\dfrac{d_{\rm max}}{d_{\rm min}}\right) & \mathcal{M} > 1. \end{cases} \tag{2.48}$$

Here $d_{\rm min}$ is the distance of closest approach between gas molecules and the perturber, and $d_{\rm max}$ is the length scale of the wake left behind as the object traverses the medium. For macroscopic objects, $d_{\rm min}$ is cut off by the size of the perturber itself. In our case, working with compact objects, the size of the perturber can be very small: a black hole of mass $10^{-9}\,{\rm M}_\odot$ has a Schwarzschild radius on the order of $3\,\mu{\rm m}$. Depending on the black hole mass and gas density, the Schwarzschild radius $R_{\rm PBH}$ may be smaller than the typical spacing of the gas molecules, $d_{\rm mol} \equiv (\rho/m_{\rm mol})^{-1/3} \approx 0.5\,\mu{\rm m}\left[\rho/(10^{-8}\,{\rm kg/m^3})\right]^{-1/3}\left[m_{\rm mol}/m_{\rm H}\right]^{1/3}$, where $m_{\rm H}$ is the mass of Hydrogen. We take $d_{\rm min}$ to be the larger of these two scales, $d_{\rm min} = \max\{R_{\rm PBH}, d_{\rm mol}\}$. We set $d_{\rm max} = vt$ a

time t after the perturber enters the cloud, and we neglect times for which $d_{\rm max} < d_{\rm min}$.

Note that the dynamical friction force is proportional to $M_{\rm PBH}^2$, so the acceleration of the perturber is linear in the perturber's mass. However, if the mass density of PBHs is held fixed, the number density scales as $M_{\rm PBH}^{-1}$, so the capture rate should be approximately independent of mass in this case. This independence is not exact due to the weak logarithmic dependence on $M_{\rm PBH}$ via $d_{\rm min}$ in the regime where the latter is set by the Schwarzschild radius.

Now we specialize to a uniform spherical cloud of radius R, and assume that the perturber travels through the center of the cloud. The energy lost over the course of the encounter is $\Delta E = \int_{-R}^{R} {\rm d}s\, F_{\rm DF}(s)$, where s parametrizes the trajectory. Anticipating that $\Delta E/E$ is small, we neglect any change in $\mathcal{M}$, so the only r-dependence in $F_{\rm DF}$ comes from setting $d_{\rm max} = vt = s + R$. Then the fractional energy loss is

$$\frac{\Delta E}{E} = \frac{8\pi^2 \rho G^2 M_{\rm PBH} R}{\mathcal{M}^4 c_s^4} \begin{cases} \log\left(\frac{1+\mathcal{M}}{1-\mathcal{M}}\right) - \mathcal{M} & \mathcal{M} < 1 \\ \log\left(1 - 1/\mathcal{M}^2\right) + 2\left[\log\left(\frac{2R}{d_{\rm min}}\right) - 1\right] & \mathcal{M} > 1. \end{cases} \tag{2.49}$$

In the far subsonic and supersonic regimes, this reduces to

$$\frac{\Delta E}{E} \simeq \frac{8\pi^2 \rho G^2 M_{\rm PBH} R}{c_s^4} \begin{cases} \mathcal{M}^{-3} & \mathcal{M} \ll 1 \\ 2\left[\log\left(\frac{2R}{d_{\rm min}}\right) - 1\right] \mathcal{M}^{-4} & \mathcal{M} \gg 1. \end{cases} \tag{2.50}$$

Note that $\mathcal{M}$ is bounded below due to acceleration by the cloud itself: an object with $v_\infty > 0$ will enter the cloud with velocity above $\mathcal{M}_{\rm min} \equiv [8\pi G\rho/3]^{1/2}\, R/c_s$.

A typical nebula has a number density of order $10^3\,{\rm cm}^{-3}$, an extent of at most $1\,{\rm pc}$, and a temperature $T \sim 10^4\,{\rm K}$ [107, 108]. The speed of sound is then $c_s \approx \sqrt{(5/3)k_{\rm B}T/m_{\rm H}} \sim 10\,{\rm km/s}$ for Hydrogen, while $\mathcal{M}_{\rm min} \sim 0.09$ for the same config-

uration. Since $\mathcal{M}_{\rm min} \ll 1$, we first consider the subsonic regime. In the subsonic limit, the capture condition $\Delta E/E > 1$ becomes $\mathcal{M} < \mathcal{M}_{\rm cap} \equiv 2\left(\pi^2 \rho R G^2 M_{\rm PBH}/c_s^4\right)^{1/3}$, and imposing $\mathcal{M}_{\rm cap} > \mathcal{M}_{\rm min}$ leads to the requirement

$$\begin{aligned} 1 < \frac{\mathcal{M}_{\rm cap}}{\mathcal{M}_{\rm min}} &= \left(\frac{27\pi}{8}\frac{G M_{\rm PBH}^2}{\rho c_s^2 R^4}\right)^{1/6} \\ &\approx 0.002 \left(\frac{M_{\rm PBH}}{1\,{\rm M}_\oplus}\right)^{1/3} \left(\frac{\rho}{10^3 m_{\rm H}/{\rm cm}^3}\right)^{-1/6} \left(\frac{R}{1\,{\rm pc}}\right)^{-2/3} \left(\frac{c_s}{10\,{\rm km/s}}\right)^{-1/3}. \end{aligned} \quad (2.51)$$

Thus, far-subsonic captures require unrealistically large masses, small system dimensions, or low sound speeds. Note that as $R, \rho \to 0$, although $\mathcal{M}_{\rm cap}/\mathcal{M}_{\rm min}$ becomes large, $\mathcal{M}_{\rm cap}$ vanishes, so small or low-density systems can only capture objects in subsonic transits for an extremely narrow range of initial velocities.

In the supersonic limit, $\Delta E/E$ is suppressed by $\mathcal{M}^4$, so the most promising regime for captures is the transonic regime, $\mathcal{M} \approx 1$. The dynamical friction force peaks here, with a divergence at $\mathcal{M} = 1$. The force in the transonic regime is well approximated by taking $F_{\rm DF}$ to be symmetric about $\mathcal{M} = 1$ and expanding the $\mathcal{M} < 1$ expression about that point. This gives

$$F_{\rm DF} \simeq -\frac{2\pi\rho G^2 M_{\rm PBH}^2}{v^2}\left[-1 + \log\left(\frac{2}{|\mathcal{M}-1|}\right)\right], \quad (2.52)$$

with $\Delta E \simeq 2R F_{\rm DF}$. In the transonic limit, one can solve for the maximal value of $|\mathcal{M}-1|$ leading to capture, finding

$$\begin{aligned} |\mathcal{M}-1| &< 2\exp\left(-1 - \frac{c_s^4}{8\pi\rho G^2 R M_{\rm PBH}}\right) \\ &\approx 2\exp\left[-3\times 10^{11}\left(\frac{c_s}{10\,{\rm km/s}}\right)^4 \left(\frac{R}{1\,{\rm pc}}\right)^{-1} \left(\frac{M_{\rm PBH}}{1\,{\rm M}_\oplus}\right)^{-1} \left(\frac{\rho}{10^3 m_{\rm H}/{\rm cm}^3}\right)^{-1}\right]. \end{aligned} \quad (2.53)$$

Thus, for any realistic parameter values, transonic captures are viable only for a vanishingly narrow range of initial velocities.

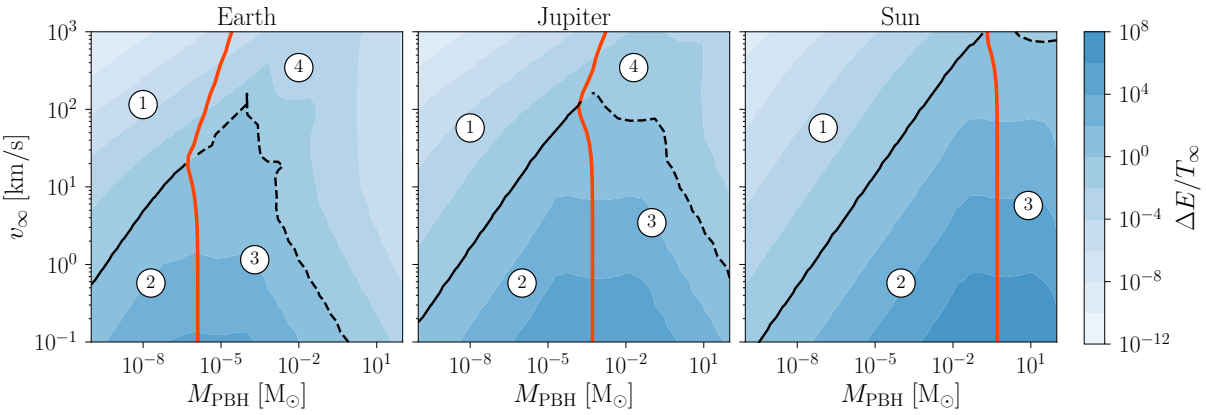

Figure 2.9: Energy loss in transiting through a body as a fraction of the kinetic energy at infinity, as a function of the PBH mass and velocity at infinity. Panels show three benchmark cases with parameters of Earth, Jupiter, and the Sun. In each panel, points below the black curve result in captures ($\Delta E > T_\infty$), and points to the right of the orange curve destroy the target object (ΔE exceeds the binding energy). The four labeled points correspond to the following scenarios: **(1)** the transit neither captures the PBH nor destroys the target; **(2)** the transit captures the PBH without destroying the target; **(3)** the target is destroyed, and the PBH is bound to the system; **(4)** the target is destroyed and the PBH remains free.

2.4.2 Collisions with stellar and planetary bodies

In the previous subsection, we have considered PBH capture by dissipation in a gas cloud, where the geometric cross section is large compared to the three-body close encounters of Section 2.3, but the typical energy losses are much smaller. However, thus far, we have neglected a mechanism for large dissipative losses with a small cross section: transit of a PBH through a planet or star. In this scenario, the PBH dissipates energy by the same dynamical friction mechanism that drives losses in a gas cloud, but the higher density of a planet or star leads to much more significant energy losses during such a transit. We now consider the population of objects that would be captured by this particular mechanism.

The energy lost to dynamical friction in passing through a star or planet can be computed by a similar procedure as in the previous section, but our treatment now differs in two ways. First, since the typical parameter values are quite different from those of gas clouds, we do not assume that the PBH velocity is constant throughout the encounter. Instead, we compute the instantaneous energy loss by numerical solution of the equations of motion. Second, since there may be observational implications, we are motivated to consider the destruction of objects by the passage of PBHs in addition to capture. We perform a simplified treatment of planet and star destruction: we say that the target is destroyed if the energy lost by the PBH to dynamical friction, $E_{\rm DF}$, exceeds the binding energy of the target, $U_G = -3GM_{\rm target}^2/(5R_{\rm target})$.

We study energy losses in three benchmark systems: Earth, Jupiter, and the Sun. The results are shown in Fig. 2.9. For each of these cases, capture of a PBH without destruction of the target is possible for sufficiently light PBHs with low initial velocities, i.e., in region **(2)** of each panel. For higher initial velocities, the encounter takes place with $\mathcal{M} \gg 1$, and the dynamical friction force is suppressed. Similarly, large PBH masses $M_{\rm PBH} \gg M_{\rm target}$ accelerate the target and guarantee a highly supersonic encounter, so the energy loss at large PBH masses is negligible and capture is impossible. Observe that in all three cases, destruction of the target requires $M_{\rm PBH} \gtrsim M_{\rm target}$: upon collision, the kinetic energy of a PBH falling from rest is given by $GM_{\rm target}M_{\rm PBH}/R_{\rm target}$, which only exceeds U_G if $M_{\rm PBH} > (3/5)M_{\rm target}$. Thus, whenever destruction is possible, the ratio of the PBH number density to the target number density is bounded above by the ratio of their volume-averaged mass densities, $n_{\rm PBH}/n_{\rm target} \lesssim \rho_{\rm PBH}/\rho_{\rm target}$. This factor in turn is $\mathcal{O}(100)$ for stars, which suggests that stellar destruction events take

place on a timescale at most a factor of $\mathcal{O}(100)$ shorter than that for collisions of stars, which are exceedingly rare. Given that the number density of planets is parametrically close to the number density of stars, the maximum rate of destruction events in these systems at first appears to be higher by a factor $M_{\rm planet}/M_{\rm star}$, but the geometric cross section suffers a comparable suppression.

Thus, destruction events of any kind are rare. Explicit computation confirms that the destruction rate is comparable across our three benchmark systems, and is no higher than $10^{-26}\,{\rm s}^{-1}$ for any PBH mass, well below the Hubble rate. Captures are also extremely rare and do not occur in excess of $10^{-24}\,{\rm s}^{-1}$. Captured objects can undergo subsequent transits, which in principle enhances the destruction rate, but not above the very low capture rate. Note that destruction by PBH encounters has been previously considered by Ref. [109] in the context of luminous signatures of PBH collisions with stellar and planetary objects, with qualitatively similar results. Stellar capture of DM has likewise been studied previously e.g. by Refs. [110, 111].

2.4.3 Adiabatic contraction

A third possibility is that dissipation of gravitational energy of the gas itself provides a mechanism for the capture of dark compact objects. As gas collapses during star formation, the potential well deepens, and nearby objects can thus be captured—not by a loss of kinetic energy, but by a reduction in potential energy. A key element of the process is that as the gas density increases, the local DM density is gravitationally enhanced, a process known as adiabatic contraction [112]. Thus, during the process of star formation, DM particles—or equivalently, dark compact objects such as PBHs—can

be efficiently captured.

This mechanism has been considered in detail by Ref. [113] for its effects on the population of luminous evaporating black holes captured around stars, and more recently by Refs. [48, 114, 115] in the context of stellar destruction. Following Ref. [114], a gas cloud of density $\rho_{\rm g}$ and radius $R_{\rm g}$ captures a DM halo with density of order

$$\rho_{\rm bound} \simeq \rho_{\rm DM} \times \frac{4\pi}{3} \left(\frac{6 G \rho_{\rm g} R_{\rm g}^2}{\frac{3}{2} v_0^2} \right)^{3/2}, \tag{2.54}$$

where v_0 is the characteristic DM velocity dispersion of Eq. (2.2). Due to adiabatic contraction, the bound DM particles (compact objects) assume an equilibrium distribution with a power-law profile $\rho_{\rm bound}(r) \sim r^{-3/2}$. We assume that the extent of the bound halo is the cloud radius R, so that the number density within any particular radius can be readily calculated.

The baryonic gas that forms stars is at first found in giant molecular clouds, with masses as large as $10^6\,{\rm M}_\odot$ and radii as large as 10 pc [116]. These clouds fragment and form many prestellar cores, with typical masses of 1–10 $\rm M_\odot$ and typical radii of 3000–6000 au [117]. Even for a dense system with a total mass of $10\,{\rm M}_\odot$ and a radius of 3000 au, with $\rho_{\rm DM} = 0.3\,{\rm GeV/cm^3}$, the density of bound DM is negligible, $\rho_{\rm bound} \approx 6 \times 10^{-7}\,{\rm GeV/cm^3}$. This is mainly due to the sharp dependence on the velocity dispersion v_0: if the system under consideration forms in a small DM substructure with a small dispersion, then the bound density can be considerable. In particular, in globular clusters, constraints can be derived from the absence of stellar destruction, as in Ref. [114]. However, for a generic stellar system, capture due to adiabatic contraction is negligible.

2.5 Discussion

In the preceding sections, we have studied several distinct mechanisms by which PBHs can be captured in stellar systems. Some of these mechanisms give rise to bound orbits which are potentially short-lived, ending in ejection from the system or accretion into another body. Others produce stable, long-lived orbits, partially compensating for smaller capture cross sections. We have also evaluated the rate of destruction of planetary or stellar bodies by PBH encounters, and this is guaranteed to be negligible, requiring fairly high PBH masses and thus low number densities.

The most interesting captures are those which give rise to clear observables. More massive PBHs, particularly in the OGLE mass range, $\sim 10^{-6}\,\mathrm{M}_{\odot}$, would be more easily detected in extrasolar systems. Such PBHs are comparably massive to planets, so any observable must discriminate between such light PBHs and planets of the same mass. It is possible that such objects could be distinguished from planets based on the *absence* of stellar occultations. Occultation events, in which a star is dimmed by the transit of a planet, are a key non-gravitational signature used to detect planets in exoplanet searches. If *gravitational* Doppler shifting is observed to occur with a statistical excess compared to occultations, this would signal the presence of compact objects that do not block light.

Still, this strategy can only be effective in a class of systems where the expected number of captured objects is at least comparable to the number of planets. On the contrary, our results indicate that the capture of such massive PBHs is exceedingly rare. The rate of captures is suppressed by the number density of PBHs, which is very small

even in the most optimistic cases. Even if objects with mass as low as $10^{-8}\,\mathrm{M}_\odot$ could be reliably detected by gravitational means, and even if they accounted for 10% of the DM density, the equilibrium values in Fig. 2.8 would still be suppressed by at least 10^{-6}. This implies that $\langle N \rangle \ll 1$ even in the widest and most massive binaries, making this an unlikely probe of the PBH population.

Rather, the capture rate is inevitably highest for DM composed of light PBHs, in the open window in the constraints for $10^{-16}\,\mathrm{M}_\odot \lesssim M_{\mathrm{PBH}} \lesssim 10^{-12}\,\mathrm{M}_\odot$. We have shown that such objects can be frequently captured in realistic stellar systems, particularly massive wide binaries. However, in most such systems, it is improbable that such a light object would produce a distinctive observable signature: objects with non-negligible capture rates would be comparable in mass to asteroids or even lighter.

For these lighter objects, there are still two observables of interest. First, there are potential implications for pulsar timing signatures. It is known that the timing signature can be observably perturbed by short-lived PBH flybys [118]. In our scenario, it is also possible to witness a short-lived capture. Here, a PBH has a close encounter with a binary companion of a millisecond pulsar, and is captured into a short-lived bound orbit before being ejected from the triple system. Such captures are almost always short-lived because of the small semimajor axis of pulsar binaries, and the cross section for such captures is extremely small. Indeed, the assumptions of Section 2.3 are typically violated in such systems, and simulations suggest that the rate is an order of magnitude smaller than the prediction of Eq. (2.22). Nonetheless, such temporary captures would have a distinctive signature in pulsar timing. In particular, the signatures and population statistics of pulsar planets have been studied previously

in the astronomy community (see e.g. Refs. [119, 120]). A captured PBH would result in the temporary appearance of a pulsar planet, which would vanish on the timescale of $\mathcal{O}(1\,\mathrm{yr})$ once the object is ejected.

Second, at the very lightest end of the allowed mass range, such black holes would be actively evaporating today. Thus, systems which capture PBHs are a promising environment in which to search for PBH evaporation. As extrasolar systems are probed by a new generation of telescopes as part of the rapidly accelerating exoplanet program, it is possible in principle to see evidence of PBH evaporation using the same instruments. PBHs at the lowest masses consistent with present-day constraints would produce radiation at MeV energies and below, with significant emission even down to optical wavelengths (see e.g. Ref. [26]). We conclude that the best prospect for the observation of a captured PBH would be the detection of evaporation signatures by an exoplanet search. However, at present, we can draw no additional constraints on low-mass PBHs.

Chapter 3

Direct detection of primordial black holes

3.1 Introduction

Our discussion thus far has been largely concerned with black holes above $10^{14}\,\mathrm{g}$. This leaves the extremely light, microscopic black holes produced at lower masses. We now turn to the viable scenarios and corresponding observables in this mass range.

Such small black holes are expected to be unstable due to Hawking radiation: they should completely evaporate within the lifetime of the universe. The evaporation process has been used to draw constraints on the population of light black holes today [28, 121, 122]. However, evaporation is not well-understood at masses of order the Planck scale. It has been suggested that Hawking radiation in fact halts near this scale, leaving a relic black hole of mass $\sim M_{\mathrm{Pl}}$ [123–126], and these relics could constitute

the entirety of dark matter [127–129]. Such a relic would be almost completely inert, interacting only via gravity, but with a mass far too small to be detected as an individual object. From an experimental viewpoint, dark matter in the form of Planck-scale relics is a "nightmare" scenario, in that dark matter is effectively a particle with no non-gravitational interaction with the standard model. As such, it is extremely difficult to constrain relic black holes as dark matter.

However, there is another possibility: suppose that such relic black holes were electrically charged. Then these objects might be detectable by existing means. Interestingly, as we will discuss here, there is reason to believe that relic black holes could *typically* carry non-zero charge. The scenario is as follows: as the black hole evaporates, it emits charged particles of both signs, and it does so stochastically. Thus, during the evaporation process, non-zero electric charges are generic. If evaporation is cut off sharply at some mass scale of order $M_{\rm Pl}$, the black hole might be frozen with leftover electric charge of random sign. Alternatively, as we will also discuss, the impact of the spontaneous charge itself on the black hole geometry may act as a stabilizing mechanism. Regardless of their origin, we call such objects *Charged Planck-scale Relics* (CPRs). In this chapter, we show that such objects, if they exist, would be detectable terrestrially.

Generally, electric charges of order e are considered to be incompatible with dark matter. However, experimental constraints on the charge of dark matter (e.g. in the context of millicharged dark matter) are always placed on some combination of the charge and mass of the dark matter species. In our case, we will be interested in objects with a charge-to-mass ratio of order $\sim e/M_{\rm Pl}$. Such objects behave as dark matter

in every respect: their self-interactions are dominated by gravity; their interactions with standard model particles impart no appreciable change in their momentum; and, since they must be extremely sparse due to their large masses, they have no impact on baryonic dynamics apart from their bulk gravitational potential.

CPRs are similar to charged massive particles (CHAMPs [130]) in that they possess integer-valued electric charges. CHAMPs have been studied as a dark matter candidate for decades, but direct detection prospects differ significantly between CHAMPs and CPRs, due mainly to the difference in the typical masses of the two objects. CHAMPs are depleted in the galactic disk due to their interactions with magnetic fields [131, 132], and a survey of other CHAMP probes by Ref. [133] yielded null results. However, these results apply only to CHAMPs with masses *below* $10^8\,\mathrm{TeV}$. We expect CPRs to be found at $\sim 10^{16}\,\mathrm{TeV}$, well above this threshold, so the CHAMP literature is largely inapplicable to our case.

The strongest cosmological constraint on the charge-to-mass ratio of dark matter comes from the CMB power spectrum [134], which requires

$$q_{\mathrm{DM}} \lesssim 2.24 \times 10^{-4} \left(\frac{m_{\mathrm{DM}}}{1\,\mathrm{TeV}}\right)^{1/2} e. \tag{3.1}$$

Our fiducial mass scale is M_{Pl}, for which this translates to $q_{\mathrm{DM}} \lesssim 2.5 \times 10^3 e$. This constraint is thus also irrelevant for our scenario, in which, as detailed below, we predict charges of order e. Indeed, as we will discuss, the cosmic censorship conjecture imposes a much stronger constraint on the electric charge of Planck-scale black holes. Constraints on q_{DM} from terrestrial experiments are also ineffective at the large masses we consider.

Thus, there are two major motivations to search for CPRs experimentally.

First, despite being electrically charged, CPRs could constitute the entirety of dark matter if evaporation halts near the Planck scale. Second, even if CPRs constitute only a small fraction of dark matter, the confirmed detection of even one such object would be of incredible value to black hole physics: it would confirm that black hole evaporation does indeed halt, and pave the way for the experimental study of quantum gravity. Remarkably, the first constraints on the abundance of CPRs can already be placed with existing experimental results, and future experiments offer the opportunity to considerably tighten these bounds.

The structure of this chapter is as follows. In Section 3.2, we show how CPRs can form, and quantitatively estimate their abundance given realistic formation scenarios. In Section 3.3, we study the interaction of CPRs with matter and evaluate mechanisms for the terrestrial detection of CPRs. In Section 3.4, we derive constraints from non-detection in existing experiments and project constraints that can be obtained from proposed or upcoming experiments. We discuss the implications in Section 3.5 and conclude in Section 3.6.

Unless otherwise indicated, we work in units with $c = \hbar = k_B = G = 1$, and $\varepsilon_0 = 1/4\pi$. In these units, the elementary charge e is given by $\sqrt{\alpha} \approx 1/11.7$. We take $M_{\rm Pl} = (\hbar c/G)^{1/2} = 1$. In these units, a black hole with charge-to-mass ratio Q/M has $Q \approx (Q/M)(11.7e)$. Additionally, note that e corresponds to "positive charge" in these units.

3.2 Evaporation and spontaneous charge

Ref. [135] showed that black holes radiate, or *evaporate*, as thermal blackbodies. A black hole's temperature is related to its surface gravity κ via $T = \kappa/(2\pi)$, and according to no-hair theorems [136], κ can only depend on three parameters: the black hole's mass M, electric charge Q, and angular momentum L. As a benchmark, a Schwarzschild black hole ($Q = L = 0$) of mass M has temperature $T = 1/(8\pi M)$ as measured by a faraway observer. Since evaporation tends to discharge angular momentum rapidly, a black hole with some initial spin is unlikely to have appreciable angular momentum once it reaches the Planck scale. In particular, Ref. [137] showed that black holes with mass below $\sim 10^{14}\,\mathrm{g}$ today should have a spin parameter very near zero, so the impact of angular momentum on the black hole metric should be negligible.

Thus, we will only consider non-rotating ($L = 0$) black holes with charge Q. Such black holes are described by the Reissner-Nordström (RN) metric:

$$\mathrm{d}s^2 = \left(1 + \frac{2M}{r} + \frac{Q^2}{r^2}\right)\mathrm{d}t^2 - \left(1 - \frac{2M}{r} + \frac{Q^2}{r^2}\right)^{-1}\mathrm{d}r^2 - r^2\,\mathrm{d}^2\Omega. \tag{3.2}$$

The radial component of the RN metric diverges at two values of r, namely

$$r_\pm = M \pm \sqrt{M^2 - Q^2}. \tag{3.3}$$

The outer horizon radius r_+ defines the surface of the black hole for our purposes, and thus plays an important role in determining the properties of particle emission. Note that we only have two distinct horizons when $Q < M$. When $Q = M$, the black hole is extremal, and its surface gravity vanishes. If $Q > M$, the black hole is super-extremal. Such states are generally thought to be non-physical. We will discuss extremality in

more detail in Section 3.2.3. The temperature of an RN black hole is given by

$$T = \frac{\left(M^2 - Q^2\right)^{1/2}}{2\pi\left(M + \left(M^2 - Q^2\right)^{1/2}\right)^2}. \tag{3.4}$$

Hawking radiation has yet to be directly observed, due mainly to the fact that all known black holes have large masses, and are therefore extremely cold. An astrophysical black hole cannot form below the Chandrasekhar limit [138] of $\sim 1.4 M_\odot$, for which the corresponding temperature is $T \sim 4 \times 10^{-12}\,\mathrm{eV}$. Thus, the effects of Hawking radiation on astrophysical black holes are negligible even on cosmological timescales. Since all known black holes are cold, with temperatures much lower than the masses of any known massive particles, black hole evaporation is often treated by considering only the emission of neutral massless particles. But in our scenario, we are interested in black holes of primordial origin, which may form with much lower masses, and thus radiate with much higher temperatures. Such black holes can produce massive charged particles at an appreciable rate.

Since there is no need for such particles to be emitted in pairs of opposite sign, a neutral black hole can spontaneously acquire an electric charge by emission of a charged particle. On the other hand, a charged black hole is more likely to emit particles of like sign [139], so the spontaneous charge of a sufficiently small black hole fluctuates rapidly around neutrality. Ref. [140] studied the distribution of black hole charges numerically, and found that if a black hole is small enough to emit charged leptons rapidly, the equilibrium charge distribution is approximately Gaussian,

$$P(Q) \sim \exp\left(-4\pi\alpha(Q/e)^2\right), \tag{3.5}$$

with rms value of Q/e given by $(8\pi\alpha)^{-1/2} \approx 2.34$. The numerical calculations in that

work show that if the product of the black hole mass and the emitted particle mass is small in Planck units, then the rms value of Q/e increases to ~ 6.

In our scenario, we envision that evaporation is halted near the Planck scale, and that any remaining charge is thus "stuck" on the black hole, leaving a charged Planck-scale relic (CPR). Of course, black hole evaporation is not well understood at masses near the Planck scale, and the outstanding issues in the study of black hole evaporation are beyond the scope of this work. Ultimately, we must neglect these problems in order to study the basic plausibility of our scenario. However, we will first review what the problems are, and discuss which ones can be ameliorated in our context and which ones cannot.

The spontaneous emission of charge by black holes has been studied analytically, e.g. by Ref. [139], and one might hope that such analytical work could serve as a guide for our study. However, such analytical techniques break down when the black hole horizon becomes smaller than the emitted particle's Compton wavelength. Thus, we must retreat to numerical techniques. In the ultra-low-mass regime, near $M_{\rm Pl}$, there are several additional issues that confound an exact calculation of the charge distribution. Of course, the behavior of gravity itself is poorly understood in this regime: quantum gravity corrections should be significant, and it is not known how this influences the charge distribution. But even treating gravity as a classical background, several problems remain.

The first problem is the treatment of backreaction from emitted charges on the rate of subsequent emissions. The relevant quantity here is the timescale separating distinct emission events. For massive black holes, with low temperatures, this timescale

is quite long [140], and backreaction can be neglected. But for small black holes, the emission rate is much higher, so it may be inappropriate to treat consecutive emission events as independent processes. The nature of backreaction and its connection to black hole stabilization is subject to ongoing discussion in the literature [see e.g. 141], and the impact on the charge distribution is unclear.

The second problem is that as the mass becomes very small, the charge-to-mass ratio becomes appreciable, and the impact of the charge on the black hole geometry cannot be neglected. This is manifested most clearly in the case that $Q \sim 12e$ for a black hole of $M \sim M_{\rm Pl}$, in which case the black hole is *near-extremal:* the charge-to-mass ratio is nearly as great as possible, and the surface gravity of the black hole drops nearly to zero. An exactly extremal black hole has a temperature of exactly zero, and emits no thermal Hawking radiation. (It may still radiate athermally, as we will discuss shortly.) The calculation of Ref. [140] assumed that $Q/M \ll 1$, a condition we may very well violate in our scenario.

The third problem concerns the role of the electromagnetic coupling α. At large black hole masses, the width of the equilibrium charge distribution in Eq. (3.5) is sensitive to α. The calculation is perturbative, so it is critical that the back-reaction of emitted particles on the metric should be higher-order in α. But this is not necessarily the case at extremely small length scales. To make matters worse, the temperature is also of order the Planck scale, meaning that the relevant value of α is subject to renormalization all the way to the Planck scale, and thus is sensitive to potentially all of BSM physics.

In light of all these issues, it is impractical to attempt a first-principles calcu-

lation of the charge distribution of relic Planck-scale black holes. Thus, in this chapter, we only perform an extremely naive estimate of the charge fraction as a plausibility argument, and then outline how such massive charged objects could be detected.

3.2.1 Emission from black holes

Ref. [142] showed that for a species with charge q and mass m, the emission rate at a frequency ω in each angular mode (ℓ, m) and polarization p is given by

$$\frac{\mathrm{d}N_{\ell,m,p}}{\mathrm{d}t\,\mathrm{d}\omega} = \frac{\Gamma_{\ell,m,p}(\omega, T, q\Phi)/2\pi}{\exp\left[(\omega + q\Phi)/T\right] \pm 1} \tag{3.6}$$

where Φ is the electrostatic potential at the surface of the hole ($-Q/r_+$ in our case), and $\Gamma_{\ell,m,p}$ is an absorption coefficient specific to that mode. The emission rate of Eq. (3.6) has the form of a thermal spectrum with a chemical potential proportional to the black hole's charge. It is sometimes useful to take a different viewpoint, and consider the emission rate to result from a combination of two mechanisms, one thermal and one athermal.

Heuristically, Hawking emission can be viewed as the separation of spontaneous virtual particle-antiparticle pairs by the black hole horizon. In the absence of charge, this process is mediated by gravity alone. This is the "thermal" component of black hole evaporation, which deviates from a blackbody spectrum only by virtue of the greybody factors $\Gamma_{\ell,m,p}$. However, if the black hole has a significant charge, then the picture must be modified: now, in addition to strong curvature near the horizon, there is a strong electric field. A strong electric field, even in the absence of curvature, can separate particle-antiparticle pairs in much the same way. This particle production

due to vacuum polarization is just the familiar Schwinger mechanism [143]. It enters into Eq. (3.6) in two ways: first, as a chemical potential in the exponential factor, and second, via the dependence of $\Gamma_{\ell,m,p}$ on the black hole's charge. Note that Eq. (3.6) is compatible with the operation of the Schwinger mechanism even when the black hole's temperature is exactly zero. We refer to the component of radiation associated with the Schwinger mechanism as *athermal* emission, and we refer to the remainder as *thermal* emission. We will discuss the consequences of these two components further in Section 3.2.3.

In a sense, these two mechanisms compete: thermal emissions drive the black hole away from neutrality in a random walk, but the athermal emissions are always of like sign to the black hole, and tend to discharge it. Equivalently, the black hole emits charges of both signs as long as $|Q| < M - e$, but as $|Q|$ increases, the emissions are increasingly biased to have the same sign as Q. This fact led Ref. [139] to observe that a small black hole cannot maintain even one elementary charge for an appreciable length of time, so long as evaporation remains active. We are interested in the characteristic lifetime of both neutral and charged black holes, where any charge implies a significant charge-to-mass ratio due to the small mass, so we cannot neglect either the thermal or the athermal component. Thus, it is important for us to compute the absorption coefficient $\Gamma_{\ell,m,p}$ for charged leptons to the extent possible in our framework. The absorption coefficient is calculated by solving the Dirac equation for an incoming wave with the appropriate boundary conditions in the Reissner-Nördstrom geometry [144]. The solution can be resolved into ingoing and outgoing waves, from which transmission and absorption coefficients can be extracted. This has been done numerically for several

particle species by Ref. [137, 140, 145], resulting in the distribution of Eq. (3.5).

In principle, this distribution applies even to black holes with large masses. However, the time to reach equilibrium grows with the timescale of lepton emission. For black holes whose Hawking temperatures are below the lowest lepton mass—and certainly for astrophysical black holes—this timescale is extremely long, and we should expect the charge distribution of such black holes to be dominated by accretion of charged particles instead of evaporation. On the other hand, in the low-mass regime, where the Hawking emission timescale is very short, any charge acquired due to accretion will quickly be erased by evaporation processes, and the equilibrium distribution will be maintained.

In light of the issues discussed in this section, it is inappropriate to directly extrapolate the charge distribution of Ref. [140] to the Planck scale. Instead, in an effort to account for as many low-mass effects as possible, we implement a similar numerical calculation, and extract an order-of-magnitude estimate of the fraction of black holes with non-zero charge. We describe this calculation in the following section.

3.2.2 Estimating the charged fraction

In light of the problems discussed in Section 3.2, it is infeasible to perform a first-principles calculation of the fraction of stalled relics with spontaneous charge. However, we can perform a naive estimate by applying results developed for massive black holes, discarding approximations wherever possible. This result will not be a robust prediction of the charged fraction, but will instead represent a semi-classical guess. In this section, we make such an estimate, and then determine the implied

abundance of CPRs today.

We can perform a first estimate of the charged fraction in the relic population by evaluating two timescales: the characteristic timescale $\tau_{\rm neutral}$ for a neutral black hole to acquire a non-zero spontaneous charge, and the characteristic timescale $\tau_{\rm charged}$ for a black hole with charge $Q = e$ to discharge and become neutral. Then the fraction of objects which are charged at the moment that evaporation stalls can be estimated as

$$f_{\rm charged} = \left(1 + \frac{\tau_{\rm neutral}}{\tau_{\rm charged}}\right)^{-1}. \tag{3.7}$$

The timescale $\tau_{\rm charged}$ can be bounded below by neglecting time spent at higher charges during the black hole's semi-random walk, and evaluating only the minimum time to discharge, i.e.,

$$\tau_{\rm charged} \gtrsim e \Big/ \left.\frac{{\rm d}Q_-}{{\rm d}t}\right|_{Q=e}. \tag{3.8}$$

Here we use the notation ${\rm d}Q_-/{\rm d}t$ to denote the rate of emission in *positive* charge only, i.e., the rate at which the spontaneous charge decreases, as though by addition of negative charge. Likewise, ${\rm d}Q_+/{\rm d}t$ denotes the rate of emission in negative charge only. The overall evolution of the charge is governed by $\langle {\rm d}Q/{\rm d}t\rangle = {\rm d}Q_+/{\rm d}t - {\rm d}Q_-/{\rm d}t$ on timescales that are long compared to the emission rate. It is critical to distinguish between the signed emission rates and the average, since $\langle {\rm d}Q/{\rm d}t\rangle = 0$ for $Q = 0$, while ${\rm d}Q_\pm/{\rm d}t$ are individually non-zero. Since a neutral black hole can decay to a state with either sign with equal probability, we have ${\rm d}Q_+/{\rm d}t = {\rm d}Q_-/{\rm d}t \equiv {\rm d}Q_\pm/{\rm d}t$, and the lifetime of the neutral state can be estimated as

$$\tau_{\rm neutral} \simeq \frac{1}{2}\, e \Big/ \left.\frac{{\rm d}Q_\pm}{{\rm d}t}\right|_{Q=0}. \tag{3.9}$$

Then the charged fraction can be estimated as

$$f_{\text{charged}} \gtrsim \left(1 + \frac{1}{2}\frac{\mathrm{d}Q_-/\mathrm{d}t|_{Q=e}}{\mathrm{d}Q_\pm/\mathrm{d}t|_{Q=0}}\right)^{-1}, \tag{3.10}$$

so our task is to compute $\mathrm{d}Q_\pm/\mathrm{d}t$ for a black hole near the Planck scale, with a charge of either zero or e. We neglect higher charges since such black holes should neutralize more rapidly. The effect of including them would only be to increase the final charged fraction.

Ref. [140] evaluates $\mathrm{d}Q_-/\mathrm{d}t$ for massive black holes following Eq. (3.6), computing the absorption probability $\Gamma_{\ell,m,p}$ numerically. The relevant information is found in fig. 4 of that work, which shows the emission rate ($\mathrm{d}N/\mathrm{d}t$) of charged leptons from a black hole as a function of the black hole charge for $-25e \leq Q \leq 25e$. Ref. [140] has calculated this rate separately for values of $M\mu$ in increments of 0.1 between 0.00 and 0.40 in Planck units, where M is the black hole and μ is the mass of the emitted lepton. In our case, $M \sim 1$ and $\mu \sim m_e/M_{\text{Pl}}$, so $M\mu = 0$ is the appropriate choice. Extrapolating those results to our regime, we find that

$$\frac{\mathrm{d}Q_-/\mathrm{d}t|_{Q=e}}{\mathrm{d}Q_\pm/\mathrm{d}t|_{Q=0}} \approx 1.02. \tag{3.11}$$

In short, this indicates that the athermal Schwinger emissions are at most comparable in rate to thermal Hawking emissions. If this is indeed the case, and given that we have neglected higher charges, the charged fraction is $f_{\text{charged}} \gtrsim 1/2$.

However, even insofar as we are neglecting the failure of various approximations in the limit $M \to M_{\text{Pl}}$, the results of Ref. [140] cannot be directly applied. In that work, the numerical calculations themselves were always performed with $M \gg M_{\text{Pl}}$. The label $M\mu = 0.00$ does not suggest that the result applies to our case, in which $M\mu$

is vanishingly small: $M\mu \sim 10^{-23}$. But even though we cannot assess the impact of Planck-scale physics on these results, we can still remove the uncertainty associated with the numerical computation by re-implementing the calculation and inserting this actual value of $M\mu$. Further, we solve eqs. (15) and (16) of Ref. [140] without neglecting the charge, as was done in that reference. The results are shown in Fig. 3.1. The implication is qualitatively unchanged for $Q \sim e$, with $\tau_{\text{neutral}}/\tau_{\text{charged}} \lesssim 1.04$.

Generally speaking, as shown in fig. 4 of Ref. [140], this ratio approaches unity as the mass of the black hole decreases. This is to be expected: the emission rate in same-sign particles scales with the absorption coefficient for modes that discharge the black hole, and in the low-mass limit, this coefficient is strongly suppressed [139]. The black hole still tends to discharge rapidly in the absence of charge fluctuations, but in our case, the timescale for discharge becomes comparable to the timescale of an upward charge fluctuation. This estimate of the charged fraction does assume that the evaporating black hole rarely enters the near-extremal regime ($Q \sim M$). However, as discussed in Section 3.2.3, we expect near-extremal states to have a very short lifetime due to athermal emission. Further, the potential stalling of evaporation due to extremality makes no significant difference to the outcome: since the rms charge distribution has a width of order 1–10e, and an extremal hole must have $Q/e \sim 10M$, we do not expect extremality to be an important consideration unless $M \sim 1$ already. In any case, extremality effects can only increase the charged fraction.

None of this discussion overcomes the fundamental difficulties with calculations near the Planck scale. However, this calculation establishes that in the absence of some new physics or new phenomenology, we should generically expect at least $\sim$50% of black

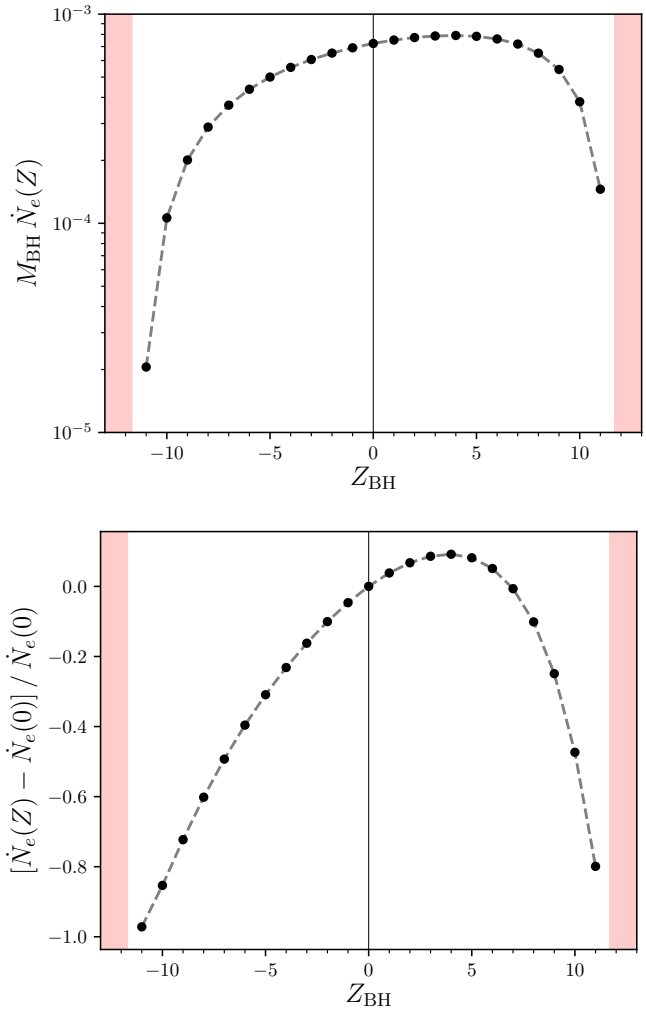

Figure 3.1: **Left:** the emission rate of (positive) particles of charge $+e$ from a black hole with charge Ze. Note that the charge number Z is shown in electron units, not Planck units, so that an extremal black hole has $Z \approx 11.7M/M_{\rm Pl}$. The mass of the black hole is fixed to $M_{\rm Pl}$ for the purposes of this calculation. This corresponds to the line "$M\mu = 0.00$" in fig. 4 of Ref. [140], but differs in that we carry out the computation for a mass which is not orders of magnitude larger than the charge. In particular, a substantial modification to the emission rate is observed when $|Q| \sim M$. The red shaded regions indicate where the black hole is super-extremal. **Right:** the fractional difference between the emission rate at $Q = Ze$ and the emission rate at $Q = 0$, computed at $M_{\rm Pl}$. Note the linear scale. The asymmetry in both the left and right plots is due to the electrostatic potential, which behaves like a chemical potential, and enhances the rate of emissions which tend to neutralize the black hole. However, for $|Q| \sim e$, the emission rate is modified by only $\sim$5%.

holes to be charged at the end of their evaporation. If their evaporation is halted, it is thus plausible that the relic population has a significant charged fraction.

3.2.3 Near-extremal regime

As the black hole mass approaches $M_{\rm Pl}$, there is an additional (classical) complication: the charge-to-mass ratio approach unity, changing the geometry of the black hole significantly. If the black hole's charge undergoes $\mathcal{O}(e)$ fluctuations, then our scenario involves charge-to-mass ratios of at least $e/M_{\rm Pl} \approx 1/11.7$. If the charge fluctuates even briefly to $\mathcal{O}(10e)$, then we may expect to have $Q/M \sim 1$ at some point during the black hole's evolution. This is the *near-extremal* regime. In this section, we review the properties of extremal Reissner-Nordström black holes and discuss the implications for the relic population.

An extremal black hole is a black hole with vanishing surface gravity, or equivalently, one whose mass is the smallest possible for its charge and angular momentum. The self-energy of the electric field and the angular momentum can both be thought of as contributing to the mass, so at fixed charge, the mass cannot be decreased arbitrarily. Super-extremal black holes, i.e., those with charge beyond the extremal limit, violate the cosmic censorship conjecture and are generally considered unphysical. In the Reissner-Nordström case, an extremal black hole has $Q/M = 1$, meaning that $r_+ = r_-$. This corresponds to a charge of $1/\sqrt{\alpha} \approx 11.7e$ for a black hole of mass $M_{\rm Pl}$.

Since the temperature of a black hole is proportional to its surface gravity, an *exactly* extremal black hole does not produce any *thermal* Hawking radiation. Indeed, this is required by the cosmic censorship conjecture: any neutral emission would reduce

the mass of the black hole without a commensurate reduction in charge, leading the black hole to be super-extremal. The temperature decreases smoothly in the near-extremal regime (see Eq. (3.4)). At first glance, this seems to provide a potential mechanism for the stability of CPRs, completely apart from Planck-scale physics: if $T = 0$, it is tempting to conclude that the black hole does not radiate. In this case, we could suppose the charge of the black hole originally fluctuates rapidly, following a distribution with some width. Then, as the mass decreases, the black hole might become extremal, stalling the evaporation process. There are two problems with this idea: first, extremal black holes are not necessarily stable, even though the Hawking temperature vanishes. Secondly, even if a near-extremal geometry stabilizes the black hole, this can only manifest if the black hole is extraordinarily close to extremality. Note that we only consider near-extremal states due to the third law of black hole thermodynamics, which states that a black hole cannot evolve to an exactly-extremal state. We will discuss each of these issues in turn.

Regarding stability, the thermodynamics of near-extremal black holes is still not fully understood. They certainly cannot emit neutral particles if the cosmic censorship conjecture holds. But all known charged particles satisfy $|q| > m$, so it may be possible for an extremal black hole to athermally emit charged particles. If not, then there exists an infinite set of stable extremal states, one for each charge number. The existence of such an infinite tower of stable states without any corresponding gauge symmetry is believed to be incompatible with string-theoretic UV completions [146]. This is a major motivation for the weak gravity conjecture [WGC, 147], which requires that for *any* gauge symmetry, there exists a state with charge $|q| > m$. The WGC

was posited in part to allow extremal black holes to be unstable, precisely in order to prevent the appearance of such an infinite tower in the spectrum of a theory.

Note that athermal emission in the extremal state is consistent with the emission rate in Eq. (3.6): in the limit of small T, the rate vanishes if the argument of the exponential is positive. On the other hand, for a sufficiently large and negative chemical potential $q\Phi$, the argument of the exponential is negative, and the low-temperature limit is proportional to $\Gamma_{\ell,m,p}(\omega, T, q\Phi)$. Recall that $\Phi = -Q/r_+$ in our case, and for an extremal RN black hole, $r_+ = M = Q$. Then the argument of the exponential is negative as long as $q\Phi < -\omega$, so the requirement for the black hole to produce a particle of charge q and mass m is exactly that $|q| > m$. Thus, resolving the question of stability depends on the calculation of the coefficient $\Gamma_{\ell,m,p}$. Heuristically, the black hole could still radiate because the electric field at the surface of the black hole remains strong, so particle-antiparticle pairs could still be separated by the Schwinger mechanism, even though the surface gravity vanishes. Several authors [148–150] discuss emissions from extremal and near-extremal states in more detail.

Regarding the physicality of the extremal state, the third law of black hole thermodynamics is analogous to the ordinary third law of thermodynamics, which implies that no statistical system can attain a temperature of exactly zero (see e.g. Ref. [151] for an extensive discussion). The applicability of statistical laws to small black holes remains an active area of research [see e.g. 152], but the situation is readily understood heuristically: as a black hole approaches extremality, its temperature decreases according to Eq. (3.4), and its emission rate decreases. This may stall evaporation temporarily, but the black hole cannot evaporate to an extremal state by this mechanism. The next

question, then, is whether such stalling is significant on cosmological timescales. The stalling of evaporation near extremality was studied numerically for large black holes by Ref. [153]. They conclude that although the third law is satisfied at all times, the reduction of the emission rate in the near-extremal limit can prolong the black hole lifetime considerably.

However, our scenario is substantially different in that the evaporation cannot be treated smoothly: since we are interested in a phase of black hole evolution which involves extremely high temperatures, the black hole can lose charge by emission of charged leptons on timescales that may be relevant to mass loss. The appropriate analogue of the analysis of Ref. [153] would be to solve a system of differential equations for the evolution of the joint probability distribution of mass and charge, $\mathcal{P}(M, Q)$, treating charge as discrete—that is, a system of the form

$$\frac{\mathrm{d}\mathcal{P}(M,Q)}{\mathrm{d}t} = -\sum_{Q'\in\mathbb{Z}e}\int_0^M \mathrm{d}M'\,\mathcal{R}(M\to M';\ Q\to Q') + \sum_{Q'\in\mathbb{Z}e}\int_M^\infty \mathrm{d}M'\,\mathcal{P}(M',Q')\mathcal{R}(M'\to M;\ Q'\to Q), \quad (3.12)$$

where $\mathcal{R}(M \to M';\ Q \to Q')$ gives the differential rate for black hole of mass M and charge Q to decay to a black hole of mass M' and charge Q'. Further, while Ref. [153] use an approximate form of the emission rate which is valid only for sufficiently massive black holes, $\mathcal{R}$ must be the full rate in our case, as computed from Eq. (3.6). Under these conditions, this system is difficult to solve numerically, especially since any numerical evaluation must be sensitive to extremely small values of $M - Q$.

Thus, a simpler estimate of the near-extremal behavior is called for. First, observe that in our case, in order for the charge to stabilize, we must at least have $T \lesssim$

m_e, or else thermal emissions alone will cause charge fluctuations. But in our scenario, black holes only have an appreciable probability of being within $1e$ of extremality when $M \sim M_{\rm Pl}$—and even then, the probability is $\mathcal{O}(1\%)$ if we naively extrapolate the distribution of Eq. (3.5). For a black hole with $M \sim M_{\rm Pl}$, to have a temperature of $T < m_e$, the charge-to-mass ratio must be extremely close to unity. For fixed Q, the mass $M^{(e)}_{\rm min}$ at which thermal electron production is frozen out ($T = m_e$) is given by

$$M^{(e)}_{\rm min}(Q) = Q + 2\pi^2 Q^3 m_e^2 + \mathcal{O}(m_e^3), \tag{3.13}$$

where we recall that, in Planck units, $m_e \ll 1$. If $Q = 12e$, then we must have $M^{(e)}_{\rm min} - Q \simeq 4 \times 10^{-44}$, so the hole's mass must depart from its extremal value by no more than $\delta M \simeq 5 \times 10^{-16}\,\text{eV}$. This is an extremely small "target" to hit: outside of this region, the power of emission in charged particles is comparable to that in neutral particles, so the charge is likely to fluctuate on the same timescale that governs the shrinking of the mass.

This alone does not make it impossible for the hole to enter the near-extremal regime, but it does make this unlikely to take place during a typical evaporation. To estimate the probability, we give the following argument: the hole is most likely to be near-extremal when the mass is lowest. Thus, suppose that the hole passes through the mass range $12e < M < M^{(e)}_{\rm min}(12e)$ at some time. What is the probability that the charge takes the value $12e$ at some moment during this interval, i.e., before the hole evaporates further to $M < 12e$? Since $M^{(e)}_{\rm min}(12e) - 12e < m_e$, the black hole charge cannot change without the mass dropping to $M < 12e$. This means that the black hole must already have charge $12e$ when $M = M^{(e)}_{\rm min}(12e)$. If the spontaneous charge

distribution for such a small hole is at all similar to that of its larger counterparts, which have rms charges of order $6e$, this situation is quite unlikely, happening with a probability of a few percent.

However, there are three scenarios in which extremality may be significant. First, suppose that evaporation of a neutral black hole does not stall, or that it stalls at a scale $M_{\rm CPR} \ll M_{\rm Pl}$, as is possible in the context of the generalized uncertainty principle [154]. In this case, the black hole may enter a mass regime where the maximal charge is comparable to or smaller than the width of the spontaneous charge distribution, and the chance of freezing out charged leptons can no longer be neglected. Note that in the sub-Planckian regime, evaporation can behave very differently. For instance, in the case of Ref. [154], $T \sim M$ for $M < M_{\rm Pl}$, which may substantially modify the results of our analysis. Second, if a sufficiently large population of near-Planck-scale black holes is produced in the early universe, and only near-extremal holes are stable, then even a tiny fraction of this initial population could account for a significant fraction of dark matter. The hot evaporation products of the remainder would redshift away like radiation. Third, any charge associated with a new U(1) symmetry would influence the metric in the same way as electric charge. In particular, the fine structure constant associated with the new symmetry could be much smaller than $\alpha_{\rm EM}$, smoothing out the discrete spontaneous charge distribution. Alternatively, the lightest charged state of the new symmetry could be much more massive than the electron, reducing its athermal production rate, and making near-extremal states long-lived [155].

For the remainder of this chapter, we will not need to assume that CPRs are extremal. However, we note that if even a small fraction of the initial PBH population

does evolve to a state sufficiently near extremality to freeze out thermal lepton emissions, and if this state is stable to athermal emissions as well, then this makes the CPR scenario viable even if PBH evaporate completely. In this case, the dark matter density is fixed by the initial number of evaporating PBH and the fraction that freeze, and all surviving PBH are near-extremal. However, we note that if extremality is the only stabilizing mechanism, and if these objects accrete opposite charge at any time, then they are unlikely to stabilize again into a charged state. Instead, they will completely evaporate. If such destabilization events are still ongoing in the late universe, this will result in potentially observable bursts of high-energy particles.

3.2.4 Cosmic history of CPRs

The existence of CPRs today requires a primordial origin for the original generation of black holes. In this section, we examine the feasibility of producing a detectable population of CPRs through such a mechanism, starting with their formation in the early universe.

In the simplest scenario, the progenitors of CPRs are produced near the Planck scale with a monochromatic mass function. However, black holes need not be dominantly produced near the Planck scale in order to leave behind CPRs today. Multi-modal mass functions have been invoked to account for all of dark matter while avoiding constraints. More generally, primordial black holes can be produced with an extended mass function, e.g. with a lognormal or power-law mass function [29], and such a broad initial spectrum will typically produce a small abundance of relics by evaporation. Any primordial black hole produced with a mass below $M_{\rm evap} \sim 5 \times 10^{14}\,{\rm g}$ will evaporate to

the Planck scale by the present day [140]. Further, any black hole produced with a mass below $\sim 10^{16}\,\mathrm{g}$ has an initial temperature of order 1 MeV, and thus produces charged leptons rapidly enough to acquire a spontaneous charge, even though it will continue to evaporate actively today. As such, even if they are dominantly produced at even higher mass scales, we generically expect to find a low-mass tail that evaporates to the Planck scale—and might leave behind CPRs.

There are two major constraints on such a scenario: first, even if a relic is left behind at the end, the total radiation produced by evaporating black holes is constrained by CMB observables and light element ratios [28]. Second, if the CPRs originate from a population of black holes with a component above M_{evap}, then they do not constitute all of dark matter, and may indeed account only for a small fraction. In this case, other probes can constrain the population at higher masses. To investigate the plausibility of such a scenario, we suppose that all dark matter is in the form of PBHs, and suppose that the black holes above M_{evap} do not lose a significant amount of their mass. Then, given the initial mass function $\mathrm{d}n/\mathrm{d}M$, the density of CPRs today is given by

$$\frac{\Omega_{\mathrm{CPR}}}{\Omega_{\mathrm{DM}}} \approx \frac{M_{\mathrm{Pl}} \int_{M_{\mathrm{Pl}}}^{M_{\mathrm{evap}}} \mathrm{d}M \, \frac{\mathrm{d}n}{\mathrm{d}M}}{M_{\mathrm{Pl}} \int_{M_{\mathrm{Pl}}}^{M_{\mathrm{evap}}} \mathrm{d}M \, \frac{\mathrm{d}n}{\mathrm{d}M} + \int_{M_{\mathrm{evap}}}^{\infty} \mathrm{d}M \, M \frac{\mathrm{d}n}{\mathrm{d}M}}. \tag{3.14}$$

For example, as a toy model, consider a power-law mass function $M\,\mathrm{d}N/\mathrm{d}M \propto M^{\gamma-1}$. Assuming $\gamma < 0$, the mass fraction in CPRs is

$$\frac{\Omega_{\mathrm{CPR}}}{\Omega_{\mathrm{DM}}} \approx \frac{M_{\mathrm{Pl}} M_{\mathrm{evap}}^{\gamma-1} - M_{\mathrm{Pl}}^{\gamma}}{M_{\mathrm{evap}}^{\gamma-1} \left[M_{\mathrm{Pl}} - (1-1/\gamma) M_{\mathrm{evap}}\right] - M_{\mathrm{Pl}}^{\gamma}}. \tag{3.15}$$

A CPR fraction $f \sim 1$ is produced when $\gamma \lesssim -0.1$, whereas e.g. $f \sim 10^{-2}$ for $\gamma \sim -10^{-2}$.

Next, we must consider the survival of such charged objects over cosmic time. It is unlikely that a Planck-scale black hole would neutralize by accretion of charged

particles, since the geometric cross section is extremely small. In other words, we expect the accretion rate to be suppressed by $M_{\rm Pl}^2$. But even if a CPR does accrete, the consequent increase in mass may restart the evaporation process, and in the low-mass regime, the emission power in charged particles is comparable to that in neutral particles. If we treat accretion as an excitation of the black hole remnant to a neutral state with a higher mass, this excited state can simply decay again to a charged state. Thus, we expect neutralization of the CPR population to take place very slowly if it happens at all.

A more plausible scenario is that the black hole forms bound states with particles of opposite sign. For positively-charged CPRs, electron capture would take place alongside the same process for hydrogen atoms during the epoch of recombination. For negatively-charged CPRs, capture of protons is even more energetically favorable, since the energy of the bound state scales with the reduced mass. At first glance, this could interfere with detection: a net-neutral bound state may be invisible to a terrestrial detector. We will show in Section 3.3.3 that such objects are still detectable. But it is still important to understand the typical charge state of CPRs far from Earth, in part because accretion of a bound charge might be possible on cosmological timescales. Thus, we now examine their ionization history.

It is easily checked that reionization proceeds almost identically for CPRs as for hydrogen, even for negatively-charged CPRs with bound protons. In this case, following Ref. [156], the CPR population will be fully ionized when emission rate of

ionizing photons per unit volume matches the rate of recombinations, that is,

$$\dot{n}_\gamma \gtrsim \frac{n_{\rm CPR}}{t_{\rm rec}}, \tag{3.16}$$

where $n_{\rm CPR}$ is the number density of CPRs and $t_{\rm rec}$ is the characteristic timescale for recombination with a free proton. We can estimate this timescale as $t_{\rm rec} \simeq 1/(n_{\rm p}\alpha_A)$, where α_A is the recombination coefficient. We calculate the recombination coefficient α_A for a bound proton following Ref. [157][1]. At a typical nebular temperature of $10^4\,{\rm K}$, we find that $\alpha_A \simeq 2\times10^{-21}\,{\rm cm}^3/{\rm s}$, versus $\alpha_A \simeq 4\times10^{-13}\,{\rm cm}^3/{\rm s}$ for hydrogen.

The much smaller recombination coefficient and number density imply that reionization should proceed much more rapidly for CPR atoms than for hydrogen: if CPRs at the Planck mass constitute all of dark matter, then their number density is still lower by a factor of $\sim10^{-19}$ compared to that of protons, and $t_{\rm rec}$ is reduced by $\sim10^{-8}$ compared to hydrogen. We can also account for the fact that the binding energy of a proton with a negatively-charged CPR is $\sim25\,{\rm keV}$ in the ground state, so the only ionizing photons are those with wavelength $\lambda \lesssim 0.5\,{\rm \AA}$, compared with $\lambda < 911\,{\rm \AA}$ for hydrogen. If we extrapolate the quasar spectrum of Ref. [156] to small wavelengths, where $L(\lambda) \sim \lambda^{1.8}$, the emission rate is suppressed by a factor of $\sim10^{-6}$ compared with photons that ionize hydrogen. This suppression is insignificant compared to the changes in the number density and recombination timescale, so we conclude that CPRs will still reionize much more efficiently than hydrogen.

[1]Note that Ref. [157] contains two typographical errors in eqs. (2) and (3). Correct versions of these equations can be found in Ref. [158].

3.3 Detecting charged black holes terrestrially

In this section, we analyze the interaction of CPRs with matter, and investigate mechanisms for direct detection. For the purposes of our calculations, we assume that CPRs account for a fraction $f_{\rm CPR}$ of dark matter by mass (i.e., $f_{\rm CPR} = \Omega_{\rm CPR}/\Omega_{\rm DM}$). We assume that evaporation is halted at a mass $M_{\rm CPR}$ which we allow to differ from $M_{\rm Pl}$, and we assume that $M_{\rm CPR}$ is independent of the black hole charge. However, we require that $M_{\rm CPR} \geq e \approx M_{\rm Pl}/11.7$, since a (classical) black hole with a mass below this threshold cannot have even one elementary charge without being super-extremal, which we prohibit. Note that in some models, the relic mass is far below the Planck scale, e.g. as in Ref. [154]. Such relics may yet evade the constraints we set here.

Even when $f_{\rm CPR} = 1$ and $M_{\rm CPR}$ is minimal, direct detection of such charged CPRs is limited primarily by the flux of these objects: at such high masses, the number density of CPRs is much lower than that of typical particle dark matter candidates. The flux is $\Phi_{\rm CPR} \simeq (\rho_{\rm DM}/M_{\rm CPR})v_{\rm DM}$, so taking $v_{\rm DM} = 300\,{\rm km/s}$ and $\rho_{\rm DM} = 0.3\,{\rm GeV/cm^3}$ gives the event rate as

$$N = 0.23\,{\rm yr}^{-1} f_{\rm CPR} \left(\frac{M_{\rm CPR}}{M_{\rm Pl}}\right)^{-1} \left(\frac{A_{\rm detector}}{1\,{\rm m}^2}\right) \mathcal{E}_{\rm detector} \tag{3.17}$$

where $\mathcal{E}_{\rm detector}$ is the fraction of CPRs that will register an event in the detector. Since we are considering electrically-charged objects, there are detectors for which $\mathcal{E}_{\rm detector} \sim 1$, as we will detail in the following subsection. However, for CPRs, we expect to have $M_{\rm CPR}/M_{\rm Pl} \sim 1$. Thus, in order to achieve a detection rate of $1\,{\rm yr}^{-1}$, a detector must have $A_{\rm detector} \gtrsim 4.3\,{\rm m}^2$.

It is clear from this calculation that a typical dark matter direct detection

experiment is unlikely to encounter more than one such object during its operational lifetime. However, as we will discuss in the next subsection, the passage of even one CPR through a detector has the potential to produce an extremely clear signature.

3.3.1 Signatures of CPR transits

In this section, we discuss the interactions of CPRs with particle detectors. The interactions of a CPR with matter are similar to those of slow-moving heavy ions. A CPR with $M_{\rm CPR} \sim M_{\rm Pl}$ does not slow down appreciably during its transit through a detector: its kinetic energy is $\sim \frac{1}{2} M_{\rm Pl} (300\,{\rm km/s})^2 \approx 6 \times 10^{21}\,{\rm eV}$. This is to be compared to atomic binding energies, which are typically of order 1 eV. Indeed, a CPR is so massive that an object with a downward trajectory will *gain* $\sim 130\,{\rm eV/\AA}$ from gravitational acceleration. Deflection is also negligible, even in a strong electromagnetic field, so a CPR will deposit energy along a very straight track. In this respect, a CPR transit can be distinguished from any standard background: energy will be deposited at a constant density at a low speed ($\sim 0.3\,{\rm m/\mu s}$) along a straight track.

While detection prospects for CPRs in any given experiment must ultimately be studied with more detailed modeling, we can still use general methods to estimate signatures of a CPR transit. As a CPR transits through a detector, it loses energy via Coulomb interactions with the target electrons and nuclei. The transferred energy may be detectable in the form of heat, ionization, or scintillation. Each of these signatures scales with the energy deposited during the transit, typically expressed in terms of the "stopping power", that is, the energy loss of the incident particle per unit distance traversed through the material. We will identify the stopping power, ionization

yield, and scintillation yield for particular experimental configurations using numerical simulations, but we begin with a simple estimate of the stopping power.

Regardless of whether they constitute a significant fraction of dark matter, CPRs should be highly non-relativistic ($\beta \sim 10^{-3}$). Since this is slower than the outer electrons of the target atoms, the calculation of stopping power in this regime differs greatly from the relativistic regime. In particular, the characteristics of CPR interactions with matter are similar to those of heavy ions. The stopping power for $\beta < 0.05$ is well-described by Lindhard-Scharff-Schiott theory [159], in which it is linear in the velocity and charge of the incident particle, and has no dependence on its mass. Thus, we can estimate the stopping power in our scenario by comparing to empirical results for the stopping of non-relativistic muons. In copper, the stopping power per unit target density for incident muons with $\beta = 10^{-3}$ has been measured as $(\mathrm{d}E/\mathrm{d}x)/\rho_{\mathrm{target}} \sim 30\,\mathrm{MeV}/(\mathrm{g/cm^2})$ [fig. 23.1 of 160]. For our purposes, it is useful to quantify the stopping power in transit through Earth and in semiconductor detectors, and silicon is a representative material for both. The linear stopping power in silicon is then

$$\frac{\mathrm{d}E}{\mathrm{d}x} \sim 70\,\mathrm{MeV/cm}. \tag{3.18}$$

Note that since we may have $\langle |Z_{\mathrm{CPR}}| \rangle > 1$, this is a conservative estimate.

This stopping power is likely too small to register in a typical calorimeter. Thus, if a CPR is to be detectable, it must produce an ionization or scintillation signature. Given the stopping power, the ionization yield depends again on the particle velocity and material properties. The stopping power has two components, corresponding to interactions with electrons (electronic stopping power) and with nuclei (nuclear

stopping power). For highly relativistic particles, nuclear stopping is typically negligible compared to electronic stopping, but this is not the case in the highly non-relativistic regime. Indeed, the maximum energy that a CPR can transfer to a recoiling electron with mass m_e is given by

$$\Delta E_{\max} = 2m_e v_{\mathrm{DM}}^2 \approx 1\,\mathrm{eV}. \tag{3.19}$$

This maximum energy transfer is smaller than typical ionization energies, so direct ionization via electronic interactions is not likely to be efficient. On the other hand, the maximum energy transfer in a recoil with a nucleus of mass number A is

$$\Delta E_{\max} = 2m_A v_{\mathrm{DM}}^2 = 186\,\mathrm{keV} \times \left(\frac{A}{100}\right)\left(\frac{v_{\mathrm{DM}}}{300\,\mathrm{km/s}}\right)^2, \tag{3.20}$$

much higher than that of electronic recoils. Thus, we expect interactions between the CPR and nuclei to dominate in a typical detector.

While ionization and scintillation are most efficiently produced by electronic interactions, nuclear stopping can also produce these signals, since the recoil energy of the nucleus can be partially transferred to bound electrons. The attendant loss of efficiency, or *quenching*, is expressed via the ratio of yields from nuclear and electronic scattering. Such *quenching factors* are dependent on the target material, and values are typically measured experimentally [see e.g. 161, 162]. To estimate the stopping power, ionization yield, and scintillation yield due to nuclear recoils, we used the Monte Carlo code `SRIM` [163], which simulates the passage of ions through matter. We simulated "hydrogen" ions with `SRIM`'s maximum allowable particle mass of 10 000 u, where u is the atomic mass unit, and with a velocity of 300 km/s. This corresponds to a kinetic energy of 4.7 MeV, still very large compared to the binding energies relevant for the

interaction. In any case, we performed our simulations in 1 micron-thick layers of detector material, so the change in the momenta of the simulated ions was negligible. In the following section, we discuss the results of our simulations and the implications for different experimental modalities.

3.3.2 Detection mechanisms

Here we briefly survey several detector technologies to evaluate whether they would be suitable for detecting CPRs.

Bubble chambers. Ref. [164] noted that charged Planck-mass black holes would leave tracks in bubble chambers, and speculated that unidentified tracks in previous experiments could be explained by the presence of these objects. A bubble will form in superheated fluid if the energy deposited within a critical radius exceeds a given threshold energy. For concreteness, we consider the response of the PICO experiment [165, 166], whose bubble chamber has a threshold energy of 3.3 keV for a critical radius $r_c = 2 \times 10^{-8}\,\mathrm{m}$.

From the `SRIM` output, we integrated the energy deposited in a sliding window of width r_c, and found that the deposited energy was sufficient to form a track with a linear bubble density of $\sim 10^5\,\mathrm{m}^{-1}$. This is not surprising: PICO is highly sensitive to α decays, which generate nuclear recoils of similar energy to those from CPRs. Further, a straight bubble track would not be expected from weakly interacting massive particles (WIMPs), which are expected to only interact once in the detector volume, and would be distinct from background signals from neutrons, which leave jagged tracks. However, even large bubble chambers have insufficient area to place strong constraints on the flux

of CPRs. The proposed 500L version of PICO [167] would require several decades of continuous exposure to place any constraint on the abundance of CPRs.

Atmospheric fluorescence detectors. Ultra-high energy cosmic rays incident on the atmosphere generate hadronic and electromagnetic showers which ionize nitrogen molecules that subsequently fluoresce, emitting visible light. Arrays of photomultiplier tubes, such as the High Resolution Fly's Eye (HiRes) observatory [168], are capable of detecting this fluorescence and reconstructing the track of the cosmic ray. Since HiRes can detect emissions over an area of order $1\,\mathrm{km}^2$, this seems like an attractive way to detect a particle with a very low flux.

We evaluated the potential of atmospheric fluorescence detectors to observe the energy deposition from the passage of a CPR. In dry air at sea level, our `SRIM` calculation yielded an energy deposition of $\sim 12\,\mathrm{MeV/m}$. Assuming an average of $3.5\,\mathrm{eV}$/photon and a (generous) fluorescence efficiency of 5% [169], the photon yield is about $5 \times 10^4\,\mu\mathrm{s}^{-1}$ for a relic moving at $300\,\mathrm{km/s}$. At the surface, background light from stars, light pollution, and other sources is about $5 \times 10^5\,\mathrm{m}^{-2}\,\mathrm{sr}^{-1}\,\mu\mathrm{s}^{-1}$. Given that a typical photomultiplier tube in HiRes observes 1 square degree with a $5\,\mathrm{m}^2$ mirror, the background event rate is $\sim 100\,\mu\mathrm{s}^{-1}$. A CPR must then pass within a few meters of the mirror for the signal to overcome background photons, reducing the effective area of this class of detectors considerably. Atmospheric detectors are thus unlikely to place strong constraints on CPRs even in a decade of operating time.

Cherenkov detectors. An attractive possibility is to search for CPRs with neutrino detectors (e.g. IceCube [170], Super-Kamiokande [171]) or imaging atmospheric Cherenkov

detectors (e.g. VERITAS [172], HAWC [173]), since they have extremely large ($\sim$km^2) effective areas. However, regardless of their origin, we expect CPRs to be highly non-relativistic. Thus, we do not expect any Cherenkov radiation to be emitted as they traverse these detector media. Instead, light would be produced only by scintillation and ionization processes. Such a signal is distinguishable from those produced by relativistic particles in that light would be emitted isotropically from the track rather than in the cone shape characteristic of Cherenkov light. But a CPR transit would be extremely slow, taking place on the order of several ms, compared to typical targets observed over a duration of order µs. Thus, even if the light from ionization is observable, detecting it would require a non-trivial triggering mechanism.

Dark matter searches. CPRs are highly penetrating and ionizing, so a CPR transit would leave a distinct signal in semiconductor and liquid xenon detectors, including dark matter direct detection experiments and neutrinoless double beta decay searches [174]. Since CPRs are negligibly slowed by their interactions with materials, a CPR would produce a straight track in a detector with a transit time of order 1 µs in a typical experiment, a unique signature. Further, while the thick layer of earth (overburden) covering these experiments significantly reduces background, it does not affect the flux of CPRs. However, the flux itself is small, and these detectors typically have small cross-sectional areas. Even next-generation experiments are unlikely to detect more than a few CPRs in ten years, so they cannot produce significant constraints on the CPR population.

Monopole searches. Magnetic monopoles are expected to be found at very large masses,

and due to their high magnetic charges, monopole transits share some characteristics with CPR transits. As such, it is possible that monopole searches can impose constraints on CPRs as well. In Section 3.4.1, we investigate this possibility in detail in the context of the MACRO experiment [175].

Liquid argon detectors. Liquid argon time-projection chambers [176] have recently been employed in several neutrino experiments [177–183], some of which have much larger cross-sectional area than is typical for dark matter experiments. Since the transit of a CPR has a unique signature, backgrounds are of no concern, so existing neutrino experiments have the potential to detect CPRs. We expect that liquid argon detectors can be used to place strong constraints on the CPR population, and we elaborate on detection prospects in the ICARUS experiment in Section 3.4.2.

Paleo-detectors. Recently, Ref. [184] proposed the detection of WIMP dark matter through small tracks left in ancient minerals by dark matter recoils [see also 185, 186]. The major strength of such "paleo-detectors" is their exposure time, of order 10^9 yr. This is uniquely well-suited to our case of interest, where detection is primarily limited by flux rather than detector efficiency. Moreover, we need not await the results of WIMP searches in ancient minerals: the recent paleo-detector proposals are extensions of similar searches already performed e.g. by Ref. [187] to constrain the flux of supermassive magnetic monopoles. These results should already be applicable to CPRs, and we discuss them in Section 3.4.3.

3.3.3 Detection of CPR "atoms"

In Section 3.2.4, we argued that if CPRs form net-neutral bound states with electrons or atomic nuclei, they should be fully reionized by the present day. This applies to the astrophysical population of CPRs, but not necessarily to the terrestrial population relevant for direct detection. While enroute to a detector, an originally bare CPR may recombine with electrons or atomic nuclei to form a bound CPR "atom" (or "neutraCHAMP", in the language of Ref. [133]).

First we consider the possibility of recombination in the atmosphere, following Ref. [133]. In this case, positively-charged CPR with bound electrons will be rapidly ionized by solar UV radiation. The probability of recombination with an atmospheric electron is given by

$$P_{\rm rec} = \sigma_{\rm rec}\rho_{e^-}L, \tag{3.21}$$

where $\sigma_{\rm rec}$ is the recombination cross section, ρ_{e^-} is the free electron density, and L is the depth of the atmosphere. We take $\sigma_{\rm rec} = 10^{-5}\pi a_0^2$, $L = 100\,{\rm km}$, and conservatively estimate $\rho_{e^-} = 10^6\,{\rm cm}^{-3}$, which yields $P_{\rm rec} \approx 10^{-3}$. Negatively-charged CPRs bound to a nucleus of atomic number A will not be ionized by solar UV, as the binding energy is $\sim Z^2\,(25\,{\rm keV})$. In this case, Ref. [133] calculated the mean free path for recombination with a ^{14}N nucleus to be $\lambda_{\rm rec} = 4\times10^{10}\,{\rm g\,cm^{-2}}\beta^2/\rho_{\rm atm} \gtrsim 300\,{\rm km}$, corresponding to a recombination cross section of $\sigma_{\rm rec} \simeq 6\times10^{-34}\,{\rm cm}^2$.

However, it is not clear that the latter cross section is applicable to interactions with atmospheric gas. Experimentally, cross sections for charge transfer onto slow ions in gaseous CO and CO_2 are much larger, of order $10^{-16}\,{\rm cm}^2$ [188]. The corresponding

mean free path in the atmosphere is microscopic. Additionally, in passing through solid earth (overburden) enroute to a detector, a bare CPR would be very likely to acquire bound charge: experiments and simulations show that the recombination timescale of molecular hydrogen ions in carbon is $\mathcal{O}(10\,\mathrm{fs})$ [189–191]. Thus, we must consider the possibility of detecting CPRs bound into neutral "atoms".

The stopping of partially-ionized or neutral atoms in materials differs from that of bare ions in that the screening effect of bound charges substantially modifies the potential. This generally diminishes but does not eliminate the stopping power. Numerous efforts have been made to model the stopping of neutral atoms [192–197], and much work in particular has been devoted to the development of an "effective charge" for such objects. In many circumstances, the effective charge of a neutral atom of atomic number Z_1 is $Z_{\mathrm{eff}} \simeq 0.7 Z_1$ [193]. However, the effective charge is generally a function of parameters beyond the nuclear charge and ionization state, including the atomic number Z_2 of the target and the velocity of the projectile [196].

We are interested in the low-velocity, low-Z_1 regime for a wide variety of target materials—from xenon ($Z_2 = 54$) to argon ($Z_2 = 18$) to the silicates, chlorides, and sulfates found in paleo-detectors. We typically have $Z_1 \ll Z_2$, which reduces the impact of screening and enhances the stopping power [196]. However, these results generally hold for velocities much greater than the Bohr velocity, $e^2/(4\pi\varepsilon_0\hbar) \simeq 7 \times 10^{-3} c$. Our fiducial velocity is $\sim 10^{-3} c$, and the effective charge may be different in this limit. Ref. [193] find that the effective charge increases at low velocities, so we expect that the aforementioned effective charge of $Z_{\mathrm{eff}} \simeq 0.7 Z_1$ is an estimate of a lower bound rather than an upper bound.

Since the stopping power of bare CPRs is very large for detection purposes, screening effects of this order have no qualitative impact on detection signatures. Thus, in subsequent calculations, we will ignore the distinction between bare and atomic CPRs. In particular, we will assume that the ionization and scintillation yields are comparable for the two projectiles. Note that we do not consider excitation or ionization of the CPR atom, as these effects would tend to increase the stopping power still further.

3.4 Constraints and future prospects

3.4.1 MACRO experiment

The MACRO experiment [175] performed a search for GUT-scale magnetic monopoles, placing an upper limit on the monopole flux at $5.5 \times 10^{-4}\,\mathrm{m}^{-2}\,\mathrm{yr}^{-1}$ (90% CL). This is significantly lower than the flux we estimate in Eq. (3.17). Since monopole transits are similar in many respects to CPR transits, we investigate the extent to which this bound can be applied to the CPR population.

The MACRO experiment consists of six independent analyses with different experimental properties. At velocities $\beta \sim 10^{-3}$, there are two applicable constraints: the Wave Form Digitizer (WFD) analysis of the liquid scintillator system, and an analysis of streamer tubes filled with a mixture of helium and n-pentane. The streamer tube analysis relies on the Drell effect [198]: an incident magnetic monopole excites helium atoms, which then ionize n-pentane molecules. The Drell effect is specific to magnetic monopoles, and we do not expect a comparable phenomenon to occur in the case of CPR transits unless they are also magnetically charged. Thus, we turn our attention

to the WFD search.

The WFD analysis was based on a data from a liquid scintillator detector equipped with photomultiplier tubes. The trigger was designed for slowly-moving monopoles, and was sensitive to photoelectrons emitted in sequence over several microseconds. Transient signals with a duration below 100 ns were discarded. The WFD search set an upper limit to the flux at $9.9 \times 10^{-4}\,\mathrm{m}^{-2}\,\mathrm{yr}^{-1}$ (90% CL). If we assume that the search was fully sensitive to CPRs, this corresponds to a bound on the charged fraction of $f \lesssim 0.5\%$, sufficient to rule out CPRs as the dominant component of dark matter. A dedicated search in comparable hardware would certainly be capable of establishing a bound at least this strong.

3.4.2 ICARUS experiment

We now restrict our attention to liquid argon detectors, of the type pioneered by Ref. [176]. For concreteness, we consider the ICARUS detector, which operated for three years at Gran Sasso Laboratory [199] and is currently being installed as a short baseline neutrino detector at Fermilab [182].

ICARUS consists of two chambers, each containing approximately 480 tons of liquid argon. The chambers are equipped with photomultiplier tubes and wire cages, so they are capable of detecting both scintillation light [200] and ionization [201] from keV-scale nuclear recoils. For the majority of the data collection at Gran Sasso, ICARUS was only triggered in time coincidence with the CERN neutrino beam, and therefore would likely have ignored any potential signal from a CPR transit.

The mean scintillation efficiency for nuclear recoils in liquid argon is about 0.25

[200], corresponding to a scintillation yield of approximately 13 photons per keV of deposited energy. We simulated the passage of a CPR through liquid argon with `SRIM` and found a stopping power per unit target density of approximately $93\,\mathrm{MeV}/(\mathrm{g/cm^2})$. Even if we only include recoils above $10\,\mathrm{keV}$, the resulting scintillation yield is $\sim 7 \times 10^5\,\mathrm{photons/cm}$. Using a conservative estimate of $6\,\mathrm{e^-/keV}$ for the secondary ionization yield [201], we find that the transit yields $\sim 3 \times 10^5\,\mathrm{e^-/cm}$. Both signals are well above the detection thresholds of ICARUS' readout electronics. Indeed, these signals are much more significant than the scintillation and ionization signals from minimum-ionizing particles like muons, which have a stopping power of about $1.5\,\mathrm{MeV}/(\mathrm{g/cm^2})$.

The combination of high scintillation yields, high ionization yields, and a long crossing time (several µs) in the detector would be a smoking-gun signature of a CPR transit. A dedicated search should be able to use these factors to discriminate against cosmic ray backgrounds, even with ICARUS at the surface. In Fig. 3.2, we show the expected upper limits on the CPR density that could be obtained from a dedicated search with ICARUS on various time scales. The limits are presented as a function of the fraction of dark matter in the form of detectable charged relics.

While we evaluate fiducial constraints using the parameters of the ICARUS experiment, there are several other similar liquid argon detectors that may be also be usable for CPR searches. Assuming that the readout electronics of each detector are sensitive to CPR transits, the maximum abundance of CPRs compatible with non-detection scales inversely with the effective area A_{eff} of the detector, defined as the cross-sectional area of the detector averaged over the arrival direction. For a rectangular prism with side lengths L_i, we have $A_{\mathrm{eff}} = \frac{1}{2}(L_x L_y + L_x L_z + L_y L_z)$, and for a cylinder

ICARUS	ArgoNeuT	LArIAT	SBND	ProtoDUNE	DUNE	MicroBooNE
80.5 [177]	1.01 [178]	1.01 [179]	22.4 [180]	104 [181]	792 [182]	28.3 [183]

Table 3.1: Effective areas of current and future liquid argon detectors in m^2 (see text for details). The strength of constraints scales linearly with the detector area. Of these, ICARUS is the largest detector currently operational.

of radius r and length L, we have $A_{\rm eff} = \frac{1}{2}\pi r(L + r)$. We summarize the effective areas of various detectors in Table 3.1.

3.4.3 Paleo-detectors

We now examine the prospects for detecting CPRs with paleo-detectors [184–186]. For WIMP detection, the key prediction is the spectrum of lengths of the ionization tracks produced by recoiling nuclei. For our purposes, the CPR itself takes the role of the nucleus. Since a CPR is negligibly slowed even by macroscopic volumes of matter, an ionization track from a CPR transit will be extremely long. Recall that we are interested in typical kinetic energies of order $10^{21}\,$eV. If we take the typical stopping power in rock to be $\mathrm{d}E/\mathrm{d}x \sim 100\,\mathrm{MeV/cm}$, as in Eq. (3.18), then CPRs should pass through Earth entirely without losing more than a fraction $\sim 10^{-4}$ of their energy. In principle, the resulting tracks would cross the entire planet, although they would be disrupted over time by geological effects.

Furthermore, the exposure time is such that a paleo-detector with cross-sectional area A should have a number of ionization tracks given by

$$N \sim 200 \left(\frac{A}{1\,\mathrm{mm}^2}\right)\left(\frac{t_{\rm obs}}{10^9\,\mathrm{yr}}\right)\mathcal{E}. \tag{3.22}$$

where the efficiency $\mathcal{E}$ accounts for the probability of track production, track survival,

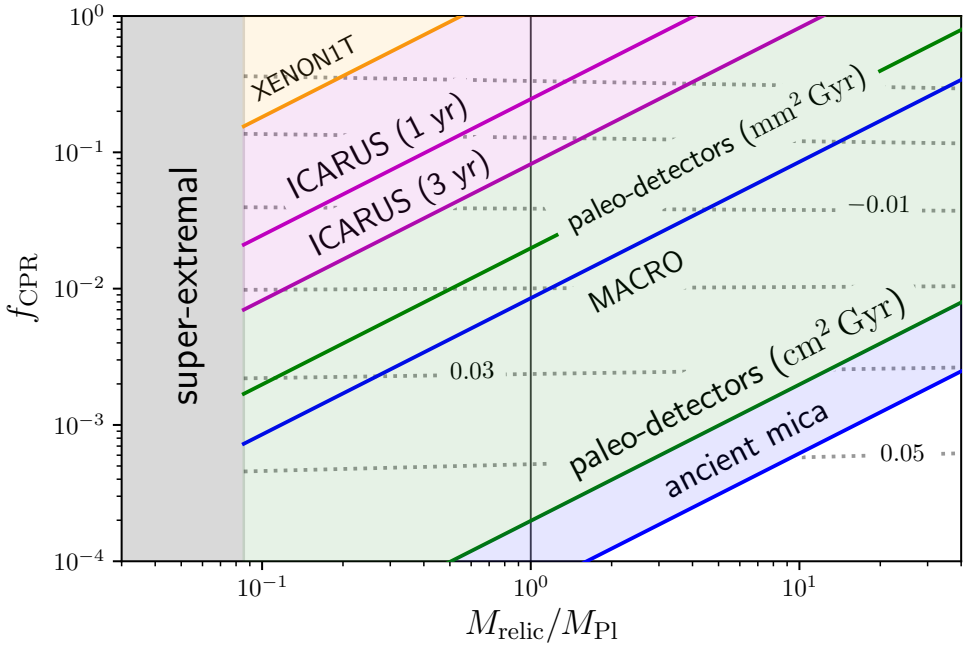

Figure 3.2: Projected 99% CL upper limit on the mass and density of CPRs with experiments of several classes. See text for details. **Orange:** a 10 yr exposure of XENON1T [3]. **Magenta:** solid: a 3 yr exposure of ICARUS. The dashed line shows a 1 yr exposure. **Green:** solid: estimated limits from a paleo-detector with $\mathcal{E} = 1$ and a 1 cm^2 Gyr exposure. The dashed line shows a 1 mm^2 Gyr exposure. **Blue:** strongest possible limits from monopole searches, including a direct search by MACRO and a search for tracks in ancient mica [187]. (Section 3.4.1). **Dotted gray:** relic fractions produced assuming an initial a power-law mass function with index γ. Contours step from $\gamma = -0.05$ to $\gamma = 0.05$ from top to bottom in increments of 0.02. **Shaded gray:** region prohibited by super-extremality for a charge of $1e$.

and track detection. If $\mathcal{E} \sim 1$, this extremely high track density means that paleo-detectors should be capable of placing very stringent constraints on the CPR population. For this reason, track searches in ancient minerals have already been used to constrain the flux of supermassive magnetic minerals, which have a similar detection signature [187]. If these searches are taken to be sufficiently sensitive to detect a CPR track, then we can already infer a limit of $\sim 4 \times 10^{-6}\,\mathrm{m}^{-2}\,\mathrm{yr}^{-1}$ on the flux of CPRs.

The disadvantage of paleo-detectors is that a particular track cannot be identified as a CPR in a small volume of material. Paleo-detectors do not directly measure the speed of the transit, which is an important experimental signature for our purposes. In a large piece of material undisturbed by geological processes, a CPR transit may be identifiable by the length of the track, but this may require additional technological development and other modifications to paleo-detector-based searches. Thus, paleo-detectors can constrain CPRs, but may not easily furnish a confirmed detection. Still, we show prospective constraints from paleo-detectors in Fig. 3.2.

3.5 Discussion

In the foregoing sections, we have argued that Planck-scale relics of evaporating primordial black holes may generically have charges of order e, and we have shown that plausible forms of the PBH mass spectrum lead to a significant CPR population today. The formation of these objects is inextricably connected to quantum gravity, and the process is sensitive to new physics at extremely high energies. We have further shown that if CPRs constitute a significant fraction of dark matter, then they can be detected

terrestrially. Indeed, not only are such objects detectable, but the detection signature would be a smoking gun with few alternative possibilities.

The implications of such a detection cannot be overstated. In addition to furnishing a direct detection of dark matter, this would confirm the PBH paradigm, providing great insight into the conditions of the early universe. An abundant population of such objects would furnish the first system for the direct laboratory study of gravity in the quantum regime. Of course, most immediately, even a single detection would establish that black holes do not evaporate completely, but leave behind a relic. Even non-detection may provide significant information: if dark matter is one day found to be composed of PBHs, and their mass function is established, the fraction of CPRs produced is easily calculated. Non-detection at that level would establish that evaporating black holes leave no relics, or that such relics cannot be charged.

With these objectives in mind, we have tentatively derived existing constraints from the MACRO experiment and ancient mineral searches, which already exclude $f_{\mathrm{CPR}} = 1$ at the 99% confidence level across the entire mass range we consider. However, there are numerous scenarios which predict a much smaller CPR fraction. If dark matter is composed mainly of PBHs at a higher mass scale, but produced with a broad mass spectrum, then a low-mass tail evaporates to the Planck scale, producing a smaller abundance of CPRs. Further, the uncertainties in the estimation of the charge fraction imply that only a small fraction of Planck-scale relics may be charged. Thus, there is ample motivation to search for smaller CPR fractions.

Fortunately, performing such a search requires no new equipment apart from possible modifications to experimental triggers. We have projected stringent constraints

from the ICARUS experiment, which can set a bound $f_{\rm CPR} \lesssim 10^{-2}$ if $M_{\rm CPR} \simeq M_{\rm Pl}$. Our projected constraints from paleo-detectors strengthen the bound to $f_{\rm CPR} \lesssim 10^{-4}$ at $M_{\rm Pl}$, and can constrain the CPR dark matter fraction at the per cent level even if $M_{\rm CPR}$ lies an order of magnitude above $M_{\rm Pl}$. Taken together, these bounds would be sufficient to exclude CPRs as a significant fraction of dark matter even if they lie at a different mass, or acquire spontaneous charge with a somewhat smaller probability. They are also extremely inexpensive to obtain.

3.6 Conclusions

Primordial black hole dark matter remains a viable and parsimonious dark matter candidate. If dark matter is in the form of Planck-scale relics, such objects would be effectively sterile with respect to the standard model if neutral. In this chapter, we have argued that such relics may in fact carry charges of order e, in which case dark matter could be composed largely of Charged Planck-scale Relics (CPRs). We have shown that CPR dark matter is detectable terrestrially, with initial constraints already set by the null results of monopole searches. Moreover, upcoming experiments can be used to conduct a much more sensitive search for CPRs with little or no modification. Even a single detection would come with significant implications for black hole physics, the behavior of gravity in the quantum regime, and the nature of dark matter.

The interpretation of non-detection is more subtle. Optimistically, null results can constrain the overall population of relic black holes, with implications for either the PBH mass function or the quantum gravity mechanisms that stabilize them. Realisti-

cally, however, the argument we present in this chapter only motivates the possibility of a substantial charged fraction—it is not a rigorous prediction. As such, we cannot immediately draw conclusions regarding the abundance of black hole relics in general.

However, up to some of the other uncertainties discussed in this chapter, it is conceivable that the charge distribution could be rigorously predicted within the context of a candidate quantum gravity theory. In this case, the abundance and charge distribution of relic black holes would become testable predictions of such a theory. Remarkably, as we have shown, we may already have experimental access to such a scenario. The constraints we draw on the abundance of CPRs may thus translate into constraints on the structure of physics at the Planck scale.

Part II

Black holes, gravitational waves, and cosmology

Invitation

At this point, let us recall the goal of this thesis: we wish to establish new probes of DM at new scales, away from the weak scale. In the previous part, we focused on the most massive candidates: we saw that significant parameter space remains open for primordial black holes (PBHs) as a DM candidate, and that both new and existing tools can explore important limits of this scenario. In this part, we use the phenomenology of PBHs as a starting point to explore windows below the weak scale. In particular, we will identify new connections between astronomical observables, terrestrial experiments, and low-mass particle species.

So far, we have stayed mostly quiet about one key observable associated with PBHs: gravitational waves, both as transients and as a cosmological background. This will be our entry point into new astronomical probes, and we will segue from PBHs to other applications of gravitational wave astronomy for new physics. After discussing uses of the stochastic gravitational wave background, we turn to other cosmological observables, and connect these to probes of particular dark matter scenarios.

This part makes a surprising transition, in the language of Fig. 0.1. In the previous part, we focused entirely on the upper reaches of that spectrum. Now, we will begin by studying gravitational waves from massive PBHs, but we will see thereafter

that gravitational waves from massive objects can probe the physics of ultralight particle species. The nature of gravitational wave astronomy in the cosmological context will lead us all the way around to the other side of the spectrum. At that point, we will turn to other opportunities in cosmological observables to probe particle physics at masses that are somewhere in between: well above the ultralight masses we will probe with gravitational waves, but still well below the weak scale. Here we will see the emergence of complementarity between cosmology and upcoming terrestrial tools.

Of the three parts of this thesis, the probes of this part are the most varied, and potentially the most broadly applicable. Many of the analyses here can be repurposed for models at other scales, or set the stage for new approaches to DM across an even wider range of masses. Still, for the time being, we will focus on two distinct regimes: ultralight bosons at the very bottom of the DM spectrum and even below, and "light" DM between 1 keV and 1 GeV. Even though these two ranges are quite far apart, they are two of the most interesting windows in the spectrum today. In particular, they lie to either side of the mass range favored for axion DM, and thus heavily probed by axion experiments. These are windows in the BSM parameter space where new tools stand to deliver probes with the greatest new breadth.

In this part, we will focus on three such tools: gravitational wave observatories, terrestrial colliders, and direct detection experiments. With the exception of gravitational wave backgrounds, the cosmological observables we will discuss here have already been heavily mined for many years. The coming years will also see new cosmological observables as a product of the next generation of surveys, and leveraging these surveys for new tests of DM physics is one of the goals of my ongoing work.

The coming chapters take the following route through parameter space. First, in Chapter 4, we will continue our exploration of PBHs to their gravitational wave signatures, and determine the prospects for their discovery at present-day gravitational wave observatories. In Chapter 5, we will step to gravitational wave signatures from black holes that are *known* to exist: supermassive black holes at the centers of galaxies. We will see how the gravitational wave background can signal the presence of new long-range forces mediated by ultralight particle species. Next, in Chapter 6, we will leverage cosmological observables for a rather different purpose: we will see how cosmological restrictions introduce new connections between cosmic chronology and low-energy observables in the sub-GeV regime. Finally, in Chapter 7, we will focus on the implications of cosmology for direct detection experiments in the sub-GeV and especially the sub-MeV regime. This will lead us into a more detailed discussion of low-mass direct detection in Part III.

Chapter 4

Discovery prospects for primordial black holes at LIGO

4.1 Introduction

The detection of black hole binaries with LIGO [202–209] has heralded a new era in probing the nature and behavior of compact objects in our Universe. In the past several years, gravitational wave detectors have directly confirmed the existence of black holes [210], provided powerful tests of general relativity [211], and ushered in the era of multimessenger astronomy [212, 213]. But as gravitational wave observatories continue to probe the black hole population, they are poised to make yet another significant discovery: mergers may provide direct evidence for the existence of primordial black holes (PBH).

The primordial-origin scenario for the black holes observed at LIGO has thus been discussed heavily in the literature. Several authors have proposed that stellar- and

primordial-origin models might be distinguished statistically in the coming years by the distributions of binary masses, spins, and eccentricities [see e.g. 214–218]. However, an extensive literature shows that the binaries observed to date are compatible with a stellar origin [219, 220], and efforts to attribute any future discrepancies to a primordial origin will be complicated by uncertainties in stellar evolution models [see e.g. 221, 222]. Thus, even in the most optimistic case, it will be difficult to positively establish a non-stellar origin for the LIGO black holes, especially if such a history applies only to a subcomponent of the merging population.

But there is one clean signal that could clearly indicate the primordial origin of a specific black hole: a low mass. Stellar evolution models predict that black holes form only when a star's mass is sufficient for the gravitational force to overcome degeneracy pressure. Thus, black holes with a stellar origin must have a mass no lower than the Chandrasekhar limit of $1.4\,M_\odot$ [138]. A black hole with a mass below $\sim 1\,M_\odot$ must have a non-stellar origin, and the detection of even one such object would be a clear smoking gun of new physics, as was already pointed out by [223]. In principle, LIGO may be sensitive to mergers of black holes well below this scale, so LIGO and other gravitational wave observatories are uniquely capable of directly establishing the existence of PBH.

Indeed, some gravitational wave detections have already come tantalizingly close to furnishing such a discovery. The latest hint comes from the recently announced LIGOScientific:2016aoc90814 [224], apparently involving a compact object at $2.6\,M_\odot$, in what was expected to be a mass gap in the population of neutron stars and stellar-origin black holes [225–228]. Additionally, the nature of the compact objects involved in LIGOScientific:2016aoc70817 [229] is uncertain: while the identification of an associated

kilonova [230] strongly indicates that one of the merging objects was a neutron star, the second compact object might also be a light $\mathcal{O}(1\,M_\odot)$ black hole, with likelihood as large as 40 per cent [see e.g. 231].

Given the potential for discovery, the LIGO Collaboration has conducted initial searches for mergers of light PBH [67, 69], with null results thus far. But interpreting these null results as constraints on the PBH population requires a model to connect the abundance and mass distribution of PBH to the rate of observed mergers. Theoretical uncertainties in the merger rate complicate such an analysis, with notable recent progress by [232]. Even so, most previous work has assumed that the PBH mass function is monochromatic, i.e., that all PBH have a single mass. This greatly simplifies the problem, but is likely unrealistic: in most formation models, PBH have an extended mass distribution with a lognormal or power-law shape [29]. In some scenarios, the mass distribution can even be multimodal [233].

A bias-free interpretation of LIGO results requires that we allow for some freedom in the shape of the mass function. This motivates the approach taken by [234], who analyze prospects for the detection of light black holes under the assumption that the mergers observed thus far have a primordial origin. To further complicate matters, the mass distribution is subject to various observational constraints across the mass spectrum, which impose additional restrictions on the space of mass functions. The uncertainty in the merger rate arising from the shape of the mass function means that it is difficult to describe prospects for either constraints on or discovery of a PBH population at gravitational wave observatories in a model-independent fashion.

In this chapter, we use numerical methods to translate null searches at grav-

itational wave observatories into constraints on the properties of the PBH population and discovery prospects for light black holes. In particular, we show that if only a small fraction of the PBH population lies in the mass window of interest, then freedom in the mass function translates to a significant gap between the *constraint* potential and the *discovery* potential, corresponding to the most pessimistic and optimistic calculations of the merger rate, respectively. We show that LIGO can establish the existence of PBH even if the abundance of such objects in the mass range of interest is far below the level of the prospective constraint. Our results provide the first model-independent gravitational wave constraints on the light black hole population, and show that there is considerable opportunity for their discovery at LIGO.

This chapter is organized as follows. In Section 4.2, we review the calculation of the merger rate and establish analytical expectations for the shapes of mass functions that maximize and minimize the merger rate. In Section 4.3, we introduce our numerical procedure and detail the inclusion of other observational constraints. We present our numerical results in Section 4.4, and we discuss the implications and conclude in Section 4.5.

Notationally, we will say that a black hole is 'light' if it has a mass below $1\,\mathrm{M}_\odot$, and we will say that a black hole is 'detectable' if it has a mass large enough to be observable by LIGO, a condition we will detail in subsequent sections. We will refer to light black holes as 'primordial', although, as we have mentioned, there are other new physics scenarios that may also result in the formation of observable black holes without stellar progenitors. Our interest is in light black holes that are 'detectable and primordial', which we will abbreviate as 'DPBH'. We will say that such black holes lie

in the 'DP' mass range.

4.2 The merger rate of DPBH

To establish the most optimistic discovery prospects, and the most pessimistic constraint potential, it is necessary to consider, respectively, the maximum and minimum merger rates that can be produced with a fixed abundance of PBH. We will perform this optimization numerically in the following sections, but first, we discuss the calculation of the merger rate and explore a few benchmark cases analytically.

The merger rate of PBH has been studied by many authors [11, 12, 53, 54, 232, 235, 236], and while predictions of the rate are still subject to some uncertainties, the theoretical formalism has improved considerably in recent years. In particular, Ref. [236] and Ref. [54] have studied the merger rate for extended mass functions, and established predictions for the merger rate as a function of the component masses. The formation of merging primordial black hole binaries is quite different from the stellar case, so we will shortly review the derivation of the merger rate and the attendant physics.

Throughout the following sections, we will denote the mass function by ψ. Denoting the PBH number density for masses up to m by $n(m)$, the mass function is defined by $\psi(m) \propto m\,\mathrm{d}n/\mathrm{d}m$ with the normalization condition

$$\int \mathrm{d}m\,\psi(m) = \frac{\Omega_{\mathrm{PBH}}}{\Omega_{\mathrm{DM}}} \equiv f_{\mathrm{PBH}}. \tag{4.1}$$

4.2.1 The detectable mass range

A key component of estimating the DPBH merger rate is defining exactly what is meant by detectability. In previous studies of DPBH mergers, a threshold is generally set at a mass of order $\sim 0.1\,\mathrm{M}_\odot$, and mergers of black holes below the threshold mass are assumed to be undetectable. We must do the same in this chapter, for reasons we will explain shortly. For the moment, note that this is a reasonable approximation, especially because gravitational wave detectors trigger on the basis of a bank of template waveforms. Thus, even if LIGO is potentially sensitive to mergers of lighter objects, a detection will not be made if no matching template has been computed. In typical operation, LIGO uses no templates with combined binary masses below $2\,\mathrm{M}_\odot$ [68], and even past searches for light black hole mergers have used a minimum template mass of $0.4\,\mathrm{M}_\odot$ [67, 69].

Neglecting templates, LIGO is potentially sensitive to mergers of very light black holes, with one important caveat: the lighter the binary, the closer it must be in order for the merger to be detectable. Thus, LIGO probes a different effective volume $V_{\mathrm{eff}}(m_1, m_2)$ for each pair of component masses (m_1, m_2). Given a particular mass function ψ, the total DPBH merger rate R_{DP} must then be written in the form

$$R_{\mathrm{DP}}(\psi) = \int_{\mathrm{DP}^2} \mathrm{d}m_1\, \mathrm{d}m_2\, \mathcal{R}(m_1, m_2) V_{\mathrm{eff}}(m_1, m_2), \tag{4.2}$$

where $\mathcal{R}(m_1, m_2)$ is the differential merger rate per unit volume for binaries with component masses (m_1, m_2), and $\int_{\mathrm{DP}^2}$ denotes an integral only over pairs of masses in the DP regime. We make the simplifying approximation that the sensitive volume depends only on the chirp mass of the binary, M_c, and not on the individual component masses,

so that $V_{\rm eff}(m_1, m_2) = V_{\rm eff}(M_c(m_1, m_2))$. This approximation has been explicitly validated by [69]. Note that we neglect any impact of binary spin on detectability.

We determine the sensitive volume $V_{\rm eff}(m_1, m_2)$ for the merger of a given binary using the maximum sensitive distance for the scenario considered in Ref. [68]. This sensitivity is already achievable with the Advanced LIGO instrument, but does involve optimistic assumptions about the template bank used to identify merger events. The number N of templates required for a given search depends strongly on the minimum mass $m_{\rm min}$ and starting frequency $f_{\rm min}$ included in the template bank, scaling as $N \propto (m_{\rm min} f_{\rm min})^{-8/3}$. We follow the optimistic benchmark of Ref. [68], choosing $f_{\rm min} = 10\,{\rm Hz}$ and $f_{\rm max} = 2048\,{\rm Hz}$, and we likewise reduce the maximum sensitive distance by a factor of 2.26 to account for variations in the location and orientation of the binary. (See fig. 2 and eq. (12) of that reference.) As noted in that work, there are significant computational costs to producing an appropriate template bank for detection of these very low-mass binaries at the greatest sensitive distances. LIGO searches completed to date use slightly less generous template banks, and in particular are completely insensitive to black holes below $0.2\,{\rm M}_\odot$ [69]. Thus, our results apply directly under the assumption that LIGO carries out a search with these parameters.

We still need to define the domain of the integral in Eq. (4.2). To meaningfully probe the abundance of light PBH, we will ultimately be interested in speaking of the abundance in a narrow mass range, neither too massive to be clearly primordial, nor too light to be typically detectable, but just right [see e.g. 237]. To that end, we will define two thresholds $m_{\rm DP}^{\rm min}$ and $m_{\rm DP}^{\rm max}$. For single masses, we will say $m \in {\rm DP}$ if $m_{\rm DP}^{\rm min} \leq m \leq m_{\rm DP}^{\rm max}$. For pairs of masses, we will say that $(m_1, m_2) \in {\rm DP}^2$ if $m_{\rm DP}^{\rm min} \leq$

$\min\{m_1, m_2\} \leq m_{\rm DP}^{\rm max}$, i.e., if

1. *both* m_1 and m_2 are above $m_{\rm DP}^{\rm min}$, and

2. *at least one* of m_1 and m_2 is below $m_{\rm DP}^{\rm max}$.

We will fix $m_{\rm DP}^{\rm max} = 1\,{\rm M}_\odot$ and $m_{\rm DP}^{\rm min} = 0.1\,{\rm M}_\odot$ throughout our analysis. We have investigated the consequences of choosing $m_{\rm DP}^{\rm min} = 0.01\,{\rm M}_\odot$, and found that there is very little impact on the qualitative outcomes of our analysis: while choosing a lower threshold threshold extends the opportunity for discovery if PBH only exist at lower masses, extant gravitational wave detectors are relatively poorly suited to probe such a population.

In order to meaningfully discuss constraints on the DPBH population, we define the DP ratio by

$$r_{\rm DP} = \frac{1}{f_{\rm PBH}} \int_{m_{\rm DP}^{\rm min}}^{m_{\rm DP}^{\rm max}} {\rm d}m\, \psi(m). \tag{4.3}$$

This is the mass fraction of PBH with masses between $m_{\rm DP}^{\rm min}$ and $m_{\rm DP}^{\rm max}$. Note the use of r (ratio) rather than f to avoid confusion with $\Omega_{\rm DP}/\Omega_{\rm DM}$, as with $f_{\rm PBH}$. We instead define $f_{\rm DP} \equiv r_{\rm DP} f_{\rm PBH}$.

Ultimately, we will evaluate maximum and minimum merger rates as a function of both $f_{\rm PBH}$ and $r_{\rm DP}$ simultaneously. This is a convenient parametrization for discussing constraints on the mass function, since despite the very simple form of the two parameters, they encode key information about the abundance of PBH in general and the abundance of light black holes in particular. This is also one of the reasons for imposing a strict cut-off at low masses: one might contend that lighter black holes, with masses below our $m_{\rm DP}^{\rm min}$, are also detectable, albeit in a smaller volume. This may indeed

be the case, but including such mergers would make the parametrization discussed here difficult to interpret in relation to the merger rate: black holes just below $m_{\rm DP}^{\rm min}$ would contribute to the DP merger rate, but not to $r_{\rm DP}$.

We reiterate that LIGO is not equipped with templates for our entire DP window during its regular operation, and a search with black hole masses below $0.2\,{\rm M}_\odot$ has not been conducted to date. Moreover, previous searches have targeted mergers between two light black holes, with templates only below $4\,{\rm M}_\odot$ in total binary mass. Thus, the constraints we draw in this chapter are prospective, assuming that an extended search is conducted on archival or future data. As we will show, such searches are well motivated both for pairs of light black holes and for mergers of light black holes with heavy partners. There is ample opportunity to discover PBH even at abundances that cannot be fully constrained.

4.2.2 Estimating the merger rate

We now review the derivation of the merger rate in Ref. [236] and Ref. [54]. First, we note that PBH binaries can form in two epochs: in the early Universe, during radiation domination, and in the late Universe, where close approaches can produce enough gravitational radiation to bind two black holes. The latter contribution is generally small, since typical relative velocities are large, meaning that the energy loss to gravitational radiation must be quite significant. There is a possible exception to this rule if the density contrast in the late Universe is exceptionally large, $\delta_0 \gtrsim 10^{10}$, but this is much larger than most estimates, so we neglect that possibility. Thus, we consider only binaries formed in the early Universe. Note that our calculation may not

accurately describe new physics scenarios in which light black holes themselves form in the late Universe.

We first review the merger rate as estimated by Ref. [236]. Consider a PBH pair with masses m_1 and m_2. First, in order for the pair to decouple from the Hubble flow and have interactions dominated by their mutual gravitation, the average mass of the black holes should exceed the background mass in the volume between them, i.e., we require $\frac{1}{2}(m_1 + m_2) > \frac{4\pi}{3}\rho_{\rm bg} r^3$. Translating this into a condition on the separation of the two black holes, one finds that the comoving distance between them must fall below the scale

$$\tilde{x}(m_1, m_2)^3 = \frac{3}{4\pi} \frac{m_1 + m_2}{a_{\rm eq}^3 \rho_{\rm eq}}, \tag{4.4}$$

where $a_{\rm eq}$ and $\rho_{\rm eq}$ are the scale factor and the density at matter–radiation equality. A binary with comoving separation $x < \tilde{x}$ thus decouples from the Hubble flow when the scale factor is

$$a_{\rm dc} = a_{\rm eq} \left(\frac{x}{\tilde{x}}\right)^3. \tag{4.5}$$

After this point, the black holes' gravity dominates the evolution of the system. Barring a close approach, gravitational interactions between these two black holes and some third body of comparable mass are necessary to move the pair into a bound configuration. Thus, we suppose that there is third PBH with mass m_3 at a comoving distance y from the first two. In this scenario, we form a binary with semimajor and semiminor axes given by

$$r_a = \alpha x a_{\rm dc}, \qquad r_b = \beta \frac{2m_3}{m_1 + m_2} \left(\frac{x}{y}\right)^3 r_a, \tag{4.6}$$

for two $\mathcal{O}(1)$ constants α and β. We take $\alpha = \beta = 1$ for the remainder of this discussion.

Assuming that there is no mechanism for hardening the binary apart from gravitational radiation, the coalescence time can then be estimated as

$$\tilde{\tau}(m_1, m_2, m_3) \left(\frac{x}{\tilde{x}(m_1, m_2)} \right)^{37} \left(\frac{y}{\tilde{x}(m_1, m_2)} \right)^{-21}, \tag{4.7}$$

where $\tilde{\tau}$ is the maximal coalescence time, given by

$$\tilde{\tau}(m_1, m_2, m_3) = \frac{348}{85} \frac{\alpha^4 \beta^7 a_{\rm eq}^4 m_3^7 \tilde{x}(m_1, m_2)^4}{G^3 m_1 m_2 (m_1 + m_2)^8}. \tag{4.8}$$

Here we have established the coalescence time for a single binary assuming a set of masses and initial separations. Distributions of these parameters can be derived from the mass function, leading to a distribution of coalescence times as a function of the component masses. Differentiating this distribution leads to the merger rate at the present time. For brevity, we define $\tilde{m}(\psi) = 1/\int \mathrm{d}m\, \psi(m)/m$. The number density of PBH at the mass of interest is accounted for through the factor $\tilde{N}(\psi;\ m_1, m_2) = \delta_{\rm dc} \Omega_{\rm DM,eq} (m_1 + m_2)/\tilde{m}(\psi)$, where $\delta_{\rm dc}$ is the density contrast at the time of decoupling. We then define

$$\mathcal{G}(\psi;\ m_1, m_2, m_3) = \Gamma \left(\frac{58}{37},\ \frac{\tilde{N}(\psi;\ m_1, m_2)\, t^{3/16}}{\tilde{\tau}(m_1, m_2, m_3)^{3/16}} \right) - \Gamma \left(\frac{58}{37},\ \frac{\tilde{N}(\psi;\ m_1, m_2)\, t^{-1/7}}{\tilde{\tau}(m_1, m_2, m_3)^{-1/7}} \right), \tag{4.9}$$

and the present-day differential merger rate between black holes with masses m_1 and m_2 is given by

$$\mathcal{R}(m_1, m_2) = \frac{9 \bar{m}(\psi)^3 \tilde{N}(\psi;\ m_1, m_2)^{\frac{53}{37}}}{296 \pi \delta_{\rm dc} \tilde{x}(m_1, m_2)^3 t^{34/37}} \times \frac{\psi(m_1)\psi(m_2)}{m_1 m_2} \int \mathrm{d}m_3\, \frac{\mathcal{G}(\psi;\ m_1, m_2, m_3)}{\tilde{\tau}(m_1, m_2, m_3)^{3/37}} \frac{\psi(m_3)}{m_3}. \tag{4.10}$$

Notably, this estimate of the merger rate considers only the tidal torque due to one additional PBH external to the binary. This may present a problem when dealing with mass functions that span many decades, for which lighter black holes have relatively high number densities. Ref. [54] follow a similar line of argument, but the authors estimate the torque by integrating over the entire PBH population. It might be expected that this form of the merger rate is more reliable for extremely broad or multimodal mass functions, which we may well encounter in the course of our analysis. Thus, we use their merger rate in the course of our calculation, and we now briefly summarize their result. We define

$$\mu = \frac{2m_1 m_2 \left(\psi(m_1) + \psi(m_2)\right)}{(m_1 + m_2)\big(m_1\psi(m_1) + m_2\psi(m_2)\big)} \tag{4.11}$$

and $n_T = \rho_c \Omega_{\rm DM,eq} \int \mathrm{d}m\, \psi(m)/m$, where the lower limit of integration is $\min\{m_1, m_2\}$. We additionally take $\langle x \rangle$ to be the average separation between black holes of mass m_1 and m_2, and define γ_X by

$$\gamma_X = \left(\frac{85}{3} \frac{t m_1 m_2 (m_1 + m_2)\big(\psi(m_1) + \psi(m_2)\big)^4}{10^{-4} \left(\frac{3}{8\pi} \frac{m_1 + m_2}{\rho_{\rm eq}(\psi(m_1) + \psi(m_2))}\right)^{4/3} X^{16/3}} \right)^{1/7} \frac{2(\psi(m_1) + \psi(m_2))\Omega_{\rm M}}{\Omega_{\rm DM} X}. \tag{4.12}$$

Then the probability distribution for the coalescence time is given by

$$\frac{\mathrm{d}P}{\mathrm{d}t} = \frac{1}{7\mu t} \int \mathrm{d}X\, \exp\left(-\frac{X}{\mu}\frac{4\pi}{3}\langle x \rangle^3 n_T\right) \frac{\gamma_X^2}{\left(1 + \gamma_X^2\right)^{3/2}}, \tag{4.13}$$

and the present-day merger rate per unit volume is given differentially in the component masses by

$$\mathcal{R}\left(m_1, m_2\right) = \rho_c \Omega_{\rm M} \min\left(\frac{\psi(m_1)}{m_1}, \frac{\psi(m_2)}{m_2}\right) \frac{\mathrm{d}P}{\mathrm{d}t}. \tag{4.14}$$

While we use this form of the merger rate in our subsequent analysis, it is not the only such calculation to take into account the torques from the entire population

of black holes. The calculation of Ref. [236] was updated and extended by Ref. [238] to include this effect via a suppression factor multiplying the rate. The suppression factor S has the form

$$S = \frac{e^{-\bar{N}}}{\Gamma(21/37)} \int \mathrm{d}v \, v^{-\frac{16}{37}} \exp\left(-\mathcal{F}(\psi, v) - \frac{3\sigma_{\mathrm{M}}^2 v^2}{10 f_{\mathrm{PBH}}^2}\right), \tag{4.15}$$

$$\mathcal{F}(\psi, v) = \bar{N}\langle m \rangle \int \frac{\psi(m)\,\mathrm{d}m}{m} {}_1F_2\left(-\tfrac{1}{2}; \tfrac{3}{4}, \tfrac{5}{4}; -\left(\tfrac{3mv}{4\langle m\rangle \bar{N}}\right)^2\right), \tag{4.16}$$

where $\langle m \rangle$ is the average PBH mass, ${}_1F_2$ is the generalized hypergeometric function, $\bar{N}$ counts the number of PBHs in the vicinity of a given binary, and $\sigma_{\mathrm{M}} = (\Omega_{\mathrm{M}}/\Omega_{\mathrm{DM}})^2 \langle \delta_{\mathrm{M}}^2 \rangle$ for δ_{M} the matter density perturbation. Note that in this merger rate calculation, the suppression effect factorizes away from the dynamics that determine the binary formation rate, so the suppression factor can really be evaluated separately from the unsuppressed rate. This is not the case in the calculation of Ref. [54]: while there is no explicit suppression factor, a comparable suppression enters the rate itself via the exponential factor in Eq. (4.13). However, the calculation does not include a suppression of the merger rate from binary disruption in later close encounters.

We note that predicting the PBH merger rate from first principles is extremely challenging, and it is likely that these estimates will be refined in the coming years. In particular, Ref. [239] recently showed that the inclusion of all three-body encounters in PBH clusters can dramatically reduce the merger rate in the late Universe. We will return to this possibility in Section 4.5, and we will also assess the robustness of our results to differences between the calculations of Ref. [54] and Ref. [238].

4.2.3 Analytical behavior of the merger rate

To establish constraint and discovery prospects, we will need to minimize and maximize the merger rate over the possible mass function shapes with some characteristic abundances held fixed. In particular, we will optimize the merger rate with $r_{\rm DP}$ and $f_{\rm PBH}$ held constant. For fixed $f_{\rm PBH}$, if the maximum merger rate falls below the LIGO sensitivity for a given value of $r_{\rm DP}$, this means that values of $r_{\rm DP}$ this low cannot be probed by LIGO, regardless of the form of the mass function. Alternately, if the minimum merger rate is detectable by LIGO, then this and higher values of $r_{\rm DP}$ can be ruled out by LIGO.

In general, the merger rate must be maximized or minimized numerically. However, to understand the dependence of the merger rate on the shape of the mass function, it is useful to consider a few simple benchmark cases in the absence of any observational constraints. For the moment, we neglect the mass dependence of the detector's sensitive volume.

First, consider a monochromatic mass function, $\psi(m) = f_1\delta(m - m_1)$. Formally, the quantities entering Eq. (4.14) are not independently well defined in this case, but we can take a mass function of the form

$$\psi_1(m) = f_1\Delta^{-1}\Theta(m - m_1)\Theta(m_1 + \Delta - m) \tag{4.17}$$

and work in the limit $\Delta \to 0$. In this case, the total DP rate is simply the overall rate, as long as m_1 lies within the DP window. If m_1 is sufficiently small, the integrand of Eq. (4.13) is dominated by values of X where the exponential is very nearly 1. As pointed out by [54], the integral can then be evaluated approximately, which gives a rate

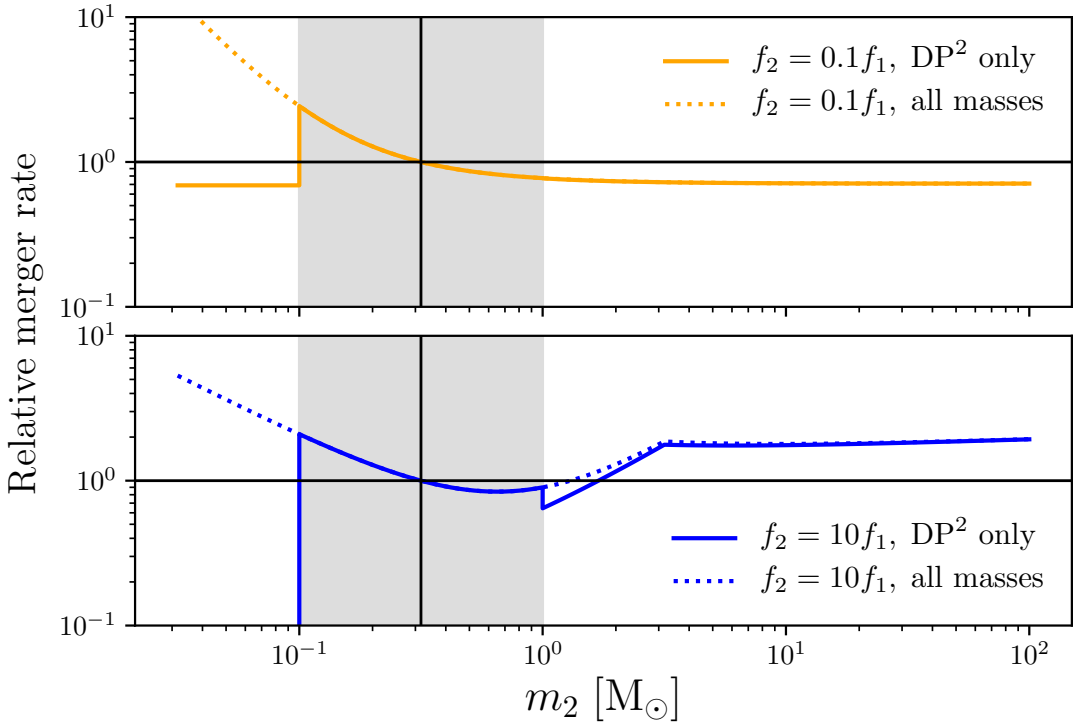

Figure 4.1: Merger rate for a dichromatic mass function, $\psi(m) = f_1\delta(m - m_1) + f_2\delta(m - m_2)$, relative to the monochromatic mass function $(f_1 + f_2)\delta(m - m_1)$. We fix $m_1 = 10^{-1/2}\mathrm{M}_\odot$, indicated by the black vertical line. This lies in the middle of the DP window, indicated by the shaded region. The dashed curves show the merger rate for pairs of all masses, while the solid curves include only mergers in DP^2. The blue curve shows the case $f_2 = 10f_1$, i.e., where mergers of black holes of mass m_2 naively dominate. The orange curve shows the case $f_2 = 0.1f_1$, where the reverse is true. Depending on the relative amplitudes and positions of the two peaks, separating them can either enhance or suppress the merger rate (see text).

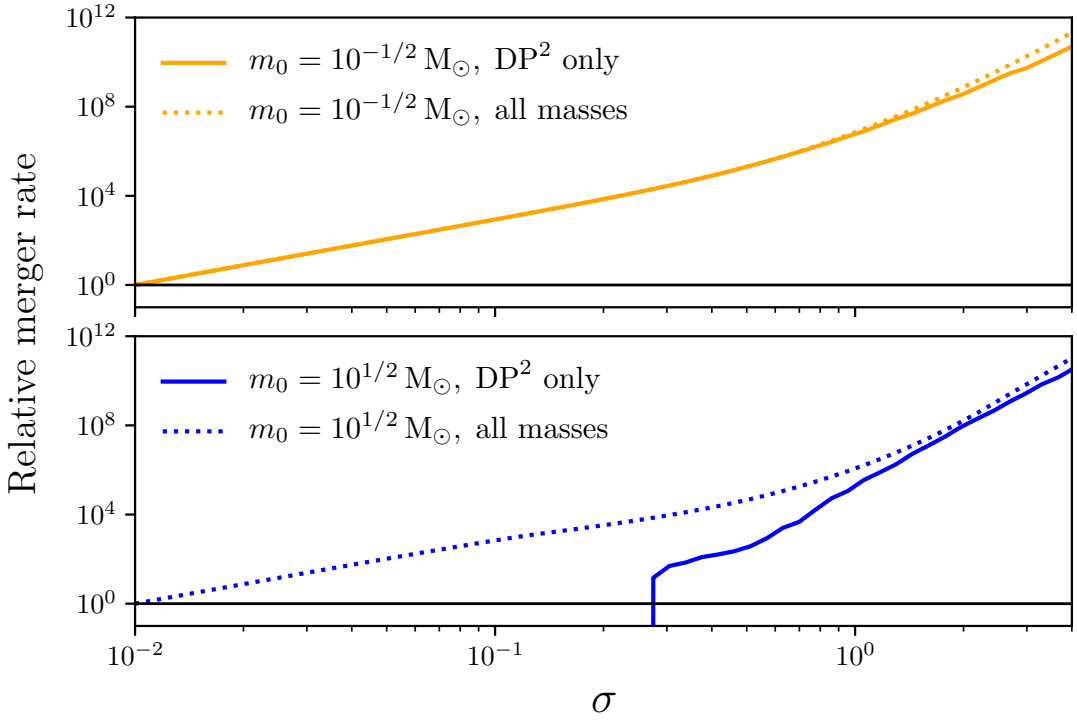

Figure 4.2: Relative merger rate as σ is increased for a lognormal mass function with two different central masses. The dashed lines include all mergers, while the solid lines include only DP^2 mergers. The curves are normalized relative to the all-inclusive merger rate at the lowest value of σ, i.e., the dashed lines are fixed to 1 at the left edge of the figure. For much of the range of σ shown here, a large fraction of mergers lie in DP^2, so the solid and dashed lines lie very close. They begin to diverge at large σ in both subplots, since much of the mass lies outside the DP window in this case. For the blue curve, the central mass lies outside the DP window, so the DP merger rate vanishes at small σ. In the limit $\sigma \to 0$, the lognormal mass function reduces to the monochromatic case.

$\mathcal{R} \propto m_1^{-32/37}$. A similar result can be derived for large m_1 by splitting the X integral into two regimes, one in which the constant term in the denominator is dominant, and one in which it is subdominant. The integral can be evaluated analytically in each of the two regimes, and it can then be shown that $\mathcal{R} \propto m_1^{-26/21}$ at large m_1.

In particular, for a monochromatic mass function, decreasing m_1 increases the merger rate. Physically, this is simply because decreasing m_1 while holding f_1 constant increases the number density of black holes. In the absence of observational constraints, we therefore expect that the merger rate will be maximized when the mass function is peaked near the bottom of the DP window, and minimized when it is peaked near the top.

This is the simplest way in which the mass function can influence the merger rate. However, monochromatic mass functions are tightly constrained by observational bounds, so it is useful to understand the behavior of the merger rate for mass functions with non-negligible width. We first consider the simplest extension of the previous case: a bimodal mass function constructed as the sum of two monochromatic mass functions. We define

$$\psi_2(m) = f_1\delta(m - m_1) + f_2\delta(m - m_2), \tag{4.18}$$

where the Dirac delta is understood to be defined as in Eq. (4.17).

For such a 'dichromatic' mass function, there are three contributions to the merger rate, corresponding to mergers of black holes with masses $\{m_1, m_1\}$, $\{m_2, m_2\}$, and $\{m_1, m_2\}$. This gives rise to complicated behavior as the peaks are separated. Two benchmark cases are shown in Fig. 4.1, with one peak fixed in the middle of the DP range and the other varying freely. In each panel, the merger rate is enhanced if the

second peak is positioned at a low mass within the DP window, due to the enhanced number density. The DP merger rate (the solid line) drops sharply as the mass of the second peak falls below the DP window, while the all-inclusive merger rate (the dotted line) continues to increase.

Notice that as the second peak rises above the DP window, the drop in the DP merger rate is much less significant. This is because the presence of these more massive black holes still affects the DP merger rate in two ways: first, more massive black holes can still participate in the formation of light PBH binaries, and secondly, mergers of binaries with masses $\{m_1, m_2\}$ themselves contribute to the DP merger rate. These effects lead to non-trivial behavior of the dichromatic merger rate as a function of the two masses. For our purposes, we note that separating peaks in a dichromatic mass function can either increase or decrease the merger rate.

Finally, we consider a lognormal mass function, which is unimodal, but has a non-vanishing width. The lognormal mass function has the form

$$\psi_L(m) = \frac{f_{\rm PBH}}{\sqrt{2\pi}\sigma m} \exp\left(-\frac{1}{2}\left[\frac{\log\left(m/m_0\right)}{\sigma}\right]^2\right), \tag{4.19}$$

where m_0 corresponds to the central mass and σ is the width of the distribution. Holding σ fixed, the merger rate is increased by reducing m_0, as in the monochromatic case. If m_0 is fixed, and σ is varied, then the merger rate is enhanced by increasing σ, i.e., broadening a sharp distribution locally increases the merger rate. This behavior is shown in Fig. 4.2.

These benchmark scenarios indicate that a fairly broad mass function favouring lower masses will generally produce a higher merger rate, but in general, observational

constraints will impose severe restrictions on the allowed shape of the mass function. Thus, the mass function that minimizes or maximizes the merger rate might indeed be a complicated multimodal function. In particular, the analysis of [30] demonstrates that the maximum value of $f_{\rm PBH}$ consistent with observational constraints is attained by a multimodal mass function, corresponding to a superposition of monochromatic mass functions. Thus, it would not be surprising to find a similar behavior for mass functions that maximize the merger rate, particularly if $f_{\rm PBH}$ is fixed at a value where observational bounds strongly constrain the mass function. The benchmark scenarios in this section suggest that the merger rate will be minimized with sharply peaked and potentially multimodal mass functions.

To go further and to incorporate observational constraints will require numerical methods, which we take up in the following sections. Again, note that the discussion above has not accounted for any characteristics of the detector. In particular, we have neglected the mass dependence in the effective volume that an instrument such as LIGO can probe. Mergers of more massive black holes are observable in a larger effective volume, and this enhances the effective merger rate at higher PBH masses, competing with the enhancement in number density at lower masses. We will include this effect in our numerical treatment.

4.3 Constraints and optimization

In this section, we detail the numerical procedure that we use to optimize the merger rate. First, we discuss the constraints that we impose on the black hole mass

function.

Since the allowed values of the merger rate depend on the allowed forms of the PBH mass function, observational constraints that restrict the form of the mass function correspondingly restrict the merger rate. Thus, the minimum or maximum merger rate is dependent not only on $f_{\rm PBH}$ and $r_{\rm DP}$, but also on the chosen set of observational constraints. The full set of observational constraints we use in this chapter is shown in Figs. 4.4 and 4.5, with descriptions in the captions. We demonstrate the behavior of the maximum and minimum merger rates both with and without the constraints imposed on the mass function by these observational bounds.

Note that there are many other observables that may place constraints on the PBH population, such as supernova lensing [49], dynamical effects [37, 45, 46], and destruction of white dwarf stars [47] and neutron stars [48]. (For reviews see [29, 31, 33].) These constraints are subject to additional uncertainties, and including them does not change our qualitative conclusions. The qualitatively important features are the relative strength of the constraints at masses above and below the DP window, and the fact that there is a gap in the constraints at low masses. The latter allows for large values of $f_{\rm PBH}$ when $r_{\rm DP}$ is small. This gap has attracted considerable attention since lensing constraints at low masses were shown to be ineffective [14, 15, 240–242], and it is possible that new constraints developed in this region will influence our results.

We introduce one important observable beyond the constraints plotted in Figs. 4.4 and 4.5: the stochastic gravitational wave background [SGWB; 243, 244]. A population of black holes merging over cosmic time produces an accumulated background of gravitational radiation that can be detected by LIGO. Since the SGWB

depends in detail on the shape of the mass function, it must be treated differently from the other constraints. However, it is essential that we include this constraint, since it has been shown that merging DPBH in particular can make a significant contribution to the SGWB [see e.g. 245]. Further, when we maximize the merger rate, we also maximize the contribution to the SGWB, so our optimal mass functions might run afoul of SGWB constraints at PBH abundances well below those excluded in other analyses.

4.3.1 Applying constraints to the mass function

In order to translate gravitational wave observables to discovery prospects and constraints on the population of DP black holes, we must alternately minimize and maximize the merger rate subject to particular constraints. This is similar to the problem of maximizing the overall abundance of black holes subject to observational constraints, as discussed by Ref. [30]. In that reference, the general form of the optimal mass function is derived analytically, and it is shown that the exact global optimum can be found semi-analytically with arbitrary precision. Since we will use some of the same methods and terminology, we briefly review this result. However, as we will explain, this formalism cannot be adapted to optimize the merger rate semi-analytically.

We treat observational constraints on the black hole population following [39]. In general, observational constraints on the black hole population are derived from some measured quantity $A_{\rm obs}$. The value of $A_{\rm obs}$ is predicted to be A_0 in the absence of any PBH, whereas in the presence of PBH, one has $A_{\rm obs} = A_0 + A_1$. For most observables,

black holes at different mass scales contribute independently, so we can write

$$A_1 = \int \mathrm{d}m\, \psi(m)\, K_1(m) \tag{4.20}$$

for some kernel $K_1(m)$. Provided that the constraining observable has this form, the constraint condition can be written in the form $\mathcal{C}(\psi) \leq 1$, where $C(\psi)$ is the functional

$$\mathcal{C}(\psi) \equiv \int \mathrm{d}m\, \frac{\psi(m)}{f_{\max}(m)}. \tag{4.21}$$

Here $f_{\max}(m)$ is the maximum allowed fraction of dark matter in the form of PBH assuming a monochromatic mass function at mass m. For the case of N independent constraints $f_{\max,j}(m)$, corresponding to a vector $\mathcal{C}_j(\psi)$, this generalizes to

$$\left\|\boldsymbol{\mathcal{C}}(\psi)\right\|^2 \equiv \sum_{j=1}^{N} \left(\int \mathrm{d}m\, \frac{\psi(m)}{f_{\max,j}(m)} \right)^2 \leq 1. \tag{4.22}$$

Note, in particular, that $\psi(m) > f_{\max}(m)$ is perfectly admissible for a subset of masses—i.e., the mass function can cross through the curves on constraint plots—as long as the condition above is still met. This is simply because constraint curves, as typically drawn, are only applicable to monochromatic mass functions.

Since the total density in PBH scales linearly with the normalization of the mass function, any mass function can be normalized to saturate observational constraints, yielding the *normalized mass*, $\mathcal{M}(\psi) = \|\boldsymbol{\mathcal{C}}(\psi)\|^{-1} \int \mathrm{d}m\, \psi(m)$. Finding the mass function that maximizes the PBH density subject to observational constraints is thus equivalent to maximizing the functional $\mathcal{M}$, which can be done semi-analytically.

One might hope that a similar method might apply to the optimization of the merger rate. But even if the merger rate functional were as simple as the normalized mass, it would still not be possible to apply the preceding formalism. As we have

discussed, it is essential to consider constraints from non-detection of a SGWB signal, but this constraint cannot be cast in the form of Eq. (4.22). We now briefly review the nature and calculation of the SGWB constraint.

A population of PBH produces a SGWB from mergers at higher redshifts [245–249]. While such backgrounds do not furnish a smoking-gun signal of a primordial origin for a particular black hole, they do constrain the PBH mass function. A differential merger rate $\mathcal{R}$ produces a stochastic background at frequency ν with density

$$\Omega_{\rm GW}(\nu) = \frac{\nu}{\rho_c} \int \mathrm{d}z\, \mathrm{d}m_1\, \mathrm{d}m_2\, \frac{\mathcal{R}(z; m_1, m_2)}{(1+z)H(z)} \frac{\mathrm{d}E_{\rm GW}}{\mathrm{d}\nu_s}(\nu_s; m_1, m_2), \tag{4.23}$$

where ρ_c is the critical density and $\mathrm{d}E_{\rm GW}/\mathrm{d}\nu_s$ denotes the spectrum of the radiation emitted during a merger, with $\nu_s = (1+z)\nu$ denoting the frequency at the source. We follow the computation of the spectrum and the resulting $\Omega_{\rm GW}(\nu)$ in [248] and [245]. LIGO is most sensitive to the SGWB at a frequency of $\nu_p \sim 20\,\mathrm{Hz}$, and the sensitivity is sharply peaked around ν_p. Thus, we determine whether a mass function is ruled out by SGWB production by simply comparing $\Omega_{\rm GW}(\nu_p)$ with LIGO constraints at that frequency, translating to the requirement that $\Omega_{\rm GW}(\nu_p) \lesssim 2 \times 10^{-9}$ [245].

Since the calculation of the SGWB is dependent on the shape of the entire mass function, this constraint cannot be expressed in the form of Eq. (4.20). In particular, note that the strength of the constraint is not linear in the normalization of the mass function, since the merger rate itself depends on the normalization in a highly non-linear fashion. There is no simple closed-form rescaling of the mass function that will saturate SGWB constraints. Practically, this is not an issue since the optimal mass functions and the corresponding constraints on the black hole population must ultimately be derived

numerically rather than analytically. In our numerical procedure, we can incorporate SGWB constraints on a nearly equal footing with other observational constraints, as we will explain in the following section.

4.3.2 Numerical procedure

We now detail the numerical procedure that we use to optimize the merger rate. We minimize and maximize the merger rate using simulated annealing [250]. In simulated annealing, at each step of the algorithm, a random modification to the state of the system is generated. Each modification is probabilistically accepted or rejected, and steps that decrease the cost function are preferrentially accepted—i.e., simulated annealing is a Monte Carlo Markov chain (MCMC) optimization algorithm. Simulated annealing is structurally similar to the Metropolis–Hastings algorithm [251, 252] for drawing samples from a distribution, but the probability of accepting a given step changes over time.

Heuristically, simulated annealing is based on an analogy to the physical process of annealing, in which a material is heated and then cooled slowly to relieve internal stresses. Heating allows the material to return to an equilibrium configuration, and since the cooling is slow, the material is likely to be in or near its equilibrium state once frozen. In simulated annealing, the system is first 'heated' in the sense that random steps are accepted with a high probability. Then the temperature is slowly reduced, so that the system increasingly disfavours departure from equilibrium. This procedure locates global optima relatively efficiently: at first, while the system is 'hot', the algorithm can generate a chain that explores the parameter space broadly, with little chance of being

stuck at a local optimum. As the system cools, the chain is less likely to depart from a local optimum, so it tends to locate that optimum more precisely with subsequent steps.

4.3.2.1 The annealing algorithm

The simulated annealing procedure is outlined in Algorithm 1. The mass function $\psi(m)$ is binned into a set of values ψ_i with bin widths Δm_i, so that $f_{\rm PBH} = \sum_i^N \psi_i \, \Delta m_i$.

The number of mass bins, N, must be large enough to allow for sufficient flexibility in the mass function, but must not be so large as to make the calculation intractable. The computational cost of the merger rate calculation scales asymptotically as a power law in N, but more importantly, each additional mass bin constitutes an additional dimension for the optimization problem. Naively, since the size of a discretized search space scales exponentially with the number of dimensions, one expects a similar behavior for the number of steps to convergence of the optimization algorithm, i.e., $n_{\rm steps} \sim b^N$. If the exponential base b were large, the numerical optimization we attempt here would be extremely challenging. Pragmatically, since values of the mass function in adjacent bins are highly correlated, b is manageably small: in direct numerical experiments, by subdividing mass bins, we find that $b \sim 1.5$. We choose $N = 21$, dividing the bins into three regions. We use 13 bins for $m < m_{\rm DP}^{\rm min}$, 5 bins for $m_{\rm DP}^{\rm min} < m < m_{\rm DP}^{\rm max}$, and 3 bins for $m > m_{\rm DP}^{\rm max}$, subdividing each region into equally sized logarithmic bins. This makes it feasible for us to generate the numerical results in this chapter with $\mathcal{O}(10^4)$ processor hours.

The probability of accepting ('jumping' to) the candidate step, $P_{\rm jump}$, is defined by

$$P_{\rm jump} = \left[\texttt{cost}(\psi')/\,\texttt{cost}(\psi)\right]^{-1/T}, \tag{4.24}$$

where `cost` represents the functional to be minimized, and T is the 'temperature'. In the simplest case, `cost` is the DP merger rate (for minimization) or its negative (for maximization). In our case, where the optimization problem is constrained, it is convenient to implement these constraints by adding terms to the cost function. Constraints appear in the cost function with a factor of $1/T$ so that, as the temperature drops, constraints become more important. Thus, our cost function is defined by

$$\texttt{cost}\,(\psi) = \pm R_{\rm DP}(\psi) + \frac{\beta}{T}\max\{0, \mathcal{P}(\psi)\}, \tag{4.25}$$

where the penalty functional $\mathcal{P}$ is defined by

$$\mathcal{P}(\psi) = \exp\left(\max\left\{\left\|\boldsymbol{\mathcal{C}}(\psi)\right\|,\ \Omega_{\rm GW}/\Omega_{\rm GW}^{\rm max}\right\} - 1\right) - 1, \tag{4.26}$$

with $\boldsymbol{\mathcal{C}}(\psi)$ defined as in Eq. (4.22). We choose $\beta = 10^3\,{\rm yr}^{-1}$ so that even when the merger rate is at its maximum, the penalty functional still dominates the cost at the lowest temperatures we consider.

In addition, there are three components of the simulated annealing algorithm that must be implemented in a manner specific to each application: the selection of the initial point, the generation of new steps, and the cooling rate (annealing schedule).

To start new chains, we determine the initial mass function ψ_0 by choosing a random value in each mass bin from the log-uniform distribution on $[1, 10^3]$. The resulting mass function is then rescaled to match the input values of $r_{\rm DP}$ and $f_{\rm PBH}$.

Algorithm 1 Annealing procedure

$k \leftarrow 0$, $\psi \leftarrow \psi_0$

while $k < k_{\max}$ **do**

  $\psi' \leftarrow \textsc{neighbor}(\psi)$ ▷ Generate modification

  if $P_{\text{jump}}(\psi', T(k)) > \texttt{random}((0,1))$ **then**

    $\psi \leftarrow \psi'$ ▷ Accept modification

    $k = k + 1$

  end if

end while

The generation of new steps is represented by the NEIGHBOR function, which mutates the current state of ψ to obtain a candidate ψ'. The behavior of NEIGHBOR is specified in Algorithm 2. Schematically, a step is generated by modifying the value of ψ in a randomly selected bin i. The modification is drawn from a normal distribution with mean ψ_i and standard deviation $\sigma\psi_i/\Delta m_i$. Appropriate sections of the resulting mass function are then rescaled to match the input r_{DP} and f_{PBH}.

We use a modified exponential cooling schedule, with a lower limit of $T = 1$. The temperature at the kth step is thus

$$T(k) = 1 + (T_0 - 1)(1 - \alpha)^k, \tag{4.27}$$

where we set $\alpha = 10^{-2}$, and T_0 is the initial temperature. In general, T_0 must be chosen empirically to allow the algorithm to explore a wide parameter space initially. We choose the initial temperature so that 80 per cent of steps from the initial position that increase the cost are accepted. Such a temperature is high enough to 'melt' the

Algorithm 2 Neighbor generation

procedure NEIGHBOR(ψ)

  $\psi' \leftarrow \psi$

  $i \leftarrow \texttt{random}(\{1, 2, \ldots, N_{\text{bins}}\})$ ▷ Choose bin

  $\psi'_i \leftarrow \psi_i \times \texttt{normal}(\psi_i, \sigma\psi_i/\Delta m_i)$ ▷ Modify bin

  $I \leftarrow \sum_i \psi_i \Delta m_i$ ▷ Fix r_{DP}

  $I_{\text{DP}} \leftarrow \sum_{i \in \text{DP}} \psi_i \Delta m_i$

  for $i \in$ DP **do**

    $\psi'_i \leftarrow \psi'_i \times r_{\text{DP}}(I - I_{\text{DP}})/[I_{\text{DP}}(1 - r_{\text{DP}})]$

  end for

  $I \leftarrow \sum_i \psi_i \Delta m_i$ ▷ Fix f_{PBH}

  for $i = 1, \ldots, N_{\text{bins}}$ **do**

    $\psi'_i \leftarrow \psi'_i \times f_{\text{PBH}}/I$

  end for

  return ψ'

end procedure

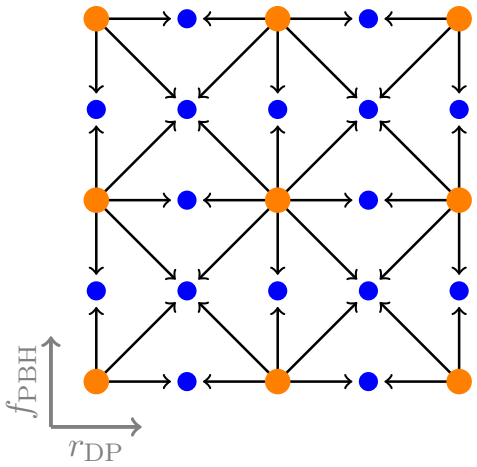

Figure 4.3: Illustration of the refinement procedure. The orange points represent a subset of the initial grid. The blue points are those added during the refinement step. An arrow from an orange point to a blue point denotes that a chain is initialized at the blue point using the optimal mass function across all chains previously evaluated at the orange point.

system, allowing almost any step to proceed, while still being low enough that steps will be constrained as the temperature is lowered.

To verify convergence, we optimize the mass function five times, i.e., with five independent chains, at each parameter point. We evolve each chain for 10^7 steps. Each of these chains begins with its own random mass function and with a high temperature, so convergence to the same optimum merger rate and mass function provides reassuring evidence that the algorithm is not stochastically settling into local optima. We find empirically that the merger rate typically converges across the chains within $\mathcal{O}(10^5)$ steps.

4.3.2.2 Two-parameter optimization

Our goal is to determine the maximum and minimum merger rates as a function of the total abundance of PBH, $f_{\rm PBH}$, and the fraction of those PBH that lie in the

DP mass range, $r_{\rm DP}$. Thus, in principle, we must perform the optimization described in the previous section at every point in this parameter space, independently. However, the optimization process is computationally expensive, so it cannot be applied directly to a fine grid in $(r_{\rm DP}, f_{\rm PBH})$. Instead, we use the simulated annealing algorithm on a coarse grid, and then use an alternative technique to interpolate between the resulting optima.

First, we note that this interpolation process is not simply an aesthetic matter. In principle, a small displacement in the $(r_{\rm DP}, f_{\rm PBH})$ plane can produce a sharp discontinuity in the shape of the optimal mass function, leading to discrete regions in which the optimal mass function evolves very differently with $r_{\rm DP}$ and $f_{\rm PBH}$. This is especially difficult to forecast when observational constraints are included. The situation is analogous to the behavior of the order parameter in a first-order phase transition: in this case, a small displacement in temperature discontinuously changes the location of the minimum of the free energy. In our case, rather than a sharp transition between two minima of the free energy, there could be a sharp transition between two shapes of the mass function. A naive interpolation of a coarse grid of points risks missing any such structure.

Therefore, we extend our coarse grid to a finer subgrid using the following procedure. First, each interval in the grid is halved to produce a refined grid. The initial mass functions for each new point on the grid are borrowed from its nearest neighbours: that is, for each neighbour, we run an independent chain at the new parameter point starting from the neighbour's optimal mass function across all of the neighbour's chains. One step of the process is illustrated in Fig. 4.3. Even if there is a transition of the kind

described above between the new point and some of its neighbours, it is still likely that the optimum at the new point is close in shape to that of at least one neighbour. Thus, one expects that only mild adjustment of these mass functions is needed to converge to the optimum at the new point.

Since we assume that at least one of the optimal mass functions drawn from the neighbouring points is close to the global optimum of the new point, there is no need for the variable temperature of simulated annealing: we need only locate the nearby optimum more precisely. We perform this adjustment by producing a chain of 10^5 steps with the Metropolis–Hastings algorithm, which is structurally similar to simulated annealing, but with a fixed temperature. We perform the entire grid refinement procedure twice to obtain a sufficiently fine grid.

Finally, the optima from all points are 'mixed' as follows. For each point on the refined grid, we generate another set of 10^5-step Metropolis–Hastings chains, each with a different initial mass function. One chain begins with the optimal mass function from the point itself. Another chain is initialized from the optimal mass function of each nearest neighbour. The mixing process is performed four times, so an optimal mass function shape found at any point in a block between initial grid points can propagate to other points in the same block.

After the refinement and mixing processes are performed, the result is a non-trivial interpolation of the initial grid, which forms the basis for our results in the following sections.

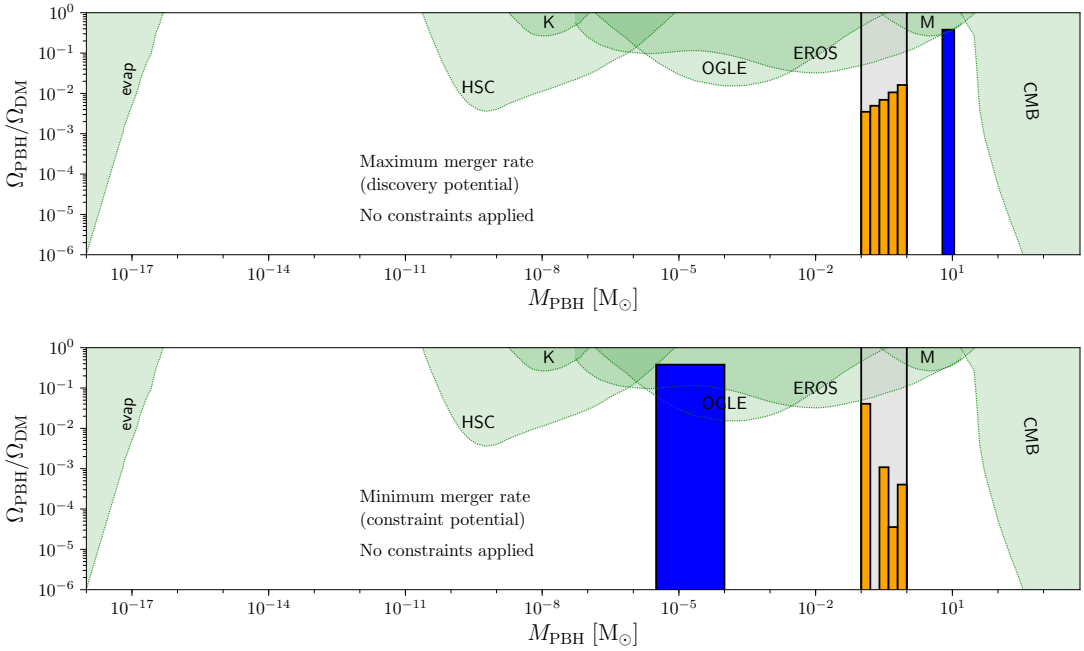

Figure 4.4: Optimal maximizing (top) and minimizing (bottom) mass functions with $f_{\rm PBH} = 0.5$ and $r_{\rm DP} = 0.1$ in the absence of observational constraints. Each mass function is shown as a set of discrete bars with height $\Omega_{\rm PBH}(m_i)/\Omega_{\rm DM} \equiv \psi(m_i)\Delta m_i$, i.e., the height of each bar indicates the total mass in the bin. The maximum merger rate corresponds to the most optimistic discovery potential, and the minimum merger rate to the most pessimistic constraint potential. The constraint curves are *not* used to constrain the mass function, and are shown here only for reference and comparison with Fig. 4.5. The DP window is indicated by the shaded gray region, and the mass function is colored orange therein. The labelled constraints are from BH evaporation [`evap`; 28], Hyper Suprime-Cam [`HSC`; 13, 15], Kepler [`K`; 34], OGLE [`Ogle`; 65], EROS-II [`EROS`; 253], MACHO [`M`; 10], and CMB observables [`CMB`; 36, 39]. Other constraints may also apply, but their inclusion does not influence our qualitative conclusions (see Section 4.3).

4.4 Results

We now examine the results of our numerical optimization. First we show results for individual parameter points, and compare the shapes of optimal mass functions to our analytical expectations. Then we show minimal and maximal merger rates with and without observational constraints.

4.4.1 Shape of the mass function

To understand the shapes of the mass functions that optimize the merger rate, we first neglect observational constraints to facilitate comparison with the analysis in Section 4.2.3. Figure 4.4 shows mass functions that minimize and maximize the merger rate without regard to observational constraints for the parameter point $(r_{\rm DP}, f_{\rm PBH}) = (0.1, 0.5)$ (i.e. 50 per cent of DM is in PBH of any mass, and 10 per cent of the PBH density is accounted for by the DP window). The two mass functions are mostly distinguished by two features. First, they have clearly different behavior outside the DP window: the DP merger rate is enhanced when the remainder of the PBH are placed at a higher mass, above the top of the window, and it is reduced when they are placed at lower mass, below the bottom of the window. Secondly, as expected from our simplified analysis in Section 4.2.3, the maximizer is broad within the DP window, while the minimizer is sharply peaked and multimodal.

Contrary to our naive expectation, the maximizer prefers higher masses within the DP window, while the minimizer prefers lower masses. This is because the full numerical calculation accounts for detectability, and the mergers of heavier black holes

are detectable in a larger volume. This also accounts for the behavior of the mass function outside the DP window. Recall that the mergers of DP black holes with *heavier* black holes are generally observable, and we assume that the lighter black hole is identifiably primordial in such a merger. However, the merger of a DP black hole with a *lighter* black hole may not be observable, or may be observable only in such a small effective volume that our assumptions for calculating the merger rate are not valid. Thus, if the 90 per cent of PBH which lie outside the DP window are positioned at higher masses, the observable merger rate is enhanced.

Having noted the behavior of the optimal mass functions in the absence of observational constraints, we now turn to the results of constrained optimization in Fig. 4.5. The general features of these optima are similar to their unconstrained counterparts, and observational constraints modify the shapes of the optimal mass functions in a comprehensible manner. The maximal merger rate is still obtained with additional PBH positioned above the DP window, but observational bounds now strictly constrain the position of this peak. The mass function which minimizes the merger rate is not subject to strong constraints within the DP window, but the additional PBH at lower masses must now be repositioned to the mass range where constraints are weaker.

4.4.2 Constraints and discovery prospects

Minimum and maximum merger rates with all constraints applied are shown as a function of $r_{\rm DP}$ and $f_{\rm PBH}$ in Figs. 4.6 and 4.7. The minimum merger rate corresponds to LIGO's constraint potential: even given complete freedom in the mass function, there is no way to obtain a lower observable merger rate. The maximum merger rate

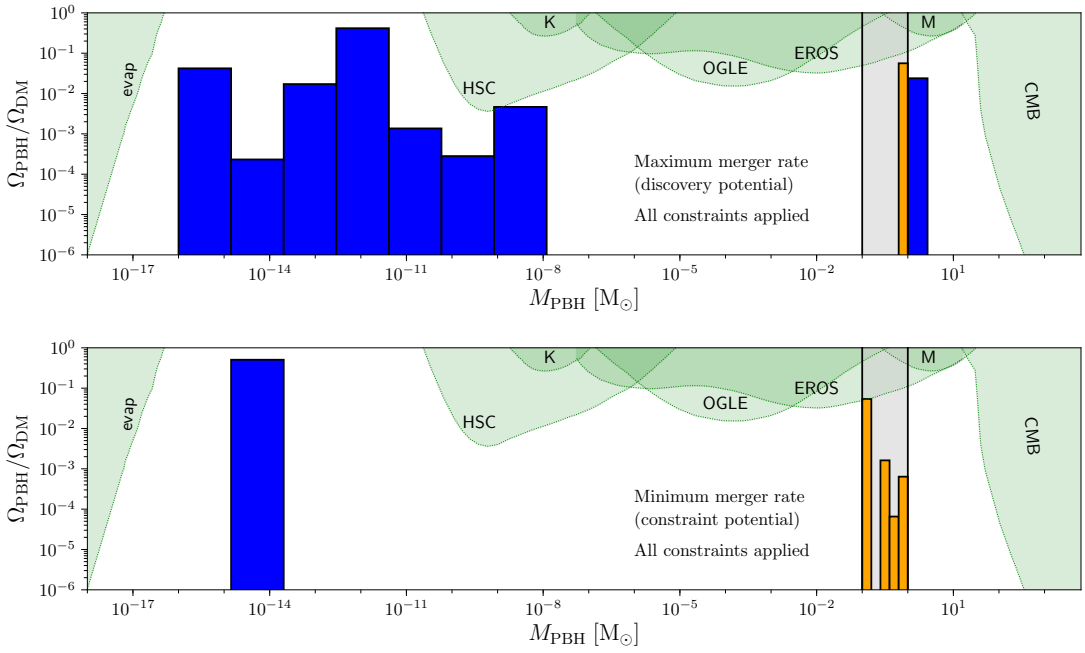

Figure 4.5: Optimal maximizing (top) and minimizing (bottom) mass functions with $f_{\rm PBH} = 0.5$ and $r_{\rm DP} = 0.1$. Each mass function is shown as a set of discrete bars with height $\Omega_{\rm PBH}(m_i)/\Omega_{\rm DM} \equiv \psi(m_i)\Delta m_i$, i.e., the height of each bar indicates the total mass in the bin. All observational constraints are applied. The maximum merger rate corresponds to the most optimistic discovery potential, and the minimum merger rate to the most pessimistic constraint potential. The DP window is indicated by the shaded gray region, and the mass function is coloured orange therein. The labelled constraints are from BH evaporation [`evap`; 28], Hyper Suprime-Cam [`HSC`; 13, 15], Kepler [`K`; 34], OGLE [`Ogle`; 65], EROS-II [`EROS`; 253], MACHO [`M`; 10], and CMB observables [`CMB`; 36, 39]. Other constraints may also apply, but their inclusion does not influence our qualitative conclusions (see Section 4.3).

corresponds to the discovery potential, i.e., the most optimistic scenario given any mass function.

The light gray region in the top right corner of each panel indicates parameter points where the numerical procedure was unable to locate any mass function consistent with constraints. This is the portion of parameter space that is already ruled out by other observables, including the SGWB. The extent of this region can be estimated using the semi-analytical procedure of [30], which can give the maximum allowed value of $f_{\rm PBH}$ for fixed $r_{\rm DP}$ if SGWB is neglected. This bound is the triangular dark gray region. Since the SGWB depends non-linearly on the mass function, it cannot be treated within the same semi-analytical framework. Thus, one expects the light gray region to extend slightly further than the dark gray region, which is exactly the behavior shown in Figs. 4.6 and 4.7.

Observe that there is a small gap between the minimum and maximum merger rates when $r_{\rm DP}$ is near 1. This is simply because there is very limited freedom in the mass function under these conditions. On the other hand, when $r_{\rm DP} \ll 1$, the minimum and maximum merger rates are radically different. In particular, while LIGO can only rule out mass functions with $r_{\rm DP} \gtrsim 0.1$, it can potentially discover PBH with only $r_{\rm DP} \gtrsim 10^{-4}$ with $\mathcal{O}(1\,\mathrm{yr})$ of data. The effect of observational constraints is evident from Fig. 4.8: in the absence of constraints, LIGO would potentially be sensitive to mergers of a subcomponent as small as $r_{\rm DP} \sim 10^{-6}$.

Finally, we note that the strength of the constraints is dependent on $r_{\rm DP}$ and $f_{\rm PBH}$ separately. One might expect the constraints to scale mainly with the product $f_{\rm DP} = r_{\rm DP} f_{\rm PBH}$, i.e., the total abundance of black holes in the DP window. This is

indeed the case for small values of $f_{\rm DP}$. However, at larger values of $f_{\rm DP}$, there are three effects that cause the constraints to depend on each of $r_{\rm DP}$ and $f_{\rm PBH}$ beyond their product. First, there is the uneven role of the observational constraints themselves. These have a complicated mass dependence, and thus introduce such dependence in the optimization results by limiting freedom in the mass function. This is mainly important at large $f_{\rm DP}$. Secondly, there is the difference between the DP window for single PBHs and the DP^2 window for binaries: a PBH outside the DP window can still contribute to the DP^2 merger rate by merging with a lighter PBH. Thirdly, even PBHs that do not participate in DP^2 mergers contribute to the formation and disruption of binaries in the DP^2 window. This holds true for both the merger rate of Ref. [54] and that of Ref. [238].

4.4.3 Convergence

For an optimization problem of this kind, which is not generally convex, there is no reliable test of algorithmic convergence. In principle, it is always possible that the loss landscape has not been fully explored, and that in some corner, there is a point that outperforms the optima that we have discovered numerically. The best defense against this issue is to compare the numerical results against simplified analytical benchmarks, as we have carried out above.

However, we also perform two more direct tests of convergence. First, we have verified that we locate the global optimum in a low-dimensional example, where the features of the loss function can be analysed by inspection; and secondly, we perform a purely numerical test of convergence by comparing the results of many MCMC chains

initialized in random configurations. We thus check directly that at benchmark points, all of our chains converge to the same merger rate within our fixed step count.

Numerical convergence is also supported qualitatively by comparison of nearby parameter points. Since we perform the optimization procedure on a grid of points in the $(r_{\rm DP}, f_{\rm PBH})$ plane, nearest neighbours in this plane should converge to similar optima. Since the contours in Figs. 4.6 and 4.7 are smooth, one might conclude that this constitutes evidence of convergence. However, note that in Figs. 4.6 and 4.7, optima from an initial run have been mixed between parameter points, as described in Section 4.3.2.2. In particular, if a global optimum is discovered at only one point, it will subsequently propagate to the rest of the parameter space, even if chains originally produced elsewhere located very different optima. Thus, smoothness of the contours is only meaningful before mixing. Since the initial grid with random priors is relatively sparse, smoothness is difficult to assess quantitatively. However, we have verified that the qualitative features of the contours in Figs. 4.6 and 4.7 are not affected by the mixing procedure, suggesting that each of the points in the initial grid is locating nearly the same optimum as that produced after mixing. Note that the sharp behavior at the top of Fig. 4.7 is entirely due to observational constraints, and disappears in their absence.

4.5 Discussion and conclusions

The discovery of PBH would be a tremendous step forward in our understanding of cosmology. If PBH exist, they encode information about cosmic history in an epoch that we have yet to probe observationally. They also provide an empirical test

of physics at extremely high scales and early times. Moreover, despite all observational constraints, PBH remain a viable and extremely simple candidate for cosmological dark matter.

Conveniently, any black hole with a mass below $\sim 1\,\mathrm{M}_\odot$ cannot have an astrophysical origin. Gravitational wave observatories are well suited to identify black holes and to measure their masses precisely, so these instruments can detect a smoking-gun signature of the existence of PBH. Even one detection event involving a light black hole would provide unambiguous evidence for new physics. Subsequent exploration of the abundance and distribution of such black holes would test the possible formation scenarios, and potentially provide a direct handle on physics at very early times.

The problem lies in the interpretation of a null observational result. In principle, experimental results at LIGO constrain the population of light PBH, and in principle, again, LIGO may be sensitive to a very small abundance of such objects. However, both of these statements have a non-trivial dependence on the shape of the PBH mass function. Since the merger rate has a complicated non-linear dependence on the mass function, it is difficult to directly assess the significance of this uncertainty. In particular, the semi-analytical analysis of [30] cannot accommodate the merger rate as a constraint on the PBH abundance.

In this chapter, we have quantified the uncertainty in the PBH merger rate that arises from freedom in the mass function, while accounting for observational constraints that restrict its shape. This uncertainty is reflected in the gap between the minimum-rate and maximum-rate contours in Fig. 4.8. While the two bands are not far apart for $r_{\mathrm{DP}} \sim 1$, they are significantly different when $r_{\mathrm{DP}} \ll 1$. Thus, it is necessary to

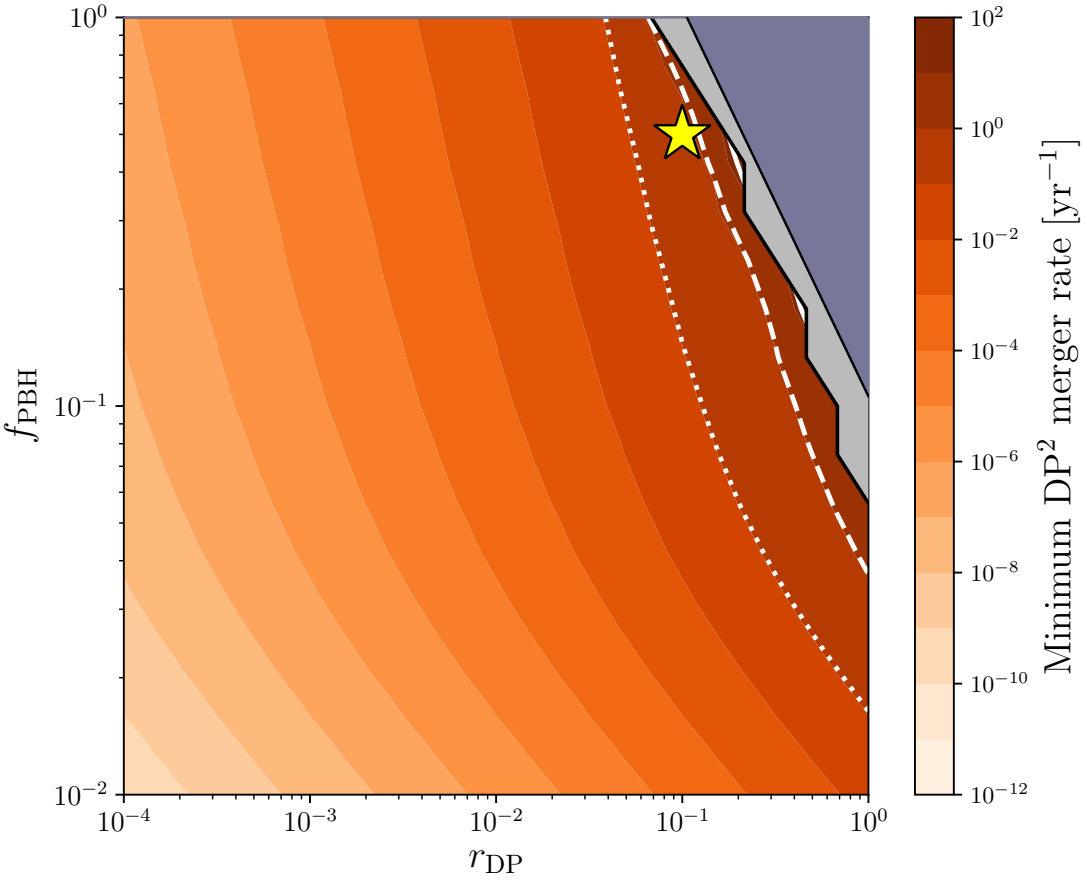

Figure 4.6: Minimum DP merger rate for mass functions constrained by all observables, including SGWB. The triangular region at the top right is ruled out by non-GW observables. The light region is ruled out by the combination of all observables. The solid, dashed, and dotted curves show contours with an *observed* DP merger rate of 10, 1, and 0.1 yr^{-1}, respectively. The star ($\star$) indicates the point shown in the bottom panel of Fig. 4.5.

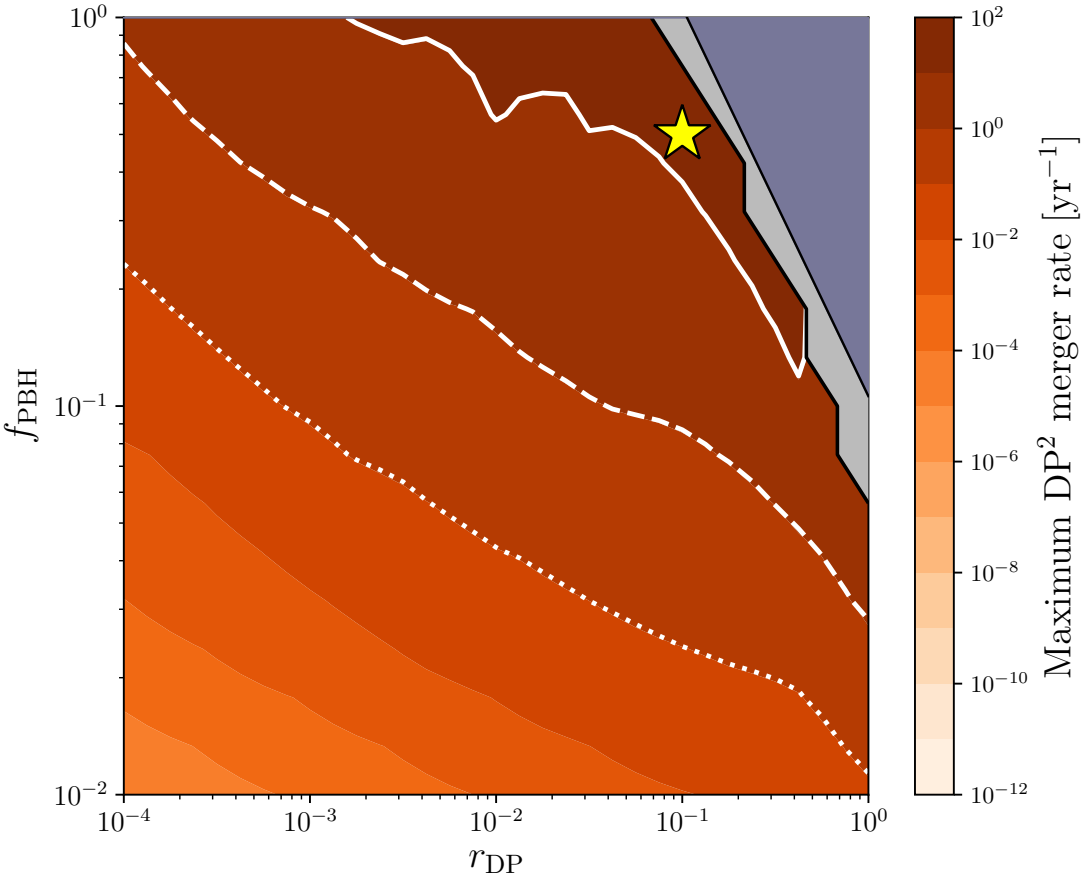

Figure 4.7: Maximum DP merger rate for mass functions constrained by all observables, including SGWB. The triangular region at the top right is ruled out by non-GW observables. The light region is ruled out by the combination of all observables. The solid, dashed, and dotted curves show contours with an *observed* DP merger rate of 10, 1, and 0.1 yr^{-1}, respectively. The star ($\star$) indicates the point shown in the top panel of Fig. 4.5.

consider the two contours as reflecting different notions of experimental sensitivity at LIGO. The minimal merger rate determines the extent of constraints that LIGO can set, if the mass function is allowed to vary freely. Conversely, the maximal merger rate determines the extent of the parameter space that can be probed by LIGO in the most optimistic scenario.

Our numerical results indicate that LIGO's constraint potential is limited to parameter space with $r_{\rm DP} \gtrsim 0.1$, and the prospects for improving this bound with binary black hole mergers are limited. On the other hand, LIGO's discovery potential extends as low as $r_{\rm DP} \sim 10^{-4}$, meaning that even a very small subcomponent of the PBH population that lies in our DP window can potentially yield a detection. This also establishes the relevance of constraints provided by other observables: in the absence of observational constraints, LIGO would be sensitive to $r_{\rm DP} \sim 10^{-6}$. Our results highlight the importance of evaluating detection prospects for specific PBH models using the full apparatus of the merger rate for extended mass functions—a small subcomponent of DPBH cannot be neglected.

One might wonder whether the optimal mass functions we consider in this chapter are realistic. Generally, there is good motivation to consider only specific forms of the mass function, particularly monochromatic, lognormal, or power-law shapes. However, most of the behavior that characterizes our optimal mass functions is captured by doubly or triply peaked mass functions, and note that a population of PBH with a multimodal mass function can easily be generated after inflation [233]. Thus, while the exact form of our optimal mass functions might require fine-tuning of initial conditions, approximate forms that retain a high or low merger rate are much more

generic. The non-trivial requirement is that a peak should fall near the DP window to maximize discovery prospects. As yet, there is no direct evidence for such placement, but only circumstantial evidence from the distribution of mergers observed thus far.

Our results are inherently subject to theoretical uncertainties in the computation of the merger rate. While the form of the merger rate employed here reflects one of the most comprehensive estimates currently available, such formulae are best suited only to computations at the order of magnitude level. For instance, one potential issue in the rate calculation is the effect of other black holes in disrupting the formation of a binary. In our calculation, as discussed by [54], we assume that two black holes of mass m_i and m_j do not form a binary if another black hole of mass $m_k \geq \min\{m_i, m_j\}$ is present in the volume between them. However, even if this were always the case, it is also possible that somewhat lighter black holes would have a similar effect. This would provide a mechanism for suppression of the merger rate, reducing the discovery potential and weakening the constraint we draw in this chapter.

As a cross-check, we have also computed the merger rate for each of our optimal mass functions using the formalism of Ref. [238]. Here, the influence of perturbing black holes on the binary is calculated by an entirely different method, as discussed in Section 4.2. For our optimal mass functions, the merger rates obtained in each of the two formalisms are comparable, generally differing by an $\mathcal{O}(1)$ factor, but the difference can be as large as $\mathcal{O}(10)$ for some points. On the one hand, this is quite good agreement, given that these are two structurally different calculations with many inherent uncertainties, applied to complicated mass functions which differ substantially from standard benchmarks. On the other hand, the disagreement requires that we limit

the interpretation of our results to the order-of-magnitude level.

Along similar lines, Ref. [239] recently showed numerically that including all subsequent three-body encounters after binary formation can significantly reduce the merger rate. The suppression described in that work can be as small as a $\mathcal{O}(2\text{–}20)$ factor, or as large as a $\mathcal{O}(10^3)$ factor, depending on the clustering properties of PBH. We thus consider a reduction of our calculated merger rate by at least a factor of $\mathcal{O}(10)$ to be physically well motivated. Thus, even at the order-of-magnitude level, it is possible that we overestimate the merger rate somewhat.

In light of the differences between Ref. [54] and Ref. [238], and the uncertainties suggested by Ref. [239], it is clear that any qualitative interpretation must include these substantial systematics. We therefore include contours with a merger rate of $10\,\mathrm{yr}^{-1}$ in Figs. 4.6 and 4.7. In this case, LIGO's discovery potential is reduced to $r_{\mathrm{DP}} \gtrsim 10^{-3}$, and constraint potential is lost completely: the $10\,\mathrm{yr}^{-1}$ contour in Fig. 4.6 is covered almost entirely by the existing non-merger constraints. Note, however, that if mergers of binaries formed in the early Universe are suppressed, binaries formed in the late Universe may make an important contribution to the rate, particularly if the density contrast in the late Universe is larger than expected. Ultimately, barring extreme modifications to the merger rate, our qualitative results stand. In particular, the gap between the maximal and minimal merger rates is very large at small r_{DP}, and is robust to adjustments in the calculation of the merger rate. However, further refinement in the prediction of the merger rate is certainly motivated.

In this chapter, we have focused on the direct observation of DP black holes as a smoking gun of the primordial-origin scenario. In the absence of such a direct signature,

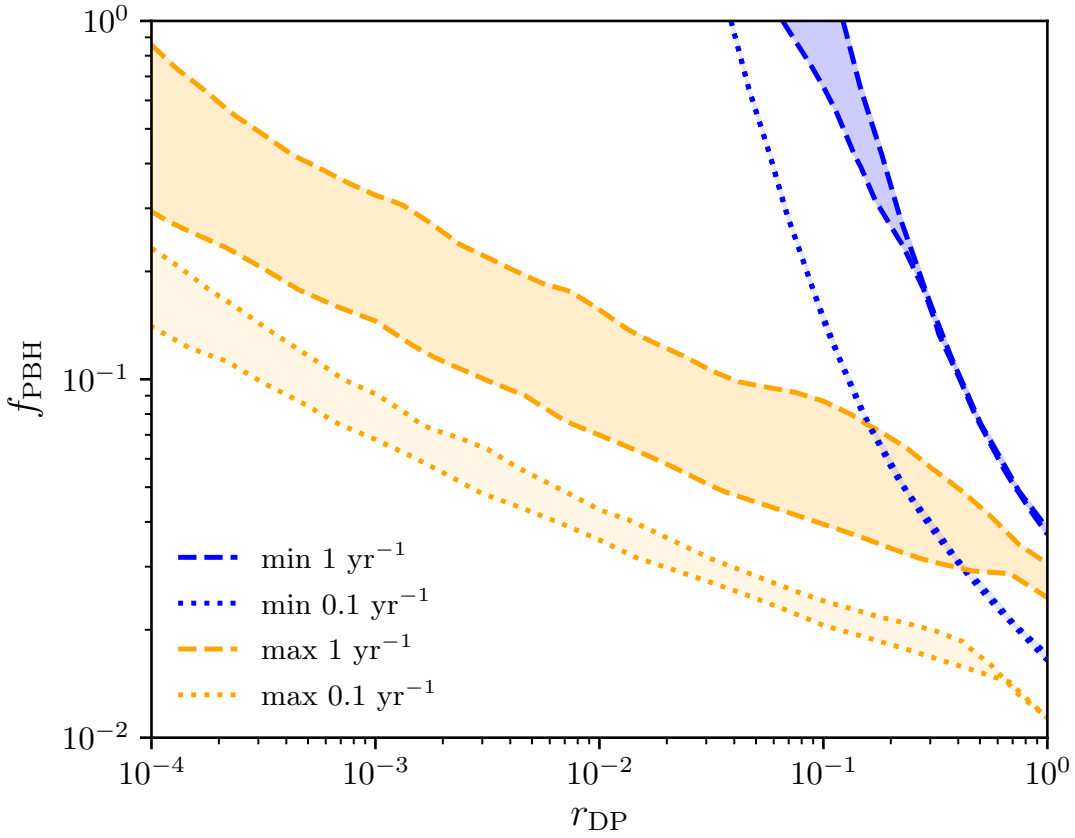

Figure 4.8: Contours with DP merger rate fixed to $1\,\mathrm{yr}^{-1}$ (dashed) and $0.1\,\mathrm{yr}^{-1}$ (dotted), with and without observational constraints. The bottom of each band shows the sensitivity with complete freedom in the mass function, and the top shows the sensitivity when observational constraints are included. The blue curves show the minimum merger rate, corresponding to the constraint potential. Note that the 10-yr minimum curves with and without constraints are essentially identical.

the SGWB associated with mergers over cosmic time may provide an additional probe of the PBH abundance. We do not evaluate SGWB as a discovery mechanism simply because such a detection would not constitute unambiguous evidence of new physics. It is possible that features of the SGWB may be connected to the features of the PBH population with enough precision to empirically test specific models, but since other physical mechanisms might also contribute to the SGWB, significant additional work would be required to confirm the existence of a population of PBH. However, we emphasize that the SGWB is still a sensitive probe of black holes in the DP window. In particular, mass function shapes that greatly enhance the merger rate can be ruled out by SGWB limits. In our framework, while we do not examine the SGWB as a tool for discovery, we do consistently include this observable as one of our constraints on the mass function: all of the constrained optima we consider, including those with maximal merger rates, are compatible with existing SGWB constraints. Significant improvement of observational bounds on the SGWB might limit freedom in the mass function, and might thus limit the discovery potential we infer in this chapter.

Our results show that while LIGO has limited power to constrain the abundance of light PBH, it nonetheless has significant discovery potential. The major obstruction to such sensitivity is not the sensitivity of the LIGO instrument, but the analysis pipeline. There are significant computational costs to conducting searches for mergers of light black holes, as discussed extensively by [68], and these costs increase further if one searches for mergers of light black holes with heavier black holes. However, the freedom in the mass function and the associated uncertainty in the merger rate provides ample motivation for the refinement of methods for such searches, and even

for the dedication of additional computational resources. A single observation of this type would have immense value, and gravitational wave observatories are in a unique position to make such a discovery.

Chapter 5

Probing new forces with supermassive black holes

5.1 Introduction

Beyond gravitational waves from individual mergers, pulsar timing arrays (PTAs) are on the verge of a historic discovery: the detection of a stochastic gravitational wave background (SGWB) produced by supermassive black hole (SMBH) binaries. PTAs use the extremely stable timing of successive light pulses from pulsars to detect gravitational waves (GWs) in the form of correlated timing distortions. In the presence of GWs, the observed time between pulses deviates from the stable rhythm in the frame of the source. The correlation of these deviations between pulsars exhibits a characteristic dependence on their angular separation, known as the Hellings and Downs curve [254], and this is considered the hallmark of a GW detection.

Three major pulsar timing collaborations are currently searching for the GW

background: NANOGrav [255], the European Pulsar Timing Array (EPTA) [256], and the Parkes Pulsar Timing Array (PPTA) [257]. The sensitivities of these experiments vary based on their observational samples. Recently, the NANOGrav experiment has found evidence for a statistically significant correlated signal among a collection of $\mathcal{O}(50)$ pulsars in its 12.5-year dataset [258]. This may be the first signal of the SGWB from SMBH mergers, which would mark a monumental event in the history of GW astronomy. While the signal does not yet conclusively exhibit the Hellings and Downs angular dependence, upcoming datasets from NANOGrav and the other collaborations will be able to definitively confirm or refute the prospective discovery. Regardless of the fate of this particular signal, upcoming radio telescopes such FAST [259] and SKA [260] will be sensitive to stochastic backgrounds well below even the most pessimistic predictions for the SGWB amplitude [261], so a conclusive detection is expected in the near future.

In this chapter, we point out that beyond astrophysical and cosmological applications, the study of the SGWB from SMBH mergers will open an entirely new observable for particle physics: the spectral shape of the SGWB. If binaries are driven to merge by gravitational radiation alone, then the frequency dependence of the SGWB is cleanly predicted to be a power law with a known index. We show that physics beyond the Standard Model can modify this prediction of the spectral shape, and we thus evaluate the possibility of using forthcoming observations of the SGWB from SMBHs to test fundamental physics.

Previous work on using GW emission to detect new forces in binaries has been focused on pulsar systems [262–264] such as the Hulse–Taylor binary [265] and individual binaries with $\mathcal{O}(\mathrm{M}_\odot)$ masses [263, 264, 266–270] using recent detections made

by the LIGO/Virgo Collaborations [224, 229]. However, the unique evolution and environmental properties of SMBH binaries make them a rich laboratory to search for physics beyond the Standard Model. SMBHs copiously accrete nearby matter as they grow and they can develop powerful jets that accelerate particles to energies well above the electroweak scale [271]. This may result in accretion or production of new particles, resulting in these SMBHs and their surroundings acquiring exotic quantum numbers. Moreover, SMBHs are known to be surrounded by a diffuse dark matter (DM) halo, and if this halo influences the dynamics of SMBH binaries, then these objects may be sensitive to dark matter interactions.

Such scenarios illustrate a range of new physics that may be discovered in SMBH mergers. Thus, the imminent measurement of the SGWB opens up an opportunity to access a wide range of physics inaccessible to terrestrial experiments. In this chapter, we focus on the simple possibility that SMBHs in merging binaries carry a charge under a new ('fifth') force with a long but possibly finite range. We study how the standard power-law prediction for the SGWB spectrum is modified in the presence of such a new force, presenting the spectral index as a robust prediction.

For simplicity, we assume that the SMBHs themselves are charged, but our results also hold if a bound cloud of charge surrounds each SMBH. Our main assumption is that the charge distribution near each SMBH is pointlike on the scale of the binary separation. Detailed mechanisms for the accumulation of charge on and near SMBHs will be the subject of future work. We emphasize that any new physics which impacts the dynamics of merging binaries is potentially observable via the SGWB spectrum. We study a new long-range force as a benchmark scenario because this case is

easily parametrized and demonstrates the key implications for the SGWB spectrum, but similar techniques can be applied in a variety of other scenarios.

This chapter is organized as follows. In Section 5.2, we review the dynamics of a single binary in the presence of a new force. In Section 5.3 we present a calculation of the stochastic spectrum, highlighting various potential systematic uncertainties. We present our results in Section 5.4 in light of current constraints and the recent NANOGrav measurement.

Throughout this chapter, we denote the binary component masses by M_1 and M_2. We use ω for the orbital angular frequency of the binary, $f_{\rm s}$ for the GW frequency in the frame of the source, and f for the observed GW frequency.

5.2 The spectrum of SMBH mergers

The SGWB from SMBH mergers has been studied extensively in the absence of new physics [261, 272–277]. Despite the complex astrophysical environments of SMBH mergers, the shape of the resulting SGWB spectrum can be predicted cleanly for a simple reason: the SGWB is dominated by contributions from binaries in the final stages of inspiral, where binary evolution is dominated by the emission of gravitational radiation with little pollution from environmental influences. Thus, there is a tight relationship between the radiated power and the hardening of the binary, leading in turn to a robust prediction for the shape of the GW spectrum.

The amplitude of the SGWB spectrum at a given frequency f can be described in several ways. To facilitate comparison with existing literature, we discuss the spec-

trum in terms of the characteristic strain h_c. This is related to the energy density $\Omega_{\rm GW}$ by [278]

$$h_c^2(f) = \frac{3H_0^2}{2\pi^2 f^2}\Omega_{\rm GW}(f) \equiv \left[A_{\rm GW} \times \left(\frac{f}{\rm yr^{-1}}\right)^{\beta}\right]^2 , \tag{5.1}$$

where H_0 is the Hubble parameter, $A_{\rm GW}$ is the dimensionless amplitude (the value of the characteristic strain evaluated at an inverse-year), and β is the spectral index. For a spectrum that is not a power law, we allow β to be frequency-dependent.

The full SGWB spectrum can then be computed by combining the spectra of individual mergers over cosmic time. Following Ref. [279], the characteristic strain of the SGWB observed at a frequency f is given by

$$h_c^2(f) = \frac{3H_0^2}{2\pi^2 \rho_c f^2} \int {\rm d}z\, {\rm d}\mathbf{X}\, \frac{{\rm d}n_{\rm s}}{{\rm d}z\, {\rm d}\mathbf{X}} \frac{f_{\rm s}}{1+z} \left.\frac{{\rm d}E_{\rm GW}}{{\rm d}f_{\rm s}}\right|_{\mathbf{X}} , \tag{5.2}$$

where $n_{\rm s}$ is the comoving number density of GW sources, $f_{\rm s} = (1+z)f$ is the frequency in the frame of the source, and ${\rm d}E_{\rm GW}/{\rm d}f_{\rm s}$ is the energy spectrum produced by a single source. Here $\mathbf{X}$ denotes the state variables needed to determine the spectrum of a single source. If the sources are circular SMBH binaries, then in the absence of new physics, $\mathbf{X}$ simply denotes the component masses.[1] As shown in the Appendix, the dominant contribution to the integral arises from redshifts of $z \lesssim 0.3$ and SMBH masses between $10^8\,{\rm M}_\odot$ and $10^9\,{\rm M}_\odot$.

In the frequency range accessible to PTAs, the observable SGWB signal is expected to be dominated by SMBH binaries in the late stages of inspiral, where gravitational radiation is the primary mechanism for the binary to lose mechanical energy. As we review below, a merger driven by gravitational radiation alone produces a GW

[1] The binary separation is not an additional parameter of SMBH binary sources, since the spectrum ${\rm d}E_{\rm GW}/{\rm d}f_{\rm s}$ is obtained by integrating over all stages of binary evolution.

spectrum with shape $\mathrm{d}E_{\rm GW}/\mathrm{d}f_{\rm s} \propto {f_{\rm s}}^{-1/3}$. This means that the frequency dependence in Eq. (5.2) can be factored out of the integral, and we obtain $h_c \propto f^{-2/3}$. Thus $\beta = -2/3$ independent of the properties of the binary population.

Our central result is that new forces between the binary components can modify this spectral shape by modifying the single-merger spectrum $\mathrm{d}E_{\rm GW}/\mathrm{d}f_{\rm s}$. We now turn to the calculation of this spectrum in the presence of new physics, starting with the calculation of the spectrum $\mathrm{d}E_{\rm GW}/\mathrm{d}\omega$ as a function of the orbital frequency ω.

The shape of the spectrum $\mathrm{d}E_{\rm GW}/\mathrm{d}\omega$ is caused by the rise in orbital frequency ω as the SMBH separation r falls over time, according to

$$\frac{\mathrm{d}E_{\rm GW}}{\mathrm{d}\omega} = \frac{\mathrm{d}E_{\rm GW}}{\mathrm{d}t}\frac{\mathrm{d}t}{\mathrm{d}r}\frac{\mathrm{d}r}{\mathrm{d}\omega}\,. \tag{5.3}$$

The orbital frequency ω is fixed by the SMBH separation through central forces acting between the binary components, via

$$\mu\omega^2 r = F(r)\,, \tag{5.4}$$

where $\mu = M_1 M_2/(M_1 + M_2)$ is the reduced mass of the system. Orbital decay then occurs as the mechanical energy $E_{\rm mech} = \frac{1}{2}\mu r^2\omega^2 + U(r)$ of the binary is lost to radiation, with energy per unit time $P_{\rm rad}$. Conservation of energy gives

$$0 = \frac{\mathrm{d}E_{\rm mech}}{\mathrm{d}t} + P_{\rm rad} = \mu r\omega^2\left(2 + \frac{r}{\omega}\frac{\mathrm{d}\omega}{\mathrm{d}r}\right)\frac{\mathrm{d}r}{\mathrm{d}t} + P_{\rm rad}\,. \tag{5.5}$$

After solving Eq. (5.5) for $\mathrm{d}r/\mathrm{d}t$, substituting into Eq. (5.3), and expressing the orbital frequency in terms of the radiation frequency as $\omega = \pi f_{\rm s}$, the spectrum of GWs produced by the merger of a single binary is

$$\frac{\mathrm{d}E_{\rm GW}}{\mathrm{d}f_{\rm s}} = -\pi^2\mu r^2 f_{\rm s}\left(\frac{2f_{\rm s}}{r}\frac{\mathrm{d}r}{\mathrm{d}f_{\rm s}} + 1\right)\frac{P_{\rm GW}}{P_{\rm rad}}\,, \tag{5.6}$$

where r is determined as a function of f_s by Eq. (5.4).

In the case that the evolution of the binary is dominated by gravity, the central force is given by Newton's law, $F(r) = GM_1M_2/r^2$, so that Eq. (5.4) yields the well-known Kepler relation

$$\omega^2 = \frac{G(M_1 + M_2)}{r^3}\,. \tag{5.7}$$

Meanwhile, orbital decay occurs predominantly through quadrupole radiation of gravity waves, with a power given by (see, e.g., Ref. [280])

$$P_{\rm rad} = P_{\rm GW} = \frac{32}{5} G\mu^2\omega^6 r^4\,. \tag{5.8}$$

The spectrum of GWs radiated in a single merger, Eq. (5.6), becomes

$$\frac{{\rm d}E_{\rm GW}}{{\rm d}f_s} = \frac{\mu}{3}\Big[\frac{\pi^2 G^2 (M_1 + M_2)^2}{f_s}\Big]^{1/3}\,, \tag{5.9}$$

which exhibits a spectral index of $-1/3$. This gives rise to the spectral index of $\beta = -2/3$ for the characteristic strain spectrum h_c when integrated over cosmic time, according to Eq. (5.2).

We now consider the effects of a new force mediated by a particle of mass m on the SGWB spectrum, similar to the treatment of individual neutron star binaries in Ref. [266]. We emphasize that our main assumption is that the charge distribution remains pointlike relative to the binary separation, which is of order $10^{-2}\,$pc in the PTA window. Thus, we do not require that the SMBHs themselves are charged. Still, we note that the particle nature of the additional species has important implications for charge stability in any concrete model where the SMBHs are directly charged. Firstly, charged black holes can neutralize by emission of charged particles. For Standard Model

electric charge, this process is very slow and can be neglected for SMBHs with masses of order $10^9\,\mathrm{M}_\odot$ [281]. This may or may not be the case for the new charge as well, depending on the mass and coupling of the lightest charged state. Secondly, for a vector mediator with $m > 0$ or for a scalar mediator, no-hair theorems suggest that charge deposited directly onto an SMBH is not stable. For a massive vector, the effective charge of the SMBH decays on a timescale of order m^{-1} [282]. We will be interested in extremely light mediators, corresponding to a relatively long timescale for decay. If the SMBHs are charged by a mechanism that remains active throughout the evolution of a binary, then no-hair theorems imply a reduction in the equilibrium charge, but do not necessarily preclude significant charges on the SMBHs themselves. On the other hand, a force mediated by a scalar can act directly on the SMBHs only in exceptional circumstances.

We now proceed to compute the GW spectrum in the presence of the new force, regardless of the particle nature of the interaction or whether the SMBHs are directly charged. The net force between the SMBH binary components is modified by the addition of a short-range contribution given by

$$F = \frac{GM_1M_2}{r^2}\Big(1 - \alpha e^{-mr}(1+mr)\Big)\,, \tag{5.10}$$

where the potential-strength parameter α parametrizes the strength of the new force. We use the convention that the force is repulsive if $\alpha > 0$. The potential-strength parameter is given by

$$\alpha = \frac{Q_1Q_2}{GM_1M_2}\,, \tag{5.11}$$

where Q_1 and Q_2 are effective dark charges on the SMBHs. The normalization is chosen

so that $\alpha = 1$ for two extremal black holes if they were directly charged. Note, however, that the effective charges might arise from a charged cloud of particles surrounding the SMBHs. We defer a more detailed treatment of this scenario to future work.

The new force can supply another contribution to energy loss in the form of dipole radiation, $P_{\rm dip}$. The precise dependence on the frequency depends on the spin of the new mediator [262],

$$P_{\rm dip} = \tfrac{1}{3} G\gamma^2\mu^2 r^2\omega^4\sqrt{1-\tfrac{m^2}{\omega^2}} \times \begin{cases} 1-\frac{m^2}{\omega^2}, & \text{(scalar)} \\ \\ 2+\frac{m^2}{\omega^2}, & \text{(vector)} \end{cases} \tag{5.12}$$

where the dimensionless dipole-strength parameter γ characterizes the strength of radiation and is given in terms of the SMBH charges and masses by

$$\gamma^2 = \frac{1}{G}\left(\frac{Q_1}{M_1} - \frac{Q_2}{M_2}\right)^2 . \tag{5.13}$$

Since nonzero γ sources dipole radiation, its effect on energy loss is parametrically enhanced relative to the quadrupole gravitational radiation. The enhancement is given by

$$\frac{P_{\rm dip}}{P_{\rm GW}} = \frac{5\gamma^2}{48 r^2\omega^2} \simeq 20\ \gamma^2\Big(\frac{10^9\,{\rm M}_\odot}{M_1+M_2}\Big)^{2/3}\Big(\frac{{\rm yr}^{-1}}{\omega}\Big)^{2/3}, \tag{5.14}$$

where, in the second approximation, r has been traded for ω using Eq. (5.7), assuming $\alpha = 0$ and $m = 0$. This shows that for $\gamma = 1$, the power lost to dipole radiation is about 20 times larger than that lost to gravitational radiation for the GW frequencies probed by pulsar timing experiments.

5.3 New forces in the SGWB spectrum

We now predict the SGWB spectrum in the presence of new forces and compare the novel spectral features to astrophysical systematics.

We compute the observed strain using Eq. (5.2) in combination with Eq. (5.6). This calculation requires an estimate of the number density of merging SMBH binaries, $n_{\rm s}$. SMBH binaries form when their host galaxies merge and their central BHs sink to the center by dynamical friction [283], so the SMBH merger rate depends on the galactic merger rate. The abundance and properties of galaxy pairs can be inferred from astronomical observations, and empirical scaling relations can then be used to connect galaxy properties to the properties of their resident SMBHs. To compute the SGWB spectrum including the normalization, we follow the procedure detailed in Ref. [276], taking the galaxy mass function from Ref. [284, 285], the black hole–bulge mass relation from Ref. [286], and the pair fraction from Ref. [287]. We give the full details of this calculation in the Appendix. Different choices of observational data from the literature produce variations in the normalization of the spectrum, resulting in a factor of $\mathcal{O}(10)$ uncertainty in the prediction of $A_{\rm GW}$. However, again, these uncertainties pertain to the normalization of the SGWB, and not its spectral shape.

5.3.1 Spectral features from new forces

SGWB spectra with nonzero values of α and γ are shown in the left and right panels of Fig. 5.1, respectively. In both cases, the spectral index of the SGWB is modified from the gravity-only prediction of $h_c \propto f^{-2/3}$ shown in black, and non-power-

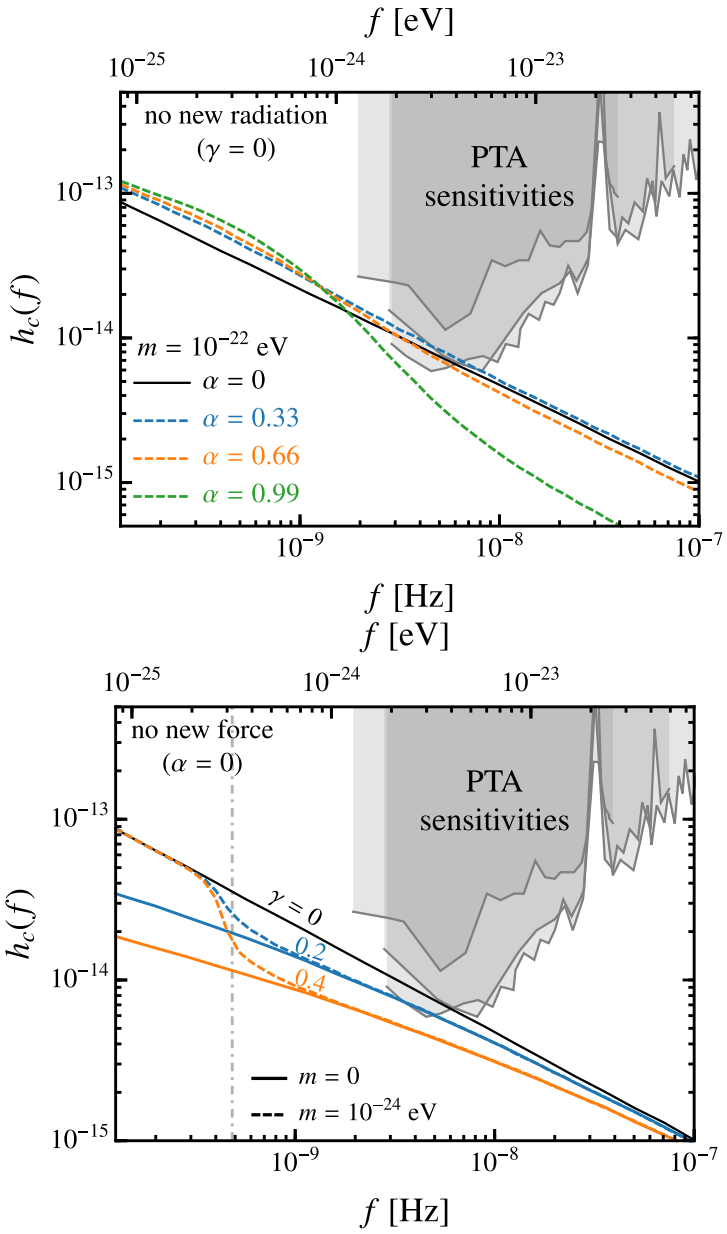

Figure 5.1: Predicted SGWB produced by a population of uniformly charged SMBH binaries. The gray regions show current PTA sensitivities [255–257]. In each panel, the black line shows the gravity-only prediction. **Left:** SGWB produced with nonzero potential-strength parameter α, defined in Eq. (5.11), for a mediator mass of $m = 10^{-22}\,\mathrm{eV}$. The location of the feature in the spectrum corresponds to black hole radial separations of order m^{-1}, and thus the location of the feature is offset from the mediator mass, as given by Eq. (5.15). **Right:** SGWB produced with nonzero dipole-strength parameter γ, defined in Eq. (5.13), for a vector mediator with mass of $10^{-24}\,\mathrm{eV}$. The dot-dashed vertical line at $f_{\mathrm{th}} = m/\pi$ indicates the threshold for dipole radiation in the source frame.

law behavior may be directly observable in certain regimes. Note that we assume that all binaries have the same values of α and γ, but nontrivial distributions can be studied by taking $\mathbf{X} = \{M_1, M_2, Q_1, Q_2\}$ in Eq. (5.2). Below we discuss effects on the SGWB spectrum from nonzero α and nonzero γ separately.

In the case with $\alpha \neq 0$ and $\gamma = 0$ (left panel of Fig. 5.1), each SMBH carries the same nonzero dark charge. Thus, there is a new force between the two objects in addition to gravity. This modifies the usual form of Kepler's law relating the binary separation to its orbital frequency. For a massless mediator $m = 0$ and $\gamma = 0$, the effect of the new force (Eq. (5.10)) leads to a rescaling of Newton's gravitational force law by $G \to G(1 - \alpha)$, and therefore preserves the shape of the SGWB spectrum while modifying only its normalization $A_{\rm GW}$. In this situation, it would be difficult to differentiate between new physics effects and astrophysical uncertainties in $A_{\rm GW}$. However, for a mediator with nonzero mass m, a distinctive feature emerges, as shown in Fig. 5.1. Since the nongravitational force is ineffective at separations $r > m^{-1}$, the spectrum departs from a power law at a frequency corresponding to this separation, given by

$$\begin{aligned} f_* &= \sqrt{G(M_1 + M_2)m^3}/\pi + \mathcal{O}(\alpha) \\ &\simeq 10^{-24}\,\mathrm{eV} \left(\frac{M_1 + M_2}{10^9\,\mathrm{M}_\odot}\right)^{1/2} \left(\frac{m}{10^{-22}\,\mathrm{eV}}\right)^{3/2} . \end{aligned} \tag{5.15}$$

In the case with $\alpha = 0$ and $\gamma \neq 0$ (right panel of Fig. 5.1), only one of the two SMBHs is charged, so there is no modification to the force law which relates binary separation and orbital frequency. However, since the dark charge distribution now has a sizable dipole moment, the binary can lose energy to dipole radiation of the light media-

tor (here we assume a vector mediator such that we employ the lower case of Eq. (5.12)). The losses to dipole radiation can easily exceed those to gravitational radiation, which is sourced by the quadrupole moment. This leads to significant modification of the spectral index.

As with the $\alpha \neq 0$ case, the $\gamma \neq 0$ case exhibits additional features when the mediator mass m is nonzero. In this case, the spectrum reveals a threshold $\omega = m$ above which dipole radiation becomes significant. The binary rapidly loses energy above this threshold, producing a steplike feature in the spectrum around $f \sim m/\pi$. For a single merger, this feature is sharp, arising from the square root in Eq. (5.12). It is slightly smoothed out in Fig. 5.1 by integration over the SMBH binary population across different redshifts.

If $\gamma \neq 0$ and the mediator is massless, then the spectrum sourced by a single binary can be parametrized as

$$\frac{\mathrm{d}E_{\mathrm{GW}}}{\mathrm{d}f_{\mathrm{s}}} \propto f^{-1/3} \frac{1}{1 + (f/\kappa)^{-2/3}}\,, \tag{5.16}$$

where κ is given by

$$\kappa \simeq 2 \times 10^{-8}\,\mathrm{Hz} \left(\frac{\gamma}{0.2}\right)^{3} \left(\frac{M_1 + M_2}{10^9\,\mathrm{M}_\odot}\right)^{-1}\,. \tag{5.17}$$

Thus, for a given binary, the spectrum transitions between spectral indexes $\beta = -2/3$ and $\beta = -1/3$ at a frequency of order κ. This transitional behavior is also visible in the integrated SGWB shown in Fig. 5.1, in which the curves with $\gamma > 0$ exhibit a shallower power-law index at frequencies $f \ll 10^8\,\mathrm{Hz}$.

In general, an SMBH binary can have both a nonzero α and a nonzero γ such that the spectrum will be a combination of the two limiting cases presented here.

5.3.2 Distinguishing new forces from astrophysics

Predictions of the amplitude and shape of the SGWB spectrum from SMBH mergers are sensitive to astrophysical uncertainties associated with galactic mergers. Even in the gravity-only calculations, the details of the cosmological population of such mergers determines the normalization of the spectrum. We seek to probe spectral features which vary between different binaries at different redshifts. Thus, it is essential to assess the full SGWB spectrum obtained by convolving the single-merger spectrum of Eq. (5.6) with the statistics of galactic mergers.

The prediction of Eq. (5.2) depends on the validity of the single-merger spectrum. In the standard scenario, the single-merger spectrum of Eq. (5.6) is valid only for circular binaries that are driven to merge purely by emission of gravitational radiation. In the absence of new physics, these assumptions are robust in the late stages of inspiral, for which the SGWB is most prominent, and it is this very simplicity that makes the SGWB spectrum such a powerful probe of new forces. However, even in the absence of new forces, there are other astrophysical processes that modify parts of the SGWB spectrum. These modifications represent possible systematics, or, if they are well understood, they may provide a new set of features for extensions of our analysis. We thus briefly outline these processes and the implications for the SGWB spectrum. Most importantly, these processes mainly influence the spectrum of GWs outside the window probed by pulsar timing arrays.

Typically, the dynamics of a binary are governed by a single energy loss mechanism at any given time. At very large separations, corresponding to low frequencies

in the SGWB, gravitational radiation is inefficient. The inspiral is instead driven by dynamical friction from stars and Eq. (5.9) is invalid: the binary shrinks much faster than would be expected from gravitational radiation alone, and thus the SGWB is suppressed at these frequencies. At small separations, stellar dynamical friction becomes inefficient as the binary depletes the region of stellar phase space that can remove energy from the binary (the "loss cone" [288]). At this point, in the absence of any other energy loss mechanism, gravitational radiation dominates the evolution of the binary for the remainder of the inspiral. The transition from stellar dynamical friction to gravitational radiation domination takes place at a characteristic separation corresponding to a frequency of order [283, 289]

$$f_{\rm GR} \simeq 10^{-9}\,{\rm Hz}\left(\frac{M_1 M_2}{(5\times10^8\,{\rm M}_\odot)^2}\right)^{-3/8}\left(\frac{M_1+M_2}{10^9\,{\rm M}_\odot}\right)^{1/8}, \tag{5.18}$$

assuming that gas and stars shrink the binary on a timescale $|r/\dot{r}| \sim 10^8\,{\rm yr}$. The merger itself imposes an upper cutoff on the frequency corresponding to the separation of the binary at the innermost stable circular orbit (ISCO), given by

$$f_{\rm ISCO} = \frac{1}{2\pi}\frac{1}{6GM_1} \simeq 11\,\mu{\rm Hz}\left(\frac{M_1}{5\times10^8\,{\rm M}_\odot}\right)^{-1}, \tag{5.19}$$

where it is assumed that $M_1 = M_2$.

However, the spectrum of Eq. (5.9) is not typically valid at frequencies immediately above $f_{\rm GR}$ for two reasons: stalled mergers and eccentric orbits. At distances of order a parsec, energy loss from gravitational radiation is not efficient enough to merge a binary within the lifetime of the Universe. As a result, the evolution of merging binaries from the end of star-driven dynamical friction until the era where gravitational radiation becomes an efficient energy loss mechanism (separations below $\sim$0.01 pc) is not known.

This is known as the "final parsec problem" [288]. Candidate mechanisms include gas dynamics [276, 277, 290–299] and asymmetry of galactic mergers [300]. In particular, efficient gas infall can dominate the evolution of the binary up to a frequency of several times $f_{\rm GR}$ [283]. In the absence of any other mechanisms, binaries may even stall until a subsequent galactic merger supplies a third SMBH, at which point few-body dynamics can shrink the binary [261, 273]. Such processes have a significant effect on the normalization of the SGWB, and gas dynamical processes may have a slight impact on the spectral shape as well at the lowest frequencies in the pulsar timing window [301, 302]. (See also Refs. [291–293, 295, 303–305] for further discussion of the role of gas dynamics in SMBH mergers.)

A more significant modification potentially arises from eccentricity of the orbits. The spectrum of Eq. (5.9) holds only for a circular binary. However, during the stage of inspiral driven by stellar dynamical friction, stellar encounters tend to enhance the eccentricity of the binary, so typical binaries in simulations have substantial eccentricities at $f_{\rm GR}$ [289]. These eccentricities are quickly reduced by gravitational radiation through a process termed "circularization" [274, 275]. Nevertheless, there is still a range of separations [and frequencies as given by Eq. (5.7)] in which binaries are driven by gravitational radiation and yet deviating from Eq. (5.9). This may change the spectral index of the SGWB in a range of frequencies above $f_{\rm GR}$. This spectral feature may extend to frequencies of order $10^{-9}\,$Hz, or even as high as $10^{-8}\,$Hz in some projections. In principle, this may mimic the effects of new physics. Claiming a discovery of a new force may require restricting analysis to GW frequencies above $10^{-8}\,$Hz.

Finally, we note that realistic predictions of the SGWB are influenced by Pois-

son noise at frequencies above $10^{-7}\,\mathrm{Hz}$, as the SGWB is expected to be dominated by relatively few sources in this regime [301]. We neglect this effect in our analysis, as a modification to the spectral shape will still produce a significant modification to the spectrum of a small number of sources. However, a full statistical treatment in this regime should be performed using a Monte Carlo simulation rather than by direct measurement of the spectral index.

5.4 Discussion

We have argued that the spectral index of the SGWB can be robustly predicted in the absence of new physics and have explored how a new force can modify the spectral index. We now discuss the implications of our results in light of the recent observation of a stochastic process among the pulsars in the NANOGrav 12.5-year dataset [258] as well as other GW detection experiments.

The NANOGrav Collaboration fits the spectrum to two types of power laws: one present in only the five lowest frequencies ($2 \times 10^{-9}\,\mathrm{Hz} \lesssim f \lesssim 1 \times 10^{-8}\,\mathrm{Hz}$) and one present among the thirty lowest frequencies ($2 \times 10^{-9}\,\mathrm{Hz} \lesssim f \lesssim 7 \times 10^{-8}\,\mathrm{Hz}$). [2] While SMBH mergers are expected to produce GWs across the entire pulsar timing frequency band, pulsar-intrinsic noise may contribute at high frequencies and mask the GW signal [258], and, as such, we focus on the five frequency analysis. The measured amplitude and spectral index for a power-law signal are depicted in Fig. 5.2. The solid black line shows the predicted index for uncharged supermassive black holes. The

[2] The analysis is also carried out with a broken power law whose results are qualitatively similar to those of using just the five lowest frequencies.

spectrum of charged black holes is not generally a power law and can have various shapes depending on the mediator mass m, the potential-strength parameter α, and the dipole-strength parameter γ. In the limit where the mediator mass vanishes and the SMBHs carry a nonzero γ, the spectrum is approximately a power law across the pulsar-timing window and we show the index evaluated at a frequency of $(5\,\mathrm{yr})^{-1}$ in Fig. 5.2. We conclude that the additional dipole radiation will soften the spectrum, and the current dataset can potentially constrain $\gamma \sim 1$.

While the uncertainties on the values of the amplitude and spectral index are still significant, pulsar timing arrays are rapidly improving in sensitivity. For identical pulsars, the signal-to-background ratio of a pulsar timing array analysis scales $\propto A_{\mathrm{GW}}^2 T^{13/3} N_p$, where T is the observation time and N_p is the number of pulsars [306]. The large scaling with observation time suggest that NANOGrav will be able to significantly improve the estimate of the spectral index and amplitude as it continues observing the current pulsar set. Furthermore, combining the 12.5-year NANOGrav data with the EPTA and PPTA datasets may be enough to detect the Hellings and Downs correlation function between pulsars, which, if observed, would confirm the first detection of a stochastic GW background. Once a discovery is made, the measurement of the spectral index will be critical to measure the charges of the SMBHs and search for additional forces.

Pulsar timing arrays are particularly well suited to measure stochastic GW spectra at frequencies of order nHz–100 µHz. SMBHs that are emitting GWs in this frequency band are near the start of their merger. As they progress toward the inspiral phase, the emission continues, building a falling characteristic strain spectrum until the

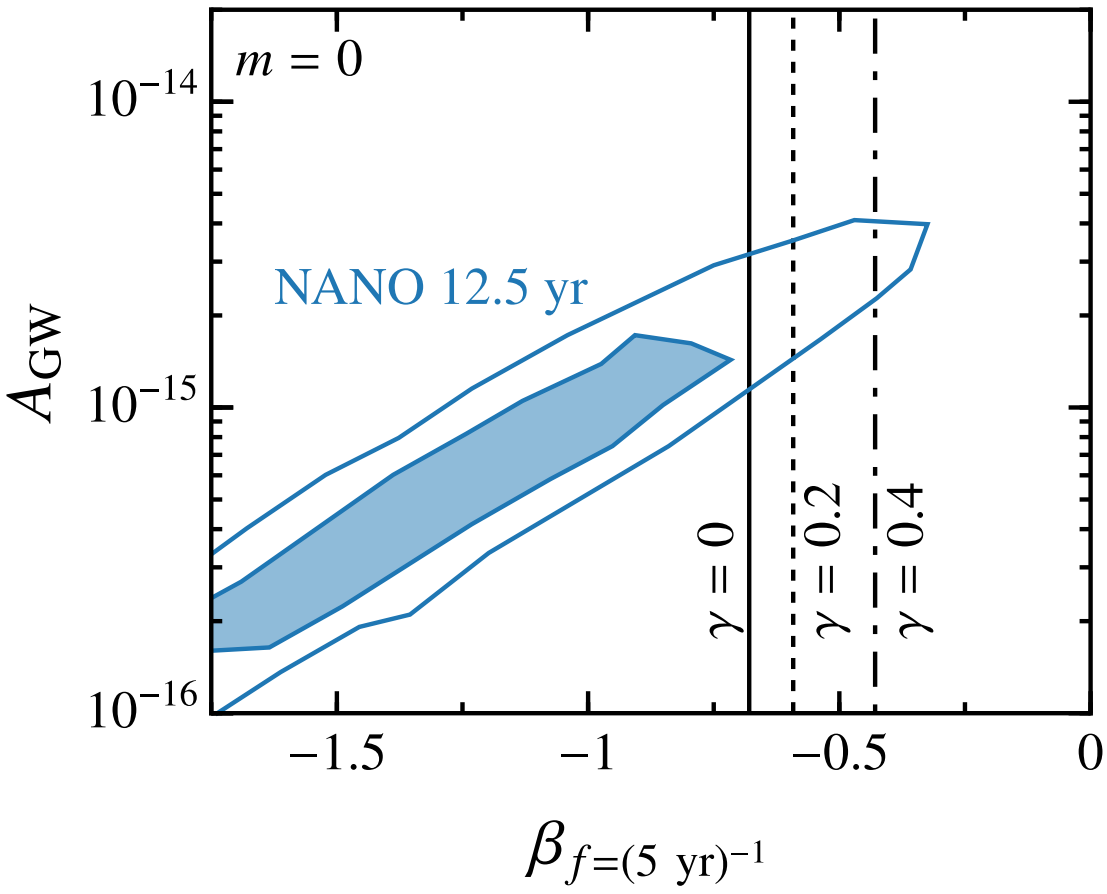

Figure 5.2: A comparison of the spectral index as measured in the NANOGrav 12.5-year data set to the value predicted by merging supermassive charged black hole binaries. The shaded and bounded regions correspond to the 1σ and 2σ posteriors derived by the NANOGrav Collaboration [258]. The black lines correspond to charged binaries under a new long-range vector force with different values of the dipole-strength parameter γ, assuming the potential-strength parameter is negligible ($\alpha = 0$). Since this spectrum is not strictly a power law, we evaluate the spectrum at roughly the peak sensitivity of NANOGrav, $f = (5\,\mathrm{yr})^{-1}$.

ISCO frequency of the heaviest black holes, $\sim 10^{-5}$ Hz [see Eq. (5.19)]. The measurement of the spectrum might be extendable using space-based interferometers such as the Laser Interferometer Space Antenna (LISA) [307, 308] or astrometry [309]. Confirmation of a consistent spectral index and amplitude across this wide range of frequencies would be a remarkable confirmation of gravity-only mergers. On the other hand, if a new force is present with a mediator mass above the pulsar timing range and below that of higher frequency detectors, it would show up as an observable break in the spectrum. This displays the critical complementarity between the different GW searches.

Motivated by the imminent discovery prospects of a stochastic background of GWs in pulsar timing arrays, we have focused our discussion on the detection of new forces in SMBH binaries. However, other GW experiments may also detect stochastic binary merger backgrounds. In particular, LISA is expected to see a stochastic background of white dwarf, neutron star, and lighter black hole binary mergers [310, 311]. While these backgrounds are highly anisotropic, it is also possible to look for new forces in these new environments by incorporating directionality in Eq. (5.2). The white dwarf background (as well as other stochastic merger backgrounds observed in the future) will provide complementary searches for dark forces in different astrophysical environments.

Finally, we note that since the GW spectrum from SMBH binaries is yet to be discovered, it is possible that SMBHs have charges so large that new force is strong relative to gravity. In this case, we may uncover additional signals in the SGWB. First, for sufficiently large dark charges, a repulsive force will stall the merger on cosmological timescales. This could reduce the SGWB amplitude below lower bounds estimated for gravity-only mergers [261]. Second, while gravitational radiation tends to rapidly

circularize binaries, dipole radiation can have the opposite effect as the binary passes through the mediator mass threshold and can have a dramatic effect on the spectrum. Such phenomena do require a mechanism for the accumulation of large charges near or on the SMBHs. We leave the study of such mechanisms and their consequences for future work.

Chapter 6

Cosmological implications of the KOTO excess

6.1 Introduction

In the previous chapter, we showed how a cosmic background of gravitational waves may provide new probes of ultralight species. We now turn to a somewhat different use of cosmological observables: in particular, we show that with existing observables, rare kaon decays in terrestrial experiments have significant implications for cosmic history.

The rare kaon decays $K^+ \to \pi^+ \nu\bar{\nu}$ and $K_L \to \pi^0 \nu\bar{\nu}$ are widely recognized as very sensitive probes of new physics (NP). In the Standard Model (SM), the branching ratios of these decays are strongly suppressed, and can be precisely predicted [312, 313]

to be

$$\mathrm{BR}(K^+ \to \pi^+ \nu\bar{\nu})_{\mathrm{SM}} = (8.4 \pm 1.0) \times 10^{-11} \ , \tag{6.1}$$

$$\mathrm{BR}(K_L \to \pi^0 \nu\bar{\nu})_{\mathrm{SM}} = (3.4 \pm 0.6) \times 10^{-11} \ . \tag{6.2}$$

On the experimental side, several $K^+ \to \pi^+ \nu\bar{\nu}$ candidate events have been observed by the E787/E949 experiment [314–316] and the NA62 experiment [317], but a discovery of $K^+ \to \pi^+ \nu\bar{\nu}$ has still to be established. The current best limit on the branching ratio is from a preliminary analysis of NA62 data and reads [318]

$$\mathrm{BR}(K^+ \to \pi^+ \nu\bar{\nu})_{\mathrm{exp}} < 2.44 \times 10^{-10} \quad (95\% \text{ C.L.}), \tag{6.3}$$

not far above the SM prediction. The NA62 experiment aims to measure the SM branching ratio with $\mathcal{O}(10\%)$ uncertainty. In the case of $K_L \to \pi^0 \nu\bar{\nu}$, the current most stringent bound on the branching ratio comes from the KOTO experiment [319], and is still two orders of magnitude above the SM prediction:

$$\mathrm{BR}(K_L \to \pi^0 \nu\bar{\nu})_{\mathrm{exp}} < 3.0 \times 10^{-9} \quad (90\% \text{ C.L.}). \tag{6.4}$$

Interestingly, in the latest status update by KOTO [320], 4 events are seen in the signal box, with an expected number of 0.05 ± 0.01 SM $K_L \to \pi^0 \nu\bar{\nu}$ events and 0.05 ± 0.02 background events. One of the events has been identified as likely background. If the remaining events are interpreted as signal, one finds a branching ratio of $\mathrm{BR}(K_L \to \pi^0 \nu\bar{\nu}) \sim 2 \times 10^{-9}$ [321]. A branching ratio of this size would be a spectacular discovery. Not only does it imply NP, it also violates the Grossman-Nir (GN) bound [322], $\mathrm{BR}(K_L \to \pi^0 \nu\bar{\nu}) \lesssim 4.3 \times \mathrm{BR}(K^+ \to \pi^+ \nu\bar{\nu}) \lesssim 10^{-9}$, when combined with the NA62 constraint in Eq. (6.3). The GN bound is very robust in models where the

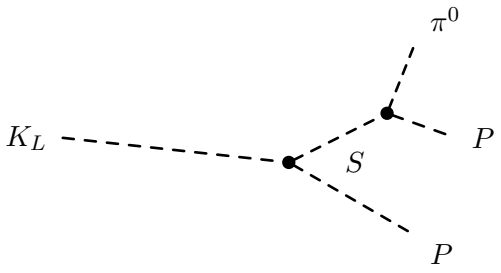

Figure 6.1: Decay chain accounting for the KOTO signal in our scenario.

$K \to \pi\nu\bar{\nu}$ decays are modified by heavy new physics well above the kaon mass. However, in the presence of light new physics, the GN bound can be violated and the observed events at KOTO may find an explanation [321, 323–333].

Here we focus on a new physics scenario first discussed in [334]. Two new light scalars S and P, neutral under the SM gauge interactions, are introduced such that K_L can decay into a pair of the new particles, $K_L \to SP$. If the decay $S \to \pi^0 P$ is allowed and P is stable on the relevant experimental scales, then the decay chain $K_L \to SP \to \pi^0 PP$ can mimic the $K_L \to \pi^0\nu\bar{\nu}$ signature (see Fig. 6.1). The corresponding chain of two-body decays does not exist for the charged kaon. A possible decay $K^+ \to \pi^+ SP$ is suppressed by three-body phase space or may be forbidden entirely by kinematics.

If P is absolutely stable, it is also a candidate for cosmological dark matter. In the minimal setup that can provide a NP explanation of the KOTO events, P couples to the SM very weakly, implying that annihilation cross sections into SM states are too small for production by freeze-out. We therefore investigate alternative scenarios for cosmological production, and interpret overproduction of P as a cosmological constraint on the structure of the low-energy theory. We show that P is readily produced non-thermally if the scale of reheating is low, close to but safely above the current obser-

vational bound. We also show that this class of models can account for the KOTO excess without requiring a low reheating temperature, but only in the presence of additional interactions. We investigate prospects for testing this model with future experiments and with additional data from KOTO, and show that much of the parameter space will be probed in the near future.

This chapter is organized as follows: in Section 6.2, we present the model and discuss how it can explain the KOTO events. In Section 6.3, we evaluate astrophysical and terrestrial constraints on the parameter space of our model. In Section 6.4, we consider cosmological production of P, and relate the production of P to the scale of reheating. We discuss the implications of our results in Section 6.5 and conclude in Section 6.6.

6.2 Model

We start with very simple kinematical considerations concerning the masses of the two scalars S and P. Figure 6.2 shows the plane of the two scalar masses m_S and m_P. As described in the introduction, we are interested in regions of parameter space where the decay $K_L \to \pi^0 PP$, which mimics $K_L \to \pi^0 \nu\bar{\nu}$, can be realized as a sequence of the two-body decay $K_L \to SP$ followed by $S \to \pi^0 P$. For m_S too large, the decay $K_L \to SP$ is kinematically forbidden, while for m_S too small, the $S \to \pi^0 P$ decay is not open, excluding the dark gray regions in the plot. In the light gray region one faces potential constraints from the charged kaon decay $K^+ \to \pi^+ SP$ that is generically expected in the models discussed below. In the white region, however, this decay is

kinematically forbidden, while $K_L \to \pi^0 \nu\bar{\nu}$ remains open.

The plot also indicates two other interesting kinematical boundaries. If $m_P < m_{\pi^0}/2$, the exotic pion decay $\pi^0 \to PP$ is possible which, as we will discuss in Section 6.4, can impact cosmological production considerably. If $m_S > 3m_P$, the decay $S \to 3P$ can be allowed, thus modifying the lifetime of S, which is a crucial parameter for beam dump constraints. Note that low P masses may be subject to constraints from supernova cooling, which we will discuss further in Section 6.3.1. A weaker lower bound on the P mass also follows from assuming a particular thermal history, a point to which we shall return in Section 6.5.

In the following sections, we will discuss four benchmark parameter points covering the most interesting regimes:

$$\begin{aligned} &\text{BM1:} \quad m_S = 400\,\text{MeV}, \quad m_P = 10\,\text{MeV}, \\ &\text{BM2:} \quad m_S = 350\,\text{MeV}, \quad m_P = 100\,\text{MeV}, \\ &\text{BM3:} \quad m_S = 300\,\text{MeV}, \quad m_P = 125\,\text{MeV}, \\ &\text{BM4:} \quad m_S = 200\,\text{MeV}, \quad m_P = 10\,\text{MeV}. \end{aligned} \tag{6.5}$$

Next we discuss in detail the interactions of S and P with SM quarks. We first focus on non-renormalizable effective couplings and identify viable regions of parameter space. Then we comment on simplified UV models that map onto the effective couplings.

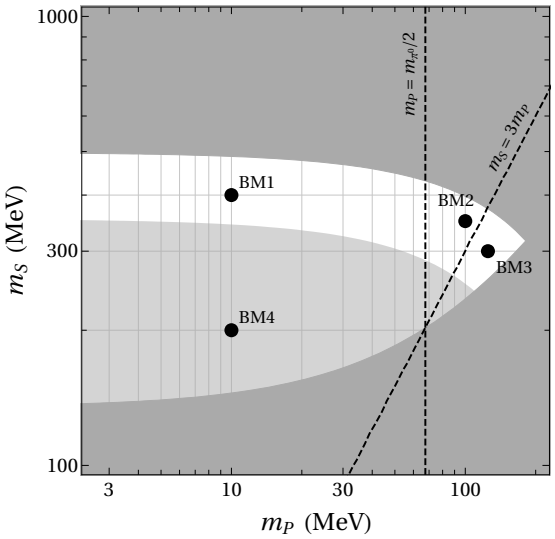

Figure 6.2: The plane of the scalar masses m_S vs. m_P. In the dark gray region the $K_L \to \pi^0 PP$ decay cannot be realized as a sequence of 2-body decays. In the light gray region the $K^+ \to \pi^+ SP$ decay is open. The black dots indicate four benchmark scenarios that we consider later (Eq. (6.5)).

6.2.1 Effective interactions of the scalars and meson decay rates

We assume that the scalars S and P interact with SM particles via the effective couplings

$$\mathcal{L}_{\rm int} \supset iSP\left(\frac{g_{dd}^{SP}}{\Lambda_{\rm NP}}(\bar{d}d) + \frac{\tilde{g}_{dd}^{SP}}{\Lambda_{\rm NP}}(\bar{d}i\gamma_5 d) + \frac{g_{ss}^{SP}}{\Lambda_{\rm NP}}(\bar{s}s) + \frac{\tilde{g}_{ss}^{SP}}{\Lambda_{\rm NP}}(\bar{s}i\gamma_5 s)\right) + iSP\left(\frac{g_{sd}^{SP}}{\Lambda_{\rm NP}}(\bar{s}d) + \frac{\tilde{g}_{sd}^{SP}}{\Lambda_{\rm NP}}(\bar{s}i\gamma_5 d) + \text{h.c.}\right). \quad (6.6)$$

The factors of i in the above Lagrangian are reminiscent of considering S to be a CP-even scalar and P to be a CP-odd pseudoscalar, a notational pattern that we will retain when matching onto low-energy QCD later on. The coefficients g_{dd}^{SP}, g_{ss}^{SP}, $\tilde{g}_{dd}^{SP}$, and $\tilde{g}_{ss}^{SP}$ are purely imaginary (by hermiticity of the Lagrangian) while the g_{sd}^{SP} and $\tilde{g}_{sd}^{SP}$ coefficients can have an arbitrary complex phase. There could also be interactions involving b quarks, but as long as they are not considerably larger than the interactions

with the light quarks, their impact on phenomenology will be negligible.

In the following, we will also entertain the possibility of additional interactions involving P^2 and S^2, of the form

$$\mathcal{L}_{\rm int} \supset P^2 \left(\frac{g_{dd}^{P^2}}{\Lambda_{\rm NP}}(\bar{d}d) + \frac{\tilde{g}_{dd}^{P^2}}{\Lambda_{\rm NP}}(\bar{d}i\gamma_5 d) + \frac{g_{ss}^{P^2}}{\Lambda_{\rm NP}}(\bar{s}s) + \frac{\tilde{g}_{ss}^{P^2}}{\Lambda_{\rm NP}}(\bar{s}i\gamma_5 s) \right) + P^2 \left(\frac{g_{sd}^{P^2}}{\Lambda_{\rm NP}}(\bar{s}d) + \frac{\tilde{g}_{sd}^{P^2}}{\Lambda_{\rm NP}}(\bar{s}i\gamma_5 d) + \text{h.c.} \right). \quad (6.7)$$

While the interactions in Eq. (6.7) are not directly relevant for the KOTO signal, they do have important implications for other meson decays and in particular for the dark matter phenomenology as we will discuss in Section 6.4 below.

The decays relevant for an enhanced KOTO signal, $K_L \to SP$ and $S \to \pi^0 P$ are induced by the couplings $\text{Re}(\tilde{g}_{sd}^{SP})$ and $\text{Im}(\tilde{g}_{dd}^{SP})$, respectively. For the corresponding decay rates we find

$$\Gamma(K_L \to SP) = \frac{1}{8\pi} \frac{f_K^2 m_{K_L}^3}{m_s^2} \left(\frac{\text{Re}(\tilde{g}_{sd}^{SP})}{\Lambda_{\rm NP}} \right)^2 \eta_{\rm QCD} \sqrt{\lambda\left(1, m_S^2/m_{K_L}^2, m_P^2/m_{K_L}^2\right)}, \quad (6.8)$$

$$\Gamma(S \to \pi^0 P) = \frac{1}{128\pi} \frac{f_\pi^2 m_{\pi^0}^4}{m_S m_d^2} \left(\frac{\text{Im}(\tilde{g}_{dd}^{SP})}{\Lambda_{\rm NP}} \right)^2 \eta_{\rm QCD} \sqrt{\lambda\left(1, m_{\pi^0}^2/m_S^2, m_P^2/m_S^2\right)}, \quad (6.9)$$

with the phase space function $\lambda(a,b,c) = a^2 + b^2 + c^2 - 2(ab + ac + bc)$. The down and strange quark masses in the above expressions should be interpreted as the $\overline{\rm MS}$ masses at a renormalization scale of $\mu = 2$ GeV. Leading-log QCD corrections are then taken into account through the factor $\eta_{\rm QCD}$

$$\eta_{\rm QCD} = \left(\frac{\alpha_s(m_t)}{\alpha_s(M)} \right)^{8/7} \left(\frac{\alpha_s(m_b)}{\alpha_s(m_t)} \right)^{24/23} \left(\frac{\alpha_s(2\text{ GeV})}{\alpha_s(m_b)} \right)^{24/25}, \quad (6.10)$$

where M is the scale of new physics that is responsible for the effective interactions of S and P with the SM quarks. Because of $SU(2)_L$ invariance we expect $M \sim \sqrt{\Lambda_{\rm NP} v}$,

where $v = 246$ GeV is the vacuum expectation value of the SM Higgs. Note that including the $\eta_{\rm QCD}$ factor is equivalent to evaluating the down and strange masses in Eqs. (6.8) and (6.9) at the scale M.

The coupling $|g_{sd}^{SP}|$ can lead to the decay $K^+ \to \pi^+ SP$, if kinematically allowed. The differential 3-body decay rate of $K^+ \to \pi^+ SP$ is given by

$$\frac{d\Gamma(K^+ \to \pi^+ SP)}{dq^2} = \frac{1}{256\pi^3} \frac{m_{K^+}^3}{m_s^2} \left(\frac{|g_{sd}^{SP}|}{\Lambda_{\rm NP}} \right)^2 \eta_{\rm QCD} \left(1 - \frac{m_{\pi^+}^2}{m_{K^+}^2} \right)^2 \times \sqrt{\lambda\left(1, m_S^2/q^2, m_P^2/q^2\right)} \sqrt{\lambda\left(1, m_{\pi^+}^2/m_{K^+}^2, q^2/m_{K^+}^2\right)}, \quad (6.11)$$

where we estimated the relevant scalar form factor as $\langle \pi^+ | \bar{s} d | K^+ \rangle \simeq (m_{K^+}^2 - m_{\pi^+}^2)/m_s$ and q^2 is the invariant mass of the SP system, with $(m_P + m_S)^2 < q^2 < (m_{K^+} - m_{\pi^+})^2$.

Similarly to the $K_L \to SP$ decay, the interactions in Eq. (6.6) also lead to the exotic eta decay $\eta \to SP$, which has been identified as a possible source of the scalar S at beam dump experiments [332]. Neglecting η–η' mixing, we find

$$\Gamma(\eta \to SP) = \frac{3}{512\pi} \frac{f_\eta^2 m_\eta^3}{m_s^2} \left(\frac{2\,{\rm Im}(\tilde{g}_{ss}^{SP}) - {\rm Im}(\tilde{g}_{dd}^{SP})}{\Lambda_{\rm NP}} \right)^2 \eta_{\rm QCD} \sqrt{\lambda\left(1, m_S^2/m_\eta^2, m_P^2/m_\eta^2\right)}. \quad (6.12)$$

For completeness, we also provide the expression for the decay $K_S \to SP$:

$$\Gamma(K_S \to SP) = \frac{1}{32\pi} \frac{f_K^2 m_{K_S}^3}{m_s^2} \left(\frac{{\rm Im}(\tilde{g}_{sd}^{SP})}{\Lambda_{\rm NP}} \right)^2 \eta_{\rm QCD} \sqrt{\lambda\left(1, m_S^2/m_{K_S}^2, m_P^2/m_{K_S}^2\right)}. \quad (6.13)$$

In the presence of the P^2 interactions in Eq. (6.7), there are additional exotic meson decays, $\pi^0 \to PP$, $\eta \to PP$, $K_{L/S} \to PP$, and $K^+ \to \pi^+ PP$, with the following

decay rates:

$$\Gamma(\pi^0 \to PP) = \frac{1}{64\pi}\frac{f_\pi^2 m_{\pi^0}^3}{m_d^2}\left(\frac{\mathrm{Re}(\tilde{g}_{dd}^{P^2})}{\Lambda_{\mathrm{NP}}}\right)^2 \eta_{\mathrm{QCD}}\sqrt{1-\frac{4m_P^2}{m_{\pi^0}^2}}\ , \tag{6.14}$$

$$\Gamma(\eta \to PP) = \frac{3}{256\pi}\frac{f_\eta^2 m_\eta^3}{m_s^2}\left(\frac{2\,\mathrm{Re}(\tilde{g}_{ss}^{P^2}) - \mathrm{Re}(\tilde{g}_{dd}^{P^2})}{\Lambda_{\mathrm{NP}}}\right)^2 \eta_{\mathrm{QCD}}\sqrt{1-\frac{4m_P^2}{m_\eta^2}}\ , \tag{6.15}$$

$$\Gamma(K_L \to PP) = \frac{1}{4\pi}\frac{f_K^2 m_{K_L}^3}{m_s^2}\left(\frac{\mathrm{Im}(\tilde{g}_{sd}^{P^2})}{\Lambda_{\mathrm{NP}}}\right)^2 \eta_{\mathrm{QCD}}\sqrt{1-\frac{4m_P^2}{m_{K_L}^2}}\ , \tag{6.16}$$

$$\Gamma(K_S \to PP) = \frac{1}{4\pi}\frac{f_K^2 m_{K_S}^3}{m_s^2}\left(\frac{\mathrm{Re}(\tilde{g}_{sd}^{P^2})}{\Lambda_{\mathrm{NP}}}\right)^2 \eta_{\mathrm{QCD}}\sqrt{1-\frac{4m_P^2}{m_{K_S}^2}}\ , \tag{6.17}$$

$$\begin{aligned}\frac{\mathrm{d}\Gamma(K^+ \to \pi^+ PP)}{\mathrm{d}q^2} = \frac{1}{128\pi^3}\frac{m_{K^+}^3}{m_s^2}\left(\frac{|g_{sd}^{P^2}|}{\Lambda_{\mathrm{NP}}}\right)^2 \eta_{\mathrm{QCD}}\left(1-\frac{m_{\pi^+}^2}{m_{K^+}^2}\right)^2 \\ \times\sqrt{1-\frac{4m_P^2}{q^2}}\sqrt{\lambda\left(1, m_{\pi^+}^2/m_{K^+}^2, q^2/m_{K^+}^2\right)}\ ,\end{aligned} \tag{6.18}$$

In the $K^+ \to \pi^+ PP$ decay width, q^2 denotes the PP invariant mass, which lies in the range $4m_P^2 < q^2 < (m_{K^+} - m_{\pi^+})^2$.

The interactions of S and P with quarks that we have introduced preserve a $\mathbb{Z}_2$ symmetry under which S and P are odd, while all SM particles are even. We assume that the $\mathbb{Z}_2$ symmetry is also respected by the scalar potential, such that P is an absolutely stable dark matter candidate. Among the allowed $\mathbb{Z}_2$ symmetric terms in the scalar potential, the SP^3 interaction

$$\mathcal{L}_{\mathrm{int}} \supset \lambda_{SP^3} S P^3\ , \tag{6.19}$$

will turn out to be relevant. When kinematically allowed, this interaction leads to the decay $S \to 3P$ with rate

$$\Gamma(S \to 3P) = \frac{3}{256\pi^3}\lambda_{SP^3}^2 m_S\ f(m_P/m_S), \tag{6.20}$$

where f is the three-body phase space integral,

$$f(y) = 2\int_{4y^2}^{(1-y)^2} \mathrm{d}x\ \sqrt{\lambda\left(1, x, y^2\right)\lambda\left(1, y^2/x, y^2/x\right)}, \tag{6.21}$$

which is normalized to 1 in the limit $y \to 0$. The $S \to 3P$ rate will modify the lifetime of S and can therefore have a crucial impact on possible constraints from beam dump experiments.

6.2.2 Events at the KOTO experiment

The model introduced in the previous section will lead to $K_L \to \pi^0 PP$ events at the KOTO experiment. We now identify the regions of parameter space in which this decay can mimic the KOTO signal.

The number of events that can be expected to be detected at KOTO can be written as

$$N = \frac{\mathrm{BR}(K_L \to SP) \times \mathrm{BR}(S \to \pi^0 P)}{\mathrm{BR}(K_L \to \pi^0 \nu\bar{\nu})_{\mathrm{SM}}} \times R \times N_{\mathrm{SM}}, \tag{6.22}$$

where $\mathrm{BR}(K_L \to \pi^0 \nu\bar{\nu})_{\mathrm{SM}} = (3.4 \pm 0.6) \times 10^{-11}$ is the SM prediction for the $K_L \to \pi^0 \nu\bar{\nu}$ branching ratio [312, 313], $N_{\mathrm{SM}} = 0.05 \pm 0.01$ is the expected number of SM signal events at KOTO [320], and

$$R = \frac{A(K_L \to SP \to \pi^0 PP)}{A(K_L \to \pi^0 \nu\bar{\nu})} \tag{6.23}$$

is the ratio of acceptances of the considered model signal and the SM signal at the KOTO detector. As has been pointed out before [321, 331, 332], an exotic contribution to the KOTO signal (in our case $K_L \to SP \to \pi^0 PP$) can have a considerably different acceptance. We determine the acceptance ratio R using a Monte Carlo simulation. Details are provided in Appendix C. The result is given in Fig. 6.3, which shows R

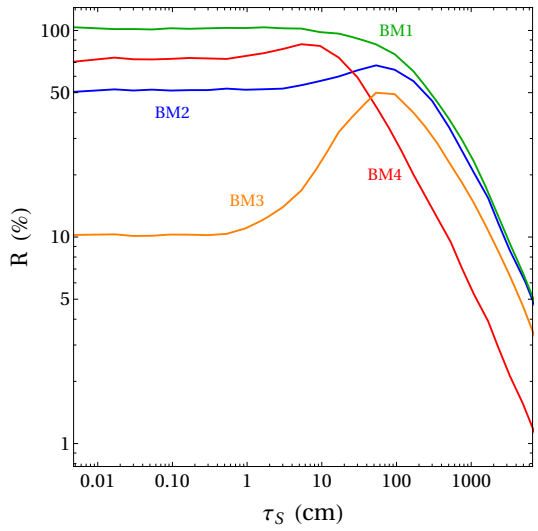

Figure 6.3: The acceptance ratio R of the $K_L \to SP \to \pi^0 PP$ signal over the SM $K_L \to \pi^0 \nu\bar{\nu}$ signal at KOTO as a function of the S lifetime τ_S for the four benchmark scenarios.

as a function of the S lifetime for our four benchmark points (Eq. (6.5)). For prompt decays, $\tau_S \to 0$, we find $\{R_{\rm BM1}, R_{\rm BM2}, R_{\rm BM3}, R_{\rm BM4}\} \simeq \{102\%, 51\%, 10\%, 73\%\}$. Once the lifetime of S becomes comparable to the size of the KOTO detector, $\tau_S \sim 1\,\text{m}$, R starts to decrease as more and more S leave the detector before decaying.

In our setup, the lifetime of S is determined by the $S \to \pi^0 P$ and $S \to 3P$ decays. In the four benchmark cases for the scalar masses defined above we find

$$\left\{\Gamma(S \to \pi^0 P)_{\rm BM1}, \Gamma(S \to \pi^0 P)_{\rm BM2}, \Gamma(S \to \pi^0 P)_{\rm BM3}, \Gamma(S \to \pi^0 P)_{\rm BM4}\right\} \simeq \left\{\frac{1}{3.3\,\text{cm}}, \frac{1}{3.4\,\text{cm}}, \frac{1}{4.4\,\text{cm}}, \frac{1}{2.7\,\text{cm}}\right\} \times \left(\frac{10^6\ \text{GeV}}{\Lambda_{dd}}\right)^2 \left(\frac{\alpha_s(10^4\ \text{GeV})}{\alpha_s(M)}\right)^{8/7}, \quad (6.24)$$

$$\left\{\Gamma(S \to 3P)_{\rm BM1}, \Gamma(S \to 3P)_{\rm BM2}, \Gamma(S \to 3P)_{\rm BM4}\right\} \simeq \left\{\frac{1}{2.0\,\text{cm}}, \frac{1}{49\,\text{cm}}, \frac{1}{4.3\,\text{cm}}\right\} \times \left(\frac{\lambda_{SP^3}}{10^{-5}}\right)^2, \quad (6.25)$$

where in the $S \to \pi^0 P$ decay width we have defined $\Lambda_{dd} = \Lambda_{\rm NP}/\,\text{Im}(\tilde{g}^{SP}_{dd})$. Note that

$S \to 3P$ is not kinematically allowed in benchmark BM3. The $S \to \pi^0 P$ branching ratio is given by $\mathrm{BR}(S \to \pi^0 P) = \Gamma(S \to \pi^0 P)/[\Gamma(S \to \pi^0 P) + \Gamma(S \to 3P)]$.

Finally, we find the following $K_L \to SP$ branching ratios

$$\Big\{ \mathrm{BR}(K_L \to SP)_{\mathrm{BM1}}, \mathrm{BR}(K_L \to SP)_{\mathrm{BM2}}, \mathrm{BR}(K_L \to SP)_{\mathrm{BM3}}, \mathrm{BR}(K_L \to SP)_{\mathrm{BM4}} \Big\} \simeq \Big\{ 1.7,\ 1.8,\ 2.3,\ 4.0 \Big\} \times 10^{-9} \times \left(\frac{10^{12}\ \mathrm{GeV}}{\Lambda_{sd}} \right)^2 \left(\frac{\alpha_s(10^4\,\mathrm{GeV})}{\alpha_s(M)} \right)^{8/7}, \quad (6.26)$$

where we have defined $\Lambda_{sd} = \Lambda_{\mathrm{NP}}/\,\mathrm{Re}(\tilde{g}^{SP}_{sd})$.

Figures 6.4 and 6.5 show the number of expected events in the Λ_{sd}–Λ_{dd} plane for our benchmark cases in the absence of the $S \to 3P$ decay (Fig. 6.4) and in the presence of the $S \to 3P$ decay induced by a coupling $\lambda_{SP^3} = 10^{-5}$ (Fig. 6.5). Along the solid green lines one expects 3 events, in the dark green regions one expects 2–4 events, and in the light green regions one expects 1–5 events. In the gray regions labeled "$K_L \to \pi^0$ inv.", the number of predicted events exceeds the limit from KOTO (see Eq. (6.4)). The right vertical axis shows the lifetime of S corresponding to Λ_{dd}. In Fig. 6.5, the lifetime is approximately constant for $\Lambda_{dd} > 10^7\,\mathrm{GeV}$, as in this region of parameter space, the lifetime is set by the $S \to 3P$ decay width.

For S lifetimes of $\tau_S \gtrsim 1\,\mathrm{m}$, existing beam dump constraints apply (see Section 6.3) as indicated in Fig. 6.4 by the dashed contours. A proposed upgrade of the SeaQuest experiment might probe S lifetimes as low as $\tau_S \gtrsim 5\,\mathrm{cm}$. In the scenarios shown in Fig. 6.5 with $\lambda_{SP^3} = 10^{-5}$, the S lifetimes are short enough throughout the parameter space that existing beam dump constraints are avoided.

In Fig. 6.5 we also show additional constraints from other meson decays. The known K_L branching fractions add up to a value compatible with 1 with very high

precision. Any additional K_L branching ratio, in particular $K_L \to SP$, is thus bounded above as $\mathrm{BR}(K_L \to SP) < 6.3 \times 10^{-4}$ [335]. In Fig. 6.5 the gray regions left of the dashed vertical lines denoted "$K_L \to$ inv." are excluded by this constraint. Note that this gives an absolute lower bound $\Lambda_{sd} \gtrsim$ few $\times\, 10^9\,\mathrm{GeV}$.

The other meson decay constraints shown in Fig. 6.5 are less robust as they depend on couplings that are in principle unrelated. If we assume that the coupling $g_{sd}^{P^2}$ (corresponding to $(\bar{s}d)P^2$) is of the same order as the coupling $\tilde{g}_{sd}^{SP}$ (corresponding to $(\bar{s}i\gamma_5 d)SP$), we find relevant constraints from the searches for $K^+ \to \pi^+\nu\bar{\nu}$. To evaluate the constraints we compare the predicted $K^+ \to \pi^+ PP$ branching ratio with the bound from NA62 given in Eq. (6.3). We correct for the different signal acceptances of $K^+ \to \pi^+ PP$ compared to $K^+ \to \pi^+\nu\bar{\nu}$ that arise due to kinematical cuts on the missing mass and the charged pion momentum. For the three P masses relevant to our benchmarks, we find the bounds $\mathrm{BR}(K^+ \to \pi^+ PP) < 2.7 \times 10^{-10}$ for $m_P = 10\,\mathrm{MeV}$, $\mathrm{BR}(K^+ \to \pi^+ PP) < 3.5 \times 10^{-10}$ for $m_P = 100\,\mathrm{MeV}$, and $\mathrm{BR}(K^+ \to \pi^+ PP) < 2.4 \times 10^{-9}$ for $m_P = 125\,\mathrm{MeV}$. Setting $\Lambda_{\mathrm{NP}}/|g_{sd}^{P^2}| = \Lambda_{\mathrm{NP}}/\,\mathrm{Re}(\tilde{g}_{sd}^{SP}) = \Lambda_{sd}$, we find that in Fig. 6.5, the regions left of the dotted vertical lines are excluded.

If we assume that the coupling $\tilde{g}_{dd}^{P^2}$ (corresponding to $(\bar{d}i\gamma_5 d)P^2$) is of the same order as the coupling $\tilde{g}_{dd}^{SP}$ (corresponding to $(\bar{d}i\gamma_5 d)SP$), we find relevant constraints from the invisible branching fraction of the neutral pion, $\mathrm{BR}(\pi^0 \to \mathrm{inv.}) < 4.4 \times 10^{-9}$ [318]. Setting $\Lambda_{\mathrm{NP}}/\,\mathrm{Re}(\tilde{g}_{dd}^{P^2}) = \Lambda_{\mathrm{NP}}/\,\mathrm{Im}(\tilde{g}_{dd}^{SP}) = \Lambda_{dd}$ in the benchmarks BM1 and BM4, the regions below the dotted horizontal lines are excluded. For benchmarks BM2 and BM3, the P mass is too large for the $\pi^0 \to PP$ decay, so the couplings are therefore completely unconstrained by $\mathrm{BR}(\pi^0 \to \mathrm{inv.})$.

6.2.3 Simplified UV models

The higher dimensional interactions in Eq. (6.6) that lead to the exotic meson decays can be UV completed by simplified models in various ways. In this section, we discuss briefly two possibilities: (1) vector-like quarks and (2) an inert Higgs doublet.

6.2.3.1 Vector-like quark model

We introduce two sets of heavy vector-like quarks D and Q which have quantum numbers of the right-handed down quark singlets, $D = (\mathbf{3}, \mathbf{1})_{-\frac{1}{3}}$, and of the left-handed quark doublets $Q = (\mathbf{3}, \mathbf{2})_{\frac{1}{6}}$, respectively. These quantum number assignments admit the following terms in the Lagrangian:

$$\begin{aligned}\mathcal{L} \supset\ & m_Q \bar{Q}_L Q_R + m_D \bar{D}_L D_R + Y_{QD}(\bar{Q}_L D_R)h + Y_{DQ}(\bar{D}_L Q_R)h^c \ +\ \text{h.c.} \\ & + X_{Dd}(\bar{D}_L d_R)S + X_{Ds}(\bar{D}_L s_R)S + Z_{Qd}(\bar{Q}_R d_L)iP + Z_{Ds}(\bar{Q}_R s_L)iP \ +\ \text{h.c.}\,. \quad (6.27)\end{aligned}$$

The first line contains the masses m_Q and m_D for the vector-like quarks, as well as interactions with the SM Higgs doublet h. The masses m_Q, m_D and the couplings Y_{QD}, Y_{DQ} are in general complex parameters. However, not all of their phases are observable. Using the freedom to re-phase the vector-like quark fields, we will choose real m_Q, m_D and Y_{QD} without loss of generality. The second line in Eq. (6.27) contains couplings of the SM down and strange quarks with S and the vector-like quark D as well as with P and the vectorlike quark Q. The couplings X_{Dd}, X_{Ds}, Z_{Qd}, and Z_{Qs} contain physical phases.

Note that the above Lagrangian is invariant under a $\mathbb{Z}_2$ symmetry under which all SM particles are even, while the vector-like quarks as well as S and P are odd. Thus

P remains an absolutely stable dark matter candidate. In addition to the couplings shown, the model could also contain $\mathbb{Z}_2$ invariant couplings involving S and Q or P and D. However, such couplings are not required to generate the desired low energy interactions and we will neglect them in the following.

Integrating out the vector-like quarks at tree level (see Fig. 6.6, left diagram), and matching onto the effective Lagrangian of Eq. (6.6), we find

$$\begin{aligned}
\frac{g^{SP}_{dd}}{\Lambda_{\rm NP}} &= \frac{-iY_{QD}v}{\sqrt{2}m_Q m_D}\,\mathrm{Im}(X_{Dd}Z^*_{Qd})\;, & \frac{\tilde{g}^{SP}_{dd}}{\Lambda_{\rm NP}} &= \frac{iY_{QD}v}{\sqrt{2}m_Q m_D}\,\mathrm{Re}(X_{Dd}Z^*_{Qd})\;, \\
\frac{g^{SP}_{ss}}{\Lambda_{\rm NP}} &= \frac{-iY_{QD}v}{\sqrt{2}m_Q m_D}\,\mathrm{Im}(X_{Ds}Z^*_{Qs})\;, & \frac{\tilde{g}^{SP}_{ss}}{\Lambda_{\rm NP}} &= \frac{iY_{QD}v}{\sqrt{2}m_Q m_D}\,\mathrm{Re}(X_{Ds}Z^*_{Qs})\;, \\
\frac{g^{SP}_{sd}}{\Lambda_{\rm NP}} &= \frac{Y_{QD}v}{\sqrt{2}m_Q m_D}\frac{1}{2}(Z_{Qs}X^*_{Dd}-X_{Ds}Z^*_{Qd})\;, & \frac{\tilde{g}^{SP}_{sd}}{\Lambda_{\rm NP}} &= \frac{Y_{QD}v}{\sqrt{2}m_Q m_D}\frac{i}{2}(Z_{Qs}X^*_{Dd}+X_{Ds}Z^*_{Qd})\;.
\end{aligned} \tag{6.28}$$

As required by $SU(2)_L$ invariance, the effective interactions $g^{SP}_{ij}/\Lambda_{\rm NP}$ and $\tilde{g}^{SP}_{ij}/\Lambda_{\rm NP}$ are proportional to the SM Higgs vev $v \simeq 246\,\mathrm{GeV}$. If all couplings X_{ij}, Y_{ij}, Z_{ij} are of $\mathcal{O}(1)$, we can expect vector-like quark masses $m_{Q,D} \sim \sqrt{\Lambda_{\rm NP} v} \sim 10^6\,\mathrm{GeV}$. The couplings above are not all independent but obey the relation

$$|\tilde{g}^{SP}_{sd}|^2 - |g^{SP}_{sd}|^2 + 2i\,\mathrm{Re}(g^{SP}_{sd}\tilde{g}^{SP*}_{sd}) = \tilde{g}^{SP}_{dd}\tilde{g}^{SP*}_{ss} - g^{SP}_{dd}g^{SP*}_{ss} + i(\tilde{g}^{SP}_{dd}g^{SP*}_{ss} + \tilde{g}^{SP*}_{ss}g^{SP}_{dd})\;. \tag{6.29}$$

One therefore expects that the flavor changing couplings are of the order of the geometric mean of the flavor conserving couplings.

The vector-like quarks also give 1-loop contributions to kaon mixing. We checked explicitly that those contributions scale as $v^2/(m_Q^2 m_D^2)$ and are completely negligible.

6.2.3.2 Inert Higgs doublet model

In a second scenario, we introduce an inert Higgs doublet H with mass m_H, which couples to down and strange quarks, the SM Higgs, and the scalars S and P through the following interactions:

$$\begin{aligned}\mathcal{L} \supset\ & m_H^2 H^\dagger H + \lambda_{SP}(H^\dagger h + h^\dagger H)SP \\ & + Y_{dd}(\bar{d}_L d_R)H + Y_{ds}(\bar{d}_L s_R)H + Y_{sd}(\bar{s}_L d_R)H \ +\ \text{h.c.}\ . \end{aligned} \tag{6.30}$$

As in the vector-like quark scenario, this inert Higgs Lagrangian is invariant under a $\mathbb{Z}_2$ symmetry: S and P are odd, while all other particles are even. Additional $\mathbb{Z}_2$ symmetric quartic couplings of the inert Higgs involving e.g. S^2 or P^2 are also possible but are not required to generate the low energy interactions in Eq. (6.6), and we neglect them in the following.

Integrating out the inert Higgs at tree level (see Fig. 6.6, right diagram), and matching onto the effective Lagrangian of Eq. (6.6), we find

$$\frac{g_{dd}^{SP}}{\Lambda_{\rm NP}} = \frac{i\lambda_{SP}v}{\sqrt{2}m_H^2}\,\text{Re}(Y_{dd})\ , \qquad \frac{\tilde{g}_{dd}^{SP}}{\Lambda_{\rm NP}} = \frac{i\lambda_{SP}v}{\sqrt{2}m_H^2}\,\text{Im}(Y_{dd})\ , \tag{6.31}$$

$$\frac{g_{ds}^{SP}}{\Lambda_{\rm NP}} = \frac{\lambda_{SP}v}{\sqrt{2}m_H^2}\frac{i}{2}(Y_{ds}+Y_{sd}^*)\ , \qquad \frac{\tilde{g}_{ds}^{SP}}{\Lambda_{\rm NP}} = \frac{\lambda_{SP}v}{\sqrt{2}m_H^2}\frac{1}{2}(Y_{ds}-Y_{sd}^*)\ . \tag{6.32}$$

In addition, integrating out the inert Higgs gives 4-fermion contact interactions of the type $(\bar{d}_L s_R)(\bar{d}_R s_L)$ that modify kaon oscillations. We find the following contributions to the kaon mixing matrix element:

$$M_{12} = \frac{m_{K^0}^3 f_K^2}{4m_s^2 m_H^2}\eta_{\rm QCD} B_4 Y_{sd}Y_{ds}^*\ , \tag{6.33}$$

where $B_4 \simeq 0.78$ [336] (see also [337, 338]) and $\eta_{\rm QCD}$ is the QCD correction factor given in Eq. (6.10), with $M = m_H$. Modifications to the mixing matrix alter the neutral kaon

oscillation frequency ΔM_K and the observable ϵ_K that measures CP violation in kaon mixing. The above contribution to M_{12} modifies these two quantities as

$$\Delta M_K = \Delta M_K^{\rm SM} + 2\,{\rm Re}(M_{12})\ , \qquad \epsilon_K = \epsilon_K^{\rm SM} + \frac{{\rm Im}(M_{12})}{\sqrt{2}\Delta M_K}\ . \tag{6.34}$$

Taking into account the SM predictions $\Delta M_K^{\rm SM}$ and $\epsilon_K^{\rm SM}$ from [339, 340], and the corresponding experimental values from [341], we find the bounds

$${\rm Re}(Y_{sd}Y_{ds}^*) < 7.3\times 10^{-9}\times\left(\frac{m_H}{1\ {\rm TeV}}\right)^2\left(\frac{\alpha_s(m_H)}{\alpha_s(1\ {\rm TeV})}\right)^{8/7}\ , \tag{6.35}$$

$${\rm Im}(Y_{sd}Y_{ds}^*) < 4.5\times 10^{-12}\times\left(\frac{m_H}{1\ {\rm TeV}}\right)^2\left(\frac{\alpha_s(m_H)}{\alpha_s(1\ {\rm TeV})}\right)^{8/7}\ . \tag{6.36}$$

Assuming $|Y_{ds}| \simeq |Y_{sd}|$ and $\mathcal{O}(1)$ CP violating phases, the kaon mixing bounds are compatible with $\Lambda_{sd} \gtrsim 3\times 10^9\,{\rm GeV}$. Also, note that the bounds are entirely avoided if either of Y_{sd} or Y_{ds} is set to zero.

6.3 Astrophysical and terrestrial constraints

We now consider extant astrophysical and terrestrial constraints that may apply to our model.

First, anticipating our treatment of P as a dark matter candidate, we note that direct detection, indirect detection, and self-interaction constraints are not relevant for our model in its minimal configuration (see Eq. (6.6)). If our P is the cosmological dark matter, but the SM is only coupled to the current SP, then direct detection is only sensitive to the inelastic scattering process $P + {\rm SM} \to S + {\rm SM}$, which is kinematically forbidden unless the dark matter is boosted. Similarly, indirect detection and self-interaction processes require two vertices, and thus the cross sections are suppressed by

$\Lambda_{\rm NP}^4$.

Extensions of our minimal model containing couplings to P^2 (see Eq. (6.7)) may be subject to these constraints due to the presence of additional interactions. However, we first treat constraints from supernova cooling and beam dump experiments, which apply directly to the minimal model.

6.3.1 Supernova constraints

Supernova cooling provides powerful constraints on new weakly-coupled light particles. Evaluating these bounds properly requires a detailed analysis that lies beyond the scope of this work, but we can perform an order-of-magnitude estimate to determine the regions of our parameter space that are likely to be subject to such constraints.

In the case of axions, the cross section for axion production $NN \to NNa$ is constrained by SN1987A to lie in the range [342]

$$3 \times 10^{-20} \lesssim \frac{\sigma}{\text{GeV}^{-2}} \lesssim 10^{-13}. \tag{6.37}$$

Below the lower limit, axions are not produced in sufficient numbers to affect the cooling process. Above the upper limit, the produced axions are trapped within the supernova environment, and are unable to cool the system more effectively than neutrinos. Many details of the calculation for axions should be modified in our case, but we will make a crude estimate of the constraints by requiring our production cross section to lie in the same range.

Since P is stabilized by a $\mathbb{Z}_2$ symmetry, it can only be produced in pairs, or in association with S. The process $NN \to NNPP$ is mediated at the loop level in

the minimal model, involving two insertions of the effective interaction vertex. Since $T_{\rm SN} \simeq 30\,{\rm MeV}$ [342], we estimate the cross section as

$$\sigma_{NN\to NNPP} \sim \frac{1}{16\pi^2}\frac{T_{\rm SN}^2}{\Lambda_{dd}^4} \simeq 6\times 10^{-34}\,{\rm GeV}^{-2}\left(\frac{T_{\rm SN}}{30\ {\rm MeV}}\right)^2\left(\frac{10^7\,{\rm GeV}}{\Lambda_{dd}}\right)^4, \tag{6.38}$$

lying below the constrained range of cross sections, even neglecting exponential suppression when $m_P \gtrsim T_{\rm SN}$. In the case of SP production, $NN \to NNSP$, since $m_S \gg T_{\rm SN}$, we estimate the cross section as

$$\begin{aligned}\sigma_{NN\to NNSP} &\sim \frac{1}{4\pi\Lambda_{dd}^2} e^{-(m_S+m_P)/T_{\rm SN}} \\ &\simeq 7\times 10^{-21}\,{\rm GeV}^{-2}\exp\left[\frac{35}{3}\left(1-\frac{m_S+m_P}{350\,{\rm MeV}}\frac{30\,{\rm MeV}}{T_{\rm SN}}\right)\right]\left(\frac{10^7\,{\rm GeV}}{\Lambda_{dd}}\right)^2.\end{aligned} \tag{6.39}$$

While parts of our parameter space are thus expected to be unconstrained by supernova limits, it is important to note that if m_P is small, or if $\Lambda_{dd} \lesssim 10^6\,{\rm GeV}$, the estimated production cross section enters the prohibited range. In particular, if $\Lambda_{dd} = 10^6\,{\rm GeV}$, then avoiding the bound requires $m_S + m_P \gtrsim 450\,{\rm MeV}$, favoring the larger P masses in Fig. 6.2. However, in this naive projection of supernova constraints, our model remains viable in a wide region of the parameter space.

6.3.2 Beam dump constraints

In minimal form, our model of the KOTO excess is potentially subject to constraints from long-lived particle searches: the partial decay width of $S \to \pi^0 P$ is bounded from below by the observed KOTO event rate, so in the absence of additional interactions, the S lifetime can be $\mathcal{O}(\text{m})$ or larger. Such lifetimes are probed very effectively by beam-dump experiments with $\mathcal{O}(100\,\text{m})$ baseline lengths. In such an

experiment, a proton beam is directed at a target, potentially producing a large number of S particles. The S particles travel unimpeded through shielding and earth over a distance L_B, reaching an instrumented decay volume with length L_D. The $S \to \pi^0 P$ events within the decay volume can be typically detected with an $\mathcal{O}(1)$ efficiency $\mathcal{E}$. Thus, the strength of the constraints is mainly determined by two factors: (1) how many S particles are produced, and (2) what fraction of these undergo $S \to \pi^0 P$ within the decay volume.

First we estimate the number of S particles produced. There are at least two channels to consider: direct production from nucleon-nucleon scattering, and secondary production from kaon and other meson decays. Observe, however, that the fraction of proton-proton collisions that produce an SP pair is of order $(s/\Lambda_{\rm NP})^2/\alpha_S^2$, which is much smaller than the branching ratios $\mathrm{BR}(K_L \to SP)$ and $\mathrm{BR}(K_S \to SP)$ implied by our interpretation of the KOTO excess. We also checked that the number of S from eta decays $\eta \to SP$ is small compared to those coming from the kaon decays in our scenarios.

Given N_p protons on target, we expect that of order $N_K \sim 10^{-2} N_p$ kaons are produced [331], and this is sufficient for kaon decays to dominate production. However, of these kaons, most will be stopped or scattered away from the axis of the beam before they decay. The dynamics of kaon energy loss and deflection in materials are complicated, but the nuclear interaction length for relativistic kaons in most materials is $L_{\rm nuc} \sim \mathcal{O}(10\,\mathrm{cm})$ [341], so we will assume that any kaons traveling this far before decaying are sufficiently slowed down or deflected such that only a negligible fraction of the S particles are directed towards the detector. Thus, the number of S particles

produced and directed towards the detector is of order

$$N_S \sim \frac{1}{2} \sum_{X=L,S} 10^{-2} N_p \frac{\Gamma(K_X \to SP)}{\Gamma_{K_X}} \left[1 - \exp\left(-\frac{\Gamma_{K_X} L_{\rm nuc}}{\gamma_{K_X}}\right)\right], \tag{6.40}$$

where γ is the boost factor. Now, accounting for the fraction of S particles which decay in the decay volume, and accounting for the efficiency of the detector, the number of events is given by

$$\begin{aligned} N_E \sim \frac{1}{2} \sum_{X=L,S} & 10^{-2} N_p \, \mathrm{BR}(K_X \to SP) \, \mathrm{BR}(S \to \pi^0 P) \mathcal{E} \\ & \times \underbrace{\left[1 - \exp\left(-\frac{\Gamma_{K_X} L_{\rm nuc}}{\gamma_{K_X}}\right)\right]}_{\text{avoid kaon deflection}} \underbrace{\exp\left(-\frac{\Gamma_S L_B}{\gamma_S}\right)}_{\text{reach decay volume}} \underbrace{\left[1 - \exp\left(-\frac{\Gamma_S L_D}{\gamma_S}\right)\right]}_{\text{decay in decay volume}}. \end{aligned} \tag{6.41}$$

In the minimal scenario, with no additional interactions, $\mathrm{BR}(S \to \pi^0 P) = 1$.

We now estimate the event counts in the CHARM [343] and NuCal [344] experiments. CHARM conducted a search for decays of axion-like particles with 2.4×10^{18} protons incident on a copper target at 400 GeV, a baseline length of 480 m, and a 35 m-long decay volume. The detector efficiency is approximately 0.5. No candidate events were observed. NuCal conducted a similar search, with 1.7×10^{18} protons incident on an iron target at 70 GeV, a baseline length of 64 m, and a 23 m-long decay volume. One candidate event was observed with an expected standard model background of 0.3. To estimate the event counts that would be produced by our model, we set $\gamma_{K_X} = \gamma_S = 10$ for CHARM and reduce these proportionally for NuCal's lower beam energy.

Assuming $\mathrm{BR}(S \to \pi^0 P) = 1$, the resulting event count is shown as a function of the S lifetime in Fig. 6.7. The minimum expected number of events at long S lifetime is large unless $\tau_S \gtrsim 10^5$ m, and lifetimes as large as 10^9 m may be excluded. This potentially rules out a significant portion of our parameter space, as indicated

in Fig. 6.4. On the other hand, the event rate cuts off sharply for $\tau_S \lesssim 1\,\mathrm{m}$, and there is indeed a region of our parameter space where $\tau_S \sim 1\,\mathrm{cm}$. These constraints can be relaxed if the coupling of the SP^3 interaction in our model is non-zero, which can shorten the S lifetime significantly if m_P is small (see Fig. 6.5). The presence of this additional interaction greatly extends the parameter space consistent with the null results at CHARM and NuCal.

Looking towards future prospects, most proposed beam-dump experiments are competitive in the same regime of S lifetimes. However, it has been suggested [345] that the SeaQuest experiment [346] may be modified to serve as a short-baseline beam dump experiment, with the instrumented area starting only $\sim 5\,\mathrm{m}$ from the beam target. Such an experiment would have sensitivity to lifetimes as short as 5 cm, and could probe most of the parameter space in which the minimal model can account for the KOTO excess. However, if the SP^3 coupling is unconstrained, the S lifetime can be shortened by many orders of magnitude, potentially evading even these experiments.

6.3.3 Direct dark matter detection

Direct detection of P can occur in the extended model via the interactions in Eq. (6.7). While the interaction terms containing $(\bar{q} i\gamma_5 q)P^2$ give rise to suppressed velocity-dependent cross sections off of nucleons, the operators $(\bar{q}q)P^2$ with $q = d, s$ produce potentially detectable scattering off of nucleons. We define the integrated nucleon form factors

$$B_q^N \equiv \langle N|\bar{q}q|N\rangle = \frac{m_N}{m_q} f_q^N, \tag{6.42}$$

where f_q^N are the form factors for nucleon N of quark q [347]. The direct detection cross section can be cast as

$$\sigma = \sum_{q=d,s} \left(\frac{2m_N}{m_P + m_N} \frac{g_{qq}^{P^2}}{\Lambda_{\rm NP}} B_q^N \right)^2 \approx \frac{4}{\Lambda_{\rm NP}^2} \left[(B_d^N)^2 (g_{dd}^{P^2})^2 + (B_s^N)^2 (g_{ss}^{P^2})^2 \right]. \tag{6.43}$$

Using the central values $B_d^p \approx 6.77$ and $B_s^p \approx 0.50$, it is clear that the dominant effect is scattering off of d quarks if $g_{ss}^{P^2} \simeq g_{dd}^{P^2}$. The scattering cross section off of protons is then

$$\sigma_p \approx 7 \times 10^{-38}\,{\rm cm}^2 (g_{dd}^{P^2})^2 \left(\frac{10^6\ {\rm GeV}}{\Lambda_{\rm NP}} \right)^2, \tag{6.44}$$

i.e., close to 0.1 pb. Cross sections of this order are above the expected neutrino background, and are within reach of future planned experimental sensitivity [348]. We will return to direct detection prospects in Section 6.5.

6.4 Cosmological production

We now turn to the question of cosmological production of the dark matter candidate P: which scenarios allow P to be produced with the observed dark matter density?

The standard thermal freeze-out paradigm is not viable in our minimal scenario. Estimating the freeze-out temperature by $n_P \sigma(PP \to {\rm SM}) \sim H(T)$, we have

$$T_{\rm FO} \sim \frac{4\pi \Lambda_{\rm NP}^2}{M_{\rm Pl}} \sim \left(\frac{\Lambda_{\rm NP}}{10^{12}\,{\rm MeV}} \right)^2 10^3\,{\rm MeV}, \tag{6.45}$$

where $\Lambda_{\rm NP}$ is the scale of new physics in question—for practical purposes, the lesser of Λ_{sd} and Λ_{dd}. For typical values of $\Lambda_{\rm NP}$ consistent with the KOTO excess, $T_{\rm FO} \gg m_P$,

so P freezes out as a hot relic, with relic abundance

$$\Omega_P h^2 \sim \frac{m_P}{\text{keV}} \left(\frac{g_\star|_{T_{\text{FO}}}}{100} \right) \sim 0.1 \left(\frac{m_P}{80\,\text{eV}} \right). \tag{6.46}$$

Thus, for the masses and couplings considered in this chapter, P is generically over-produced in the freeze-out scenario. If the P mass were small enough to be produced with the right relic abundance, then P would be ruled out as a dark matter candidate because of structure formation constraints on relativistic relics.

Departing from the minimal scenario outlined above opens up the possibility that an *additional* effective interaction with SM species keeps P in thermal equilibrium, and that the P relic abundance is set by thermal decoupling (freeze-out). Since generally thermal decoupling happens at temperatures $T \sim m_P/25$, in order to avoid possible constraints from BBN, one can assume that the effective interaction only involves SM neutrinos:

$$\mathcal{L} \supset \frac{1}{\Lambda_{\nu\nu}} \bar{\nu}\nu PP. \tag{6.47}$$

For the effective dimension five operator in the equation above, we find that the zero-velocity thermally averaged product of the pair-annihilation cross section and relative velocity is

$$\lim_{v\to 0} \langle \sigma v \rangle = \frac{1}{4\pi} \frac{1}{\Lambda_{\nu\nu}^2}. \tag{6.48}$$

A standard treatment of the relic abundance for the pair-annihilation cross section above indicates that P would be produced in the right amount if $\Lambda_{\nu\nu} \simeq 7$ TeV. This is several orders of magnitude above current limits for dark matter interactions with SM neutrinos, independent of flavor [349]. Thus, if P were in equilibrium at high temperatures, an effective interaction with SM neutrinos—which, incidentally, can be

quite naturally embedded in the UV completions described above—could suppress the P abundance to an acceptable relic density in agreement with observations.

In the absence of the additional neutrino portal described in the paragraph above, the only alternative is production via freeze-in [350]. Here the dark species is produced out of equilibrium by some standard model species, and the abundance increases until cosmological expansion halts production. It is thus possible to avoid overproduction of dark matter with extremely small couplings. Note that while other mechanisms might allow for additional production of P, the freeze-in contribution is unavoidable in the range of temperatures where our effective theory is valid.

Typically, freeze-in is applied to a UV-complete theory, where the dark matter production rate can be computed starting at very high temperatures. In the context of a renormalizable model, it can be shown that dark matter is produced primarily at lower temperatures, so the details of the UV physics are unimportant. Thus, freeze-in can be used to consistently calculate the non-thermal relic abundance, even though a formal dependence on initial conditions remains. Note that this is in contrast to the freeze-out paradigm, where equilibrium with the standard model bath erases any non-trivial initial conditions in the dark sector.

However, in our scenario, the dark matter is produced through non-renormalizable interactions, and the standard freeze-in mechanism cannot be directly applied: our effective theory cannot be applied at scales above some $\mathcal{O}(\Lambda_{\rm NP})$ cutoff. At first, this does not seem to be a problem: in standard freeze-in, production is IR-dominated, and we can apply our effective theory in this regime. But for higher-dimension operators, production is no longer IR-dominated, and it is no longer possible to self-consistently

estimate the relic abundance unless an initial condition is fixed at a temperature where the effective theory is valid.

Naively, one can place a lower bound on the relic abundance by fixing the dark matter abundance to zero at $T \sim \Lambda_{\rm NP}$ and computing the amount of dark matter produced at lower temperatures, where the effective theory is valid. However, as we shall see in the following section, this still leads to overproduction of P. Thus, in our model, it would seem that dark matter is overproduced in the freeze-in scenario, even with the most favorable initial conditions.

There is, however, a significant loophole in this argument: setting the dark matter abundance to zero at $T \sim \Lambda_{\rm NP}$ is in fact not the most favorable initial condition. If reheating occurs at a temperature $T_{\rm rh} \ll \Lambda_{\rm NP}$, then the dark matter abundance should be set to zero at this lower temperature, allowing for a much lower relic abundance. There is nothing particularly unnatural about this scenario: in general, freeze-in production of dark matter depends on the reheating temperature. This dependence is weak if the reheating scale happens to be much higher than any scale in the theory, but the convenience of this arrangement does not constitute evidence for it. Moreover, if $T_{\rm rh} \ll \Lambda_{\rm NP}$, then our effective theory can be used to self-consistently compute the dark matter relic abundance independently of any UV completion. This paradigm is known as UV freeze-in [351].

6.4.1 Computing the yield

First, we briefly review the computation of the dark matter relic abundance in the standard freeze-in paradigm. The basic technology of UV freeze-in is identical to

that of standard freeze-in, but the initial condition is fixed at the reheating temperature $T_{\rm rh}$, which becomes an important free parameter of the theory. In certain scenarios, the dark matter yield is quite sensitive to temperatures near $T_{\rm rh}$, and decreasing $T_{\rm rh}$ can significantly reduce the relic abundance.

The starting point is the Boltzmann equation,

$$\dot{n}_\chi + 3Hn_\chi = \sum_{I,F} \left[N_\chi(F) - N_\chi(I)\right] \int \mathrm{d}^{n_I}\Pi_I \, \mathrm{d}^{n_F}\Pi_F (2\pi)^4 \delta^4\left(p_I - p_F\right) \left|\mathcal{M}_{I\to F}\right|^2 \prod_{i\in I} f_i. \tag{6.49}$$

Here n_χ denotes the number density of a dark species χ, I and F index initial and final states, $N_\chi(S)$ denotes the number of χ particles in the state S, $\mathrm{d}\Pi_i = g_i \, \mathrm{d}^3\mathbf{p}_i/(2\pi)^3 2E_i$, $\left|\mathcal{M}_{I\to F}\right|^2$ is the spin-averaged squared matrix element, and f_k is the phase space density of the species k. We assume Maxwell-Boltzmann statistics, and by conservation of comoving entropy density, we rewrite the left-hand side of Eq. (6.49) as $\dot{n}_\chi + 3Hn_\chi = S\dot{Y}_\chi$, where $S = (2\pi^2/45) g_{\star S} T^3$ is the entropy density and $Y_\chi \equiv n_\chi/S$. In turn, since $\dot{T} \approx -HT$, we have $S\dot{Y}_\chi \approx xHSY'_\chi(x)$, where $x = \mu/T$ for any fixed mass μ.

In freeze-in, one assumes that the phase space density of the dark species is always small, so that any initial state with $N_\chi(I) > 0$ makes a negligible contribution in Eq. (6.49). If all of the initial-state species are now in equilibrium, the phase space densities f_i can be replaced with equilibrium distributions $e^{-E_i/T}$. Now Eq. (6.49) reads

$$Y'_\chi(x) = \frac{1}{xHS} \sum_{I\not\ni\chi,F} N_\chi(F) \int \mathrm{d}^{n_I}\Pi_I \, \mathrm{d}^{n_F}\Pi_F (2\pi)^4 \delta^4\left(p_I - p_F\right) \left|\mathcal{M}_{I\to F}\right|^2 \exp\left(-xE_I/\mu\right). \tag{6.50}$$

We will be interested in two types of processes: $1 \to 2$ decays and $2 \to 2$

scattering. In the $1 \to 2$ case, with a process $i \to \chi f$, we set $\mu = m_i$, i.e., $x = m_i/T$. We recognize the decay width $\Gamma_{i\to\chi f}$ in Eq. (6.50), which becomes

$$Y'_\chi(x) = \frac{1}{2\pi^2}\frac{g_i m_i^3}{x^2 HS} N_\chi(F)\Gamma_{i\to\chi f} K_1(x)\,, \tag{6.51}$$

where K_1 is a modified Bessel function of the second kind, and now $N_\chi(F)$ is either 1 or 2, depending on whether $f = \chi$. Substituting $H = 1.66 g_\star^{1/2} x^{-2} m_i^2 M_{\rm Pl}^{-1}$, the total yield can now be computed by performing a 1-dimensional integration of Eq. (6.51), as

$$Y_\chi(\infty) = \frac{45 N_\chi(F) g_i M_{\rm Pl}\Gamma_{i\to\chi f}}{1.66\times 4\pi^4 m_i^2}\int_{x_{\rm min}}^\infty {\rm d}x \frac{x^3 K_1(x)}{g_\star^{1/2} g_{\star S}}. \tag{6.52}$$

In particular, suppose that $f = \chi$, $m_\chi \ll m_i$, and $|\mathcal{M}_{i\to\chi\chi}|^2 = \lambda^2$. If production mainly takes place during an epoch when $g_\star$ and $g_{\star S}$ are not changing rapidly, then we can estimate the yield as

$$Y_\chi(\infty) \simeq \frac{135 N_\chi(F) g_i M_{\rm Pl}\lambda^2}{1.66\times 8(2\pi)^4 g_\star^{1/2} g_{\star S} m_i^3}\begin{cases} 1 & x_{\rm min} \ll 1 \\ \frac{1}{3}\sqrt{\frac{2}{\pi}}\, x_{\rm min}^{5/2}\exp(-x_{\rm min}) & x_{\rm min} \gg 1. \end{cases} \tag{6.53}$$

Similarly, if the abundance of χ is set by $2 \to 2$ processes of the form $ij \to \chi f$, then the integrals over the final-state phase space produce the cross section $\sigma_{ij\to\chi f}$, and Eq. (6.50) becomes

$$Y'_\chi(x) = \frac{N_\chi(F) g_i g_j}{xHS}\int \frac{{\rm d}^3\mathbf{p}_i}{(2\pi)^3}\frac{{\rm d}^3\mathbf{p}_j}{(2\pi)^3}\,\sigma v \exp(-xE_i/\mu)\exp(-xE_j/\mu)\,. \tag{6.54}$$

This remaining integrals can be reduced to a single 1d integral, following e.g. [352]. Integrating in x, the yield is then

$$\begin{aligned} Y_\chi(\infty) = \frac{\mu N_\chi(F) g_i g_j}{2(2\pi)^4}\int_{x_{\rm min}}^\infty \frac{{\rm d}x}{x^2 HS}\int_{s_{\rm min}}^\infty {\rm d}s\,\sigma(s)\, r_- r_+ \times \\ \times\left\{\frac{m_+ m_-}{s}\left(\frac{\mu}{x}+\sqrt{s}\right)\exp\left(-x\sqrt{s}/\mu\right) + \frac{r_- r_+}{\sqrt{s}} K_1\left(x\sqrt{s}/\mu\right)\right\}, \end{aligned} \tag{6.55}$$

where $m_\pm = |m_i \pm m_j|$, $r_\pm = \left(s - m_\pm^2\right)^{1/2}$, and $s_{\rm min} = \min(m_i + m_j,\, m_\chi + m_f)^2$. As in the $1 \to 2$ case, we can estimate the yield analytically for a process $ii \to \chi\chi$ when $m_i \ll m_\chi$ and the evolution of $g_\star$ and $g_{\star S}$ is negligible. If $|\mathcal{M}_{ii\to\chi\chi}|^2 = \lambda^2$, then the result is

$$Y_\chi(\infty) \simeq \frac{45 N_\chi(F) g_i^2 M_{\rm Pl} \lambda^2}{1.66 \times 512 \pi^5 g_\star^{1/2} g_{\star S} m_i} \begin{cases} (3\pi/8) m_i/m_\chi & x_{\rm min} \ll 1 \\ x_{\rm min} \exp\left(-2 x_{\rm min} m_\chi/m_i\right) & x_{\rm min} \gg 1, \end{cases} \tag{6.56}$$

where $x_{\rm min} = m_i/T_{\rm max}$. The analogous expression for $m_\chi \ll m_i$ is obtained by interchanging m_i and m_χ and taking $\mu = m_\chi$ (i.e. $x_{\rm min} = m_\chi/T_{\rm max}$). However, in our model, $2 \to 2$ processes are driven by effective 4-point vertices suppressed by a scale $\Lambda_{\rm NP}$, so we should instead set $|\mathcal{M}_{ii\to\chi\chi}|^2 = s/\Lambda_{\rm NP}^2$. In this case, the result is

$$Y_\chi(\infty) \simeq \frac{45 N_\chi(F) g_i^2 M_{\rm Pl} m_\chi^2}{1.66 \times 128 \pi^4 g_\star^{1/2} g_{\star S} m_i \Lambda_{\rm NP}^2} \begin{cases} \frac{8}{\pi} \left(m_\chi/m_i\right)^{-2} x_{\rm min}^{-1} & x_{\rm min} \ll 1 \\ x_{\rm min} \exp\left(-2 x_{\rm min} m_\chi/m_i\right) & x_{\rm min} \gg 1. \end{cases} \tag{6.57}$$

This demonstrates a key difference between standard freeze-in and UV freeze-in: a naive extrapolation of the production rate to arbitrarily high temperatures (small $x_{\rm min}$) diverges. Of course, one should not expect to accurately compute the production rate in the effective theory at $T \gg \Lambda_{\rm NP}$. But even so, if $\Lambda_{\rm NP} \gg T_{\rm max} \gg \max\{m_\chi, m_i\}$, then production can dominated by $2 \to 2$ processes, whereas $1 \to 2$ decays typically dominate in standard freeze-in. In our case, m_χ and m_i are MeV-scale, while $\Lambda_{\rm NP} \gtrsim 10^6\,{\rm GeV}$. Thus, production by $2 \to 2$ processes at high temperatures is potentially very significant.

Using the approximate forms of the yield derived above together with the dark matter abundance today, $Y_\chi(\infty) \approx 2 \times 10^{-6} (m_\chi/{\rm MeV})$, we can estimate the ranges of parameters which account for all of dark matter—or, at least, those which do not

overclose the universe. If dark matter in our model is produced dominantly by quark annihilation via an interaction of the form $\Lambda_{dd}^{-1} d(i\gamma_5)\bar{d}SP$, then the only important parameters are Λ_{dd} and $x_{\rm min}$. Note that if this is the only interaction at work, there is no contribution from decays.

First, suppose that $x_{\rm min} \ll 1$. Then the scale Λ_{dd} must satisfy

$$\Lambda_{dd} \gtrsim \left(\frac{g_\star|_{T_{\rm rh}}}{100}\right)^{-3/4} \left(\frac{T_{\rm rh}}{\rm GeV}\right)^{1/2} 3 \times 10^{10}\,{\rm GeV}. \tag{6.58}$$

Per the analysis in Section 6.2.2, this is too large to account for the KOTO excess—and this estimate accounts for only one production channel! In particular, if $T_{\rm rh} > \Lambda_{dd}$, dark matter is dramatically overproduced. At the very least, one requires $T_{\rm rh} \lesssim 100\,{\rm MeV}$, where the approximations made for this estimate are no longer trustworthy. However, suppose instead that reheating indeed takes place near the MeV scale, so that $x_{\rm min} \gg 1$. Then the situation is quite different: neglecting the difference between m_S and m_P, we have

$$\Lambda_{dd} \gtrsim \left(\frac{g_\star|_{T_{\rm rh}}}{10}\right)^{-3/4} \left(\frac{T_{\rm rh}}{10\,{\rm MeV}}\right) \exp\left[-\left(\frac{m_S}{T_{\rm rh}} - 30\right)\right] 300\,{\rm GeV}. \tag{6.59}$$

This bound poses no obstacle to accounting for the KOTO excess. When combined, these two estimates naively suggest that our model can account for all of dark matter if reheating takes place between 100 MeV and 10 MeV. While the scale of reheating is often assumed to be much higher, the strongest observational lower bound on the reheating temperature is in fact only $T_{\rm rh} \gtrsim 5\,{\rm MeV}$ [353, 354]. There is no particularly strong motivation for a very high reheating temperature, and certainly nothing inconsistent about reheating taking place at 10 MeV.

However, production at such low temperatures introduces a new complication:

our simplistic estimates above have presumed not only that production is dominated by $2 \to 2$ processes, but also that the initial state consists of free quarks. If $T_{\rm rh} < 100\,\text{MeV}$, then quarks are confined into hadrons during the entire production period. One must then modify the effective couplings to account for hadronic scattering, and since the initial and final states are all (pseudo)scalars, the matrix elements no longer carry any s-dependence. Additionally, since single hadrons can now decay to S and P, hadronic decays can dominate the relic abundance, and must be included in the calculation of the yield.

In the following section, we treat these issues in detail and calculate the relic density numerically.

6.4.2 Determining the reheating temperature

Our estimates in the previous section suggest that P can be produced non-thermally, and can account for all of dark matter, if the initial temperature of the SM bath is between 100 MeV and 10 MeV. We now refine our estimate of the yield to account for confinement and hadronic decays, and then numerically compute the yield to establish the required reheating temperature in our model.

At $T \lesssim 200\,\text{MeV}$, quarks are confined into hadrons, and the effective interactions of the hadrons with S and P are well described by chiral perturbation theory (chiPT). The effective couplings of hadrons to S and P are built from a combination of the new physics scales and QCD parameters. Since the couplings in the quark-level effective Lagrangian are proportional to $\Lambda_{\rm NP}^{-1}$, and the hadron-level $1 \to 2$ coupling must have mass dimension 1, the latter must be of order $\Lambda_{\rm chiPT}^2/\Lambda_{\rm NP}$, where $\Lambda_{\rm chiPT}$ is some

scale associated with low-energy QCD. Similarly, in the $2 \to 2$ case, the hadron-level coupling should have the form $\Lambda'_{\rm chiPT}/\Lambda_{\rm NP}$. As we will see momentarily, $\Lambda^{(\prime)}_{\rm chiPT}$ is a combination of two constants, $f_\pi \approx 92\,\text{MeV}$ and $B_0 \approx 2666\,\text{MeV}$. To determine the couplings explicitly, we match our effective quark-level Lagrangian onto the chiPT Lagrangian following [355, 356]. Our application of this method to light scalars is also similar to the treatment in section 3.1 of [329].

The interactions of QCD degrees of freedom with our light scalars can be written as the couplings of quarks to external currents s and p, respectively a scalar and pseudoscalar. These take the form

$$\mathcal{L}_{\rm QCD}[s,p] = -\bar{\mathbf{q}}\left(s(x) - i\gamma_5 p(x)\right)\mathbf{q}. \tag{6.60}$$

Interactions of hadrons with these currents enter the chiPT Lagrangian via the current $\chi = 2B_0(s+ip)$. At lowest order, we have

$$\mathcal{L}_2 \supset \frac{f_\pi^2}{4}\,\text{tr}\left(\chi U^\dagger + U\chi^\dagger\right), \qquad U = \exp\left(\frac{i\sqrt{2}}{f_\pi}\,\Phi\right), \tag{6.61}$$

where Φ is the PNGB matrix [see e.g. 355]. Now consider a quark-level interaction of the form

$$\mathcal{L} \supset \frac{1}{2}\bar{q}_i\left(g_{ij}^{\mathcal{O}_S} - i\tilde{g}_{ij}^{\mathcal{O}_S}\gamma_5\right)q_j\mathcal{O}_S + \frac{i}{2}\bar{q}_i\left(g_{ij}^{\mathcal{O}_P} - i\tilde{g}_{ij}^{\mathcal{O}_P}\gamma_5\right)q_j\mathcal{O}_P + \text{h.c.} \tag{6.62}$$

where $\mathcal{O}_S$ is a scalar (CP-even) and $\mathcal{O}_P$ is a pseudoscalar (CP-odd). We can then identify

$$s_{ij} = -\frac{1}{2}\left(g_{ij}^{\mathcal{O}_S} + g_{ji}^{\mathcal{O}_S *}\right)\mathcal{O}_S - \frac{1}{2}\left(g_{ij}^{\mathcal{O}_P} - g_{ji}^{\mathcal{O}_P *}\right)\mathcal{O}_P, \tag{6.63}$$

$$p_{ij} = -\frac{i}{2}\left(\tilde{g}_{ij}^{\mathcal{O}_S} - \tilde{g}_{ji}^{\mathcal{O}_S *}\right)\mathcal{O}_S - \frac{i}{2}\left(\tilde{g}_{ij}^{\mathcal{O}_P} + \tilde{g}_{ji}^{\mathcal{O}_P *}\right)\mathcal{O}_P. \tag{6.64}$$

Substituting these expressions into Eq. (6.61) with $\mathcal{O}_S = S^2$, P^2, and $\mathcal{O}_P = SP$ gives the interactions of S and P with the PNGBs. For instance, the interactions of S and P with π^0 are specified by

$$\mathcal{L}_2 \supset B_0 f_\pi \pi^0 \left(SP \,\mathrm{Im}\, g_{dd}^{SP} - S^2 \,\mathrm{Im}\, \tilde{g}_{dd}^{S^2} - P^2 \,\mathrm{Im}\, \tilde{g}_{dd}^{P^2} \right) - \frac{1}{2} B_0 (\pi^0)^2 \left(SP \,\mathrm{Re}\, \tilde{g}_{dd}^{SP} - S^2 \,\mathrm{Re}\, \tilde{g}_{dd}^{S^2} - P^2 \,\mathrm{Re}\, \tilde{g}_{dd}^{P^2} \right) + \cdots , \quad (6.65)$$

where the ellipsis denotes a series of higher-dimensional operators. We include all terms up to second order in the PNGB fields in our analysis, and the form of the hadron-level Lagrangian is as expected from dimensional analysis. Note that it is essential to consider complex-valued g_{ij} and $\tilde{g}_{ij}$, without which some interactions will vanish.

We can now determine the reheating temperature required to produce the observed dark matter density as a function of our model parameters. First, using the normalization factors as they appear in Eq. (6.65), we can now estimate the relative significance of decays and scattering, starting with Eqs. (6.54) and (6.56). Assuming that all dimensionless couplings are $\mathcal{O}(1)$, we set the coupling λ for 3-point vertices equal to $B_0 f_\pi / \Lambda_{\rm NP}$, and we set the coupling for 4-point vertices to $B_0/\Lambda_{\rm NP}$. In this regime, we typically have $m_i \gg \max\{m_P, T_{\rm rh}\}$, and in this limit,

$$\frac{Y_P^{1\to 2}(\infty)}{Y_P^{2\to 2}(\infty)} \simeq 32 \left(\frac{f_\pi}{m_i}\right)^2 \begin{cases} 1 & m_P \ll T_{\rm rh} \ll m_i \\ \frac{3\pi}{8} (T_{\rm rh}/m_i) \exp\left(2m_i/T_{\rm rh}\right) & T_{\rm rh} \ll m_P \ll m_i. \end{cases} \quad (6.66)$$

Our parameter space includes $1\,\mathrm{MeV} \lesssim m_P \lesssim 200\,\mathrm{MeV}$, so the ratio above can be large or $\mathcal{O}(1)$ depending on the choice of the P mass, but it is never small. Note, however, that increasing m_P can also close certain decay channels. In particular, if there exist

interactions allowing the decay $\pi^0 \to PP$, this channel naively dominates production at low temperatures, but is closed for $2m_P > m_{\pi^0}$.

Since decays dominate in most of the parameter space, we can make a first estimate of the yield by considering only production via $K_L \to SP$, the same decay process which is necessary to account for the KOTO excess. Neglecting the distinction between m_S and m_P, the yield is

$$Y_P^{K_L \to SP}(\infty) \simeq \frac{45}{1.66 \times 4(2\pi)^{9/2} g_\star^{1/2} g_{\star S}} \left(\frac{B f_\pi}{m_K \Lambda_{sd}}\right)^2 \frac{M_{\rm Pl}}{T_{\rm rh}} \exp\left(-2m_S/T_{\rm rh}\right), \tag{6.67}$$

and the resulting upper bound on Λ_{sd} is

$$\Lambda_{sd} \gtrsim \left(\frac{g_\star|_{T_{\rm rh}}}{10}\right)^{-3/4} \left(\frac{T_{\rm rh}}{15\,\text{MeV}}\right)^{1/2} \exp\left[-\left(\frac{m_S}{T_{\rm rh}} - 20\right)\right] 5 \times 10^6\,\text{GeV}. \tag{6.68}$$

For the typical parameter values selected above, this upper bound is towards the lower edge of our parameter space of interest for the KOTO excess. Thus, although hadronic decays significantly enhance production relative to the prediction of Eq. (6.59), this channel on its own does not pose an obstacle to accounting for the KOTO excess.

However, in general, it is necessary to numerically evaluate the yield to determine the extent of the viable parameter space—and, in particular, to identify the reheating temperature that produces the observed relic density at each parameter point. The resulting reheating temperatures are shown in Fig. 6.8, and are of order 10 MeV throughout the parameter space of interest. The required reheating temperature is mainly controlled by the smaller of Λ_{sd} and Λ_{dd}, with a slight bias towards Λ_{sd}, since production by η decays is suppressed compared to production by K^0 decays due to their relative masses. Note that all couplings except for g_{sd}^{SP} and g_{dd}^{SP} are neglected in Fig. 6.8, so, in particular, $\pi^0 \to PP$ does not contribute to the relic density even when

$2m_P < m_{\pi^0}$. If we suppose that all of the couplings in the effective theory are of similar order, the viable parameter space can change significantly.

We can estimate this effect by taking $g^{S^2}_{q_1 q_2} = g^{SP}_{q_1 q_2} = g^{P^2}_{q_1 q_2}$ and setting $g^{\mathcal{O}}_{sd} = \left(g^{\mathcal{O}}_{ss} g^{\mathcal{O}}_{dd}\right)^{1/2}$ to fix $g^{\mathcal{O}}_{ss}$. The resulting reheating temperatures are shown in Fig. 6.9. With these choices for the couplings, our two benchmark points with $m_P = 10\,\text{MeV}$ are incompatible with freeze-in as a production mechanism, since the required reheating temperature is below observational bounds throughout the relevant parameter space. This is due to the open $\pi^0 \to PP$ decay, which is kinematically closed for the other two benchmark points with $m_P = 100\,\text{MeV}$ and $m_P = 125\,\text{MeV}$. For these points, the required reheating temperature is again of order 10 MeV throughout the relevant parameter space. At the top-left of the corresponding panel of Fig. 6.9, the required reheating temperature *decreases* with increasing Λ_{dd}. This is just because of our assumption that $g^{\mathcal{O}}_{sd}$ is the geometric mean of $g^{\mathcal{O}}_{dd}$ and $g^{\mathcal{O}}_{ss}$: increasing Λ_{dd} corresponds to decreasing $g^{\mathcal{O}}_{dd}$, so if $g^{\mathcal{O}}_{sd}$ is held fixed, then $g^{\mathcal{O}}_{ss}$ must increase to compensate. This increases the relic density, forcing a lower reheating temperature.

Finally, we note that the reheating temperatures shown in Figs. 6.8 and 6.9 are potentially imprecise, and should be viewed as lower bounds. Our calculation of the yield assumes that all of the initial-state species are thermalized, but the mesons freeze out at temperatures of the same order considered here. In particular, π^0, K^0, and η freeze out at 3 MeV, 10.5 MeV, and 11.6 MeV, respectively. In a scenario with a high reheating temperature, this concern would be less significant: the mesons would have a thermal distribution at early times, so as long as dark matter production is not dominated by temperatures well below the mesons' freeze-out temperatures, the effect

should be small. However, we are speculating that the reheating temperature itself is lower than e.g. the kaon freeze-out temperature in parts of our parameter space, in which case the kaons may never be populated with anything resembling a thermal distribution. It is thus possible that Eq. (6.50) overestimates the dark matter relic abundance.

This does not have a significant effect on our qualitative results: we can safely predict that DM is overproduced if $T_{\rm rh} \gtrsim 15\,\text{MeV}$, in which case all of the relevant mesons are thermalized, so this is an upper bound on $T_{\rm rh}$. Likewise, we can see that DM would be underproduced for $T_{\rm rh}$ below a particular value even if the mesons have their equilibrium number densities.[1] This lower threshold is $\mathcal{O}(7\,\text{MeV})$ if $\pi^0 \to PP$ is forbidden, and $\mathcal{O}(2\,\text{MeV})$ if it is not. The only qualitative importance of out-of-equilibrium effects is that it may be possible to construct a cosmologically-viable model in which dark matter is not overproduced even if $\pi^0 \to PP$ is open. However, such a model would depend on the details of reheating, and this analysis lies beyond the scope of this work.

6.5 Discussion

In the foregoing sections, we have introduced a model to account for the KOTO excess and explored the cosmological effects. We now discuss the implications of our results and future experimental prospects.

If the KOTO excess is interpreted at face value, this suggests apparent violation

[1] Since production is dominated by decays, the dark matter relic abundance is mainly determined only by the number density of the parent mesons, and is fairly insensitive to other details of the phase space distribution.

of the GN bound. As has been discussed by several authors [321, 323–333], such a signal at KOTO can be mimicked by a decay of the form $K_L \to \pi^0 X$, where X denotes one or more invisible species. In contrast to most studies, we focus on a new physics scenario where the decay $K_L \to \pi^0$ inv. is realized through a sequence of two-body decays $K_L \to SP \to \pi^0 PP$, where S and P are light neutral scalar particles. Similar scenarios were also studied in [332] where the light particles interact with the SM through a vector or scalar portal. Here we instead analyze a setup where S and P are coupled to the SM through effective operators at a characteristic new physics scale of $\Lambda_{\rm NP} \sim 10^6$–$10^9\,{\rm GeV}$. We have stabilized P with a $\mathbb{Z}_2$ symmetry under which SM species are even and our new species are odd, and we have entertained the possibility of other interactions consistent with such $\mathbb{Z}_2$ invariance, including an SP^3 term that could mediate the decay of $S \to 3P$. Our effective theory is readily UV-completed by e.g. very heavy vector-like quarks or a TeV-scale inert Higgs doublet. Such UV completions can realize a minimal case in which only interactions between SM quarks and SP are present at low energies, as well as more generic cases that include interactions with S^2 and P^2.

If the KOTO excess persists, the GN bound heavily constrains new physics interpretations. A model of the type we consider, with new light scalars, is one of the simplest and most elegant solutions. Since the scale $\Lambda_{\rm NP} \sim 10^6$–$10^9\,{\rm GeV}$ indicated by the KOTO excess is so large, most other experiments are not substantially constraining (with the notable exception of beam dump experiments, to which we will return shortly). In particular, in our scenario, there is a large region of parameter space which can account for the KOTO excess while still unconstrained by other rare meson decays. However, it is important to consider astrophysical constraints. Supernova cooling lim-

its can potentially rule out lower P masses: as discussed in Section 6.3.1, supernova temperatures are high enough, at tens of MeV, to probe the lightest S and P masses that we consider in Fig. 6.2. These constraints are most significant for $\Lambda_{dd} \lesssim 10^6\,\mathrm{GeV}$, and it is important to note that establishing firm constraints from supernova cooling requires a much more detailed analysis beyond the scope of this work. However, the simplistic expectation is that P masses of $\mathcal{O}(10\,\mathrm{MeV})$ and below are disfavored, making our scenario easier to test.

Since the KOTO excess motivates the introduction of new feebly-coupled particles, it is natural to speculate that these new species might contribute to cosmological dark matter—and indeed, we have shown that S and P can constitute all of DM even in the most minimal scenarios needed to explain the KOTO signal. Nevertheless, this comes at a cost: in the absence of additional interactions, there is no mechanism to reduce the DM abundance, and cosmological reheating must take place at very late times, at a temperature of order 10 MeV. This requirement should be interpreted as a cosmological constraint on our model and similar models accounting for the KOTO excess. The scale of the preferred reheating temperature originates mainly from the masses of the new scalars: since the DM abundance is exponentially suppressed in $m_{\mathrm{DM}}/T_{\mathrm{rh}}$, the required reheating temperature depends only logarithmically on the couplings and other scales of new physics.

Such a thermal history is necessary because the effective coupling lies in an intermediate regime: it is too small for freeze-out to deplete the DM abundance, but large enough that UV freeze-in generically overproduces DM. Thus, an additional feature is needed to prevent overproduction. The simplest mechanism to accomplish this, without

any modification to the model, is to make a judicious choice of the reheating temperature. Since we are working with an effective theory, the DM relic density is inherently sensitive to the reheating temperature—indeed, if $T_{\rm rh} \gtrsim \Lambda_{\rm NP}$, we cannot consistently calculate the relic density, but only bound it below. Thus, since $T_{\rm rh}$ is necessarily a parameter of our model, $T_{\rm rh} \sim 10\,\text{MeV}$ is as natural as any other choice. As we have discussed, observational constraints are ineffective at temperatures above $\sim 5\,\text{MeV}$.

We note that in principle low-temperature reheating might leave an imprint on early universe probes such as BBN and CMB. Unfortunately, such potential signals are highly model dependent. Specifically, low-reheating temperature scenarios have been shown in the literature to impart a significant effect on the synthesis of light elements, primarily via (i) modifications to the Hubble rate around BBN by changing the energy density of both relativistic and matter species; (ii) changing the momentum distribution of electron-flavor neutrinos, which directly enters charged current interactions, in turn governing the neutron-proton chemical equilibrium; and (iii) by entropy exchange that can affect the ratio of neutrino to photon temperature, which in turn is testable with CMB data.

Previous studies (see e.g. [357] and references therein) relied on simple assumptions such as a single massive matter species driving reheating, and decaying primarily into neutrinos [354], or electromagnetically-interacting species [353], or hadrons [358]. Generally, testable effects arise for $T_{\rm rh} \lesssim 5$ MeV, implying that no signal is expected for the scenario discussed here, where $T_{\rm rh} \gtrsim 10$ MeV. However, it is important to point out that the reheating scenario might include features that could manifest themselves when more stringent probes of CMB become available in the future [359]. For instance, the

field driving reheating might actually be an *ensemble* of fields, with different masses; the S and P particles might be directly produced in the decay of the field(s) driving reheating, changing the predictions for $T_{\rm rh}$ made above; or new physics in the neutrino sector could make reheating temperatures in the 10s of MeV visible once constraints on $N_{\rm eff}$ significantly improve.

There are other mechanisms which prevent the overproduction of dark matter without requiring a particular temperature for reheating. One possibility is to add an interaction with the SM to restore freeze-out as a viable thermal history, as we discussed briefly in the context of a neutrino portal. This would be a heartening scenario: reheating can still take place at a very high temperature, and the coupling to leptons might allow for additional experimental probes. However, there are several other possibilities. In particular, it is possible that the DM abundance is depleted by additional interactions within the dark sector. This is not possible in our effective theory, but one can consider extensions which keep the DM in thermal equilibrium long after decoupling from the SM bath, or which allow other number-changing processes at a sufficient rate to allow for freeze-out at high temperatures. We emphasize again that our results imply cosmological constraints on models of the KOTO excess: cosmology requires either a restricted range of reheating temperatures or additional features of the low-energy theory, regardless of what fraction of cosmological DM is composed of P.

Of course, one can also consider constraints which only apply if P makes up a significant fraction of DM. The simplest of these is the Lyman-α constraint on warm DM [360], which requires the P population to be non-relativistic at temperatures of $\mathcal{O}$(keV). If P is produced non-thermally via decays at 10 MeV, typical energies will

be of order the masses of the parent states, i.e., $\mathcal{O}(100\,\mathrm{MeV})$. Thus, in order for P to be non-relativistic when $T_\gamma \sim$ keV, we require that $m_P \gtrsim 10\,\mathrm{keV}$. This is a somewhat weaker bound than one expects from supernovae, but it is not subject to the complicated physics involved in such constraints.

The annihilation cross section into visible states is much too small ($\sim 10^{-50}\,\mathrm{cm}^2$) for indirect detection to be viable, nor is there any significant self-interaction in the dark sector. However, the scattering cross section with nuclei could be as large as $\sim$0.1 pb, and thus potentially within reach of future, planned experimental sensitivity for sub-GeV direct dark matter searches. It is thus possible (albeit not guaranteed) that future experiments will probe such signatures associated with our model—particularly direct detection—but it is important to note that in the minimal scenario for the KOTO excess, these signatures are substantially suppressed even compared to the generic expectation. This is because the KOTO excess only requires SM interactions with the current SP, and not PP. Since any DM accounted for by our model is composed entirely of P, this means that any diagrams contributing to indirect detection must be suppressed by $\Lambda_{\mathrm{NP}}^{-4}$. Moreover, at lowest order, direct detection is only sensitive to the inelastic scattering process $NP \to NS$, which is kinematically prohibited for non-relativistic DM. It is thus challenging to conclusively establish that P makes up cosmological DM through direct observational means.

However, it is potentially much easier to determine whether a model like ours accounts for the KOTO excess. If the excess persists at its present size, then as KOTO reaches its design sensitivity, hundreds of events will be observed. With a sample of this size, it is possible to distinguish our model from SM three-body decays kinematically in

much of our parameter space, simply by measuring the pion's transverse momentum. In Fig. 6.10, we show the transverse momentum distributions expected at KOTO in the SM and in our model. By sampling from these distributions and applying the Kolmogorov–Smirnov test, we find that the p_T distribution in our model can be distinguished from the SM three-body decay at 5σ with $\mathcal{O}(100)$ events in much of our parameter space. Sensitivity is lost when m_P is small and $m_S \sim m_{K_L}$, and the distributions may also be too close to distinguish at smaller m_S if the S lifetime is shorter than $\mathcal{O}(10\,\text{cm})$. Still, there are good prospects for making such a determination within the next several years, as KOTO continues to collect data.

There are also discovery prospects for S particles with meter- and centimeter-scale lifetimes at future beam-dump experiments. In particular, as discussed in Section 6.3.2, the SeaQuest experiment can probe much shorter lifetimes than those to which CHARM and NuCal are sensitive. Backgrounds are relatively easy to control for experiments of this type, and they remain sensitive even in our minimal scenario. The figure of merit is the S lifetime, which is at least $\mathcal{O}(\text{cm})$ in our minimal scenario. This can be reduced by enhancing the SP^3 interaction in our effective theory, but nonetheless, searches for long-lived particles promise to be a powerful probe of our scenario in the coming decade.

6.6 Conclusions

Taken together, the anomalous KOTO events and the Grossman–Nir bound provide a strong hint for light new physics. In this chapter, we have introduced an effec-

tive theory that accounts for the excess in the $K_L \to \pi^0$ inv. channel with a metastable scalar S; a lighter, stable pseudoscalar P; and effective dimension-5 operators that mediate interactions between S, P and the d and s quarks. We provided two UV-complete models that would produce an effective theory consistent with our assumptions. We then investigated the implications of our effective theory for cosmology and vice versa. In particular, we have shown that cosmological overproduction of P places important constraints on the structure of the low-energy theory.

At face value, in our minimal scenario, P cannot account for either dark matter or the KOTO excess unless the reheating temperature is close to 10 MeV. While it is possible to escape this conclusion by augmenting the model, e.g. with couplings of P to neutrinos, a low reheating temperature is unavoidable in the model's simplest incarnation. However, unless P is very light, the required reheating temperature is compatible with current constraints from BBN and CMB, possibly even offering an observational handle on the model once CMB Stage IV experiments further probe the effective number of relativistic species.

Finally, we have discussed three experimental tests of our scenario. First, we have shown that portions of our parameter space are within reach of future dark matter direct detection experiments. Second, our metastable S may be discovered by upcoming long-lived particle searches, particularly the planned SeaQuest upgrade. Finally, if P is in our favored mass range, future KOTO data alone can discriminate between our decay chain and the SM three-body decay on the basis of the neutral pion p_T distribution. There are thus strong discovery prospects for P dark matter within the next decade.

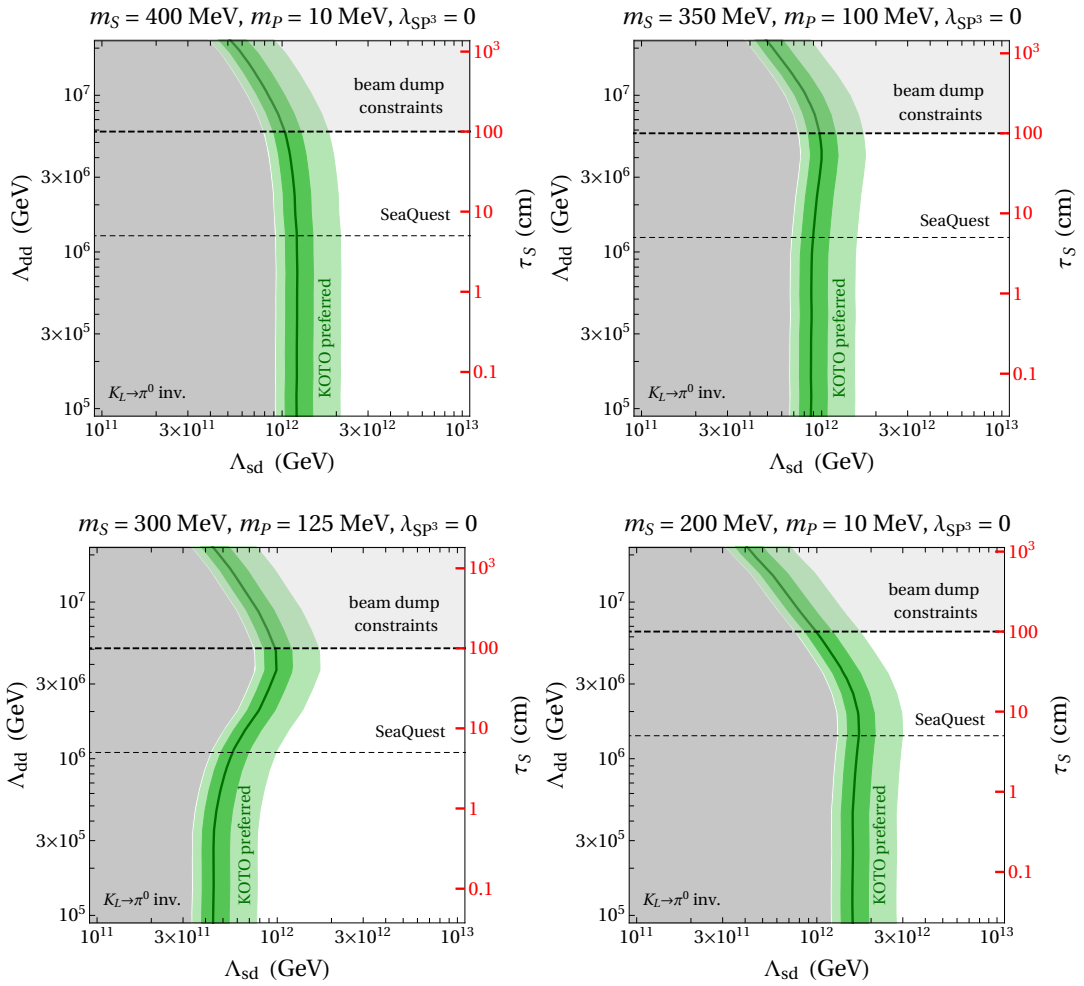

Figure 6.4: Number of expected $K_L \to SP \to \pi PP$ events at KOTO in the Λ_{sd}–Λ_{dd} plane for four benchmark points of the S and P masses. The SP^3 coupling is set to zero. The right vertical axis indicates the S lifetime. One expects 3 events along the solid dark green line, 2–4 events in the dark green region, and 1–5 events in the light green region. In the gray regions labeled "$K_L \to \pi^0$ inv.", the number of predicted events exceeds the limit from KOTO. The dashed lines show constraints from existing beam dump experiments and the potential reach of the SeaQuest upgrade.

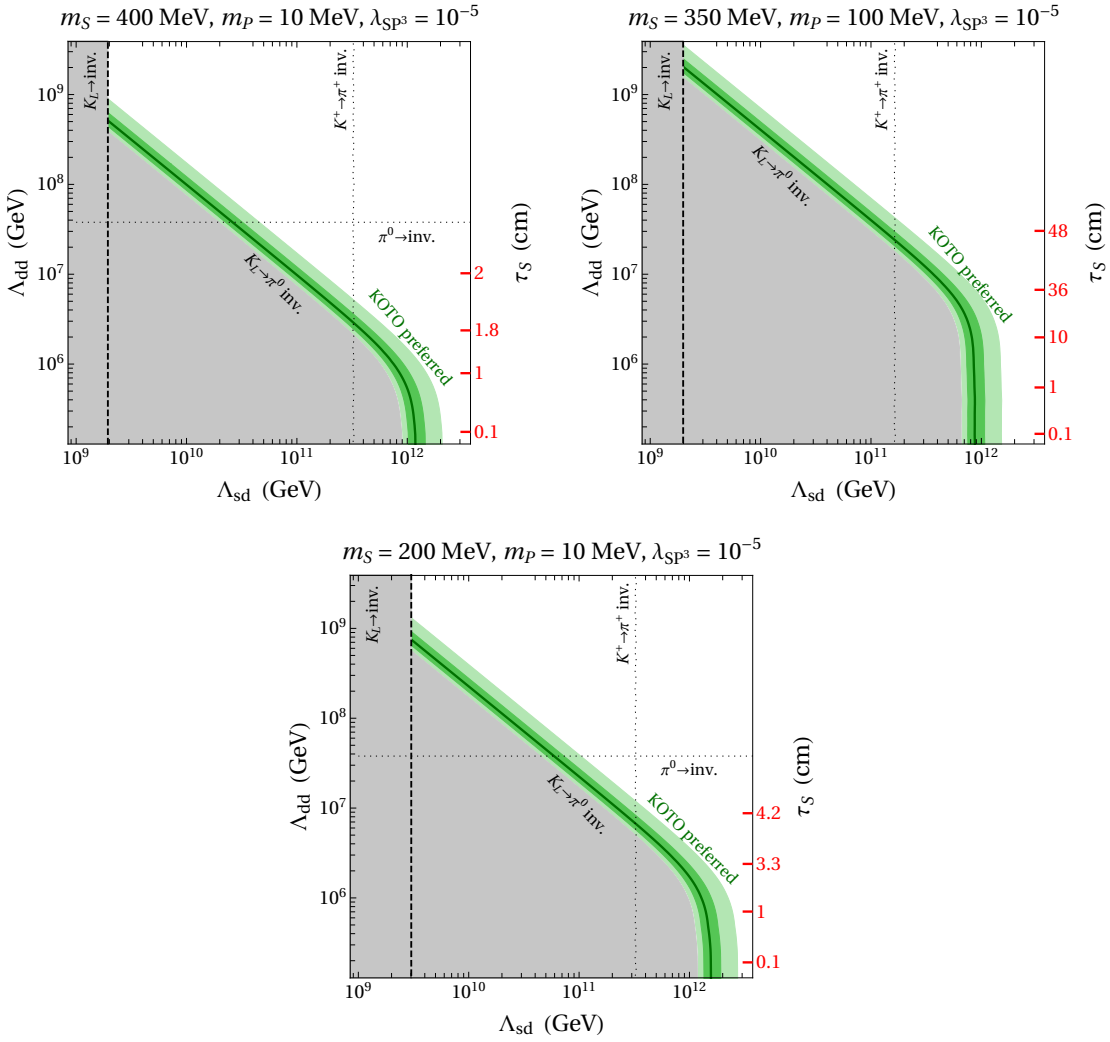

Figure 6.5: Number of expected $K_L \to SP \to \pi PP$ events at KOTO in the Λ_{sd}–Λ_{dd} plane for three benchmark points of the S and P masses. The SP^3 coupling is set to $\lambda_{SP^3} = 10^{-5}$. The right vertical axis indicates the S lifetime, which is approximately constant for $\Lambda_{dd} > 10^7$ GeV. One expects 3 events along the solid dark green line, 2–4 events in the dark green region, and 1–5 events in the light green region. The gray regions are excluded by the KOTO limit on $K_L \to \pi^0$ inv. or the bound on the invisible K_L branching ratio. The dotted lines show the generic location of other constraints that depend on additional model parameters. Benchmark BM3 is not shown, as the $S \to 3P$ decay is kinematically forbidden.

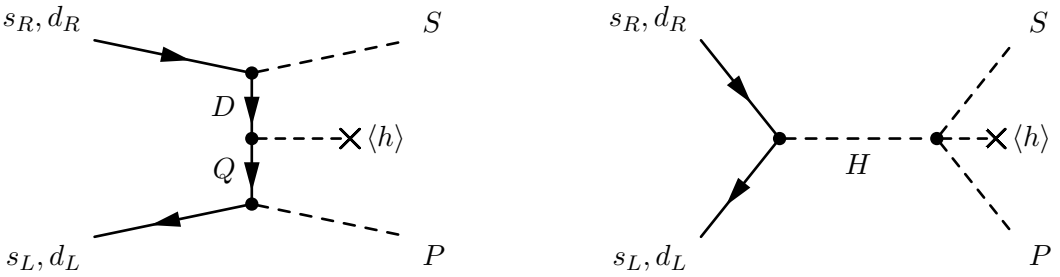

Figure 6.6: Feynman diagrams that show the matching of the vector-like quark model (left) and the inert Higgs model (right) onto the effective $SPqq'$ interactions in Eq. (6.6).

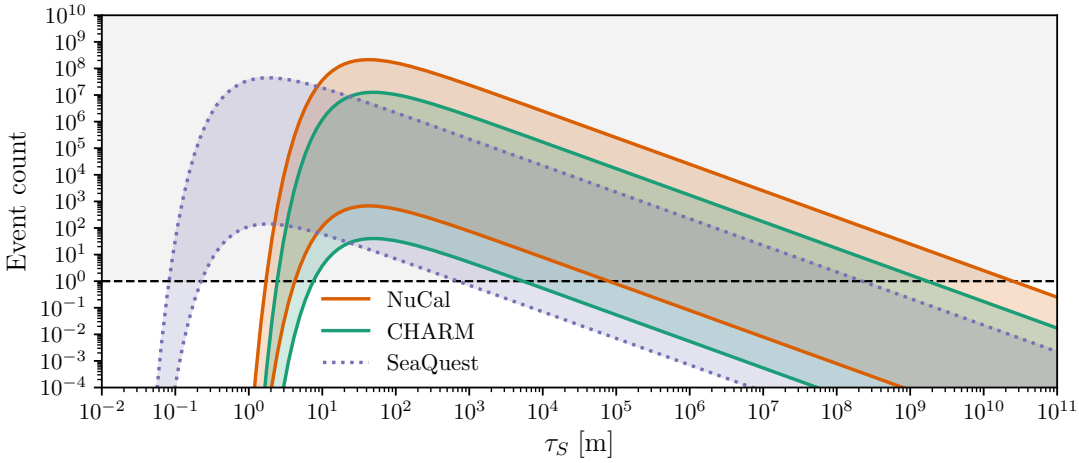

Figure 6.7: Estimated event counts at CHARM and NuCal and prospective event counts at SeaQuest as a function of the S lifetime. The top curve fixes $\Gamma(K_L \to SP)$ to saturate the experimental bound on the invisible K_L width. The bottom curve fixes $\Gamma(K_L \to SP)$ such that $\mathrm{BR}(K_L \to SP)$ is equal to the ratio inferred from the KOTO excess, i.e., $\Gamma(K_L \to SP)$ is the smallest width for which this model can account for the excess. Both curves assume that $\Gamma(K_L \to SP) = \Gamma(K_S \to SP)$ and that $\mathrm{BR}(S \to \pi^0 P) = 1$. Under these conditions, $1\,\mathrm{m} \lesssim \tau_S \lesssim 10^5\,\mathrm{m}$ is ruled out. SeaQuest may eventually probe lifetimes as short as $\tau_S \sim 5\,\mathrm{cm}$.

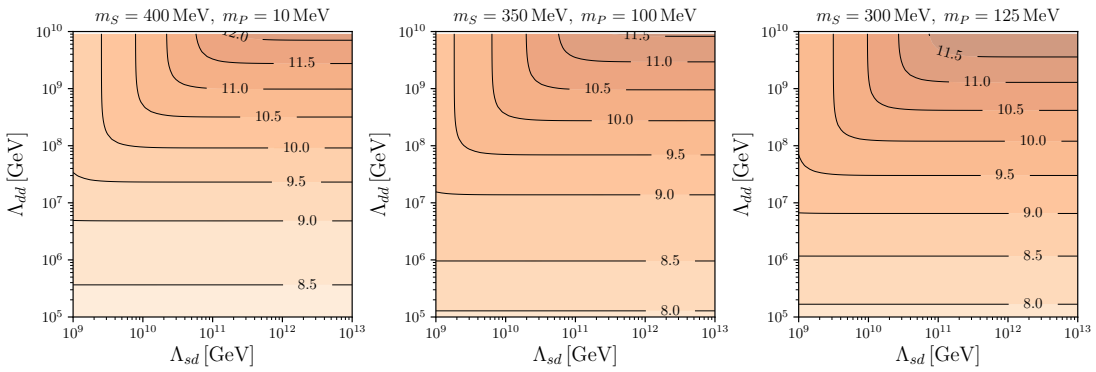

Figure 6.8: Reheating temperature in MeV to produce the observed DM relic density, including all production channels with no DM in the initial state. The couplings g_{sd}^{SP} and g_{dd}^{SP} are taken to be purely imaginary, while all other couplings are set to zero, corresponding to the minimal scenario to account for the KOTO excess. In the leftmost panel (BM1), all decay channels are open. In the middle panel (BM2), $S \to 3P$ is kinematically closed, so there are no number-changing interactions in the dark sector: S decays via $S \to \pi^0 P$. In the rightmost panel (BM3), $S \to 3P$ and $\pi^0 \to PP$ are both closed, so there is no contribution to the relic density from π^0 decays.

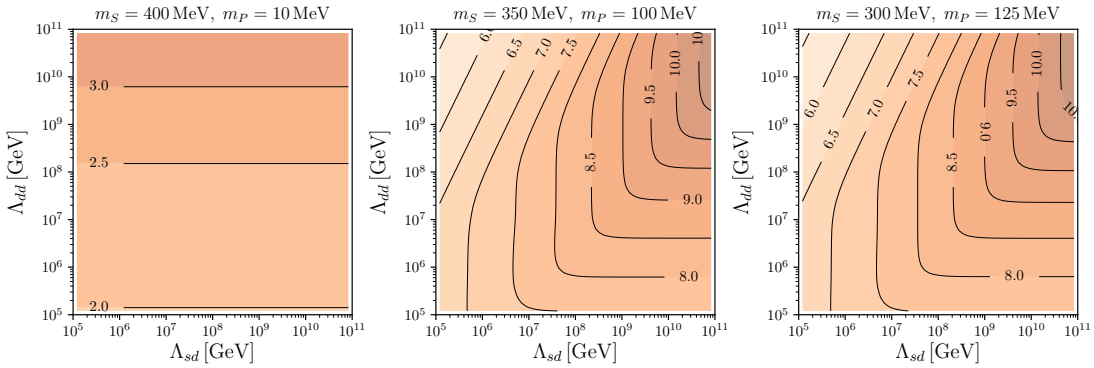

Figure 6.9: Reheating temperature (in MeV) to produce the observed DM relic density, including all production channels with no DM in the initial state, as in Fig. 6.8. Here it is assumed that S^2, P^2, and SP couple equally to light quark bilinears, and that $g^{\mathcal{O}}_{sd}$ is the geometric mean of $g^{\mathcal{O}}_{ss}$ and $g^{\mathcal{O}}_{dd}$. The real and imaginary parts of all couplings are taken to be equal. In the first panel (BM1), all decay channels are open, and production is dominated by π^0 decays. In the middle panel (BM2), $\pi^0 \to PP$ is closed, but $S \to 3P$ is still open. In the rightmost panel (BM3), both $\pi^0 \to PP$ and $S \to 3P$ are closed, so S decays only via $S \to \pi^0 P$. In the leftmost panel, since production is dominated by $\pi^0 \to PP$, the relic abundance is controlled exclusively by Λ_{dd}. In this case, the required reheating temperatures are observationally inviable throughout the parameter space. In the other two panels, production is dominated by K^0 and η decays, and their relative importance depends on Λ_{sd} and Λ_{dd}.

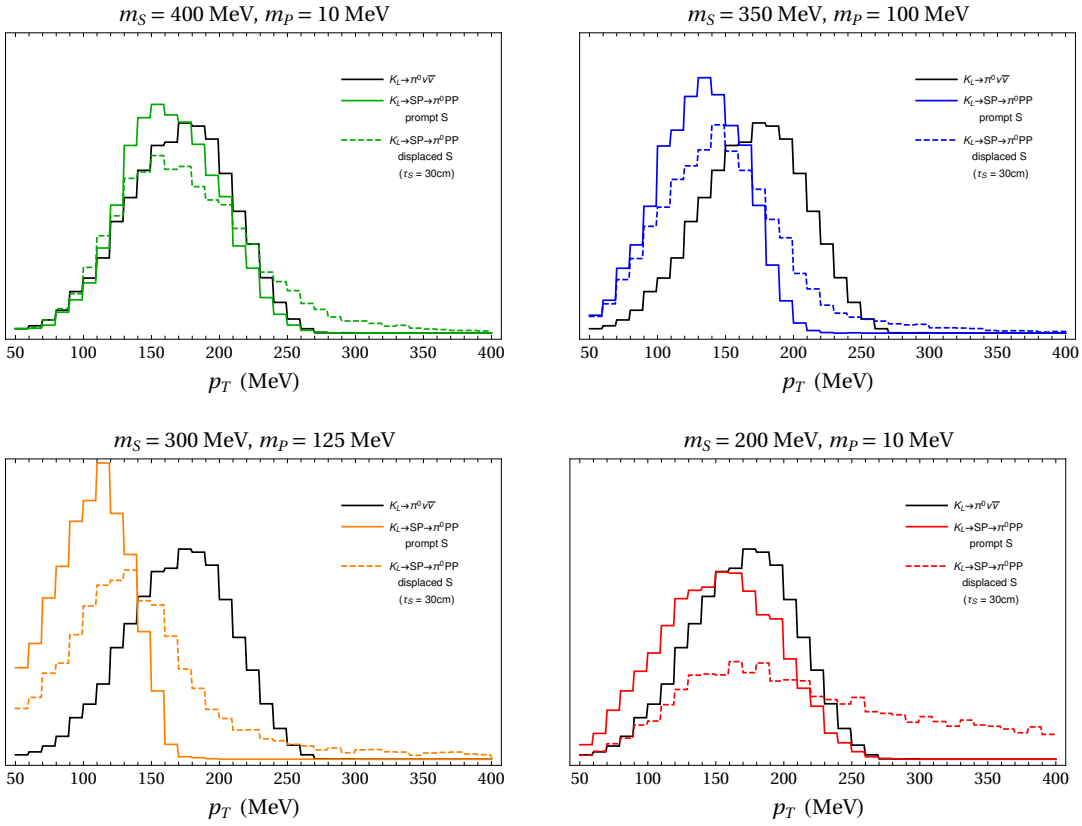

Figure 6.10: Pion p_T distributions for the $K_L \to \pi^0 \nu\bar{\nu}$ decay and the $K_L \to SP \to \pi^0 PP$ decay in our benchmark points. The distributions are shown both for prompt S decays and S decays with a lifetime of 30 cm.

Chapter 7

Complementarity between cosmology and direct detection

7.1 Introduction

In the previous chapter, cosmological observables were used to tie terrestrial experiments to cosmic history. Now we turn to a more direct use of cosmological data for the study of DM: direct constraints on DM interactions. In this chapter, we will demonstrate the powerful complementarity between cosmological constraints and the next generation direct detection experiments. We will give a cursory treatment of the direct detection experiments themselves, with more details to follow in Part III.

To date, direct searches for DM have largely targeted the weak scale. Most extant direct detection experiments are designed to detect the scattering of DM with atomic nuclei, and due to kinematic limits, they have poor sensitivity to a DM particle with mass below 10 GeV [361–363]. Analyses of the phase space distribution of DM in

dwarf spheroidal galaxies bound the mass of fermionic DM to $m_{\rm DM} \gtrsim 1\,\text{keV}$ regardless of the production mechanism [364], and the Lyman-α forest imposes a comparable constraint on thermal relic DM of any kind [360]. But beyond these bounds, DM models with mass between 1 keV and 10 GeV are poorly constrained. Several well-motivated scenarios [e.g. asymmetric DM, 365] naturally feature masses between 1 keV and 10 GeV, making this range an appealing target for future direct detection experiments [366].

This has driven much interest in novel detection methods suited to light DM particles, and several such experiments have been proposed in the last few years [367–377] (see sections IV–V of [378] for a review). These experiments are designed to be sensitive to the very small recoil energies characteristic of the scattering of light particles, and as such, many are designed to search for the scattering of DM with electrons instead of nuclei, a strategy first detailed in [367]. Several experiments now constrain DM–electron scattering at masses as low as $\sim 1\,\text{MeV}$ [379–384]. The more recent proposal of [369], based on electrons in aluminum superconductors, is sensitive to deposited energies of order 1 meV, allowing for the detection of particles as light as 1 keV.

However, although the most generic astrophysical constraints do not restrict DM at masses between 1 keV and 10 GeV, it is well known that particular models can be constrained by cosmological observables, especially for masses below 1 MeV [385]. In particular, light DM interacting with electrons risks running afoul of the following restrictions:

- The DM must not significantly alter successful predictions of the ratios of light elemental abundances produced in big bang nucleosynthesis (BBN) [386–388];

- To accord with measurements of the effective number of neutrino species ($N_{\rm eff}$), the thermal history of the DM species must not significantly alter the temperature ratio of photons and neutrinos at recombination [389];

- While a single species of DM particle may not account for the entirety of the present-day DM density, no species may be produced with an abundance exceeding that threshold.

In each case, such cosmological constraints bound the couplings between new species and Standard Model (SM) particles, which also determine the event rates in direct detection experiments. Thus, in a given model, the cosmological effects of light DM can be related to the direct detection cross section. Given an experimental proposal and a DM model, one can then determine the extent of the parameter space accessible to the experiment and consistent with cosmology. Such an approach has been applied to electron recoil experiments by [390] for a class of simplified models, and more recently in a variety of model-dependent instances [391–399].

In this chapter, we show that cosmological constraints on a new light (sub-MeV) species interacting with electrons can be greatly generalized with a small number of assumptions. Assuming a heavy mediator between DM and the SM, we study the cosmological implications of a light DM species in an effective field theory (EFT), and use the same EFT to evaluate direct detection prospects. We thus obtain model-independent cosmological limits on the scattering cross section of DM with electrons in an actual experiment. The model-independent methodology is similar in spirit to [400–402], but applied to directly connect cosmological constraints and detection prospects

in the sub-MeV regime.

This chapter is organized as follows. In Section 7.2, we describe our EFT framework for modeling light DM coupled to electrons. In Section 7.3, we derive model-independent cosmological constraints on the DM species. In Section 7.4, we evaluate the DM–SM scattering cross section in our EFT, and compare cosmological bounds with prospects in a fiducial experiment. Finally, we discuss implications for direct detection experiments in Section 7.5. A complete set of constraints and tables of cross sections are placed after the end of the text.

Throughout this chapter, we denote a scalar DM field by ϕ and a fermionic DM field by ψ. When speaking about the DM species generally, without specifying its spin, we will denote it with χ.

7.2 Effective interactions of sub-MeV dark matter

In this section, we build a theoretical framework to study the effective interactions of sub-MeV DM of spin 0 or $\frac{1}{2}$. We study DM candidates that are singlets under the SM gauge groups, and we consider both scalar and fermionic DM. We first specify the working assumptions of our EFT framework, and we thereafter develop the scalar and fermion cases separately.

7.2.1 The EFT framework

We assume that DM is dominated by a single particle species with a mass below 1 MeV. The MeV scale is cosmologically significant as the scale of big-bang nucleosynthesis (BBN). The DM annihilation and scattering processes that we consider

in this chapter always involve energy exchanges well below this scale, whether they take place in the early universe or in a laboratory today. Thus, this situation lends itself well to an effective low-energy description with an EFT that has a cutoff of order 10 MeV. In general, the EFT can be valid up to higher scales, but since cosmological history is poorly constrained at temperatures above a few MeV, we only apply the EFT at or below this scale.

At energies well below the MeV scale, the only dynamical SM degrees of freedoms are electrons and positrons ($e^{\pm}$), neutrinos (ν), and photons (γ). We assume further that there is no additional light degree of freedom besides the DM particle: all remaining new physics is presumed to lie well above the MeV scale, including any mediators between DM and SM particles. Physics at sub-MeV scales is thus well described by an EFT in which only $e^{\pm}$, ν, γ, and the DM χ are dynamical degrees of freedom. This is the theoretical framework we employ for our analysis.

Before presenting the EFT in more detail, it is instructive to take a step back and discuss the conceptual starting point of our work: a renormalizable theory with DM as well as mediator fields in the spectrum. The EFT language powerfully encodes the many UV-complete realizations which give the same low energy physics. We make three additional assumptions about the UV-complete theory, described below and graphically summarized in Fig. 7.1:

1. The DM is stablilized by a $\mathbb{Z}_2$ symmetry and is thus absolutely stable.

2. The couplings between mediators and SM fields respect electroweak gauge invariance, in the sense that the χ–e_L coupling is equal to the χ–ν coupling. We make

this assumption to clarify the impact of the DM species on $N_{\rm eff}$, as discussed in the next section. It does not influence the other constraints.

3. DM couples to the visible sector via mediator fields ζ_i, with masses satisfying $T_{\rm BBN} \ll m_{\zeta_i}$.

When writing our EFT Lagrangian, it is convenient to take $m_{\zeta_i} \ll m_{\rm weak} \simeq 100\,\text{GeV}$, so that weak-scale degrees of freedom in the SM can be integrated out before the mediators. It is then possible to define an intermediate EFT with weak scale particles integrated out and mediators in the spectrum. However, our results do not depend on this assumption—it simply clarifies how we should write the low-energy Lagrangian to accommodate lower mediator masses.

Ultimately, our EFT will contain a mass scale $\Lambda_{\rm EFT}$ which is related to the mediator masses, and each operator will appear with a coupling (Wilson coefficient) g. We ensure that we remain in the regime of validity of the EFT by enforcing $\Lambda_{\rm EFT} \gg T_{\rm BBN}$, so it is convenient to assume that $g \sim \mathcal{O}(1)$ and take $\Lambda_{\rm EFT}$ to be the free parameter in our analysis. Small deviations of g from unity can then be absorbed by rescaling $\Lambda_{\rm EFT}$. But if g is not $\mathcal{O}(1)$ in a typical UV completion, and $\Lambda_{\rm EFT}$ is not many orders of magnitude larger than $T_{\rm BBN}$, we have reason for caution: rescaling $\Lambda_{\rm EFT}$ to absorb a very small g could violate the requirement that $\Lambda_{\rm EFT} \gg T_{\rm BBN}$. Thus, when the scale of the DM–SM interaction is smaller, it is important to separate g from any non-$\mathcal{O}(1)$ coupling typical of UV completions. An intermediate EFT lying below the weak scale guides our expectations for the size of the coupling in the effective theory after integrating out the mediators.

In particular, if a scalar ζ mediates the DM–SM interaction, it is easy to generate a factor of the electron Yukawa coupling y_e. Coupling ζ to the lepton doublet L without breaking gauge invariance involves interaction terms of the form

$$\mathcal{L}_{\rm UV} \supset M_1 \zeta \phi^\dagger \phi + M_2 \zeta H^\dagger H + \zeta^\dagger \zeta H^\dagger H + y_e \bar{L} H e_R + \text{c.c.} \tag{7.1}$$

Thus, after EWSB, ζ mixes with the Higgs boson h. To construct an EFT from the Lagrangian in the broken phase, we must integrate out the mass eigenstates corresponding to (ζ, h), which will always produce a factor of y_e in addition to the inverse of the mediator mass scale.

Such a factor of y_e in the EFT is also expected on general grounds if minimal flavor violation is assumed, regardless of the nature of the mediator. However, in general, one can also write UV completions which do not generate a factor of y_e, e.g. by employing a vector mediator. Still other UV completions can be constructed to introduce other small coefficients besides y_e in the EFT. When we tabulate the EFT operators, to facilitate comparison with arbitrary UV completions, we do not normalize the operators with such any such factor. However, since a factor of y_e is well-motivated, we will give our results in a format that shows constraints both with and without a factor of y_e.

Finally, note that we ignore any renormalizable couplings between the DM and SM fields, assuming that all interactions are encoded in the EFT. Notice that no such operators exist in the fermionic case under our assumptions, since we take the DM to be a SM singlet, and the $\mathbb{Z}_2$ symmetry forbids the lepton portal operator $\phi L H$. In the scalar case, on the other hand, this is something we impose. However, as we will discuss

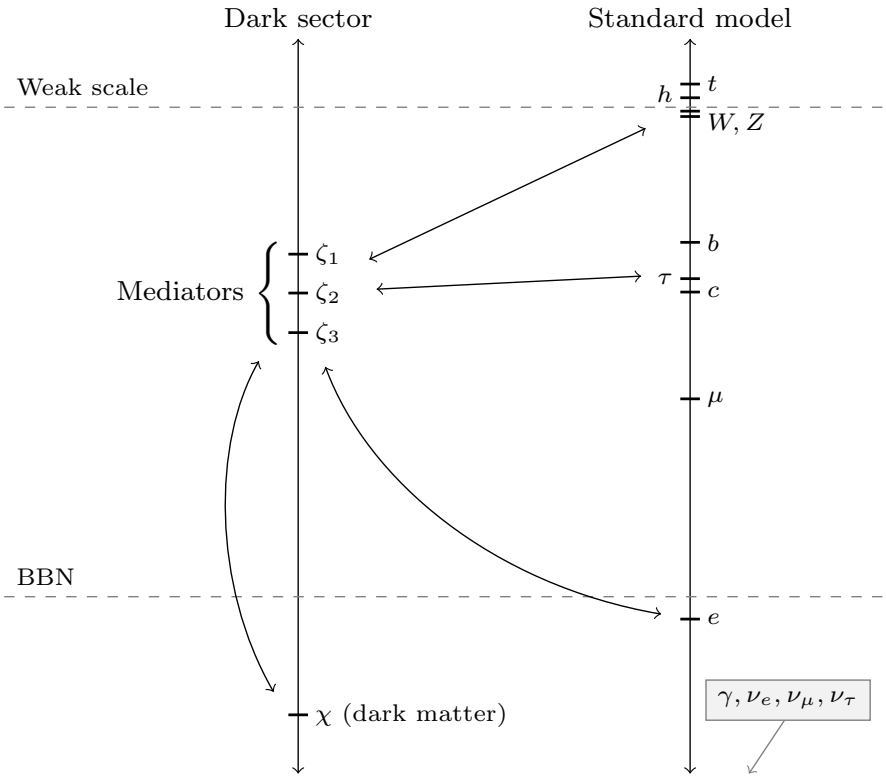

Figure 7.1: Schematic description of a UV completion of our effective theory. The vertical direction on the diagram corresponds to the mass scale. Arrows denote renormalizable couplings. Note that there is no renormalizable interaction between the DM and SM fields. The line labeled "BBN" corresponds to the scale of big bang nucleosynthesis, $T \sim 1\,\mathrm{MeV}$. Our results are unchanged if $m_{\zeta_i} > m_{\mathrm{weak}}$.

shortly, this assumption has no consequences for the results of our analysis.

At energies at or below the scale of BBN, the effective Lagrangian schematically reads

$$\mathcal{L}_{\mathrm{EFT}} = \mathcal{L}_{\mathrm{SM}} + \mathcal{L}_{\mathrm{DM}} + \sum_{d>4,\alpha} \frac{c_\alpha}{\Lambda_{\mathrm{EFT}}^{d-4}} \mathcal{O}_\alpha. \tag{7.2}$$

Here Λ_{EFT} is the mass scale associated with the EFT, which reflects the scale of the heavy degrees of freedom in the theory; $\mathcal{L}_{\mathrm{SM}}$ is the SM Lagrangian with only the $e^\pm$, ν, and γ fields; and $\mathcal{L}_{\mathrm{DM}}$ is the DM free theory contribution. The form of $\mathcal{L}_{\mathrm{DM}}$ depends

on whether the DM is a scalar ϕ or a fermion ψ. If the DM is a scalar, then

$$\mathcal{L}_{\rm DM} = \mathcal{L}_\phi = \begin{cases} \frac{1}{2}\partial^\mu\phi\partial_\mu\phi - \frac{1}{2}m_\phi^2\phi^2 & \text{real scalar} \\ (\partial^\mu\phi)^\dagger(\partial_\mu\phi) - m_\phi^2\phi^\dagger\phi & \text{complex scalar,} \end{cases} \tag{7.3}$$

and if the DM is a fermion, then

$$\mathcal{L}_{\rm DM} = \mathcal{L}_\psi = \begin{cases} \frac{1}{2}\bar{\psi}i\partial\!\!\!/\psi - \frac{1}{2}m_\psi\bar{\psi}\psi & \text{Majorana fermion} \\ \bar{\psi}i\partial\!\!\!/\psi - m_\psi\bar{\psi}\psi & \text{Dirac fermion.} \end{cases} \tag{7.4}$$

The remaining (infinite) sum over the higher-dimensional operators in Eq. (7.2) accounts for the effective interactions between DM and SM fields. In our analysis, we will retain terms up to dimension 6.

In the following subsection, we parametrize the interactions between DM and electrons. All operators consistent with a $\mathbb{Z}_2$ symmetry have the schematic form

$$\mathcal{O}^{(\chi)} \propto B_I(\chi)\,\bar{e}\,\Gamma^I e, \tag{7.5}$$

where the function $B_I(\chi)$ contains an even number of DM fields, and I denotes a set of Lorentz indices. We will eventually truncate all operators beyond dimension 6, so for our purposes, $B_I(\chi)$ always contains two DM fields. This DM bilinear is multiplied by an electron bilinear, for which the independent Dirac structures can be fully enumerated:

$$\Gamma^I \in \text{span}\left\{1,\, i\gamma^5,\, \gamma^\mu,\, \gamma^\mu\gamma^5,\, \sigma^{\mu\nu}\right\}. \tag{7.6}$$

If the electron bilinear is not a Lorentz scalar, the contraction of its free Lorentz indices with the ones of the DM bilinear ensures that the full operator in Eq. (7.5) is a Lorentz invariant. We now discuss the allowed operators for scalar and fermion DM.

7.2.2 EFT for scalar DM

To describe our EFT for scalar DM, we must enumerate all operators of the form

$$\mathcal{O}^{(\phi)} \propto B_I(\phi)\, \bar{e}\, \Gamma^I e \tag{7.7}$$

up to some mass dimension. Note that ϕ carries no Lorentz indices or spinor indices. Thus, if the index set I carried by the electron bilinear is non-empty, the only option is to insert derivatives in the scalar bilinear so that all indices are contracted.

A classification of all possible cases is provided in Table 7.1. Of the four resulting operators, two are dimension-5, while the other two include a derivative and are dimension-6. We use the notation

$$\phi^\dagger \overset{\leftrightarrow}{\partial}_\mu \phi \equiv \phi^\dagger \partial_\mu \phi - (\partial_\mu \phi^\dagger)\phi. \tag{7.8}$$

Note that we omit the operator $\left(\partial_\mu \phi^\dagger\, \phi + \phi^\dagger\, \partial_\mu \phi\right) \bar{e}\gamma^\mu e$, since it vanishes under integration by parts and application of the equation of motion:

$$\int \mathrm{d}^4x\, \left(\partial_\mu \phi^\dagger\, \phi + \phi^\dagger\, \partial_\mu \phi\right) \bar{e}\gamma^\mu e = -\int \mathrm{d}^4x\, \phi^\dagger \phi\, \partial_\mu \left(\bar{e}\gamma^\mu e\right) = 0. \tag{7.9}$$

Similarly, the operator $\left(\partial_\mu \phi^\dagger\, \phi + \phi^\dagger\, \partial_\mu \phi\right) \bar{e}\gamma^\mu\gamma^5 e$ is redundant: integrating by parts again, we obtain

$$\int \mathrm{d}^4x\, \partial_\mu\big(\phi^\dagger\phi\big)\bar{e}\gamma^\mu\gamma^5 e = -\int \mathrm{d}^4x\, \phi^\dagger\phi\, \partial_\mu \left(\bar{e}\gamma^\mu\gamma^5 e\right) \tag{7.10}$$

$$= -2im_e \int \mathrm{d}^4x\, \phi^\dagger\phi\, \bar{e}\gamma^5 e. \tag{7.11}$$

The resulting integrand is proportional to $\mathcal{O}_P^{(\phi)}$, one of the other operators in our basis. Moreover, this contribution is dimension-6 while $\mathcal{O}_P^{(\phi)}$ is dimension-5, so it is suppressed in the Lagrangian with an additional factor of $\Lambda_{\mathrm{EFT}}^{-1}$.

In some cases, renormalizable operators are allowed, and might appear in addition to the effective operators discussed above. For instance, in the context of a Higgs portal model [see e.g. 403] the operator $\phi^\dagger\phi\, H^\dagger H$ is allowed without affecting DM stability. After electroweak symmetry breaking (EWSB), this operator produces a cubic coupling $\phi^\dagger\phi\, v\, h$. Integrating out the SM Higgs boson generates an effective operator proportional to $\mathcal{O}_S^{(\phi)}$. Thus, adding renormalizable couplings does not introduce any new physical effects in our analysis. The only effect is to add a correction to the Wilson coefficient of a single operator, with a size typically smaller than the values we consider in our analysis.

At a qualitative level, we can guess at the relative prospects for direct detection in the case of each operator in Table 7.1. The operator $\mathcal{O}_S^{(\phi)}$ is easily generated by integrating out a scalar mediator, so we can expect that the relative strength of constraints and detection prospects for this operator will be comparable to results found in the context of simplified models with a scalar mediator [390]. Unlike $\mathcal{O}_S^{(\phi)}$, the other operators for scalar DM are suppressed by their momemtum dependence in the nonrelativistic limit, relevant for scattering. Each of these operators vanishes as the velocity and momentum transfer are taken to zero. Thus, for scalar dark matter, we expect from the outset that none of our operators will improve on the detection prospects of a simplified model with a scalar mediator, and we will indeed confirm these suspicions in the following sections.

With the effective operators in the scalar case now enumerated, we can consider annihilation and scattering processes for each one. Matrix elements for $2 \to 2$ annihilation and scattering are given in Table 7.3. The corresponding cross sections are

Symbol	Operator	Real case
$\mathcal{O}_S^{(\phi)}$	$g\Lambda_{\rm EFT}^{-1}\phi^\dagger\phi\,\bar{e}e$	Yes
$\mathcal{O}_P^{(\phi)}$	$ig\Lambda_{\rm EFT}^{-1}\phi^\dagger\phi\,\bar{e}\gamma^5 e$	Yes
$\mathcal{O}_V^{(\phi)}$	$ig\Lambda_{\rm EFT}^{-2}\phi^\dagger\overleftrightarrow{\partial}_\mu\phi\,\bar{e}\gamma^\mu e$	No
$\mathcal{O}_A^{(\phi)}$	$ig\Lambda_{\rm EFT}^{-2}\phi^\dagger\overleftrightarrow{\partial}_\mu\phi\,\bar{e}\gamma^\mu\gamma^5 e$	No

Table 7.1: Operators coupling the electron to a dark scalar ϕ. The third column indicates whether or not the operator survives when ϕ is taken to be a real scalar.

given in Tables 7.4 and 7.5.

7.2.3 EFT for fermion DM

If the DM is a fermion ψ, the structure of the EFT is similar to the scalar case. We again have a set of operators which are products of an electron bilinear and a ψ bilinear. Using generalized Fierz identities, it can be shown that operators of the form $(\bar{\psi}\mathcal{O}_1 e)(\bar{e}\mathcal{O}_2\psi)$ are redundant, in that they can be written as linear combinations of operators of the form $(\bar{\psi}\mathcal{O}_1'\psi)(\bar{e}\mathcal{O}_2' e)$ [404]. Thus, we can construct a complete basis of effective operators by enumerating the possible insertions $\mathcal{O}_1'$ and $\mathcal{O}_2'$. All of the electron bilinears from the scalar case appear here as well, and most of the possible ψ bilinears are obtained from these by making the replacement $e \to \psi$.

In addition to these bilinears, we can form a spin-2 current at dimension 6, e.g. of the form $\bar{\psi}\sigma_{\mu\nu}\psi$. Since $\sigma_{\mu\nu}$ is antisymmetric, the other bilinear must not be symmetric in its Lorentz indices, so it must contain another insertion of $\sigma_{\mu\nu}$. Thus, such an operator has the general form $W_{\mu\nu\alpha\beta}\bar{\psi}\sigma^{\mu\nu}\psi\bar{e}\sigma^{\alpha\beta}e$. At dimension 6, the indices

of $W_{\mu\nu\alpha\beta}$ can come only from two factors of the metric or one factor of the Levi-Civita symbol ε. In the former case, again due to antisymmetry of $\sigma^{\mu\nu}$, the only nontrivial contraction is

$$g_{\mu\alpha}g_{\nu\beta}\bar{\psi}\sigma^{\mu\nu}\psi\bar{e}\sigma^{\alpha\beta}e. \tag{7.12}$$

If W is instead formed from the Levi-Civita symbol, then the operator has the form $\varepsilon_{\rho_1\rho_2\rho_3\rho_4}\bar{\psi}\sigma^{\mu\nu}\psi\bar{e}\sigma^{\alpha\beta}e$, where $(\rho_1, \rho_2, \rho_3, \rho_4)$ is a permutation of $(\mu, \nu, \alpha, \beta)$. Up to an overall sign, the indices ρ_i can be rearranged into the latter order, so all such operators are proportional to

$$\bar{\psi}\sigma^{\mu\nu}\psi\bar{e}\left(\varepsilon_{\mu\nu\alpha\beta}\sigma^{\alpha\beta}\right)e. \tag{7.13}$$

But $\varepsilon_{\mu\nu\alpha\beta}\sigma^{\alpha\beta} = -2i\sigma_{\mu\nu}\gamma^5$, so if we simply add $i\sigma_{\mu\nu}\gamma^5$ to our list of insertions, we can assume that $W_{\mu\nu\alpha\beta}$ is a product of metric tensors. (We retain the factor of i to preserve Hermiticity.) Further, the argument above demonstrates that it is sufficient to place this insertion in only one of the two bilinears: the operator formed by inserting $i\sigma_{\mu\nu}\gamma^5$ in both bilinears is redundant. We choose to place this insertion in the electron bilinear.

The complete list of operators for fermionic DM is shown in Table 7.2. Matrix elements for $2 \to 2$ annihilation and scattering are given in Table 7.6. The corresponding cross sections are given in Tables 7.7 and 7.8.

As in the scalar case, we estimate relative prospects for direct detection among the operators in Table 7.2. The operator $\mathcal{O}_{SS}^{(\psi)}$, like $\mathcal{O}_{S}^{(\phi)}$, is naturally generated by simplified models with a scalar mediator. While many of the other operators are momentum-suppressed in the non-relativistic limit, as in the case of scalar DM, the operators $\mathcal{O}_{VV}^{(\psi)}$, $\mathcal{O}_{AA}^{(\psi)}$, and $\mathcal{O}_{TT}^{(\psi)}$ are not. These operators may be expected to compete with or exceed

Symbol	Operator	Maj.	Symbol	Operator	Maj.
$\mathcal{O}_{SS}^{(\psi)}$	$g\Lambda_{\rm EFT}^{-2}\bar{\psi}\psi\,\bar{e}e$	Yes	$\mathcal{O}_{PS}^{(\psi)}$	$ig\Lambda_{\rm EFT}^{-2}\bar{\psi}\gamma^5\psi\,\bar{e}e$	Yes
$\mathcal{O}_{SP}^{(\psi)}$	$ig\Lambda_{\rm EFT}^{-2}\bar{\psi}\psi\,\bar{e}\gamma^5 e$		$\mathcal{O}_{PP}^{(\psi)}$	$g\Lambda_{\rm EFT}^{-2}\bar{\psi}\gamma^5\psi\,\bar{e}\gamma^5 e$	
$\mathcal{O}_{VV}^{(\psi)}$	$g\Lambda_{\rm EFT}^{-2}\bar{\psi}\gamma_\mu\psi\,\bar{e}\gamma^\mu e$	No	$\mathcal{O}_{AV}^{(\psi)}$	$g\Lambda_{\rm EFT}^{-2}\bar{\psi}\gamma_\mu\gamma^5\psi\,\bar{e}\gamma^\mu e$	Yes
$\mathcal{O}_{VA}^{(\psi)}$	$g\Lambda_{\rm EFT}^{-2}\bar{\psi}\gamma_\mu\psi\,\bar{e}\gamma^\mu\gamma^5 e$		$\mathcal{O}_{AA}^{(\psi)}$	$g\Lambda_{\rm EFT}^{-2}\bar{\psi}\gamma_\mu\gamma^5\psi\,\bar{e}\gamma^\mu\gamma^5 e$	
$\mathcal{O}_{TT}^{(\psi)}$	$\frac{1}{2}g\Lambda_{\rm EFT}^{-2}\bar{\psi}\sigma_{\mu\nu}\psi\,\bar{e}\sigma^{\mu\nu}e$	No	$\mathcal{O}_{T\tilde{T}}^{(\psi)}$	$\frac{i}{2}g\Lambda_{\rm EFT}^{-2}\bar{\psi}\sigma_{\mu\nu}\psi\,\bar{e}\sigma^{\mu\nu}\gamma^5 e$	No

Table 7.2: Operators coupling the electron to a dark fermion ψ. The third column in each half of the table indicates whether or not the operator survives when ψ is taken to be a Majorana fermion.

the detection prospects associated with $\mathcal{O}_{SS}^{(\psi)}$, an expectation that we will confirm in our analysis.

7.3 Cosmological constraints

Cosmological constraints on DM are typically model-dependent. However, the broad class of models which we consider admits only a very restricted set of thermal histories for the DM species, which allows us to derive general cosmological constraints in the context of our EFT.

We divide the thermal histories into two cases: either the DM is in thermal equilibrium with the SM at high temperatures, and freezes out below some temperature; or it never attains thermal equilibrium, and the abundance is instead set non-thermally. It is possible that the dark species only enters equilibrium at late times, but this scenario mirrors the thermal freeze-out case in almost every respect.

In the freeze-out scenario, two constraints are particularly robust: first, if the DM is thermalized and relativistic during the epoch of big bang nucleosynthesis (BBN), its effect on the Hubble parameter is generally sufficient to perturb light elemental abundances [388]. Second, if at some temperature the DM is in thermal equilibrium with electrons and not with neutrinos, or vice versa, then entropy can be transferred from the DM to neutrinos alone or to electrons and photons alone. This changes the temperature ratio of the two thermal baths, which modifies the effective number of neutrino species, $N_{\rm eff}$, as determined from the cosmic microwave background (CMB) [389].

Finally, in the case of out-of-equilibrium (non-thermal) production, the DM never attains thermal equilibrium, and so may evade these two constraints. However, if the coupling to electrons is too large, DM will be overproduced even under the most generous assumptions.

Note that new light species are also subject to constraints from energy loss in stars and supernovae [342]. However, these constraints rely on complicated microphysical inputs that must be computed in detail for each model. Moreover, supernova temperatures lie up to an order of magnitude above the scale of BBN, requiring our effective theory to be valid at higher energies. Thus, we do not evaluate these constraints explicitly, but simplistic estimates suggest that they are at best comparable in strength to our cosmological constraints over the mass range of interest.

We now examine each of our constraints in more detail.

7.3.1 Freeze-out and primordial nucleosynthesis

Light element abundances today are a sensitive probe of cosmology at scales near 1 MeV. If an additional light species is assumed to be in thermal equilibrium at these scales, the standard predictions of big bang nucleosynthesis (BBN) are modified, with observable consequences. Since thermal equilibrium in turn depends on DM interactions, light element abundances translate to stringent constraints on the interaction rates.

In a broad class of models, the DM species is in thermal equilibrium with the SM bath at high temperatures, and eventually drops out of equilibrium below some freeze-out temperature, $T_{\rm FO}$. In our framework, freeze-out is a generic requirement of any scenario in which DM is in thermal equilibrium with electrons at temperatures $T \lesssim 1\,\text{MeV}$, since the EFT is valid in this regime.

If the DM species freezes out during or after BBN, and the DM species is in equilibrium at higher temperatures, then the predictions of light element abundances are generally perturbed to a degree incompatible with their measured values [386–388, 405]. The ratios of these abundances are set by the temperatures at which interconversion processes freeze out, which depend in turn on the Hubble parameter H. Since H is sensitive to the energy density, adding a new species that stays in equilibrium and remains relativistic for much of the epoch of BBN has a significant impact on the produced light element abundances. Note that in a small range of our parameter space, equilibrium during BBN is consistent with observables if the dark species *enters* equilibrium at a specific time during BBN [394]. This is a very narrow exception to our framework, so

we neglect it for the remainder of this chapter.

The temperature at which freeze-out occurs is fixed by the DM mass and the couplings. The prospect of experimental detection by any particular apparatus places a lower bound on the scattering cross section $\chi e^- \to \chi e^-$. However, for a given interaction, the scattering cross section is directly related to the annihilation cross section $\chi\chi \to e^+e^-$ which regulates the thermodynamics of the DM species in the early universe. A lower bound on the scattering cross section thus corresponds to a lower bound on the annihilation cross section, which translates to an upper bound on the freeze-out temperature.

For our purposes, we will only consider a model to be ruled out by light element abundances if it predicts that DM is in equilibrium at $T = 1\,\mathrm{MeV}$. This choice of threshold temperature is slightly different from some other treatments of BBN constraints in the literature. In particular, [389] find that sub-MeV DM is generally ruled out by elemental abundances if the DM is in equilibrium after neutrinos decouple at $2.3\,\mathrm{MeV}$. However, these constraints assume that the DM is in equilibrium with only one of electrons and neutrinos, and not both, so that the temperature ratio T_ν/T_γ is modified. We will discuss this scenario in detail in the following section, but for the moment, we note that our EFT accommodates equilibrium with both electrons and neutrinos, with decouplings taking place at different temperatures. In such situations, constraints from T_ν/T_γ can potentially be relaxed in some areas of the parameter space. Thus, there is not necessarily any connection between neutrino decoupling and BBN constraints in our model.

Given more detailed information about the dark sector and its couplings to

the SM, it is possible that BBN could place constraints on DM which decouples at even higher temperatures. Between $T \sim 10\,\mathrm{MeV}$ and $T = 1\,\mathrm{MeV}$, no SM species become non-relativistic, so the SM bath is not heated relative to a decoupled dark sector. Thus, even if the DM decouples from the SM bath at 10 MeV or above, it is possible that $T_\chi = T_\gamma$ during BBN, in which case sub-MeV DM will typically disrupt BBN. Additionally, if DM is in equilibrium with only one of neutrinos and electrons after neutrino decoupling takes place, then the constraints of [389] do apply.

We wish to place conservative constraints that are independent of these details, and also independent of cosmological modifications at $T \gg 1\,\mathrm{MeV}$ that might occur outside the context of our DM model. We regard 1 MeV as a reasonable fiducial threshold for assessing BBN constraints. However, while it is possible to avoid the constraints of [389] in our model, this takes additional tuning. Thus, we will give two versions of the BBN constraint: one with a threshold of 1 MeV, and another with a threshold of 2.3 MeV, corresponding to the constraint of [389]. This also serves to demonstrate the sensitivity of our constraints to higher thresholds.

The freeze-out temperature and relic density for a given model are found by solving the Boltzmann equation in a relatively simple incarnation. In our framework, we have only a single DM species χ which interacts with electrons exclusively through $2 \to 2$ processes. For this case, using Maxwell–Boltzmann statistics, the Boltzmann equation takes the form

$$\frac{x}{Y_{\mathrm{eq}}}\frac{\mathrm{d}Y}{\mathrm{d}x} = -\frac{n_{\mathrm{eq}}(x)\,\langle\sigma\,|v|\rangle\,(x)}{H(x)}\left(\left(\frac{Y(x)}{Y_{\mathrm{eq}}(x)}\right)^2 - 1\right), \tag{7.14}$$

where $x \equiv m_\chi/T$ parametrizes cosmic time; σ is the cross section for $\bar{\chi}\chi \to e^+e^-$;

$Y \equiv n/s$ is the abundance of χ, where n is the number density and s the entropy density of χ; and $Y_{\rm eq}$ and $n_{\rm eq}$ are the equilibrium abundance and number density of χ, respectively. We identify $\Gamma_A \equiv n_{\rm eq} \langle \sigma |v| \rangle$ as the annihilation rate of χ when in equilibrium. The thermally-averaged cross section can be obtained as [352]

$$\langle \sigma |v| \rangle = \frac{\int_{s_{\rm min}}^{\infty} {\rm d}s \left(s - 4m_\chi^2\right) \sqrt{s} \sigma K_1 \left(\sqrt{s}/T\right)}{8 m_\chi^4 T K_2 (m_\chi/T)^2}. \tag{7.15}$$

It is clear from Eq. (7.14) that the abundance will stabilize once $\Gamma_A/H \lesssim 1$. This condition gives an estimate of the temperature $T_{\rm FO}$ at which χ departs from equilibrium, and thus allows us to test whether a set of parameter values is consistent with BBN observables.

In particular, we can immediately estimate the impact of changing the threshold used for assessing BBN constraints. Since the DM is relativistic at decoupling, the freeze-out temperature can be estimated by the relation $T^3 \langle \sigma |v| \rangle \sim T^2/M_{\rm Pl}$, where σ is the DM annihilation cross section. For our operators, the cross sections scale like $s/\Lambda_{\rm EFT}^4$ or $1/\Lambda_{\rm EFT}^2$, so if we adjust $T_{\rm FO}$ and determine the corresponding value of $\Lambda_{\rm EFT}$, then $\Lambda_{\rm EFT}$ is approximately proportional to $T_{\rm FO}^{3/4}$ or $T_{\rm FO}^{1/2}$. In particular, we expect the difference between the 1 MeV threshold and the 2.3 MeV threshold to correspond to a $\mathcal{O}(1)$ factor in the constraint on $\Lambda_{\rm EFT}$.

In general, when studying the decoupling of χ, it is important to consider the coupling to neutrinos as well as electrons. If χ has a non-negligible coupling to neutrinos, it is conceivable that the DM could be kept in equilibrium at later times via thermal contact with the neutrino bath, which would tend to strengthen our constraints. However, the coupling to neutrinos can always be set to zero independent of the coupling

to electrons: we assume χ couples to the neutrino only via the $\mathrm{SU}(2)_L$ doublet, and χ can couple independently to e_R and to e_L. Thus, when evaluating BBN constraints, we ignore thermal contact with neutrinos in order to obtain the most conservative limits.

7.3.2 Effective number of neutrinos in CMB

Another powerful constraint applicable to a new light species is the effective number of neutrino species, N_{eff}, as measured from CMB. To establish constraints with the greatest possible generality, we evaluate bounds from the CMB without regard to the BBN constraints. As we will show, the bounds from BBN and the CMB are comparable in reach, but imposing each independently means that exceptional cases that escape one bound or the other can still be constrained.

N_{eff} characterizes the contributions to the radiation energy density at recombination from relativistic species apart from photons, and is defined by

$$\frac{\rho_{\mathrm{rad}}}{\rho_\gamma} \equiv 1 + \frac{7}{8}\left(\frac{4}{11}\right)^{4/3} N_{\mathrm{eff}}. \tag{7.16}$$

In the absence of any other relativistic species, $N_{\mathrm{eff}} \simeq 3$. The SM actually predicts $N_{\mathrm{eff}} = 3.046$, accounting for the three neutrino species and for small effects due to non-idealities in the decoupling process [406, 407]. This is consistent with analyses of Planck data, which find $N_{\mathrm{eff}} \simeq 3.1 \pm 0.2$ [408]. Additional species are strongly disfavored. A single additional relativistic degree of freedom, i.e., a real scalar, is weakly consistent with current limits. However, CMB stage 4 experiments are expected to measure $\Delta N_{\mathrm{eff}} \equiv N_{\mathrm{eff}} - 3.046$ to within ± 0.03, which is just sensitive enough to probe the minimum contribution from a real scalar at 1σ [409].

But a new species need not be relativistic at recombination to alter $N_{\rm eff}$. The introduction of a light DM species can change $N_{\rm eff}$ by modifying the ratio of the photon and neutrino temperatures [385, 389, 410, 411], and hence the ratio of energy densities in Eq. (7.16). In the absence of additional species, the chemical decoupling of electrons and neutrinos takes place at $T_D^0 \approx 2.3\,\mathrm{MeV}$ [412]. Any entropy transferred from DM to electrons after this decoupling leads to heating of the photon bath, and any entropy transferred to neutrinos heats the neutrino bath. If the new species transfers entropy differentially to the photon and neutrino baths at any time after the two baths decouple, the temperature ratio of the baths is modified. Note that $\Delta N_{\rm eff}$ thus depends on the relative size of the couplings to electrons and neutrinos, as pointed out in [410] and detailed extensively in [413].

Typically, the DM will transfer its entropy to one or both baths as a consequence of the conservation of comoving entropy density: when the DM becomes non-relativistic while still in thermal equilibrium, the associated entropy must be transferred to any relativistic species to which it is still coupled. Thus, these species are heated when the DM becomes non-relativistic. Now, suppose that a sub-MeV DM species is coupled to electrons and neutrinos when $T < T_D^0$. If the DM species decouples from one and only one of these two relativistic species before it becomes non-relativistic itself, then the DM will reheat only one of the two baths, changing the temperature ratio. An exception to this rule occurs when the DM *enters* equilibrium with one bath below T_D^0, so that the DM accepts entropy of the same order that it loses upon decoupling later on [391]. We will discuss this scenario further in Section 7.5.

We now examine the calculation of $N_{\rm eff}$ in detail. We will write T_{XY} to denote

the temperature at which species X and Y lose *direct* thermal contact, i.e., the temperature below which $\Gamma(X \leftrightarrow Y)/H < 1$ in our effective theory. The species X and Y might be kept in thermal equilibrium by a third species Z in our framework, i.e., through processes $X \leftrightarrow Z$ and $Z \leftrightarrow Y$ that remain active. We define T_D to be the actual temperature at which electrons and neutrinos drop out of thermal equilibrium with one another once all inter-conversion processes have frozen out, including multi-step processes involving the dark species. Thus, in the standard scenario, $T_D = T_{e\nu} \equiv T_D^0 \approx 2.3\,\mathrm{MeV}$, but the introduction of a new species can keep electrons and neutrinos in thermal equilibrium at lower temperatures.

In particular, suppose that DM decouples from electrons instantaneously at a temperature $T_{\chi e}$, and from neutrinos at a temperature $T_{\chi\nu}$. If $T_D^0 < \min\{T_{\chi e}, T_{\chi\nu}\}$, then any entropy transferred to either photons or neutrinos can be shared between the two, so DM reheats these species equally, and the standard calculation is unchanged. However, if $T_{\chi e} < T_D^0 < T_{\chi\nu}$, then χ remains in thermal contact with photons while relativistic, reheating the photon bath but not the neutrino bath. This increases the photon temperature, *reducing* N_{eff}. Similarly, if $T_{\chi\nu} < T_D^0 < T_{\chi e}$, then the reverse is true: DM reheats the neutrino bath, and N_{eff} increases.

The only other possibility is $\max\{T_{\chi e}, T_{\chi\nu}\} < T_D^0$, in which case χ acts as a thermodynamic mediator between electrons and neutrinos below T_D^0. In this situation, electrons and neutrinos remain in thermal equilibrium until the temperature falls below $T_D = \max\{T_{\chi e}, T_{\chi\nu}\}$. If the electron is still relativistic throughout this process, then the impact on N_{eff} is determined by the ordering of $T_{\chi e}$ and $T_{\chi\nu}$. But if $T_D \lesssim m_e$, the impact on N_{eff} is quite different: photons and neutrinos are still in thermal contact

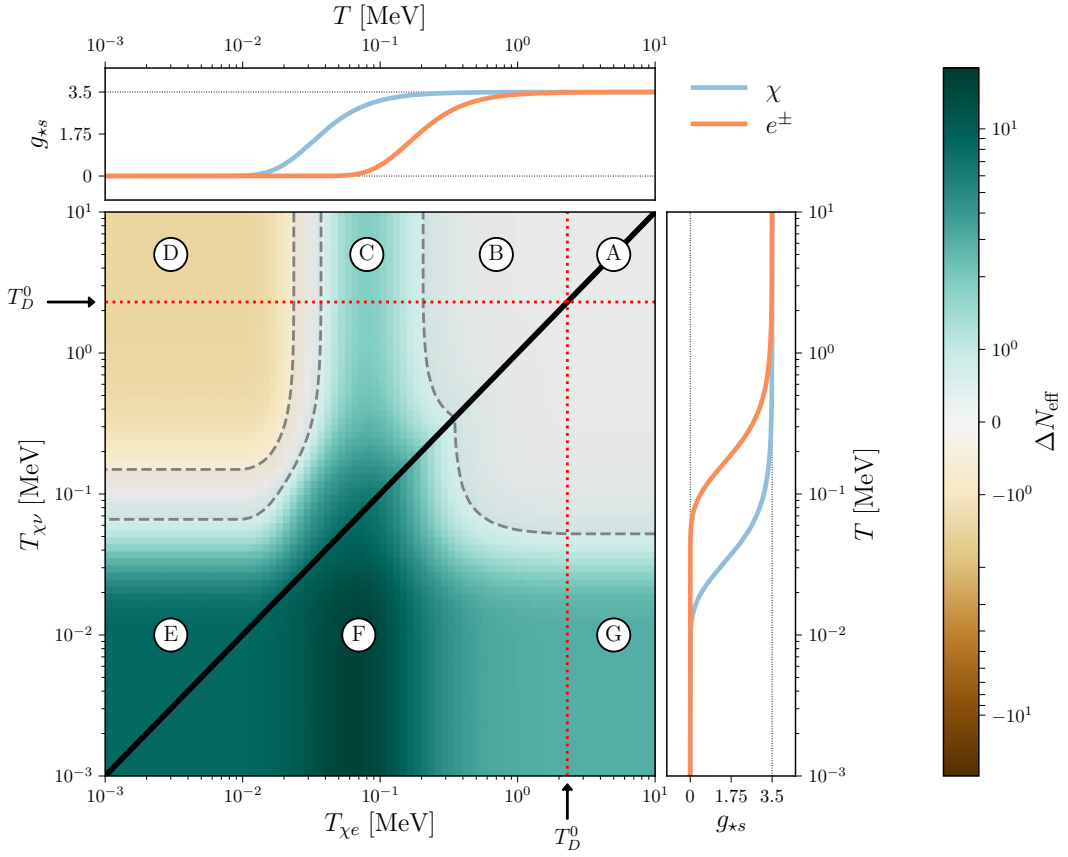

Figure 7.2: $\Delta N_{\rm eff}$ as a function of the two decoupling temperatures $T_{\chi e}$ and $T_{\chi\nu}$, assuming that χ is a Dirac fermion with mass 100 keV. Side and top panels show entropic degrees of freedom as a function of temperature. Gray shaded area indicates the region consistent with current data at 2σ. Labeled regions can be understood qualitatively as follows. *Region A:* $T_{\chi e}, T_{\chi\nu} > T_D^0$. Thus any entropy transferred by χ is shared between the γ and ν baths before they decouple. The standard calculation of $N_{\rm eff}$ is unaltered. *Region B:* $T_{\chi e} < T_D^0 < T_{\chi\nu}$. However, χ and $e^{\pm}$ are relativistic at both decoupling events, so little entropy is transferred to either the γ or the ν bath. *Region C:* Now $e^{\pm}$ becomes non-relativistic while still in thermal contact with the relativistic χ. The entropy ordinarily transferred by $e^{\pm}$ to γ is now shared with χ, so γ is reheated less efficiently, and $N_{\rm eff}$ increases. *Region D:* Here χ is relativistic below both T_D^0 and $T_{\chi\nu}$, but becomes non-relativistic before $T_{\chi e}$ is reached. Thus, χ reheats the γ bath exclusively upon becoming non-relativistic, decreasing $N_{\rm eff}$. *Region E:* χ becomes non-relativistic above both $T_{\chi\nu}$ and $T_{\chi e}$, so it reheats both baths. The impact on $N_{\rm eff}$ in this region comes from the delayed $e^{\pm}$–ν decoupling (see text). *Region F:* $T_{\chi e} > T_{\chi\nu}$, and χ is relativistic at $T_{\chi e}$. Thus, in addition to the delayed $e^{\pm}$–ν decoupling, χ reheats the ν bath. *Region G:* The electron and χ are relativistic at $T_{\chi e}$, so here the impact on $N_{\rm eff}$ is due to χ reheating the ν bath.

while electrons become non-relativistic, so the *electron* also transfers some of its entropy to the neutrino bath. As we will see shortly, this can have a dramatic impact on $N_{\rm eff}$.

To calculate $N_{\rm eff}$, we follow the procedure described in [414]. In our scenario, the DM species is non-relativistic at recombination, so we assume that $N_{\rm eff}$ is not modified by any additional degrees of freedom *at recombination*. Then, given the temperature ratio of the neutrino and photon baths at recombination, $N_{\rm eff}$ is given by

$$N_{\rm eff} = \left(\frac{4}{11}\right)^{-4/3} \left(\left.\frac{T_\nu}{T_\gamma}\right|_{\rm rec}\right)^4 N_\nu, \tag{7.17}$$

where N_ν is the number of SM neutrinos (3). In turn, we can determine the temperature ratio from conservation of comoving entropy density.

Recall that the entropy density of a relativistic bosonic species i with g_i internal degrees of freedom is given by $2\pi^2 g_i T^3/45$. Away from the relativistic limit, denoting the true entropy density by s_i, we say that this species has $g_{\star s} \equiv s_i/(2\pi^2 T^3/45)$ entropic degrees of freedom. Now, let $g_{\star s}^{(\gamma)}$ and $g_{\star s}^{(\nu)}$ denote the entropic degrees of freedom in equilibrium with photons and neutrinos, respectively. Then $g_{\star s}^{(\alpha)}$ is given explicitly by

$$g_{\star s}^{(\alpha)} = \sum_{i\in I} \frac{15 g_i}{4\pi^4} \int_{x_i}^\infty {\rm d}u \, \frac{\left[4u^2 - x_i^2\right]\left[u^2 - x_i^2\right]^{1/2}}{\exp(u) \pm 1}, \tag{7.18}$$

where $x_i = m_i/T_\alpha$, and I indexes all species in equilibrium with species α (γ or ν). The sign in the denominator is determined by the statistics of species i. It can be shown [414] that if no entropy leaves the photon or neutrino baths after they decouple, then

$$\left.\frac{T_\nu}{T_\gamma}\right|_{\rm rec} = \left(\left.\frac{g_{\star s}^{(\nu)}}{g_{\star s}^{(\gamma)}}\right|_{T_D} \left.\frac{g_{\star s}^{(\gamma)}}{g_{\star s}^{(\nu)}}\right|_{\rm rec} \right)^{1/3}. \tag{7.19}$$

However, in our scenario, it is possible for entropy to leave one of the two baths below T_D: suppose the DM decouples from one of the two baths above T_D, and decouples from

the other below T_D, but while still relativistic. At this second decoupling, the DM's remaining entropy leaves the bath to which it was last coupled. This only happens if $T_{\chi e} < T_D \leq T_{\chi\nu}$ or $T_{\chi\nu} < T_D \leq T_{\chi e}$.

To account for this possibility, we modify the calculation of the temperature ratio as follows. Let us assume for the moment that $T_{\chi e} < T_D \leq T_{\chi\nu}$. Conservation of comoving entropy density in a thermal bath α amounts to the assertion that $g_{\star s}^{(\alpha)}|_T T^3 a^3$ is constant, where a is the scale factor. For $T < T_D$, comoving entropy density is conserved in each bath except when $T_\gamma = T_{\chi e}$, so the temperatures of the two baths satisfy

$$\begin{aligned} T_\nu &= k_1 a^{-1} g_{\star s}^{(\nu)}|_{T_\nu}^{-1/3}, \\ T_\gamma &= \begin{cases} k_2 a^{-1} g_{\star s}^{(\gamma)}|_{T_\gamma}^{-1/3} & T_{\chi e} < T_\gamma < T_D \\ k_3 a^{-1} g_{\star s}^{(\gamma)}|_{T_\gamma}^{-1/3} & T_\gamma < T_{\chi e}, \end{cases} \end{aligned} \tag{7.20}$$

where the k_i are constants. Generally, $T_{\rm rec} < T_{\chi e}$, so

$$\left.\frac{T_\nu}{T_\gamma}\right|_{\rm rec} = \frac{k_1}{k_3}\left(g_{\star s}^{(\nu)} \Big/ \left. g_{\star s}^{(\gamma)}\right|_{\rm rec}\right)^{-1/3}. \tag{7.21}$$

Thus, to determine the temperature ratio, it is sufficient to identify the ratio k_1/k_3, which can be done in two stages. First, since T_ν and T_γ are equal at T_D, we must have

$$\frac{k_1}{k_2} = \left(g_{\star s}^{(\nu)} \Big/ \left. g_{\star s}^{(\gamma)}\right|_{T_D}\right)^{1/3}. \tag{7.22}$$

Similarly, at $T_{\chi e}$, $g_{\star s}^{(\gamma)}$ changes discontinuously while T_γ is continuous in a. Thus, k_3 must satisfy

$$\frac{k_3}{k_2} = \left(\frac{g_{\star s}^{(\gamma)}|_{T_{\chi e}^-}}{g_{\star s}^{(\gamma)}|_{T_{\chi e}^+}}\right)^{1/3}, \tag{7.23}$$

where $T_{\chi e}^{\pm}$ denotes a temperature just above or below $T_{\chi e}$. Now we have

$$\left.\frac{T_\nu}{T_\gamma}\right|_{\rm rec} = \left(\left.\frac{g_{\star s}^{(\nu)}}{g_{\star s}^{(\gamma)}}\right|_{T_D} \frac{g_{\star s}^{(\gamma)}|_{T_{\chi e}^+}}{g_{\star s}^{(\gamma)}|_{T_{\chi e}^-}} \left.\frac{g_{\star s}^{(\gamma)}}{g_{\star s}^{(\nu)}}\right|_{\rm rec} \right)^{1/3} . \tag{7.24}$$

A similar calculation applies if $T_{\chi\nu} < T_D \leq T_{\chi e}$. Note that Eq. (7.24) still assumes that χ does not *enter* equilibrium below T_D, an exception we discuss further in Section 7.5.

From Eq. (7.24), it is easy to see why low DM decoupling temperatures can have a large impact on $N_{\rm eff}$. In the standard scenario, $g_{\star s}^{(\gamma)}|_{T_D}$ includes photons (2) and relativistic electrons ($\frac{7}{8} \times 4$), which gives

$$\frac{g_{\star s}^{(\gamma)}|_{\rm rec}}{g_{\star s}^{(\gamma)}|_{T_D}} = \frac{2}{2 + \frac{7}{8} \times 4} = \frac{4}{11}. \tag{7.25}$$

But if neutrinos and photons remain in thermal contact after electrons become non-relativistic, then $g_{\star s}^{(\gamma)}|_{T_D}$ includes only photons, and the above ratio is increased to 1. This increases $N_{\rm eff}$ by a factor of $(11/4)^{4/3} \approx 3.9$, already leading to $N_{\rm eff} \approx 12$. If $T_{\chi e} < T_{\chi\nu}$, then χ reheats the photon bath when it becomes non-relativistic, reducing $N_{\rm eff}$. But if $T_{\chi\nu} < T_{\chi e}$, then χ reheats the neutrino bath, increasing $N_{\rm eff}$ even further. The impact of relative decoupling temperatures on $N_{\rm eff}$ is shown in Fig. 7.2.

This approach assumes that the decouplings take place instantaneously, which is generally a good approximation. However, the approximation is poor when the decoupling process overlaps the range of temperatures during which a species becomes non-relativistic. In this case, the entropy of the species is changing rapidly, so it is difficult to estimate the amount of entropy transferred to other relativistic species before decoupling is complete. The temperature ratio can be determined precisely by numerical methods [see e.g. 413, 415], and while that lies outside the scope of the present work, we note that instantaneous decoupling should be an effective approximation away

from a narrow range of temperatures $T_{\chi e}$ and $T_{\chi\nu}$, corresponding to a very small span of $\Lambda_{\rm EFT}$ values in our parameter space.

To translate these results into constraints on the coupling between χ and electrons, we must make an assumption about the coupling between χ and neutrinos. If the coupling to neutrinos is very small, then χ may maintain thermal contact with electrons after decoupling from neutrinos. On the other hand, if the coupling to neutrinos is very large, then χ may remain in thermal contact with neutrinos after decoupling from electrons. In our case, we will assume that χ couples to ν exclusively by coupling to the lepton doublet $(e_L, \nu_e)^{\rm T}$. That is, we will assume that the χ–ν coupling is the same as the χ–e_L coupling.

Even in this framework, the impact on $N_{\rm eff}$ depends on the relative strengths of the χ–e_L and χ–e_R couplings. A non-zero coupling to e_R tends to keep χ in equilibrium with electrons to lower temperatures, meaning that χ typically reheats the photon bath. This reduces the temperature ratio of Eq. (7.19), producing $\Delta N_{\rm eff} < 0$. However, if χ stays in equilibrium long enough to modify T_D, then we can obtain $\Delta N_{\rm eff} > 0$, as discussed above. Either way, increasing the coupling to e_R only strengthens the effect, so we neglect this coupling to obtain conservative constraints. Note that this is different from our assumption in evaluating BBN constraints, where conservative constraints are obtained by neglecting the coupling to e_L.

7.3.3 Non-thermal production

A viable model of DM must (partially) account for, but not exceed, the observed DM density of $\Omega_{\rm DM}h^2 \simeq 0.12$ [416]. If the DM is produced by thermal freeze-out,

then a larger annihilation cross section reduces the relic density, so larger couplings conducive to direct detection are less likely to overproduce DM. But in the alternative scenario, if DM is produced out of equilibrium, the relic density increases with the annihilation cross section. In this case, overproduction is an important consideration.

If the DM species never attains thermal equilibrium with the SM, the abundance of DM will evolve toward its equilibrium value, but once $\Gamma_A/H \lesssim 1$, the abundance will stay fixed. For renormalizable interactions, this out-of-equilibrium production process is the standard freeze-in mechanism [350]. Out-of-equilibrium production has also been studied for non-renormalizable operators in the context of so-called ultraviolet freeze-in [351]. For temperatures below $\sim 10\,\mathrm{MeV}$, within the constraints of our framework, such non-thermal production represents the only alternative to the freeze-out scenario.

The relic density of non-thermal DM is determined using the Boltzmann equation, much like the freeze-out case. The only difference is that the DM species χ is not in thermal equilibrium with $e^{\pm}$, and thus we cannot assume that χ has an equilibrium phase space density. Instead, we assume that the density of χ is negligible, such that the f_χ^2 term drops out of the Boltzmann equation. In other words, starting from Eq. (7.14), we approximate $Y/Y_{\mathrm{eq}} \simeq 0$, which gives $Y'(x) \simeq n_{\mathrm{eq}}(x) \langle\sigma|v|\rangle (x)/H(x)$. It follows that the out-of-equilibrium yield can be estimated as

$$Y(\infty) \simeq Y(x_{\mathrm{min}}) + \int_{x_{\mathrm{min}}}^{\infty} \mathrm{d}x \, \frac{n_{\mathrm{eq}}(x) \langle\sigma|v|\rangle (x)}{H(x)}. \tag{7.26}$$

As with freeze-out, the relic density in the non-thermal case is determined by the DM mass and couplings with SM particles. However, there is also a dependence on initial

conditions in the form of $x_{\rm min}$ and $Y(x_{\rm min})$. In the freeze-out scenario, the abundance of DM in the early universe is simply the equilibrium abundance: equilibrium effectively erases the initial condition. But in the non-thermal scenario, equilibrium is never attained, so the dependence on the initial abundance is retained. Typically, when DM is produced by SM annihilations out of equilibrium, one calculates the relic density by fixing the DM density to zero at very early times and evolving non-thermally. This procedure requires that the interactions considered are renormalizable, in order for the production process to be modeled consistently at very high temperatures. Our effective operators are non-renormalizable, so we cannot determine the relic density precisely in the non-thermal case: the result depends on the choice of UV completion.

However, we can still place a lower bound on the relic density. We require that our effective theory is valid at scales below $\sim 10\,\mathrm{MeV}$, so if we fix the abundance to some value at $10\,\mathrm{MeV}$, we can determine the resulting relic abundance. In particular, by fixing the initial abundance to zero, we necessarily underestimate the relic density. This corresponds to a choice of $x_{\rm min}$ and the condition that $Y(x_{\rm min}) = 0$. With this initial condition, we can exclude models on the basis of their relic densities even when they never attain thermal equilibrium with the SM. Further, these constraints are determined entirely by conditions below $T_{\rm BBN}$, and are thus completely independent of the UV completion.

Note that if $\Lambda_{\rm EFT}$ is sufficiently small, then even with this initial condition, the DM species will thermalize with the SM between $T_{\rm BBN}$ and the present day. In this case, the relic density is set by the standard freeze-out paradigm, and Eq. (7.26) is not valid. Even if the DM species does not quite enter thermal equilibrium, as long as

it attains a non-negligible abundance, Eq. (7.26) can significantly overpredict the relic density. Thus, while Eq. (7.26) is useful to understand the qualitative features of the non-thermal relic density, we evaluate the constraint by numerically solving Eq. (7.14).

As in the previous cases, we need to specify the coupling to neutrinos to perform these calculations consistently. Since the neutrino bath has a temperature comparable to the electron bath, a light χ can be effectively produced by neutrinos as well as electrons. Thus, a coupling between ν and χ can significantly affect the relic abundance. However, as with the coupling to electrons, the relic density is not monotonic in the coupling to neutrinos. If the DM never enters thermal equilibrium with any SM species, then a coupling to neutrinos tends to enhance the relic abundance by providing another production channel. On the other hand, if DM does enter equilibrium *with neutrinos*, then a larger coupling to neutrinos keeps it in equilibrium longer, reducing the relic abundance. However, at most of the points of interest in our parameter space, the constraint is driven by out-of-equilibrium production, so we neglect the coupling to neutrinos when evaluating the relic density.

7.4 Constraints and detection rates

The constraints we place on sub-MeV DM are relevant for direct detection experiments based on elastic electron–DM recoils. In principle, there are many such experiments, but they share several important features. Generically, electron recoil experiments prepare a low-temperature collection of electrons for scattering with galactic halo DM, and by whatever mechanism, the experiment is sensitive to deposited recoil

energies between some $E_{\min}$ and $E_{\max}$. We calculate the detector sensitivity following [369], but the results are typical of electron recoil experiments with very low thresholds.

7.4.1 Estimation of the event rate

In the proposal of [369], the detector is constructed from an aluminum superconductor. At low temperatures, electrons move through the detector with velocities of order the Fermi velocity v_F, and with the appropriate instrumentation, recoil energies as low as 1 meV may be detectable. We now review the calculation of the detection rate, following [369] and [417].

To compute the detection rate, we will consider scattering events at fixed recoil energy E_R. We label the initial and final DM momenta by $\mathbf{p}_1$ and $\mathbf{p}_3$, and the initial and final electron momenta by $\mathbf{p}_2$ and $\mathbf{p}_4$. We do the same for the energies, so that $E_R = E_1 - E_3 = E_4 - E_2$. We define the 3-momentum transfer by $\mathbf{q} = \mathbf{p}_1 - \mathbf{p}_3$. We denote 4-momenta by P_i, and we write $q = |\mathbf{q}|$ and $p_i = |\mathbf{p}_i|$. We denote the local DM number density by n_χ, and the scattering rate by $\Gamma = \langle n_e \sigma v_{\rm rel} \rangle$. The event rate per unit detector mass is

$$R = \frac{n_\chi}{\rho_{\rm detector}} \int \mathrm{d}v_\chi \, \mathrm{d}E_R \, f_\chi(v_\chi) \, \frac{\mathrm{d}\Gamma(v_\chi, E_R)}{\mathrm{d}E_R}, \tag{7.27}$$

where $f_\chi(v_\chi)$ is the local DM velocity distribution in the lab frame. We take the velocity distribution to be a Maxwell–Boltzmann distribution in the galactic frame with rms velocity 220 km/s and a cutoff at the halo escape velocity $v_{\rm esc} \simeq 500$ km/s. We then determine $f_\chi(v_\chi)$ by taking the Earth velocity to be 244 km/s in the galactic frame [418].

Now we turn to the evaluation of the scattering rate $\Gamma(v_\chi, E_R)$. Observe that Γ not only contains the scattering cross section, but also accounts for the effects of Pauli blocking, effectively controlling the available phase space for scattering events. Following [417], we estimate Γ by

$$\frac{\mathrm{d}\Gamma(E_1, E_R)}{\mathrm{d}E_R} = \int \frac{\mathrm{d}^3\mathbf{p}_2}{(2\pi)^3}\frac{\mathrm{d}^3\mathbf{p}_3}{(2\pi)^3}\frac{\mathrm{d}^3\mathbf{p}_4}{(2\pi)^3} W(\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \mathbf{p}_4) 2 f_{\mathrm{FD}}(E_2)(1 - f_{\mathrm{FD}}(E_4))\delta_E \delta_P^4, \tag{7.28}$$

Here, δ_P^4 is a Dirac delta enforcing conservation of 4-momentum; δ_E fixes the recoil energy, setting $E_1 - E_3 = E_R$; $f_{\mathrm{FD}}(E) = 1/(1 + \exp(E - \mu)/T)$ is the Fermi-Dirac distribution; and we define

$$W(\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \mathbf{p}_4) = \frac{\left\langle |\mathcal{M}|^2 \right\rangle}{16 E_1 E_2 E_3 E_4}, \tag{7.29}$$

where $\left\langle |\mathcal{M}|^2 \right\rangle$ is the matrix element for the scattering process.

In many cases of interest, W is independent of the initial and final momenta of the target ($\mathbf{p}_2$ and $\mathbf{p}_4$), in which case the rate factorizes as

$$\frac{\mathrm{d}\Gamma(E_1, E_R)}{\mathrm{d}E_R} = \int \frac{\mathrm{d}^3\mathbf{p}_3}{(2\pi)^3}\delta_E W(\mathbf{p}_1, \mathbf{p}_3) S(E_R, q), \tag{7.30}$$

where S accounts for Pauli blocking, and is given explicitly by

$$S(E_R, q) = \int \frac{2\,\mathrm{d}^3\mathbf{p}_2\,\mathrm{d}^3\mathbf{p}_4}{(2\pi)^2} f_{\mathrm{FD}}(E_2)(1 - f_{\mathrm{FD}}(E_4))\delta_P^4. \tag{7.31}$$

In our EFT, W is not generally independent of the target momenta. However, we can treat scattering in the non-relativistic limit, where such independence is guaranteed: the denominator in Eq. (7.29) is independent of the momenta to first order, and can be replaced with $16 m_\chi^2 m_e^2$. The squared matrix element depends on the momenta only

through the Mandelstam variables s and t, which have non-relativistic limits

$$s \simeq (m_e + m_\chi)^2, \qquad t \simeq 2\mathbf{p}_1 \cdot \mathbf{p}_3, \tag{7.32}$$

so $\left\langle |\mathcal{M}|^2 \right\rangle$ is also independent of $\mathbf{p}_2$ and $\mathbf{p}_4$ to first order. Thus, for the remainder of this chapter, we will consider W to be a function of $\mathbf{p}_1$ and $\mathbf{p}_3$ only, and factorize the rate as in Eq. (7.30).

We work in the low-temperature limit, where $f_{\rm FD}$ reduces to a Heaviside step function, $f_{\rm FD}(E_i) = \Theta(E_F - E_i)$, where $E_F \approx 11.7\,\text{eV}$ is the Fermi energy of aluminum. In this case, $S(E_R, q)$ can be evaluated explicitly. We perform the $\mathbf{p}_4$ integral using the 3-momentum–conservation delta function, and we use the remaining energy-conservation delta function to integrate over $\cos\theta_2$. This leaves a 1-dimensional integral,

$$S(E_R, q) = \frac{m_e}{\pi q} \int p_2 \, \mathrm{d}p_2 \, \Theta\left(1 - \left|\frac{2m_e E_R - q^2}{2p_2 q}\right|\right) \times \Theta\left(E_F - E_2\right)\left[1 - \Theta\left(E_F - E_2 - E_R\right)\right]. \tag{7.33}$$

This integral can be evaluated directly by comparing the arguments of the Heaviside functions. The result is

$$S(E_R, q) = \frac{m_e\left(m_e E_R - E_S^2\right)}{\pi q} \Theta\left(2m_e E_R - E_M^2\right), \tag{7.34}$$

where $E_M^2 = \left(2m_e E_R - q^2\right)^2/(4q^2)$ and E_S^2 is given by

$$E_S^2 = \max\left(2m_e(E_F - E_R),\, E_M^2\right). \tag{7.35}$$

To actually evaluate the rate in Eq. (7.28), we change coordinates to (E_R, q). Since there is no dependence on the azimuthal angle, we obtain

$$\mathrm{d}^3\mathbf{p}_3 = \frac{2\pi m_\chi q}{p_1} \, \mathrm{d}q \, \mathrm{d}E_R, \tag{7.36}$$

and the limits of integration are $q_- < q < q_+$, where

$$q_\pm = \sqrt{p_1^2 + p_3^2 \pm 2p_1 p_3}. \tag{7.37}$$

Under this change of coordinates, in the non-relativistic limit, $t \simeq 2p_1^2 - 2m_\chi E_R - q^2$. In particular, this means that W depends on $\mathbf{p}_1$ and $\mathbf{p}_3$ only through q, p_1, and E_R. Then the differential scattering rate $\mathrm{d}\Gamma/\mathrm{d}E_R$ in Eq. (7.27) is given by

$$\frac{\mathrm{d}\Gamma}{\mathrm{d}E_R} = \frac{m_\chi}{(2\pi)^2 p_1} \int_{q_-}^{q_+} q\,\mathrm{d}q\, W(p_1, E_R, q) S(E_R, q). \tag{7.38}$$

The limits of the E_R integral in Eq. (7.27) are set by the lower and upper thresholds of the detector, which we take to be 1 meV and 1 eV, respectively. Note that there are kinematical constraints on the minimum DM velocity (E_1) required to deliver a given recoil energy E_R. Thus, the cutoff in the velocity distribution effectively imposes a maximum E_R at fixed m_χ.

7.4.2 Detection prospects and constraints by operator

We now examine our cosmological constraints in relation to the projected experimental reach for each of the operators in Tables 7.1 and 7.2. Figures 7.4 to 7.7 show cosmological constraints alongside projected 95% CL direct detection constraints with a 1 kg yr exposure. In order to point to some general features of our results, we duplicate constraints for $\mathcal{O}_{SS}^{(\psi)}$ in Fig. 7.3. However, the following discussion applies to all of the results in Figs. 7.4 to 7.7.

All of the interactions considered for ψ a Dirac fermion can also be evaluated for ψ a Majorana fermion, and we do not consider matrix elements for Majorana fermions separately. Rather, we can directly relate our cosmological constraints

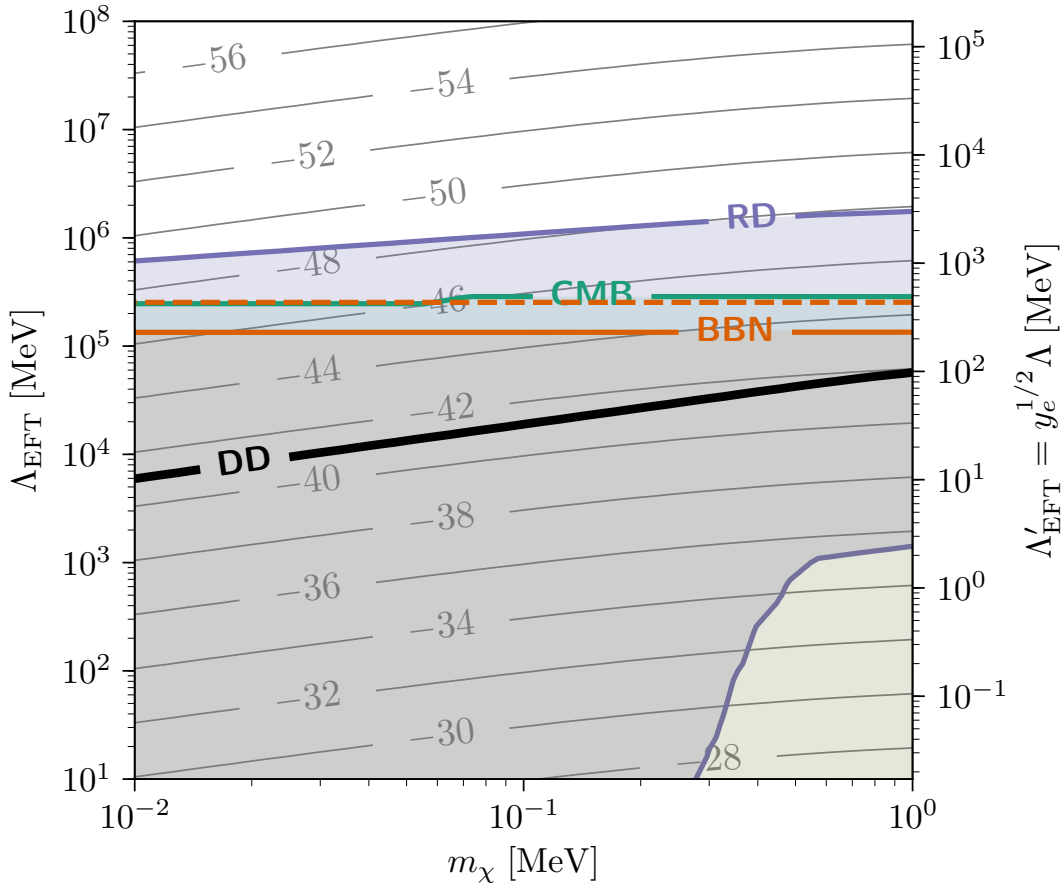

Figure 7.3: Constraints on a Dirac fermion ψ interacting via the operator $\mathcal{O}_{SS}^{(\psi)} = \Lambda_{\rm EFT}^{-2}\bar{\psi}\psi\bar{e}e$ ($g = 1$). Background contours show scattering cross section, labeled as $\log_{10}(\sigma_{\rm scat}/{\rm cm}^2)$. *Black, DD:* direct detection sensitivity (95% CL) with 1 kg yr exposure. *Green, CMB:* constraint from $N_{\rm eff}$. *Orange, BBN:* solid line: constraint from light element abundances with a threshold temperature of 1 MeV. Dashed line: constraint with a threshold temperature of 2.3 MeV (see Section 7.3.1). *Blue, RD:* constraint from relic density.

on a Dirac fermion to the Majorana case. Whereas the relic density is controlled by $n_\psi \Gamma_A = n_\psi^2 \langle \sigma |v| \rangle$ for a Dirac fermion, this expression double-counts the phase space for a Majorana fermion. Since the relic density is inversely proportional to the annihilation rate, it follows that the relic density of a Majorana fermion is simply twice that of a Dirac fermion with the same mass and interactions [352, 419].

The annihilation rate also sets the freeze-out temperature for a species in equilibrium with the SM, via the condition $\Gamma_A \simeq H$. In general, $\Gamma_A \sim \Lambda_{\rm EFT}^{4-k}$ for a dimension-k operator. All of our operators with DM a fermion are dimension-6, so to go from the Dirac case to the Majorana case, it is sufficient to make the replacement $\Lambda_{\rm EFT} \to 2^{-1/(4-k)} \Lambda_{\rm EFT} = \sqrt{2} \Lambda_{\rm EFT}$. In principle, the value of $N_{\rm eff}$ is also different in the Majorana case, but in nearly the entire excluded parameter space, $\Delta N_{\rm eff}$ is large compared with experimental uncertainty, sufficient to rule out a Majorana fermion as well as a Dirac fermion. Thus, in sum, the cosmological constraint curves in Fig. 7.3 are shifted up slightly by a factor of $\sqrt{2}$ in the Majorana case, while the direct detection projections are unchanged.

In each figure, the left vertical axis shows the suppression scale $\Lambda_{\rm EFT}$, effectively corresponding to inverse coupling. Thus, a stronger constraint line appears higher on the plot, and excludes the parameter space below. The left axis in each plot gives the value of $\Lambda_{\rm EFT}$ alone, and the coupling g is taken to be 1. This is distinct from fixing $g/\Lambda_{\rm EFT}$ or $g/\Lambda_{\rm EFT}^2$, since we must have $\Lambda_{\rm EFT} \gg T_{\rm BBN}$ at all points regardless of the value of the coupling. Otherwise, the EFT would be applied outside its regime of validity.

However, as discusssed in Section 7.2, many UV completions naturally generate

a coupling of order y_e. To account for this possibility, we show a second vertical axis on the right of each plot, corresponding to the value of $\Lambda_{\rm EFT}$ in the case that $g = y_e$. For dimension-5 operators, which appear with a factor of $\Lambda_{\rm EFT}^{-1}$, this corresponds to $\Lambda'_{\rm EFT} = y_e \Lambda_{\rm EFT}$. For dimension-6 operators, $\Lambda'_{\rm EFT} = y_e^{1/2} \Lambda_{\rm EFT}$ instead.

Where $\Lambda'_{\rm EFT} \lesssim T_{\rm BBN}$, the EFT may not be applicable. This is important, e.g., for comparing the EFT to specific UV completions, but it has little effect on our conclusions: in every case, our constraints become relevant at $\Lambda'_{\rm EFT} \gg T_{\rm BBN}$, and a significant range of direct detection cross sections can still be ruled out by cosmology. In principle, cosmological constraints on cross sections that lie below $\Lambda'_{\rm EFT} \sim T_{\rm BBN}$ can be evaded by models that have new MeV-scale degrees of freedom in addition to the DM species. However, models of this kind do not generically alleviate the constraints.

The projected direct detection reach (DD, black) is generally the lowest line in each figure, i.e., the weakest constraint. The next line, stronger at low masses by greater than an order of magnitude in $\Lambda_{\rm EFT}$, is the constraint from light element ratios (BBN, orange). In certain cases, a higher threshold temperature of 2.3 MeV is appropriate, see for instance [389] (see Section 7.3.1). The corresponding constraints are shown as dashed curves. However, in general, we can only place a constraint at the lower temperature of 1 MeV, shown with solid curves. In either case, a comparable constraint is obtained from $N_{\rm eff}$ as measured from T_ν/T_γ (CMB, green). The final constraint is from overproduction of DM (RD, blue). The final constraint is from overproduction of DM (RD, blue). Note that for some operators, there are narrow islands of parameter space where the $N_{\rm eff}$ constraint is weakened. In these regions, the impact on $N_{\rm eff}$ is transitioning between $\Delta N_{\rm eff} < 0$ and $\Delta N_{\rm eff} > 0$, as in Fig. 7.2. Similarly, some regions

with small $\Lambda_{\rm EFT}$ are not ruled out by overproduction, since the DM thermalizes and freezes out at a lower abundance.

As anticipated in Section 7.2, when comparing direct detection prospects to cosmological constraints, no operator improves on the prospects of $\mathcal{O}_S^{(\phi)}$ for scalar DM. For fermionic DM, on the other hand, we expect that the operators $\mathcal{O}_{VV}^{(\psi)}$, $\mathcal{O}_{AA}^{(\psi)}$, and $\mathcal{O}_{TT}^{(\psi)}$ will be at least competitive with $\mathcal{O}_{SS}^{(\psi)}$, and this is borne out by our results. Still, we find no region of parameter space in which the projected direct detection constraints exceed all three cosmological probes for any of our effective operators.

Simplistically, this suggests that any model with a heavy mediator detectable by such an experiment is ruled out by cosmology. However, there remain possible exceptions to these constraints, as we discuss in the following section.

7.5 Discussion and conclusions

In this chapter, we have derived cosmological constraints on a broad class of sub-MeV DM models that can be compared directly with detection prospects in electron recoil detectors. We now revisit the generality of our constraints, point out possible exceptions, and discuss the outlook for sub-MeV DM at electron recoil experiments.

Effectively, our goal has been to derive cosmological constraints on the *scattering* cross section between electrons and sub-MeV DM. Cosmology is mainly sensitive to the DM annihilation cross section, and in order to connect the two cross sections, we have produced these constraints in the context of an EFT. We have enumerated the possible thermal histories for a single DM species in this framework. If the DM is in

thermal equilibrium with electrons at high temperatures, then light element abundances and $N_{\rm eff}$ constrain the freeze-out temperature, and thereby constrain the interactions between χ and the SM. In the alternative scenario, if the DM is out of equilibrium at early times, a lower bound can be placed on the relic density, providing an independent constraint on the interactions. In both cases, a constraint is placed on the coupling between DM and electrons, assuming a specific form for the interaction.

In general, the form of the operator coupling electrons to DM affects the relationship between the annihilation cross section at early times and the scattering cross section today. Typically, then, constraints obtained by these methods are model-dependent. However, if the DM–SM mediator has a mass above $\sim 10\,\mathrm{MeV}$, then our approach is quite general: our results are only sensitive to physical processes at lower temperatures, where the EFT is valid and cosmological history is well-established. Still, beyond the mediator mass, there are a few possible exceptions to the constraints derived here.

First, some of these constraints can be evaded with an extended dark sector. In principle, the overproduction constraint can be weakened: such models provide mechanisms to deplete the DM relic density, although we will discuss caveats to this scenario shortly. However, even in this case, the existence of a light DM species is enough for the BBN and $N_{\rm eff}$ bounds to remain effective—adding additional dark degrees of freedom does nothing to improve the situation. One could still escape these constraints by assuming that a phase transition takes place in an extended dark sector between $T_{\rm BBN}$ and the present day, such that the EFT is not valid in both epochs.

Another class of exceptions consists of models in which the dark species *enters*

thermal equilibrium with the SM below $T_{\rm BBN}$, and thus below T_D, the temperature of neutrino-photon decoupling. In this case, the entropy transferred to the SM bath upon freeze-out can be comparable to the entropy accepted upon equilibration, so the constraint from $N_{\rm eff}$ can be circumvented [391]. This scenario is possible only in a very limited segment of the heavy-mediator parameter space, which we estimate as follows. We set the abundance of DM to zero at 1 MeV, and then determine the minimum value of $\Lambda_{\rm EFT}$ below which DM thermalizes before the temperature drops to 0.5 MeV, thus still influencing BBN. Above this value of $\Lambda_{\rm EFT}$, it is possible to evade bounds from BBN and $N_{\rm eff}$, depending on initial conditions. Typically, this minimal value of $\Lambda_{\rm EFT}$ is about one decade weaker than the BBN limit, and still out of reach of direct detection projections across most of our mass range.

Note that the overproduction bound already assumes an initial condition with zero DM abundance, so it cannot be evaded in this way. This is an example of the utility of the several overlapping constraints: the most conservative assumptions are different for each constraint, and correspondingly, exceptions apply differently as well. It is thus necessary to consider all of our constraints simultaneously, even in cases where one constraint appears to dominate. Our goal is to generalize the constraints to the broadest possible class of models, and even though many regions of parameter space are ruled out by multiple observables, it is important to carefully evaluate each constraint independently.

Still, the fact that the overproduction constraint exceeds the constraints from BBN and the CMB is itself a notable result. In general, there are many mechanisms that can influence the dark matter density, so constraints from the relic density are

typically confounded by significant model dependence. However, in the scenario of interest, the model dependence is quite limited. To evade the constraint, one would need a mechanism of depleting the dark matter density at temperatures well below 1 MeV.

There are some simple methods of accomplishing this depletion, e.g., entropy dilution [420], a late phase transition in the dark sector, or late-time decay of a heavy species into sub-MeV DM today. However, each of these can also be used to evade constraints from BBN and the CMB, so they do not bestow any additional model-dependence on the overproduction bound. It is conceivable that number-changing interactions in the dark sector (e.g. $4 \to 2$ processes) could be used to deplete the DM density without modifying the other constraints, and this model dependence is unique to the overproduction bound. But even this strategy would only work in a narrow region of parameter space, and in that sense, it is comparable to known exceptions in the usual BBN and CMB bounds [391, 394].

The overproduction constraint thus sets a new target for future direct detection proposals. Considering only BBN and CMB constraints motivates direct detection experiments that probe scattering cross sections a few orders of magnitude beyond the projections in this chapter. However, overcoming the overproduction bound requires experimental proposals to reach several orders of magnitude beyond the BBN and CMB constraints.

Finally, we note that it might be possible to evade our constraints by taking some arbitrary linear combination of the effective operators in Tables 7.1 and 7.2. In principle, in this high-dimensional parameter space, there might be points for which

interference of the matrix elements in Tables 7.3 and 7.6 conspires to reduce the DM annihilation or production cross section while preserving the scattering cross section. Then each of our cosmological constraints would be weakened, while the projected direct detection constraints would be maintained. However, in order for this to work, the Wilson coefficients would have to be engineered to produce such a cancellation.

In light of these constraints, the outlook for extant electron recoil detection proposals is brightest for DM masses $1\,\mathrm{MeV} \lesssim m_\chi \lesssim 1\,\mathrm{GeV}$ or for mediator masses $m_\zeta \ll 10\,\mathrm{MeV}$. In order to access parameter space which is viable in our framework, and in particular to surpass the overproduction bound, future proposals must probe scattering cross sections at least six orders of magnitude beyond current proposals. A light mediator certainly remains a possibility, but is subject to additional constraints [see e.g. 397]. The case of a light mediator is thus best studied in the context of simplified models, as in the analysis of [390]. Inelastic scattering may also improve direct detection prospects relative to cosmological constraints, and, of course, DM masses above $\sim 1\,\mathrm{MeV}$ remain an interesting target. However, if DM is dominantly composed of a single light species, and interacts dominantly with electrons via a heavy mediator, then cosmological constraints compromise the prospects of proposed experiments.

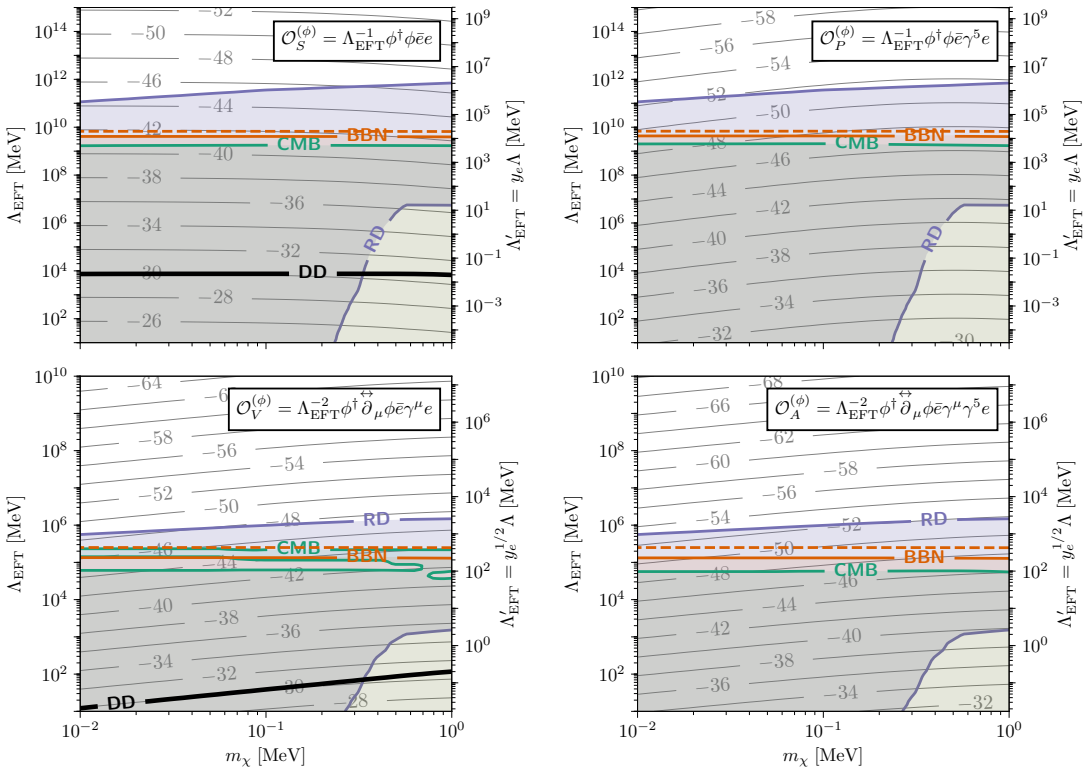

Figure 7.4: Constraints by operator for DM a scalar ϕ. Background contours show scattering cross section, labeled as $\log_{10}(\sigma_{\rm scat}/{\rm cm}^2)$. *Green, CMB:* constraint from $N_{\rm eff}$. *Orange, BBN:* solid line: constraint from light element abundances with a threshold temperature of 1 MeV. Dashed line: constraint with a threshold temperature of 2.3 MeV (see Section 7.3.1). *Blue, RD:* constraint from relic density. *Black, DD:* direct detection sensitivity (95% CL) with 1 kg yr exposure. Note that the direct detection contour does not appear for $\mathcal{O}_P^{(\phi)}$ or $\mathcal{O}_A^{(\phi)}$. For these operators, direct detection can constrain smaller values of $\Lambda_{\rm EFT}$ than shown on the plot, but our framework requires that $\Lambda_{\rm EFT} \gtrsim 10\,{\rm MeV}$.

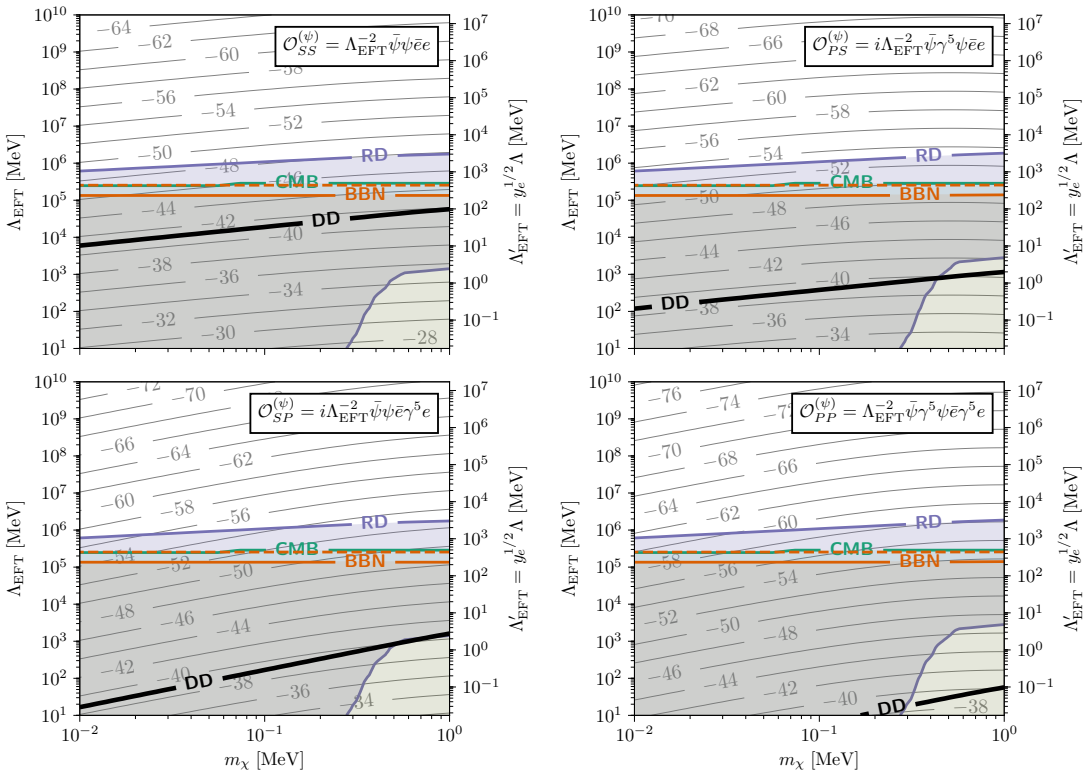

Figure 7.5: Constraints by operator for DM a fermion ψ, for operators composed of scalar or pseudoscalar bilinears. Background contours show scattering cross section, labeled as $\log_{10}(\sigma_{\rm scat}/{\rm cm}^2)$. *Green, CMB:* constraint from $N_{\rm eff}$. *Orange, BBN:* solid line: constraint from light element abundances with a threshold temperature of 1 MeV. Dashed line: constraint with a threshold temperature of 2.3 MeV (see Section 7.3.1). *Blue, RD:* constraint from relic density. *Black, DD:* direct detection sensitivity (95% CL) with 1 kg yr exposure.

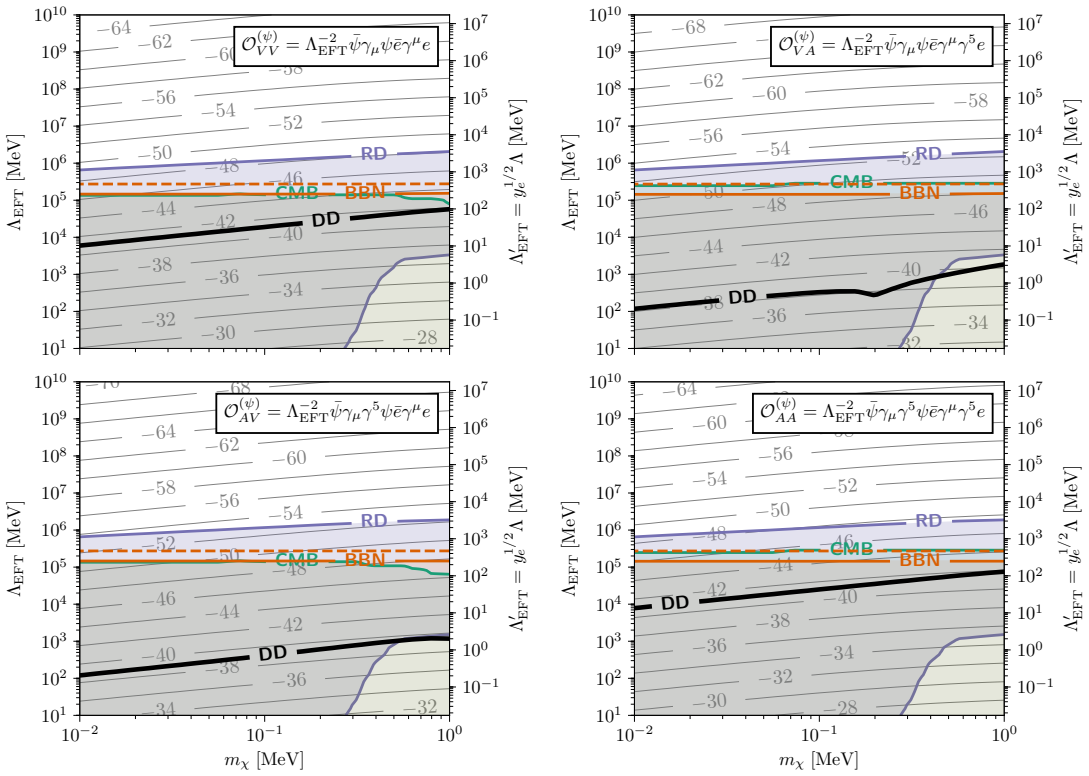

Figure 7.6: Constraints by operator for DM a fermion ψ, for operators containing a vector or axial vector current. Background contours show scattering cross section, labeled as $\log_{10}(\sigma_{\rm scat}/{\rm cm}^2)$. *Green, CMB:* constraint from $N_{\rm eff}$. *Orange, BBN:* solid line: constraint from light element abundances with a threshold temperature of 1 MeV. Dashed line: constraint with a threshold temperature of 2.3 MeV (see Section 7.3.1). *Blue, RD:* constraint from relic density. *Black, DD:* direct detection sensitivity (95% CL) with 1 kg yr exposure.

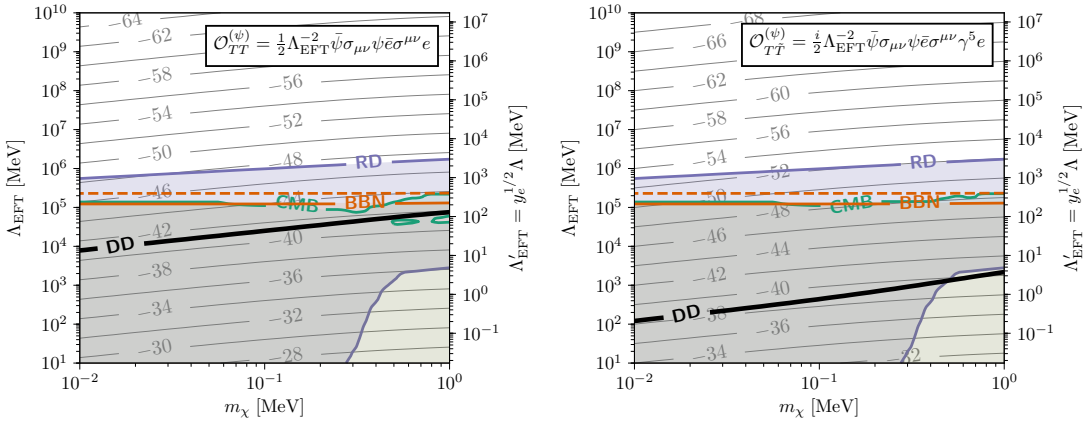

Figure 7.7: Constraints by operator for DM a fermion ψ, for operators containing a spin-2 current. Background contours show scattering cross section, labeled as $\log_{10}(\sigma_{\rm scat}/{\rm cm}^2)$. *Green, CMB:* constraint from $N_{\rm eff}$. *Orange, BBN:* solid line: constraint from light element abundances with a threshold temperature of 1 MeV. Dashed line: constraint with a threshold temperature of 2.3 MeV (see Section 7.3.1). *Blue, RD:* constraint from relic density. *Black, DD:* direct detection sensitivity (95% CL) with 1 kg yr exposure.

Operator	$g^{-2}\Lambda_{\rm EFT}^2 \sum_{\rm spin} \lvert\mathcal{M}\rvert^2_{\phi\bar{\phi}\to e^+e^-}$
$\mathcal{O}_S^{(\phi)}$	$2s - 8m_e^2$
$\mathcal{O}_P^{(\phi)}$	$2s$

Operator	$y_e^{-2}g^{-2}\Lambda_{\rm EFT}^4 \sum_{\rm spin} \lvert\mathcal{M}\rvert^2_{\phi\bar{\phi}\to e^+e^-}$
$\mathcal{O}_V^{(\phi)}$	$-8\left(t - m_e^2\right)\left(s + t - m_e^2\right) + 16m_\phi^2\left(t - m_e^2\right) - 8m_\phi^4$
$\mathcal{O}_A^{(\phi)}$	$-8t(s+t) + 16m_e^2 t + 16m_\phi^2\left(t + m_e^2\right) - 8m_e^4 - 8m_\phi^4$

Table 7.3: Squared matrix elements for $\phi\bar{\phi} \to e^+e^-$ with ϕ a complex scalar, summed over final spin states. The operators are as defined in Table 7.1. Note that the matrix elements for $\mathcal{O}_V^{(\phi)}$ and $\mathcal{O}_A^{(\phi)}$ vanish if ϕ is taken to be a real scalar. The matrix elements for scattering, $\phi e^- \to \phi e^-$, are obtained from these by the substitution $s \leftrightarrow t$.

Operator	$g^{-2}\Lambda_{\rm EFT}^2\sigma(\phi\bar{\phi}\to e^+e^-)$
$\mathcal{O}_S^{(\phi)}$	$\frac{1}{8\pi s}\left(s-4m_e^2\right)^{3/2}\left(s-4m_\phi^2\right)^{-1/2}$
$\mathcal{O}_P^{(\phi)}$	$\frac{1}{8\pi}\left(s-4m_e^2\right)^{1/2}\left(s-4m_\phi^2\right)^{-1/2}$

Operator	$y_e^{-2}g^{-2}\Lambda_{\rm EFT}^4\sigma(\phi\bar{\phi}\to e^+e^-)$
$\mathcal{O}_V^{(\phi)}$	$\frac{1}{12\pi s}\left(s+2m_e^2\right)\left(s-4m_e^2\right)^{1/2}\left(s-4m_\phi^2\right)^{1/2}$
$\mathcal{O}_A^{(\phi)}$	$\frac{1}{12\pi s}\left(s-4m_e^2\right)^{3/2}\left(s-4m_\phi^2\right)^{1/2}$

Table 7.4: Cross sections for $\phi\bar{\phi}\to e^+e^-$ for each effective operator in Table 7.1, summed over final spins. Note that the matrix elements for $\mathcal{O}_V^{(\phi)}$ and $\mathcal{O}_A^{(\phi)}$ vanish if ϕ is taken to be a real scalar.

Operator	$g^{-2}\Lambda_{\rm EFT}^2\sigma(\phi e^-\to\phi e^-)$
$\mathcal{O}_S^{(\phi)}$	$\frac{1}{16\pi s^2}\left[s^2+6m_e^2s-2m_\phi^2(s+m_e^2)+m_\phi^4+m_e^4\right]$
$\mathcal{O}_P^{(\phi)}$	$\frac{1}{16\pi s^2}\left[(s-m_e^2)^2-2m_\phi^2(s+m_e^2)+m_\phi^4\right]$

Operator	$y_e^{-2}g^{-2}\Lambda_{\rm EFT}^4\sigma(\phi e^-\to\phi e^-)$
$\mathcal{O}_V^{(\phi)}$	$\frac{1}{16\pi s}\left[s^2+2\left(m_e^2+m_\phi^2\right)s-\left(m_e^2-m_\phi^2\right)^2\right]$
$\mathcal{O}_A^{(\phi)}$	$\frac{1}{16\pi s}\left[s^2-6m_e^2s+2m_\phi^2\left(s+m_e^2\right)-m_e^4-m_\phi^4\right]$

Table 7.5: Cross sections for $\phi e^-\to\phi e^-$ for each effective operator in Table 7.1, averaged over initial spins and summed over final spins. Note that the matrix elements for $\mathcal{O}_V^{(\phi)}$ and $\mathcal{O}_A^{(\phi)}$ vanish if ϕ is taken to be a real scalar.

Operator	$g^{-2}\Lambda_{\rm EFT}^4 \sum_{\rm spin} \lvert\mathcal{M}\rvert^2_{\psi\bar{\psi}\to e^+e^-}$
$\mathcal{O}_{SS}^{(\psi)}$	$4\big(s-4m_e^2\big)\big(s-4m_\psi^2\big)$
$\mathcal{O}_{PS}^{(\psi)}$	$4s\big(s-4m_e^2\big)$
$\mathcal{O}_{SP}^{(\psi)}$	$4s\big(s-4m_\psi^2\big)$
$\mathcal{O}_{PP}^{(\psi)}$	$4s^2$
$\mathcal{O}_{VV}^{(\psi)}$	$8\,(s+t)^2+8t^2+16m_+^4-32m_+^2t$
$\mathcal{O}_{VA}^{(\psi)}$	$8\,(s+t)^2+8t^2+16m_-^4-32m_+^2t-32sm_e^2$
$\mathcal{O}_{AV}^{(\psi)}$	$8\,(s+t)^2+8t^2+16m_-^4-32m_+^2t-32sm_\psi^2$
$\mathcal{O}_{AA}^{(\psi)}$	$8\,(s+t)^2+8t^2+16m_+^2-32m_+^2t-32m_+^2s+2\,(8m_em_\psi)^2$
$\mathcal{O}_{TT}^{(\psi)}$	$8(s+2t)^2+32m_+^4-16\,(s+4t)\,m_+^2+(8m_em_\psi)^2$
$\mathcal{O}_{T\tilde{T}}^{(\psi)}$	$8(s+2t)^2+32m_-^4-16\,(s+4t)\,m_+^2$

Table 7.6: Squared matrix elements for $\psi\bar{\psi}\to e^+e^-$ with ψ a Dirac fermion, summed (not averaged) over initial and final spin states. The operators are as defined in Table 7.2. Note that the matrix elements for $\mathcal{O}_{VV}^{(\psi)}$, $\mathcal{O}_{VA}^{(\psi)}$, $\mathcal{O}_{TT}^{(\psi)}$, and $\mathcal{O}_{T\tilde{T}}^{(\psi)}$ vanish if ψ is taken to be a Majorana fermion. For brevity, we define $m_\pm^2\equiv m_e^2\pm m_\psi^2$. The matrix elements for scattering, $\psi e^-\to\psi e^-$, are obtained from these by the substitution $s\leftrightarrow t$.

Operator	$g^{-2}\Lambda_{\rm EFT}^4\sigma(\psi\bar{\psi}\to e^+e^-)$
$\mathcal{O}_{SS}^{(\psi)}$	$\frac{1}{16\pi}\frac{T_e^3T_\psi}{s}$
$\mathcal{O}_{PS}^{(\psi)}$	$\frac{1}{16\pi}\frac{T_e^3}{T_\psi}$
$\mathcal{O}_{SP}^{(\psi)}$	$\frac{1}{16\pi}T_eT_\psi$
$\mathcal{O}_{PP}^{(\psi)}$	$\frac{1}{16\pi}\frac{sT_e}{T_\psi}$
$\mathcal{O}_{VV}^{(\psi)}$	$\frac{1}{12\pi}\frac{T_e}{T_\psi}\left(s+2m_e^2\right)\left(s+2m_\psi^2\right)$
$\mathcal{O}_{VA}^{(\psi)}$	$\frac{1}{12\pi}\frac{T_e^3}{sT_\psi}\left(s+2m_\psi^2\right)$
$\mathcal{O}_{AV}^{(\psi)}$	$\frac{1}{12\pi}\frac{T_eT_\psi}{sT_e}\left(s+2m_e^2\right)$
$\mathcal{O}_{AA}^{(\psi)}$	$\frac{1}{12\pi}\frac{T_e}{T_\psi}\left[s^2-4\left(m_\psi^2+m_e^2\right)s+28m_\psi^2m_e^2\right]$
$\mathcal{O}_{TT}^{(\psi)}$	$\frac{1}{24\pi}\frac{T_e}{sT_\psi}\left[\left(s+2m_e^2\right)s+2m_\psi^2\left(s+20m_e^2\right)\right]$
$\mathcal{O}_{T\tilde{T}}^{(\psi)}$	$\frac{1}{24\pi}\frac{T_e}{sT_\psi}\left[\left(s+2m_e^2\right)s+2m_\psi^2\left(s-16m_e^2\right)\right]$

Table 7.7: Cross sections for $\psi\bar{\psi}\to e^+e^-$ for each effective operator in Table 7.2, averaged over initial spins and summed over final spins. Note that the cross sections for $\mathcal{O}_{VV}^{(\psi)}$, $\mathcal{O}_{VA}^{(\psi)}$, $\mathcal{O}_{TT}^{(\psi)}$, and $\mathcal{O}_{T\tilde{T}}^{(\psi)}$ vanish if ψ is taken to be a Majorana fermion. For brevity, we define $T_i^2\equiv s-4m_i^2$.

Operator	$48\pi s^3 g^{-2}\Lambda_{\rm EFT}^4\sigma(\psi e^- \to \psi e^-)$
$\mathcal{O}_{SS}^{(\psi)}$	$s^4 + 2m_+^2 s^3 + 2s^2\left(3m_e^4 - 14m_e^2 m_\psi^2 + 3m_\psi^4\right) + 2m_-^4 m_+^2 s + m_-^8$
$\mathcal{O}_{PS}^{(\psi)}$	$\left(s^2 + 4sm_e^2 + m_e^4 + m_\psi^4 - 2m_\psi^2 s_e^+\right)\left(s_e^{-2} - 2m_\psi^2 s_e^+ + m_\psi^4\right)$
$\mathcal{O}_{SP}^{(\psi)}$	$\left[m_\psi^2\left(4s - 2m_e^2 + m_\psi^2\right) + s_e^{-2}\right]\left(s_e^{-2} - 2m_\psi^2 s_e^+ + m_\psi^4\right)$
$\mathcal{O}_{PP}^{(\psi)}$	$\left(s_e^{-2} - 2m_\psi^2 s_e^+ + m_\psi^4\right)^2$
$\mathcal{O}_{VV}^{(\psi)}$	$2s^2\left(4s^2 - 10m_+^2 s + 9m_e^4 + 22m_e^2 m_\psi^2 + 9m_\psi^4\right) - 8m_+^2 m_-^4 s + 2m_-^8$
$\mathcal{O}_{VA}^{(\psi)}$	$2\left(s_e^{-2} - 2m_\psi^2 s_e^+ + m_\psi^4\right)\left[(s + s_e^+)^2 - 2m_\psi^2 s_e^+ + m_\psi^4\right]$
$\mathcal{O}_{AV}^{(\psi)}$	$2\left[2s\left(2s - m_e^2 + 2m_\psi^2\right) + m_-^4\right]\left(s_e^{-2} - 2m_\psi^2 s_e^+ + m_\psi^4\right)$
$\mathcal{O}_{AA}^{(\psi)}$	$2s^2\left(4s^2 - 4m_+^2 s - 3m_e^4 + 46m_e^2 m_\psi^2 - 3m_\psi^4\right) + 4m_+^2 m_-^4 s + 2m_-^8$
$\mathcal{O}_{TT}^{(\psi)}$	$2s^2\left(7s^2 - 13m_+^2 s + 6m_e^4 + 52m_e^2 m_\psi^2 + 6m_\psi^4\right) - 2m_+^2 m_-^4 + 2m_-^8$
$\mathcal{O}_{T\tilde{T}}^{(\psi)}$	$2\left[s\left(m_+^2 + 7s\right) + m_-^4\right]\left(s_e^{-2} - 2m_\psi^2 s_e^+ + m_\psi^4\right)$

Table 7.8: Cross sections for $\psi e^- \to \psi e^-$ for each effective operator in Table 7.2, averaged over initial spins and summed over final spins. Note that the cross sections for $\mathcal{O}_{VV}^{(\psi)}$, $\mathcal{O}_{VA}^{(\psi)}$, $\mathcal{O}_{TT}^{(\psi)}$, and $\mathcal{O}_{T\tilde{T}}^{(\psi)}$ vanish if ψ is taken to be a Majorana fermion.For brevity, we define $m_\pm^2 \equiv m_e^2 \pm m_\psi^2$ and $s_i^\pm \equiv s \pm m_i^2$.

Part III

Direct detection of light dark matter

Invitation

In the previous two parts, we have explored new probes of DM and new physics across a wide range of scales, from massive PBHs to ultralight bosons. In particular, in Chapter 7, we demonstrated major complementarity between cosmology and direct detection experiments for sub-GeV dark matter. At that time, we described the experimental landscape in simplistic terms, introducing and treating superconducting detectors only as deeply as necessary to provide a benchmark for comparison with cosmology. In this part, however, we delve into the other side of the complementary duo: we develop prospects and proposals for sub-GeV direct detection in detail.

Direct detection in the sub-GeV regime has been a rapidly developing field over the past decade, spurred by advances in quantum sensing technology. The goal of this part, reflecting the goal of this thesis, is to develop the theoretical infrastructure to leverage such new tools for DM searches. But first, we pause to consider the motivation for new probes in the sub-GeV mass range. What should be the goalposts for new searches?

Clearly defined benchmarks are invaluable for direct detection. While it is extremely useful to have experiments that can test a multitude of models simultaneously, it is just as important to have some theoretical prior for the model space that

experiments should probe. Absent any particular targets, one can always come up with a larger and more sensitive experiment that searches for nothing in particular. With this in mind, there are two important boundaries to target in the sub-GeV regime.

First, in the cross section itself, we can aspire to probe scattering cross sections that are typical of benchmark simplified models in which the DM is produced with the correct abundance while satisfying all constraints. There is always some model dependence in such targets, and the space of simplified models is still not fully explored—indeed, this is a goal of my ongoing work—but at least within a given model, freeze-in production (see Chapter 6) corresponds to a well-defined target for direct detection.

Second, in the DM mass, we can aspire to design experiments sensitive to the scattering of keV-scale DM. The guidepost arises from cosmological constraints: if DM is fermionic, then its mass must be at or above the keV scale. Of course, if DM is bosonic, lower masses are possible. But keV-scale fermions are still an excellent target: probing scattering at these masses corresponds to sensitivity to energy deposits of order 1 meV, which means that such experiments can also probe the absorption of bosonic DM with masses down to this smaller scale. For bosons with masses far below 1 meV, very different experimental approaches are required. Thus, for traditional direct detection, the keV scale is a meaningful target.

To understand how to probe keV-scale DM, it is first important to understand why the current generation of experiments does not. This can be understood in terms of kinematics: consider an elastic collision between an incoming DM particle, χ, with a stationary target, T. The maximum fraction of the kinetic energy of the DM particle that can be transferred to the target is given by $4m_\chi m_T/(m_\chi + m_T)^2$. In the limit

$m_\chi \ll m_T$, this becomes $\sim 4m_\chi/m_T \ll 1$, and the kinetic energy of the DM also scales with m_χ. This means that the maximum amount of energy deposited in a detector decreases rapidly when the DM mass becomes lower than the mass of the target particle. In typical direct detection experiments, the target particle is an atomic nucleus with an atomic number of $\mathcal{O}(10\text{–}100)$, so the maximum deposit shrinks quickly as the DM mass goes below $\sim 10\,\mathrm{GeV}$. For $m_\chi \lesssim 1\,\mathrm{GeV}$, the maximum deposit is smaller than the threshold energy in typical experiments, and all sensitivity is lost.

As this situation illustrates, sensitivity to light DM is a matter of two independent considerations. The first is kinematic matching between the DM and the target. The DM must be able to transfer a large fraction of its kinetic energy to the detector. The second consideration is the threshold of the detector itself—it makes no difference how much kinetic energy is transferred if depositing all of the kinetic energy of the DM still would not trigger a count in the detector. Thus, we need a system that is well kinematically matched to light DM while allowing for a very low detection threshold.

A convenient way to configure a system with a low threshold is to use a system with a small gap in its excitation spectrum. The system can be cooled to a temperature well below the gap, such that thermal excitations are exponentially suppressed, and then excitations above the gap can be attributed to DM interactions. Kinematic matching is a trickier issue: detectors with macroscopic volumes are ultimately built out of atoms, and nuclei have GeV-scale masses. But the electrons in such detectors offer a complementary set of opportunities to probe DM interactions with electrons at low masses, with a target mass at the MeV scale. In the eleven years since this was first pointed out by Ref. [367], experiments based on electron recoils have successfully

probed DM interactions at masses as low as 1 MeV.

Electronic systems also offer a convenient set of gapped spectra for direct detection experiments. For example, the current generation of experiments uses atomic ionization as the target process: here the gap is an ionization energy of order 1 eV. In recent years, it has become technologically feasible to register single ionization events, so the thresholds of these experiments are now limited by the physical process itself, i.e., by the size of the gap. Constructing experiments with sensitivity to meV deposits requires a system with a much smaller gap, but there are numerous experimental proposals to achieve this target. In particular, we will discuss superconducting detectors in some detail. Here the gap is the binding energy of Cooper pairs, which is $\mathcal{O}(\text{meV})$ for typical superconductors. Thus, such systems can in principle probe DM scattering all the way to the keV scale.

The future for electron recoil experiments seems bright. Does poor kinematic matching doom efforts to probe nuclear couplings? Not necessarily: for small deposits, the nuclei, being bound in a lattice, can no longer be treated as free particles. Instead of prompting the recoil of a single nucleus, a DM scattering event can excite a collective mode amongst the nuclei in the lattice, i.e., a phonon. The kinematics of phonons are entirely different from those of free nuclei, and may enable sensitivity to interactions of nuclei with keV-scale DM if single-phonon excitations can be reliably read out. Adapting superconducting detectors for this purpose is one goal of my ongoing work.

However, the collective modes that may be a blessing for DM–nucleon scattering have also been the bane of DM–electron scattering. Collective modes are inherently complicated: their properties are determined by condensed matter physics, and are dif-

ficult to predict a priori. That is the problem that will occupy us at the beginning of this part: how can one predict the scattering rate of DM with electrons when the electrons are not free particles, but bound in a condensed matter system, with highly nontrivial wavefunctions?

The answer, in Chapter 8, will turn out to be remarkably simple, and will provide substantial intuition for behavior of electron recoil experiments, with important implications for experimental design. In Chapter 9, we will go yet further with these collective excitations, and demonstrate how quasiparticle excitations in superconductors can be used to design detectors with directional sensitivity, a key feature for DM discovery. Finally, in Chapter 10, we will take a step back to the concrete, and we will give real new limits on DM interactions in otherwise unconstrained parameter space using a prototype superconducting detector. This detector is the first step towards realizing extremely low thresholds for light DM detection, and we will detail the roadmap towards achieving the goalposts we set for the future of direct detection.

Chapter 8

Dark matter–electron scattering and the dielectric function

Dark matter (DM)–electron scattering was first proposed for sub-GeV DM detection less than a decade ago [367], and there has been enormous theoretical [368–370, 372, 375, 376, 378–380, 421–447] and experimental [381–384, 448–457] progress since then. Since electrons are not free particles, but are bound in atoms or delocalized across solids, they have favorable kinematics for light DM scattering. However, the rich complexity of condensed matter systems complicates the calculation of scattering rates. Not only do bound electrons have different wavefunctions than their free-particle counterparts [458], many condensed matter systems exhibit collective electronic modes such as plasmons [459]. A formalism describing DM scattering with a single electronic state [368, 436] can potentially miss important electron interaction and correlation effects, and must carefully account for 'screening' where the electron density rearranges itself to partially cancel out DM-induced perturbations [370].

In this chapter, we propose to bypass the single-particle formulation entirely, and frame the problem of DM–electron scattering in terms of matrix elements of the many-body electron density operator. This perspective is inspired by a classic paper on collective energy loss in solids [460], and since it does not rely on a particular choice of eigenstates, it is equally applicable to *all* systems: atoms, molecules, metals, insulators, or more exotic materials. Moreover, it intrinsically accounts for all electron interactions and correlations in the target by relating the scattering rate to an experimentally-measurable quantity, the complex dielectric function $\epsilon(\mathbf{q},\omega)$. Crucially, since $\epsilon(\mathbf{q},\omega)$ is defined as a linear response function, the response of the target to a momentum transfer $\mathbf{q}$ and energy deposit ω is determined by density matrix elements which are the *same* whether measured by DM–electron scattering or by an electromagnetic probe [461, 462]. The assumption of linear response applies as long as DM interactions are weaker than electromagnetism.

The key result of this chapter is that the total scattering rate for DM with mass m_χ and velocity $\mathbf{v}_\chi$ in an arbitrary target is given by

$$\Gamma(\mathbf{v}_\chi) = \int \frac{\mathrm{d}^3\mathbf{q}}{(2\pi)^3}\,|V(\mathbf{q})|^2\,\left[2\frac{q^2}{e^2}\,\mathrm{Im}\left(-\frac{1}{\epsilon(\mathbf{q},\omega_\mathbf{q})}\right)\right], \tag{8.1}$$

where $\omega_\mathbf{q} = \mathbf{q}\cdot\mathbf{v}_\chi - \frac{q^2}{2m_\chi}$, $q = |\mathbf{q}|$, e is the electron charge, and $V(\mathbf{q})$ is the non-relativistic DM-electron potential. The full derivation can be found in Appendix D, and follows mainly from the arguments made in Ref. [460]. The target-dependent object which appears in the integrand,

$$\mathcal{W}(\mathbf{q},\omega) \equiv \mathrm{Im}\left(-\frac{1}{\epsilon(\mathbf{q},\omega)}\right) = \frac{\mathrm{Im}[\epsilon(\mathbf{q},\omega)]}{|\epsilon(\mathbf{q},\omega)|^2}, \tag{8.2}$$

is known as the loss function. The *only* assumptions we have made about the DM

interactions in deriving Eq. (8.1) are (*i*) that the non-relativistic Hamiltonian coupling DM to electrons takes the form $\hat{H}_{\rm int} = \sum_i V(\hat{\mathbf{r}}_\chi - \hat{\mathbf{r}}_i)$, depending only on the electron position operators $\hat{\mathbf{r}}_i$ and no other operators such as spin or momentum, and (*ii*) that $\hat{H}_{\rm int}$ can be treated perturbatively.

The consequences of Eq. (8.1) are of immediate importance for DM–electron scattering. Spin-independent Hamiltonians arise in many common benchmark models, including those for scattering through scalar and vector mediators. The presence of $\mathcal{W}$ implies that *all* of these interactions are screened. The importance of screening was first noted for a kinetically-mixed dark photon mediator in a solid-state target [370], and later for a scalar mediator [445]. Our results show that a scalar force which couples equally and oppositely to electrons and protons, whether short- or long-ranged, is screened *exactly* like a kinetically-mixed dark photon. Furthermore, as long as ion contributions to the loss function are negligible (as in semiconductors well above the gap), forces that couple differently to nucleons and electrons are still screened identically. All such screening effects are invisible in a single-particle picture.

Since $\mathcal{W}(\mathbf{q}, \omega)$ is directly measurable through electromagnetic scattering, DM–electron scattering experiments can be calibrated experimentally, exactly as was done for DM absorption [371, 373, 463] using the measured real conductivity $\sigma_1(\omega) = (\omega/4\pi)|\epsilon(0,\omega)|^2\mathcal{W}(0,\omega)$. The advantage of our approach is that the loss function can also be modeled semi-analytically in certain relevant energy and momentum regimes, and such models can be compared directly to data. This enables rapid assessment of candidate experimental targets, and potentially bypasses the need for numerical electron wavefunctions to determine the reach of novel detector materials.

In the following sections, we show that in a material with free carriers, the loss function scales as $\mathcal{W}(\mathbf{q},\omega) \propto q$ at small ω, which can be interpreted as the familiar screening which partially suppresses the $1/q^4$ enhancement characteristic of a light mediator. We then show that if $\mathcal{W}(0,\omega)$ is nonvanishing, a rate enhancement at small q remains whenever ω is kinematically accessible. This behavior of the loss function can arise in two qualitatively different ways: interband transitions in insulators, and long-range plasmons which are generically present in all materials. As we will show, the low-energy plasmon tail may improve the sensitivity of superconducting detectors to light DM by several orders of magnitude, and materials with Fermi velocities *slower* than v_χ may allow DM to access the bulk of the loss function rather than the tail. We illustrate these kinematic regimes in Fig. 8.1.

In this chapter, we adopt a generic form for the potential, $V(\mathbf{q}) = V(q) = \frac{g_\chi g_e}{q^2+m_{\phi,V}^2}$ which is valid for DM coupling through a scalar mediator ϕ or vector V. We compute scattering rates by integrating Eq. (8.1) over the DM velocity distribution, for which we take the Standard Halo Model (see Appendix D for details). We frame our results in terms of a reference cross section $\overline{\sigma}_e = (\mu_{e\chi}^2/\pi)|V(q_0)|^2$ where $\mu_{e\chi}$ is the electron–DM reduced mass and $q_0 = \alpha m_e \simeq 3.7$ keV is a reference momentum. We show results for a light mediator $m_{\phi,V}^2 \ll q^2$, with heavy mediator results given in Appendix D (Fig. D.6).

8.1 Conventional superconductors

Ref. [369] first proposed using superconducting metals such as aluminum (Al) as targets for DM–electron scattering. Ref. [370] soon pointed out that long-range Coulomb forces among electrons would screen DM interactions if mediated by a kinetically-mixed dark photon. This effect was incorporated by multiplying the free-particle matrix element by $1/|\epsilon_{\rm RPA}(\mathbf{q},\omega)|^2$, where $\epsilon_{\rm RPA}$ is the dielectric function of a free electron gas (FEG) in the random phase approximation (RPA) at zero temperature.

Even within RPA, our formalism identifies two important corrections to the DM interaction rate from Ref. [370]. First, *all* interactions coupling to electron density are screened, including a light scalar mediator and a non-kinetically-mixed vector mediator. This unifies the reach for all models considered in Ref. [370]. Second, the analytic structure of the loss function imposed by causality implies a particular choice of branch cut in $\epsilon_{\rm RPA}$ differing from that used in Ref. [370] (see Appendix D for details).

The latter correction improves the projected sensitivity of conventional superconductor detectors to DM scattering through a light mediator by several orders of magnitude at low masses. We can understand this by examining $\epsilon_{\rm RPA}$ in the kinematic regime $q \ll k_F$, $\omega \ll q v_F$ relevant for sub-MeV DM scattering near the Fermi surface, where k_F is the Fermi momentum and v_F is the Fermi velocity, respectively 3.5 keV and 6.8×10^{-3} in Al. The result is [468]

$$\epsilon_{\rm RPA}(\mathbf{q},\omega) \approx \frac{\lambda_{\rm TF}^2}{2q^2} + i\frac{3\pi\omega_p^2\omega}{2q^3 v_F^3}, \tag{8.3}$$

where $\lambda_{\rm TF} \simeq 3.8\,\text{keV}$ is the Thomas-Fermi screening length and $\omega_p \simeq 15\,\text{eV}$ is the plasma frequency. The imaginary part is typically smaller than the real part, so $\mathcal{W}(\mathbf{q},\omega)$ scales

as $\frac{\omega/q^3}{1/q^4} \sim \omega q$, a much softer screening than the q^4 implied from $1/|\epsilon|^2$.

Moving beyond RPA, we use the results of Ref. [464], which fits to data a model containing both a 1-loop 'local field' correction to the electron vertex and a q-dependent plasmon width $\Gamma_p/\omega_p \simeq 0.1$–$0.3$. The fit implies that the contribution from the ion polarizability in Al is small, justifying our approximation that only electrons contribute to the loss function. The projected reach for a 1 meV threshold is shown in Fig. 8.2 for a light mediator, with comparisons to previous results given in Fig. D.5 of Appendix D. The orange band reflects theoretical uncertainty in the proper form of the loss function in the energy range of interest (see Appendix D).

In most materials, the loss function features a plasmon with a Lorentzian lineshape peaked at ω_p [459, 461] and a low-energy tail (see Fig. 8.1 and Appendix D). In the parametrization of Ref. [464], $\mathcal{W}(q = 0, \omega)$ scales linearly with ω for $\omega \ll \omega_p$, and the plasmon tail dominates over the RPA contribution. Our results suggest that a kg-yr exposure of an Al target with a 1 meV threshold is sufficient to cover the entire freeze-in thermal relic target [350, 367, 466, 467] above 10 keV. However, this depends on the extrapolation of the plasmon tail to meV energies, and existing measurements only characterize the loss function at $\omega \gtrsim 100$ meV [469]. Thus, additional measurements of $\mathcal{W}$ are crucial to accurately determine the sensitivity. There may also be contributions to $\mathcal{W}$ from coherent scattering with the Cooper pair condensate for energies $\omega \simeq 2\Delta$, as well as finite-temperature effects. We leave investigation of these effects for future work [470].

8.2 Semiconductors

In a typical semiconductor like silicon (Si) with a gap $E_g \sim$ eV, an energy deposit $\omega \simeq E_g$ requires a momentum deposit $q \geq E_g/v_\chi \sim$ keV for $v_\chi \sim 10^{-3}$, independent of the DM mass, as shown in Fig. 8.1. The size of the first Brillouin zone (BZ) in Si is $2\pi/a \simeq 2\,\text{keV}$, where a is the lattice constant. Thus, for $\omega \gtrsim 2\,\text{eV}$, DM is probing interatomic distances rather than delocalized electrons, and the electrons may be modeled as an FEG with an effective $k_F \simeq 2\pi/a$ set by the total valence electron density. This approximation is an excellent match to both density functional theory (DFT) calculations [471] and data [472] for $q \simeq 5\,\text{keV}$ and $\omega \gg E_g$ in Si [473]; for sufficiently large q ($\sim 15\,\text{keV}$, see Appendix D), the bound electron orbitals give large-momentum tails not captured by the FEG.

Equation (8.1) and Fig. 8.1 show that at fixed ω, the rate receives contributions from $\mathcal{W}(\mathbf{q}, \omega)$ over many orders of magnitude in q for $m_\chi \gtrsim 10\,\text{MeV}$, so the FEG approximation is best for a light mediator, where $V(q) \propto q^{-4}$ weights the integrand most towards small q. Our formalism thus suggests a generic explanation for the behavior of the DM–electron spectrum in the 5–15 eV range (2–4 electron–hole pairs in Si [368]) from light mediator exchange in any conventional semiconductor. The projected reach in Si under the FEG approximation with a $2e^-$ threshold is shown in Fig. 8.2 for a light mediator.

The differences among various targets become most apparent when $\omega \simeq E_g$, where the band structure describing delocalized electrons with $q \lesssim 2\pi/a$ becomes important. In addition to band structure effects, there is also an irreducible contribution

from the plasmon [474], where the tail extends into the kinematically allowed region for DM. This has important implications for rate predictions in currently-operating semiconductor detectors [455–457]. DFT calculations predict a rate which peaks in the 1- or 2-electron bin, corresponding to $\omega \lesssim 8.3\,\text{eV}$, for all DM masses for which these energies are kinematically accessible [368]. Currently available measurements of $\mathcal{W}$ suggest the true rate in these few-electron bins may be somewhat larger. Near-gap effects are quite difficult to model [473], but in our formalism, they can be accounted for by making more precise measurements at $\omega \simeq E_g$ and $q \simeq E_g/v_\chi$.

On the other hand, for near-gap scattering in a narrow-gap semiconductor ($E_g \sim 10\,\text{meV}$), we have $q_{\text{min}} \simeq 10\,\text{eV} \ll 2\pi/a$, so the delocalized electrons in the uppermost valence band dominate the behavior of the scattering rate as $q \to 0$. We may understand the absence of screening in these systems through the Lindhard form of the dielectric function [468], which shows that $\epsilon(\mathbf{q}, \omega)$ has a finite limit as $\mathbf{q} \to 0$, with the imaginary part proportional to the interband transition matrix element. The lack of mobile charge carriers inhibits the screening present in metals. In the next section, we discuss an example of such a narrow-gap semiconductor: a Dirac material.

8.3 Novel Materials

Our formalism suggests that optimal materials for sub-GeV DM detection will have a loss function with large support for $\omega < v_\chi q$ (Fig. 8.1). For an ordinary metal with an electron effective mass $m^* = m_e$, the loss function is maximized at large q when $\omega = q v_F$, where $v_F = k_F/m^* \simeq 10 v_\chi$. This is outside of the kinematically-allowed region

for DM scattering. For small q, collective modes such as the plasmon will dominate, but the plasmon is damped at momenta $q > q_c \simeq \omega_p/v_F$ [468] due to decay into the particle-hole continuum. Therefore, DM can only excite the undamped plasmon if $v_\chi > v_F$ [442]. Here we explore two qualitatively different ways to achieve $v_F < v_\chi$: Dirac materials, in which v_F is not tied directly to free-electron properties, and heavy-fermion materials, where strongly-correlated electrons can create a Fermi surface with a large m^*.

Dirac materials, characterized by linear electronic dispersion $\omega(k) = v_F k$ with widely-varying v_F across materials [475], are promising targets for DM detection [375, 431, 438, 439]. Consider a gapless isotropic Dirac material with a single Dirac cone and effective background dielectric constant $\kappa \equiv \mathrm{Re}[\epsilon(0,0)]$. In typical materials, $\mathrm{Re}(\epsilon) \gg \mathrm{Im}(\epsilon)$ over the relevant $\mathbf{q}$ and ω [438], and we may write the loss function as

$$\mathcal{W}_{\mathrm{Dirac}}(q,\omega) = \frac{e^2}{12\pi\kappa^2 v_F}\Theta(\omega - v_F q)\Theta(\omega_{\mathrm{max}} - \omega). \tag{8.4}$$

The loss function with a gap 2Δ is given in Appendix D; $\mathcal{W}_{\mathrm{Dirac}}(q,\omega)$ is constant as $q \to 0$ for all $\omega > 2\Delta$, as anticipated. The loss function immediately displays two key features of scattering in Dirac materials [375]: small v_F increases the rate, and scattering is forbidden if $v_\chi < v_F$ for $\omega = \omega_{\mathbf{q}}$. In Fig. 8.2, we show the sensitivity of an isotropic Dirac material for a light mediator.

This analysis neglects many-body effects, including the plasmon contribution to the loss function. Dirac materials are expected to exhibit two tuneable plasmon modes distinct from the ordinary valence plasmon: a temperature-dependent mode which could lie in the $\mathcal{O}$(meV) range [476–478], and a zero-temperature mode tuneable with chemical potential [479]. Therefore, measurements of the loss function in real materials are

crucial to accurately estimate the scattering rate, since the plasmon contribution may dominate [480] as was the case for superconductors.

Another way to lower v_F is to find materials with ordinary quadratic dispersion but large effective masses. As an example, a number of materials containing f-electrons are known as heavy-fermion systems because they display a Fermi surface with $m^* \sim (10\text{–}100)m_e$ [481–483]. These materials are expected to have a plasmon at energy $\omega_p^* \simeq T^*$, the Fermi temperature of the heavy electrons [484]. One such material is URu_2Si_2, a heavy-fermion superconductor with $T^* = 75\,\mathrm{K} = 6.5\,\mathrm{meV}$ and $m^* \simeq 6m_e$ [485], from which one may estimate $v_F \simeq 6.5 \times 10^{-5}$, $\omega_p^* \simeq T^* = 6.5\,\mathrm{meV}$, and $q_c \simeq \omega_p^*/v_F \simeq 100\,\mathrm{eV}$. In reality, the measured loss function in URu_2Si_2 [465] shows considerable anisotropy with Lorentzian peaks at either 4 meV or 6 meV depending on the direction of $\mathbf{q}$, as well as a broad peak around 18 meV, which can also be interpreted as a heavy-fermion plasmon (see Appendix D). Despite the extremely rich electron dynamics in this material, in our formalism we may compute the DM rate unambiguously once $\mathcal{W}$ is measured in the relevant kinematic regime.

The measured data (see Appendix D, Fig. D.1) show that $\mathcal{W}(\omega) \propto \omega$ above the heavy-fermion plasmon peaks, consistent with the tail of the ordinary valence electron plasmon. However, in contrast to spectra from conventional superconductors or semiconductors, the measured loss function in URu_2Si_2 shows rich structure which could be used to separate signals from backgrounds not due to fast-particle scattering. Integrating over ω from a threshold of 1 meV up to $\omega_{\mathrm{max}} = 74\,\mathrm{meV}$, the maximum value where data exists, we obtain the projected reach in Fig. 8.2. The band spans measurements of $\mathcal{W}(\mathbf{q}, \omega)$ as $\mathbf{q} \to 0$ along two different crystal axes. We leave a full analysis

of the anistropic response to future work [470]. As expected, the reach in URu_2Si_2 can surpass Al in the mass range 5–40 keV, where the DM kinetic energy is comparable to the heavy-fermion plasmon energies. Our reach estimates motivate further study of URu_2Si_2 and similar materials as targets for light DM scattering.

8.4 Implications for experiments

The advantage of our formulation of the DM scattering rate is that no theoretical input from *e.g.* DFT is required to compute the scattering rate; the DM energy loss spectrum from spin-independent electron scattering may be precisely predicted from a measurement with an electromagnetic probe in the appropriate kinematic regime. For MeV–GeV DM, X-ray scattering covers the regime $q \sim$ keV and $\omega \sim$ eV [462], while for keV–MeV DM, momentum-resolved electron energy loss spectroscopy (EELS) can cover $q \sim$ eV and $\omega \sim$ meV [461, 486]. These techniques are standard in condensed matter physics, and a rich literature on measurements of dielectric and loss functions already exists for a number of systems of interest.

The downside of this formalism is that it does not directly predict how many electron–hole pairs are created in the material per unit deposited energy, or how the energy is down-converted from plasmon excitations to charge and phonons. However, if individual quasiparticle contributions to $\mathcal{W}(\mathbf{q}, \omega)$ can be modeled, this information can be reconstructed. Moreover, the quasiparticle contributions may be determined empirically by correlating scattering events using an electromagnetic probe with the partition of excitations read out by the detector, as has been done for nuclear recoil calibrations

at higher energy. We argue that these measurements should be considered the primary calibration mechanisms for DM–electron scattering, analogous to photoabsorption for bosonic DM absorption [371, 373, 463].

Finally, our work may be applied to unify the electronic and phonon descriptions of DM scattering with other sub-gap loss mechanisms that have not yet been explored, such as dielectric heating in insulators or coherent scattering off the superconducting condensate. Dielectric skin depth in the long-wavelength limit $\mathbf{q} \to 0$ is proportional to $\sqrt{\mathrm{Re}[\epsilon(\omega)]}/(\omega\,\mathrm{Im}[\epsilon(\omega)])$, and thus materials with a small skin depth for THz photons and calorimetric readout should respond efficiently to DM–electron scattering, even for meV-scale energy deposits below the eV-scale electronic band gaps. Many materials have THz absorption features, so high-resolution THz or infrared transmission spectra are likely fertile ground for exploring new materials for keV-scale DM scattering.

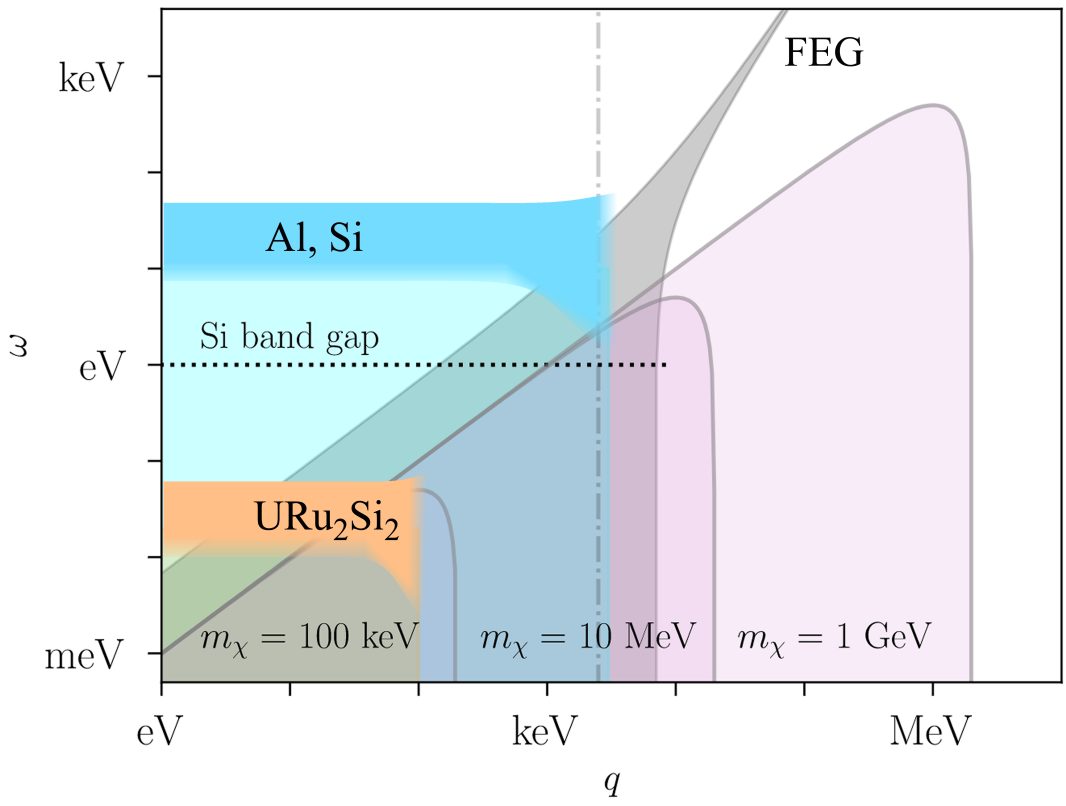

Figure 8.1: Schematic depiction of the relevant kinematics for sub-GeV DM. The shaded purple parabolas represent the kinematically-allowed region of q and ω for the labeled DM masses, for a fixed DM speed $v_\chi = 10^{-3}$, with upper boundary $\omega = qv_\chi$ independent of m_χ. The blue and orange shaded regions represent the support of the plasmon part of the loss function. The tail extends into the DM region for conventional materials such as Al and Si, and for heavy-fermion materials such as URu_2Si_2, the plasmon peak lies in the DM region. The range of support for the free electron gas (FEG) loss function is shown in shaded grey, and can be used to approximate the rate in both superconductors and semiconductors over a limited range of ω. The dot-dashed vertical line indicates the size of the Brillouin zone ($q \approx 2.3\,\text{keV}$) of Si, while the horizontal dashed line indicates the band gap above which electron scattering can produce ionization.

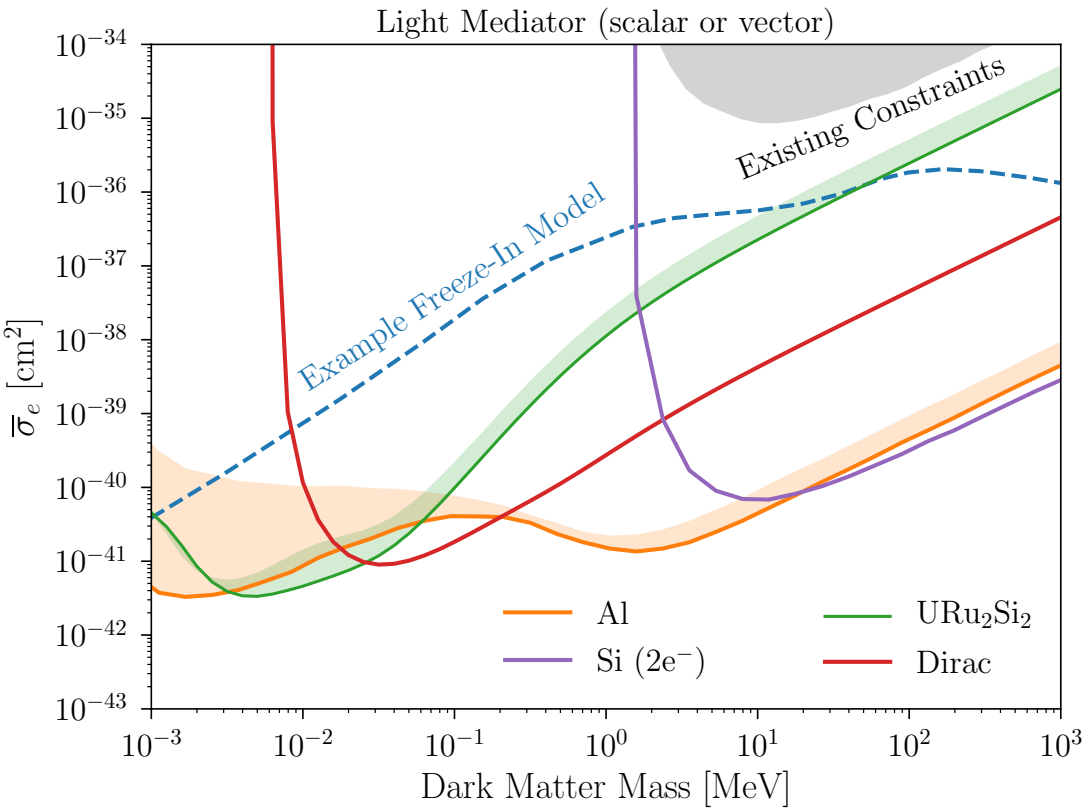

Figure 8.2: The projected 3-event reach of a 1 kg-yr exposure target of Al (orange), Si (purple), and URu_2Si_2 (green), computed for a light scalar or vector mediator using Eq. (8.1). For Al, the solid line uses $\mathcal{W}$ from Ref. [464], and the top of the shaded region uses the FEG model, both with $\omega \in [1\,\text{meV}, 1\,\text{eV}]$. The URu_2Si_2 loss function is taken from Ref. [465] with $\omega \in [1\,\text{meV}, 74\,\text{meV}]$, and the shaded region spans $\mathcal{W}$ measured along two crystal axes. Si is treated as a FEG with a $2e^-$ threshold, using the ionization model of Ref. [368]. We also show the reach for a Dirac material with density $10\,\text{g/cm}^3$, gap $2\Delta = 20\,\text{meV}$, Fermi velocity $v_F = 4 \times 10^{-4}$, background dielectric constant $\kappa = 40$, and Dirac band cutoff $\omega_{\text{max}} = 0.5\,\text{eV}$ (red); existing constraints from SENSEI [455], SuperCDMS HVeV [457], DAMIC [384], Xenon10 [380], DarkSide-50 [381], and Xenon1T [454] (shaded gray); and the theory target of a freeze-in model when the mediator is a kinetically-mixed dark photon [350, 367, 466, 467] (dashed blue). The corresponding plot for a heavy mediator is shown in Appendix D as Fig. D.6.

Chapter 9

Superconducting detectors and directional sensitivity

The identity of dark matter (DM) remains one of the most pressing open questions in particle physics and cosmology. Contrary to decades of theoretical expectations, numerous experimental probes have found no conclusive evidence of DM at the weak scale, leading to renewed interest in models of DM at much lower scales. Many new ideas have recently been proposed to search for such light DM in the laboratory [367–370, 372, 375, 376, 378–380, 421–447, 487], and several of these novel direct detection exeperiments have already begun to probe significant parameter space [381–384, 448–457]. Among the new ideas, superconducting targets stand out with the lowest possible thresholds, giving them sensitivity to the lowest DM masses through DM–electron interactions [369, 370, 435, 488]. With superconducting energy gaps of $\mathcal{O}$(meV), such detectors may eventually probe DM with mass as low as the keV scale, where cosmological constraints become significant [364, 489, 490].

Despite the impressive potential reach of superconducting targets, current projections assume that the detectors are insensitive to the *direction* of incoming DM. Directional detection has long been recognized as a powerful tool in DM experiments, including those in the keV–GeV regime [438, 487]: due to the halo wind, the local DM distribution is not isotropic in the laboratory frame, leading to a characteristic modulation of the signal that can be used to reject backgrounds and confirm a discovery. If superconducting detectors can be made sensitive to the direction of the incoming DM, then such targets will offer exceptional promise for future experiments. Such a detector would be capable of making a definitive discovery of DM as light as a keV.

In this chapter, we show that even isotropic superconductors are capable of directional detection via the angular distribution of the excitations produced by DM scattering. For such a measurement to be viable, two key features are required. Firstly, the direction of the initial excitations produced by the DM interaction should be correlated with that of the incoming DM particle. Secondly, the secondary excitations produced by the initial excitations as they down-convert in the material should exhibit directionality correlated with that of the initial excitations. As we will show, both features indeed occur in superconducting targets, paving the way for directional detection of keV-scale DM.

This chapter is organized as follows. We begin by considering the initial scattering of DM with electronic states of a superconductor into excited quasiparticles. Next we consider the down-conversion of these initial excitations in the material into secondary quasiparticles and phonons, treating the general case with a new numerical code. We then present our results for directionality, and end with a discussion of

experimental prospects.

Throughout this chapter, we use the following notation: for a 3-vector $\mathbf{q}$, we write $q = |\mathbf{q}|$. We use angles with two subscripts to denote the relative angle between the two axes specified by those subscripts. All other angles are defined relative to the DM wind axis. The Fermi energy and momentum are denoted by $E_{\rm F}$ and $p_{\rm F}$, respectively. We set $\hbar = c = 1$.

9.1 Dark matter scattering

We first demonstrate the directionality of the initial excitations produced in a DM scattering event in a superconducting target. This requires reformulating the description of the scattering process in terms of the appropriate degrees of freedom in the BCS vacuum of the superconductor [491]. The DM scattering rate in superconductors was originally computed by Ref. [370] considering only large energy deposits compared to the superconducting gap, and so the DM–detector interaction was described in terms of the DM, χ, scattering with individual electrons, $|\chi\rangle\,|e^-\rangle \longrightarrow |\chi'\rangle\,|e^-\rangle$. This is not suitable for studying the kinematics at small deposits. Here, the appropriate degrees of freedom are Bogoliubov quasiparticles (QPs) [492], which are electron–hole superpositions.

In this description, the DM excites the BCS vacuum by pair-producing QPs, as $|\chi\rangle\,|0_{\rm BCS}\rangle \longrightarrow |\chi'\rangle\,|{\rm QP}_1, {\rm QP}_2\rangle$. The total momentum of these *two* QPs is the momentum transfer $\mathbf{q}$ imparted by the DM scatter. The wave functions of the electrons in the BCS vacuum automatically account for Pauli blocking through a *coherence factor,* which has

significant support only when one of the QPs is below the Fermi momentum and the other is above. In Appendix F, we show that for energy deposits much larger than the superconducting gap energy, the scattering rate becomes identical to that for scattering with individual electrons. We use the labels 1 and 2 to refer to the two initial QPs produced in the scattering process. QPs 1 and 2 are interchangeable, so we use the label i for statements that apply to either label.

The QPs have a dispersion relation of the form

$$E_{\rm QP}(\mathbf{p}) = \sqrt{\mathcal{E}_{\mathbf{p}}^2 + \Delta^2}\,, \tag{9.1}$$

where $\mathcal{E}_{\mathbf{p}} = \mathbf{p}^2/(2m_{\rm e}) - E_{\rm F}$ is the Bloch energy relative to the Fermi surface and Δ is half of the superconducting gap energy. Counterintuitively, the energy of a QP is minimized for $p = p_{\rm F}$. The free-electron dispersion relation is recovered in the limit $p_{\rm QP} \gg p_{\rm F}$, whereas the limit $p_{\rm QP} \ll p_{\rm F}$ gives the energy of a hole far below the Fermi surface. This nontrivial dispersion relation modifies the kinematics of DM scattering near the gap, and thus significantly influences directional correlations and down-conversion. For energy deposits $\omega \lesssim$ keV, the momenta $\mathbf{p}_1$ and $\mathbf{p}_2$ will be well inside the first Brillouin zone (BZ). For deposits $\omega \gtrsim$ keV, we expect Eq. (9.1) to receive band structure corrections near the edge of a BZ of order tens of eV or less, small compared to this keV scale, so Eq. (9.1) is a valid approximation for all the energy scales considered in this chapter.

The overall DM scattering rate is given by [488]

$$\Gamma(v_\chi) = \int \frac{\mathrm{d}^3\mathbf{q}}{(2\pi)^3}\,|V(\mathbf{q})|^2 \frac{2\mathbf{q}^2}{e^2}\,\mathrm{Im}\left[-\frac{1}{\epsilon_{\rm BCS}(\mathbf{q},\omega_{\mathbf{q}})}\right], \tag{9.2}$$

where $v_\chi = |\mathbf{v}_\chi|$ is the magnitude of the DM velocity; $\mathbf{q}$ is the momentum transfer; $\omega_{\mathbf{q}} = \mathbf{q}\cdot\mathbf{v}_\chi - \mathbf{q}^2/2m_\chi$ is the deposited energy; and $\epsilon_{\rm BCS}$ is the dielectric function of

a superconductor in the BCS vacuum. In this chapter, we make the approximation $\mathrm{Im}\left(-1/\epsilon_{\mathrm{BCS}}\right) \equiv \mathrm{Im}\left(\epsilon_{\mathrm{BCS}}\right)/\left|\epsilon_{\mathrm{BCS}}\right|^2 \simeq \mathrm{Im}\left(\epsilon_{\mathrm{BCS}}\right)/\left|\epsilon_{\mathrm{L}}\right|^2$. Here ϵ_{L} is the Lindhard form of the dielectric function for a normal metal [468], which accounts for the effects of in-medium screening and collective modes in the normal metal phase [488]. We compute $\mathrm{Im}\left(\epsilon_{\mathrm{BCS}}\right)$ in terms of the QP dispersion relation Eq. (9.1) and the BCS coherence factor, which accounts for near-gap effects [493] and for Pauli blocking. Our approach interpolates between the approximate superconducting dielectric response near the gap and the normal-metal response far from the gap. A more complete treatment explicitly computing the dielectric function in the BCS vacuum will be pursued elsewhere [494]. We take $|V(\mathbf{q})|^2 = (g_e g_\chi)^2\left(\mathbf{q}^2+m_\phi^2\right)^{-2}$, where g_e and g_χ are the couplings of the mediator to the electron and the DM, respectively. This is appropriate for any spin-independent interaction. Further details on DM interactions and the computation of $\mathrm{Im}\left(\epsilon_{\mathrm{BCS}}\right)$ are given in Appendix F.

At fixed DM velocity, the parameter distribution of the initial QPs prior to down-conversion is proportional to the differential rate, which we determine by differentiating Eq. (9.2) with respect to kinematical variables. In our numerical results, we draw samples from the joint distribution $f_{\mathrm{QP}}(\cos\theta_1, \cos\theta_2, E_1, E_2)$, marginalized over the DM distribution, where θ_{i} is the angle between the QP momentum and the DM *wind* axis. In certain regimes, the angular distribution of excitations can be understood analytically by virtue of kinematical constraints. Conservation of energy yields a closed-form expression for $\cos\theta_{\mathbf{qi}}$, the cosine of the angle between QP i and the momentum transfer $\mathbf{q}$, in terms of q and p_{i}. Now consider small DM masses and light mediators, where small deposits are favored. In the limit of small deposits, to leading order in

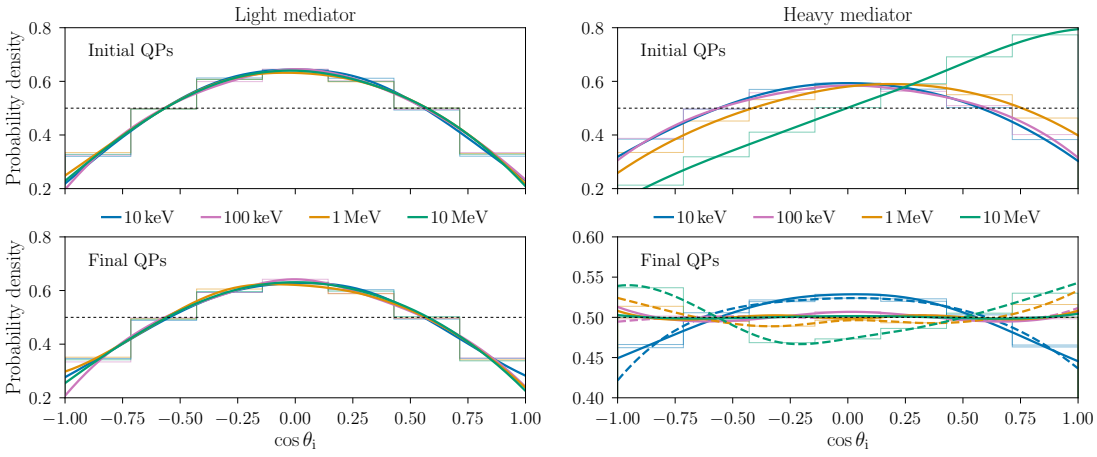

Figure 9.1: Angular distributions of QPs produced by DM scattering in Al. The angles shown are defined with respect to the axis of the DM wind. The distribution of DM orientations in the Standard Halo Model is included. The left and right column show the distributions in the light- and heavy-mediator limits, respectively. In each panel, the colors correspond to different DM masses, and a dashed horizontal line at $\cos\theta_{\mathrm{i}} = \frac{1}{2}$ indicates the isotropic distribution. Thick lines interpolate between histogram values (thin lines) for ease of visualization. The top and bottom rows show distributions of QPs before and after down-conversion, respectively. In the bottom-right panel, several solid curves overlap near the isotropic distribution. The dashed curves show angular distributions obtained by restricting to events with total deposit $\omega < 20\Delta$, for which the effects of down-conversion are less significant.

$\omega - 2\Delta$, we have $p_\mathrm{i} \simeq p_\mathrm{F}$. Then conservation of momentum requires

$$\cos\theta_{\mathbf{q}\mathrm{i}} \simeq \frac{m_\chi v_\chi - \sqrt{2m_\chi\left(\frac{1}{2}m_\chi v_\chi^2 - 2\Delta\right)}}{2m_\mathrm{e}v_\mathrm{F}}\,, \tag{9.3}$$

where $v_\chi = |\mathbf{v}_\chi|$, which implies $0 \lesssim \cos\theta_{\mathbf{q}\mathrm{i}} \lesssim 2\Delta/(v_\chi p_\mathrm{F})$. For aluminum (Al), with $\Delta = 0.3\,\mathrm{meV}$ and $E_\mathrm{F} = 11.7\,\mathrm{eV}$, this leads to the condition $0 \lesssim \cos\theta_{\mathbf{q}\mathrm{i}} \lesssim 10^{-4}$, so excitations produced near the gap are nearly orthogonal to the momentum transfer $\mathbf{q}$. In turn, the direction of $\mathbf{q}$ is correlated with that of $\mathbf{v}_\chi$, so the distribution of $\cos\theta_{\chi\mathrm{i}}$ is peaked at zero.

On the other hand, consider DM interacting via a heavy mediator, for which large deposits are favored. In particular, for $p_\chi \sim p_\mathrm{F}$, momentum transfers of order p_F are possible, corresponding to $\cos\theta_{\mathbf{q}\mathrm{i}} \sim 1$. For example, for $p_\chi = 2p_\mathrm{F}$, if the DM is fully stopped and its energy shared equally between the two QPs, then $\cos\theta_{\mathbf{q}\mathrm{i}}$ is given uniquely by $\cos\theta_{\mathbf{q}\mathrm{i}} = \sqrt{v_\mathrm{F}/(v_\mathrm{F} + v_\chi)}$. For typical materials, $v_\chi \ll v_\mathrm{F}$, so indeed $\cos\theta_{\mathbf{q}\mathrm{i}} \approx 1$. In Al, this solution corresponds to $\cos\theta_{\mathbf{q}\mathrm{i}} \approx 0.93$. Further, fully stopping the DM implies that $\cos\theta_{\chi\mathbf{q}} = 1$, so $\cos\theta_{\chi\mathrm{i}} = \cos\theta_{\mathbf{q}\mathrm{i}}$. Thus, when $p_\chi \sim p_\mathrm{F}$ and large q is favored, the angular distribution can peak in the direction of the DM wind.

The marginal distribution of $\cos\theta_\mathrm{i}$ is shown by the solid curves in the top panels of Fig. 9.1 for several DM masses in Al. We assume the Standard Halo Model with the parameters of Ref. [418]. For light DM or a light mediator, small energy transfers are favored, leading to a peak in the distribution orthogonal to the DM wind axis. For heavier DM and mediators, larger energy transfers lead to a forward-peaked distribution.

9.2 Quasiparticle relaxation

We have shown how the directions of the initial quasiparticles are related to the direction of the incoming DM. The second requirement for directional detection is that this directionality must be preserved after the initial excitations relax. Thus, we now study the down-conversion of the initial QP excitations.

Following Refs. [495–497], we model down-conversion as a repeating sequence of two distinct processes: first, energetic QPs relax by emission of phonons, and second, energetic phonons decay into QP pairs. Quasiparticle pair production eventually stops once all remaining phonons have energy below 2Δ. We treat such phonons as ballistic, including them as part of the final state that is eventually read out by the detector. As shown in Appendix G, phonon emission is kinematically forbidden for very low-energy QPs, so the QPs eventually become ballistic as well. The down-conversion process is finished when all particles are ballistic. For other approaches to the relaxation of highly energetic QPs, see Refs. [498, 499].

We study the impact of down-conversion on directionality by explicit simulation, implemented in a public code based on this chapter.[1] We begin with an ensemble of initial excitations sampled from the distribution f_{QP}, and then iterate the relaxation processes described above until all QPs and phonons are ballistic. Computing the momentum of the outgoing excitations after each relaxation process requires knowledge of the differential rate of these processes with respect to the kinematical parameters of the final state. We take the differential rates for phonon emission and QP pair production from Eqs. (16) and (27) of Ref. [500]. In each case, imposing conservation of energy and

[1] http://github.com/benvlehmann/scdc

momentum using the dispersion relation of Eq. (9.1) gives the distribution of final-state angles. The differential rate of QP pair production by a phonon of energy $\omega_{\rm ph}$ is given by

$$\frac{\mathrm{d}\Gamma}{\mathrm{d}E_{\rm i}} \propto \frac{E_{\rm i}(\omega_{\rm ph} - E_{\rm i}) + \Delta^2}{\sqrt{\left[E_{\rm i}^2 - \Delta^2\right]\left[(\omega_{\rm ph} - E_{\rm i})^2 - \Delta^2\right]}}\,, \tag{9.4}$$

where $E_{\rm i}$ is the energy of one QP and $\omega_{\rm ph} - E_{\rm i}$ is the energy of the other. This distribution is sharply peaked at $E_{\rm i} \sim \Delta$ and $E_{\rm i} \sim \omega_{\rm ph} - \Delta$, corresponding to the case in which one of the two QPs receives most of the phonon's energy. Here, the dispersion relation of Eq. (9.1) implies that the QPs are produced nearly orthogonal to the axis of the initial phonon. Thus, for small deposits, the angular distributions of QPs and phonons in the final state will peak in orthogonal directions.

We sample initial excitations and simulate the down-conversion process for several DM masses in Al. The angular distributions of the resulting QPs are shown as the solid curves in the bottom panels of Fig. 9.1. Directionality of the phonons is weaker (see Appendix G). For light mediators, the initial QPs have excellent directionality, which is well-preserved after down-conversion. For a heavy mediator, despite the directionality of the initial QPs, the down-converted distribution is much closer to isotropic, particularly for heavier DM. This is simply because heavy mediators and heavy DM favor larger deposits, which lead to a larger number of relaxation events. For this reason, the dashed curves in the bottom-right panel of Fig. 9.1 show the angular distribution including only events with total deposits $\omega < 20\Delta$. These dashed curves retain directionality associated with the low-energy part of the initial angular distribution.

9.3 Results

Figure 9.2 shows the estimated sensitivity for the detection of the directional DM wind, with impressive reach. Existing direct detection constraints are shown in gray. (Complementary model-dependent probes may also apply [501–503].) Critically, since the target is *isotropic*, there is no modulation in the overall rate, unlike many directional detection schemes. Instead, the reach must be defined in terms of the anisotropy in the distribution of the final-state excitations. The dashed lines in Fig. 9.2 show the projected reach for a detector which counts final-state quasiparticles in each of two angular bins: the "on-axis" bin, with $|\cos\theta| > \frac{1}{2}$, and the "off-axis" bin, with $|\cos\theta| < \frac{1}{2}$. Establishing that the signal is not isotropic requires a certain number of events, which translates to a minimal cross section for a fixed experimental exposure. Note that for a heavy mediator, we impose a cut to include only small deposits, which reduces the overall rate, but lessens the impact of down-conversion.

The two-bin configuration is the minimal experimental configuration for directional detection, and represents the most conservative projection. For heavy mediators, a more ambitious projection is obtained by assuming precise resolution of $\cos\theta$. In this case, the detailed shape of the angular distribution can be compared with the distribution under the null hypothesis of an isotropic background. This again translates to a minimal number of events to detect directionality, and a corresponding minimal cross section, shown by the dotted line in Fig. 9.2. Further details are given in Appendix H. For light mediators, since directionality is manifest, this procedure gives almost exactly the same result as the two-bin configuration, so the corresponding line is not shown.

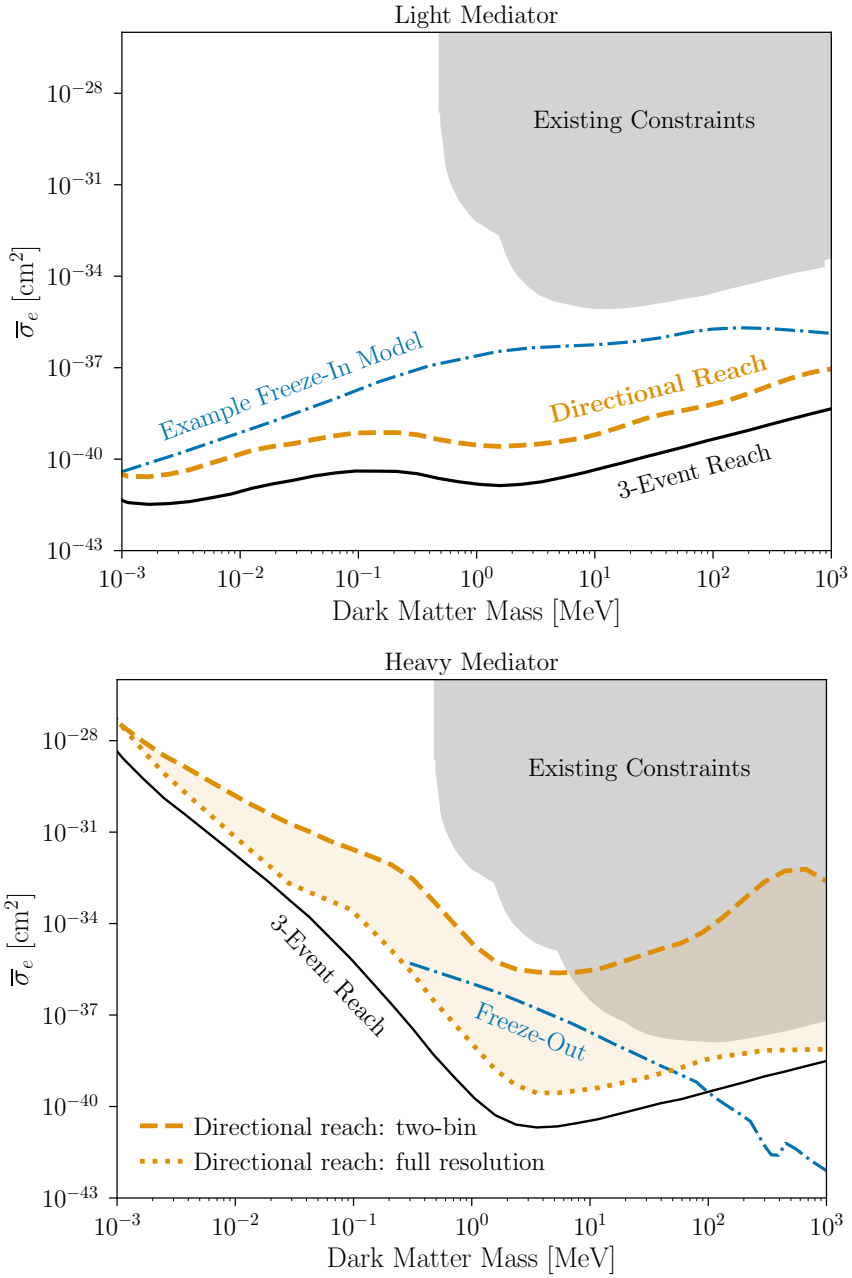

Figure 9.2: Directional detection discovery reach for DM scattering in an Al superconductor via a light (*left panel*) or heavy (*right panel*) mediator. Solid lines: 3-event reach for a kilogram-year exposure, not including directionality. Dashed lines: estimated discovery reach for directionality at 95% C.L. using only two angular bins. Dotted line: estimated discovery reach for directionality in an experiment with high-precision measurement of $\cos\theta$. Blue dot-dashed lines show cross sections for example DM models [378]. Shaded gray regions indicate existing constraints from SENSEI [455], SuperCDMS HVeV [457], DAMIC [384], Xenon10 [380], DarkSide-50 [381], and Xenon1T [454].

The dashed and dotted curves in Fig. 9.2 are representative of the directional sensitivity of an experiment in the simplest and most sophisticated configurations. A kilogram-year exposure of such an experiment would be capable of detecting directionality for DM masses keV $\lesssim m_\chi \lesssim$ GeV. Both configurations discussed here would be directionally sensitive at cross sections covering important cosmological targets that are not currently probed by any direct detection experiments.

9.4 Discussion

In this chapter, we have shown that superconductors [369, 370, 488] can probe DM *directionally*, even in the case of an isotropic medium. This has important implications for the design of directional superconducting detectors. Directionality will require detectors to push resolutions lower, with thresholds close to the superconducting gap. Detectors that can trap primary QPs are preferred as they will be able to take advantage of the directional correlations in the DM signal. This is in contrast to the weak directionality in the phonon system.

In order to scale detectors to kilogram-year exposures while retaining directional sensitivity, a massive multiplexing scheme will likely be required: typical detector volumes will be of order $1\,\mathrm{cm}^3$ or smaller in order to attain high collection efficiencies [370]. Intrinsic QP lifetimes do diverge for very low temperatures and QP occupancy, even in the "dirty" limit [500].

For realistic applications of bulk superconducting targets, characterization of mean QP diffusion length in real samples will determine which materials are best suited

for directional DM detection. Such work has been done for large niobium (Nb) crystals [504], but has largely been put aside over the last few decades. These programs will need to be restarted in order to characterize samples with the appropriate properties to detect small energy deposits in the regime relevant to directional DM detection. For materials with a known diffusion length, this directionality can be converted to a rate modulation by making a detector in which one path length to the sensor is much shorter than this diffusion length, and the orthogonal path length is much longer. Chemical Vapor Deposition–grown superconducting crystals of Nb or Al instrumented on their large surface, with cross-sections of a few mm^2 and thickness of around 100 microns, would achieve this behavior for typical diffusion lengths of a few hundred microns.

Our results demonstrate directional detection of DM in a target that is otherwise isotropic, in contrast to most directional studies. Here, directionality is inferred from the geometric properties of the excitations themselves, rather than from a rate variation. Our results strongly suggest that for a gapless material with typical acoustic phonon modes, no directional correlation is preserved between the initial DM scatter and the outgoing excitations, as phonons can always be emitted in the limit $\Delta \to 0$. In an anisotropic material, some correlations may still persist further above the gap due to an increased number of forbidden transitions. As an example, indirect-gap materials, if such superconductors exist, would be much more likely to preserve directionality even in the limit of a small gap if large energies are required for inter-valley transitions. We leave the exploration of such scenarios for future work [470].

Chapter 10

New constraints from superconducting nanowires

As highlighted in the preceding chapters, after decades of theoretical and experimental focus on DM at the electroweak scale, attention has recently shifted to lighter masses, with sub-GeV DM capturing the limelight from both the theoretical [367–370, 372, 375, 376, 378–380, 421–447] and experimental [381–384, 448–457] perspectives. Direct detection of sub-GeV DM requires detectors with much lower thresholds than traditional experiments, and this has motivated the development of many novel detection techniques. Among the proposed detectors, superconductors [369, 370, 435] stand out: due to their exceptionally small band gaps of $\mathcal{O}(\text{meV})$ and correspondingly small detection thresholds, these materials are capable of detecting light sub-MeV DM. In principle, they are sensitive to the scattering (absorption) of DM with mass as light as $\sim$1 keV ($\sim$1 meV).

Realizing the full potential of superconducting detectors for DM will require

additional technological developments [505]. However, existing devices being used for other applications can already play a meaningful role for dark matter detection. Superconducting nanowire single-photon detectors (SNSPDs) are one such established sensor technology, with numerous applications from quantum sensing to telecommunications (see *e.g.* Refs. [506–508]). These devices are sensitive to the deposit of extremely small amounts of energy, with proven sub-eV thresholds and low dark count rates [435, 509–511, 511–515] and potential to measure the spectrum of energy deposits [516]. Under certain conditions, they may even be sensitive to the direction of the deposited momentum [517]. In Ref. [435], we proposed to apply this mature technology for the first time to the DM hunt by using the SNSPDs simultaneously as the target and for readout: *i.e.*, the SNSPD is both the material with which DM interacts and the sensor that registers the deposited energy and momentum.

In this chapter we report on a 180-hour measurement performed with a prototype SNSPD device that we use to place new bounds on DM, including the strongest terrestrial constraints to date on dark matter with sub-MeV (or sub-eV) masses that scatters with (or is absorbed by) electrons. For the first time, we evaluate bounds using a novel theoretical framework that accounts for the many-body physics of the detector and includes an enhancement due to the thin-layer geometry. Our results represent novel constraints on DM interactions from a superconducting detector system, realizing prospects envisioned nearly a decade ago and providing a new driver for the development of quantum sensing technology. We present a roadmap for the development of future experiments and demonstrate the prospects for SNSPDs to lead exploration of the light DM parameter space. Throughout this chapter we use natural units, where

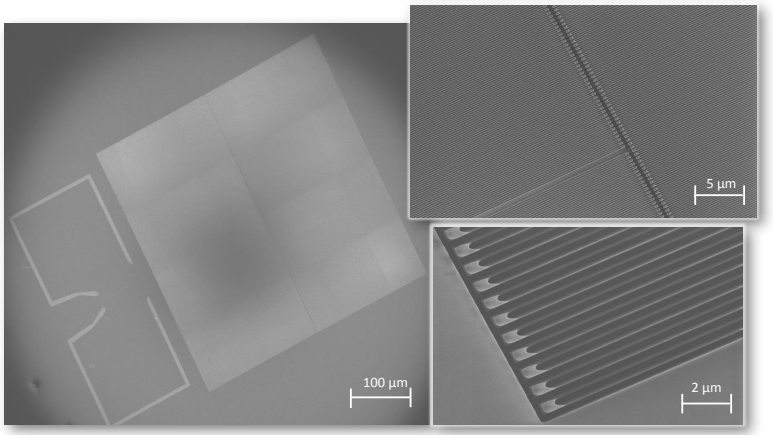

Figure 10.1: SEM images of the prototype WSi SNSPD device taken at different magnifications. *Left:* the entire device with two contact pads and active area of 400 µm by 400 µm. *Top right:* View of the detector area in the center. *Bottom right:* Several individual nanowires.

$c = \hbar = 1$.

10.1 Experimental setup

SNSPDs operate by maintaining a bias current in a superconducting nanowire, keeping the device in the superconducting phase very near the edge of the superconducting transition. Under these conditions, any deposited energy above threshold can cause a portion of the device to undergo a transition to the normal metal phase, locally increasing the resistance of the wire. This results in a brief but significant voltage pulse that can be amplified and then read out. Typical events produce pulses with an amplitude of order 1 mV lasting for several nanoseconds for absorbed energy ranging from 0.1 meV to 10 eV. Further information on energy thresholds and calibration can be found in Ref. [518].

Scanning electron microscope (SEM) images of our prototype device are shown

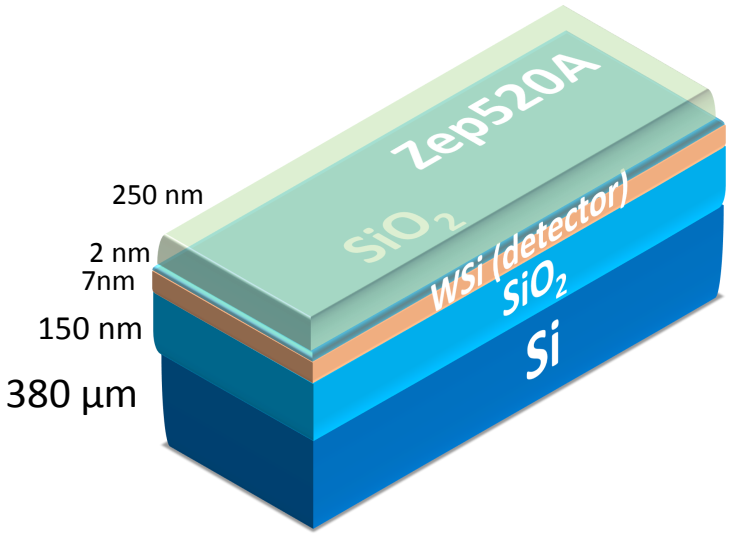

Figure 10.2: Schematic cross section of a single nanowire. Layers are not drawn to scale.

in Fig. 10.1. The device is a square array of nanowires measuring 400 µm on a side, with two contact pads for the readout electronics. Each nanowire in the array measures 140 nm in width, and the spacing between each wire and the next is 200 nm, corresponding to a pitch of 340 nm. Each nanowire consists of several layers, illustrated in Fig. 10.2. The thin tungsten silicide (WSi) layer is the active detector layer, but the other layers still modify the detector response to deposited energy and momentum, as we discuss below. The device was fabricated from a 7 nm-thick WSi film which was sputtered on a 150 nm-thick thermal silicon oxide film on a silicon substrate at room temperature with RF co-sputtering. Additionally, a thin 2 nm Si layer was deposited on top of the WSi film in-situ to prevent oxidation of the superconductor. A layer of ZEP520A, a high performance positive tone electron beam resist, was spin-coated onto the chip at 5000 rpm, which ensured a thickness of 335 nm. The ZEP520A pattern was then transferred to the WSi by reactive ion etching in CF_4 at 50 W. The ZEP520A

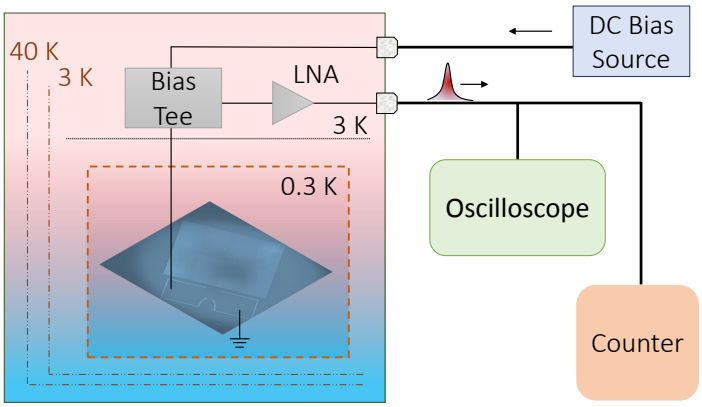

Figure 10.3: Sketch of the experimental setup. The prototype device was embedded in a light-tight box and cooled to a temperature of 0.3 K. The high-frequency signal was carried out of the cryostat though a low-noise cryogenic amplifier to the read-out, while the DC path was connected to a low-noise voltage source. A low-temperature bias tee decoupled the high-frequency path from the DC bias path at the 3 K stage.

thickness is estimated to be 250 nm after etching and is left on the top surface. The prototype device is contained inside a light-tight box at 0.3 K as shown in Fig. 10.3. The signal was amplified at the 3 K stage by cryogenic low-noise amplifiers with a total gain of 56 dB and then sent to a pulse counter. To minimize the effect of blackbody illumination, the optical path was disconnected. The cryostat also has several layers of shielding at the 3 K and 40 K stages. For the science run, the bias current was fixed to 4.5 µA, and the device was exposed for 180 hours, with four dark counts observed. The device threshold is at most 0.73 eV. The observed dark counts may be due to cosmic ray muons, Cherenkov photons generated in the optical setup, or high-energy particles excited by radioactive decay events. The data is further described in Ref. [519], which studies DM absorption in a haloscope configuration.

We use this data to set world-leading bounds on DM–electron interactions, as explained below.

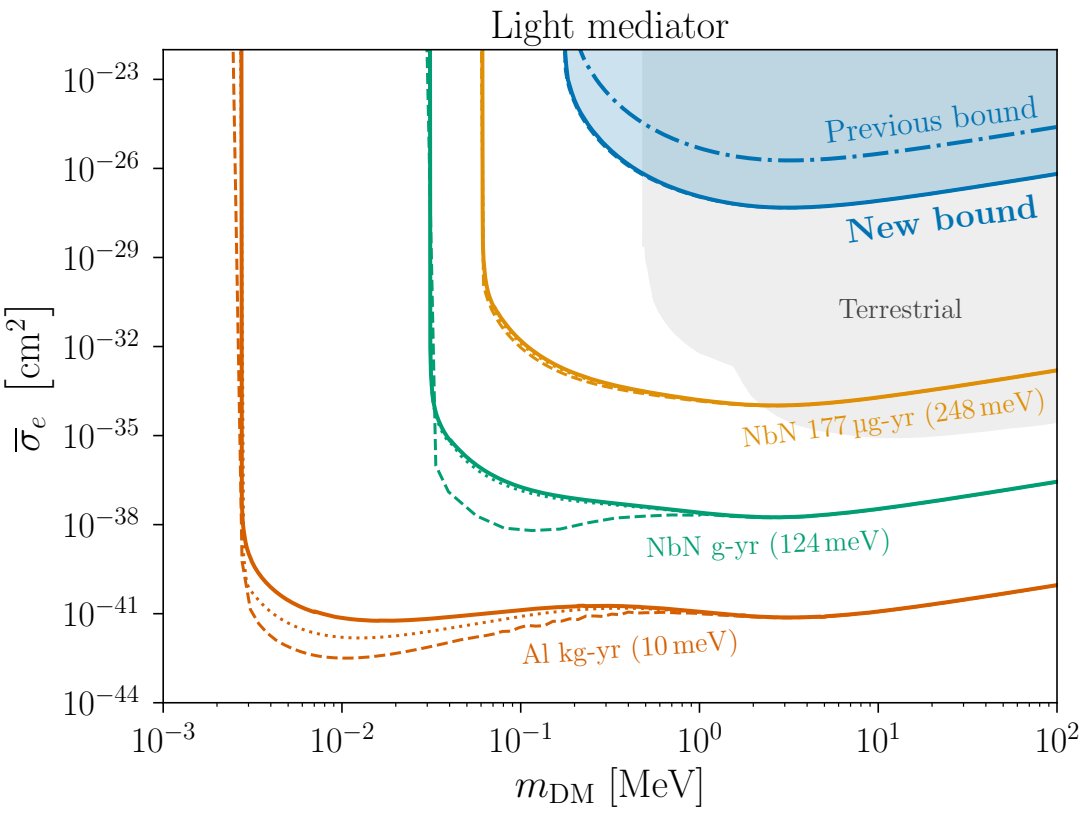

Figure 10.4: New constraints and updated expected reach for DM–electron scattering in SNSPDs via a light mediator at 95% C.L. as a function of DM mass. The shaded blue region indicates the new bound placed by our prototype device with 4.3 ng exposed for 180 hours with four dark counts observed. The dot-dashed blue curve indicates results from our previous run [435] with an exposure of 10 000 seconds, now updated to include in-medium effects. Other curves show the projected reach for WSi, NbN, or Al targets with the indicated exposures and thresholds, assuming that sources of dark counts are eliminated. Solid curves conservatively neglect thin-layer enhancements. Dashed curves include these enhancements following Ref. [520]. Dotted curves conservatively include estimated effects of dissipation in neighboring layers (see text). The 177 µg exposure corresponds to a 10 cm × 10 cm area of NbN at 4 nm thickness and a 50% fill factor, and 248 (124) meV threshold corresponds to a 5 (10) µm wavelength. In shaded gray we show the existing constraints from SENSEI [455], SuperCDMS HVeV [457], DAMIC [384], Xenon10 [380], DarkSide-50 [381], and Xenon1T [454].

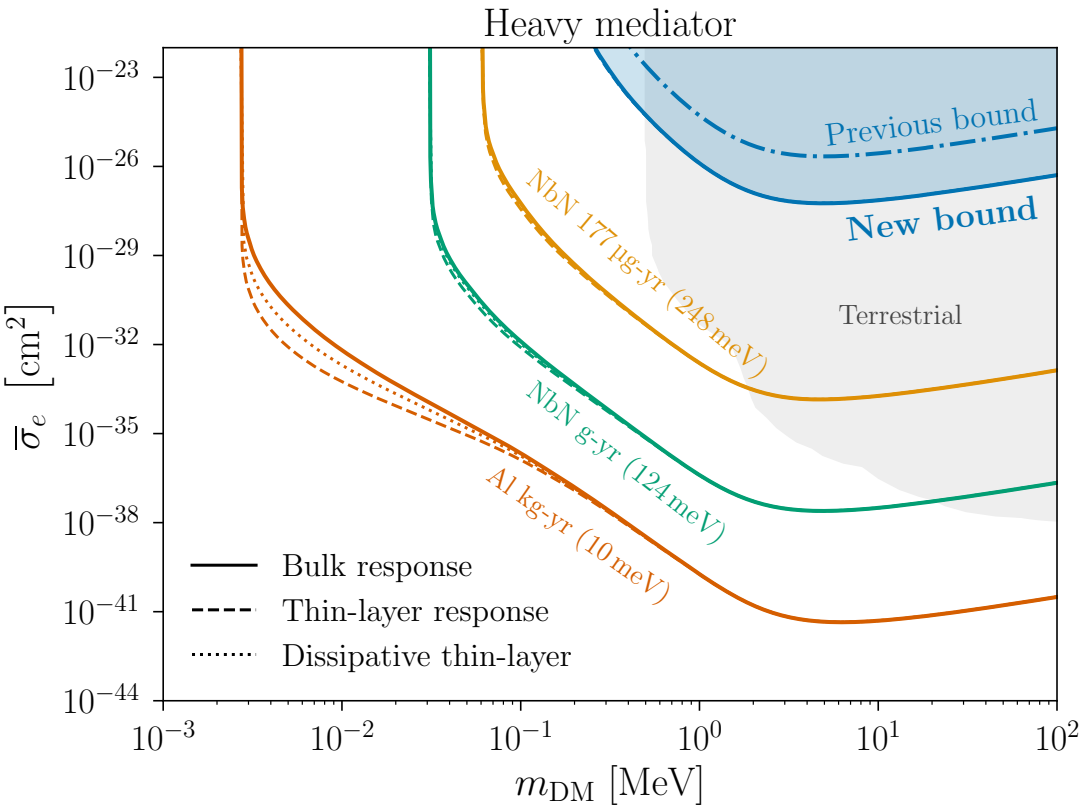

Figure 10.5: New constraints and updated expected reach for DM–electron scattering in SNSPDs via a heavy mediator at 95% C.L. as a function of DM mass. The shaded blue region indicates the new bound placed by our prototype device with 4.3 ng exposed for 180 hours with four dark counts observed. The dot-dashed blue curve indicates results from our previous run [435] with an exposure of 10 000 seconds, now updated to include in-medium effects. Other curves show the projected reach for WSi, NbN, or Al targets with the indicated exposures and thresholds, assuming that sources of dark counts are eliminated. Solid curves conservatively neglect thin-layer enhancements. Dashed curves include these enhancements following Ref. [520]. Dotted curves conservatively include estimated effects of dissipation in neighboring layers (see text). The 177 µg exposure corresponds to a 10 cm × 10 cm area of NbN at 4 nm thickness and a 50% fill factor, and 248 (124) meV threshold corresponds to a 5 (10) µm wavelength. In shaded gray we show the existing constraints from SENSEI [455], SuperCDMS HVeV [457], DAMIC [384], Xenon10 [380], DarkSide-50 [381], and Xenon1T [454].

10.2 DM interaction rate

The concept of our experiment is that local DM particles may interact with the electrons in an SNSPD. In this case, a DM particle may occasionally exchange sufficient energy with these electrons to overcome the threshold of the detector, producing a count in the device when no other sources are present. In order to translate rate measurements of an SNSPD device to bounds on the DM–electron interactions, for both scattering and absorption processes, it is necessary to compute the rates of these processes in the detector.

10.2.1 Bulk interaction rate

For small energy and momentum transfers, electrons in the detector cannot be considered free particles, and the many-body physics of the target material becomes important. We compute the DM interaction rates using a new theoretical method recently developed by Ref. [488] (see also Ref. [521]). This technique is based on the dielectric response of the target material as characterized by its loss function, and naturally incorporates the many-body physics of the detector, eliminating substantial uncertainties associated with first-principles approaches. The key input quantity, the *dielectric function,* can be either measured experimentally or computed theoretically using established models from condensed matter physics.

The loss function (or, equivalently, the dielectric function) is readily measured by X-ray or electron scattering in the relevant regime of energy and momentum transfers. However, to our knowledge, no data is yet available for the loss function in WSi at the

relevant values of $\mathbf{q}$ and ω. Therefore, in this chapter, we compute the loss function using the well established Lindhard model [468]. In the Lindhard model, also known as the random phase approximation or the free electron gas model, the dielectric function can be written in closed form in the low-temperature limit as

$$\epsilon_{\rm L}(\mathbf{q},\omega) = 1 + \frac{3\omega_{\rm p}^2}{q^2 v_{\rm F}^2}\left\{\frac{1}{2} + \frac{k_{\rm F}}{4q}\left(1-Q_-^2\right)\mathrm{Log}\left(\frac{Q_-+1}{Q_--1}\right) + \frac{k_{\rm F}}{4q}\left(1-Q_+^2\right)\mathrm{Log}\left(\frac{Q_++1}{Q_+-1}\right)\right\}, \tag{10.1}$$

where $\omega_{\rm p} = (4\pi\alpha n_{\rm e}/m_{\rm e})^{1/2}$ is the plasma frequency, for $n_{\rm e}$ the number density of electrons; $k_{\rm F}$ is the Fermi momentum; $v_{\rm F} = k_{\rm F}/m_{\rm e}$ is the Fermi velocity; and $Q_\pm = q/(2k_{\rm F}) \pm \omega/(qv_{\rm F})$. The Lindhard dielectric function exhibits a resonance at the plasma frequency $\omega_{\rm p}$. In the form above, this resonance is present but infinitely narrow. A non-zero width is obtained under the replacement $\omega \to \omega + i/\tau$, where the excitation lifetime τ can be fitted to experimental data. Such a width may enhance the loss function at deposits very far from the peak of the resonance [488]. In this chapter, we estimate $1/\tau = \frac{1}{10}\omega_{\rm p}$, a typical width for a metal.

We consider both DM scattering and absorption processes. For DM scattering, we place limits on the DM–electron scattering cross section. These hold for any spin-independent interaction that couples the DM to the electron density [488], including both scalar and vector mediators. For absorption, we consider a fiducial theory of a dark photon A'_μ, with field strength $F'_{\mu\nu} \equiv \partial_\mu A'_\nu - \partial_\nu A'_\mu$, kinetically mixed with the Standard Model photon. That is, we assume a Lagrangian of the form

$$\mathcal{L} \supset -\tfrac{1}{2}\kappa F_{\mu\nu}F'^{\mu\nu}. \tag{10.2}$$

The absorption rate per unit volume can then be written as

$$\Gamma_{\mathrm{A}} = \kappa^2 m_\chi \mathcal{W}(\mathbf{p}_\chi, m_\chi)\,, \tag{10.3}$$

where m_χ is the DM mass and $\mathbf{p}_\chi$ is the momentum of the incoming DM particle. The kinetic mixing parameter κ is the quantity that we bound in our experiment (see Fig. 10.6).

10.2.2 Interaction rate in a thin layer

Each unit of our prototype detector is composed of a stack of thin layers of different materials, as illustrated in Fig. 10.2. For a low-dimensional target system, or for heterogeneous systems with interfaces, the dielectric response of the detector is different from that of a bulk sample of material, and these differences should be accounted for in the rate. These effects are newly explored in Ref. [520], which derives the DM interaction rate in a thin layer. In particular, if the layer width is small compared to the inverse momentum transfer in the interaction, the response of the layer itself is significantly modified, and features a new resonance for small energy deposits. Thus, the DM scattering rate per unit volume for a thin layer can be enhanced significantly with respect to a bulk detector.

Preliminary estimates suggest that the absorption rate is subject to even larger enhancements, but the approach of Ref. [520] cannot be directly applied in this kinematic regime, where the deposited momentum is much smaller than the deposited energy. We do not quantify this enhancement in this chapter, but leave this as a task for future experimental characterization.

The thin-layer interaction rate derived by Ref. [520] assumes that the detector layer is the only dissipative component of the system, such that energy deposited in any other layer is eventually dissipated there. However, experimental characterization of our prototype detector suggests that dissipation in the other layers is in fact significant: only large deposits far above the threshold in the other layers produce measurable events in the WSi layer. Thus, in what follows, we also show a conservative result that includes dissipation in all layers, and neglects deposits outside the detector layer. Further details are given in Appendix E. Our treatment yields a conservative bound on DM–electron interactions compared to what could be achieved with more complete knowledge of the prototype device response. Future study of the prototype nanowire to accurately characterize sensitivity to energy deposits outside the WSi layer, as a function of their magnitude and location, will allow for even stronger DM limits.

Each nanowire contains layers of Si and SiO_2 in addition to WSi. The dielectric function of Si can be approximated using the Lindhard model with $E_F = 18.9\,\mathrm{eV}$, which originates from a phenomenological fit [488]. For SiO_2, we use the fit provided by Ref. [522]. We model the ZEP520A top layer with a constant and real dielectric function, taking the (real) index of refraction to be $\sim$1.5.

10.3 Results

Our new constraints are summarized in Figs. 10.4 and 10.5 for DM–electron scattering with light and heavy mediators (left and right panels, respectively), and in Fig. 10.6 for DM absorption. Existing terrestrial constraints are shown in shaded

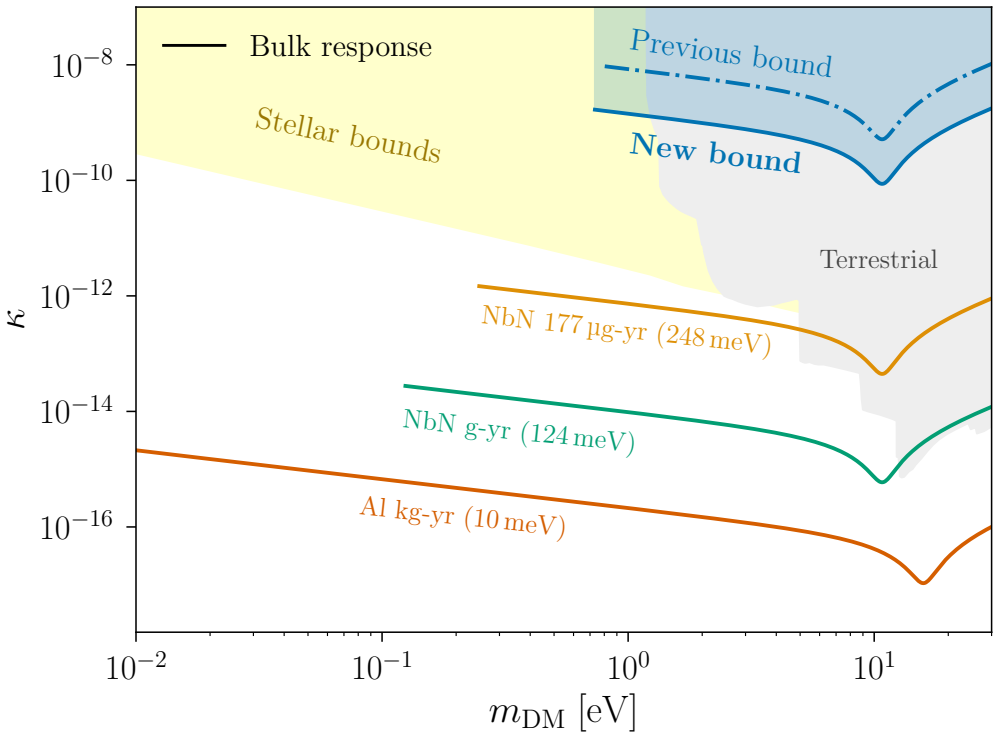

Figure 10.6: New constraints and updated expected reach for DM absorption in SNSPDs as a function of DM mass, for a relic kinetically mixed dark photon. As in Figs. 10.4 and 10.5, the shaded blue region indicates the new bound at 95% C.L., and other solid curves indicate projections for future experiments, neglecting possible geometric effects. The shaded gray region shows existing terrestrial constraints from Xenon data [523], SuperCDMS [382], DAMIC [384], EDELWEISS [456], FUNK [524] and SENSEI [455], while the yellow region indicates model-dependent stellar bounds [523, 525, 526].

gray, and model-dependent stellar constraints are shown in yellow. (Other model-dependent cosmological constraints may also apply; see *e.g.* Refs. [527–529].) Our previous nanowire bounds [435], updated to incorporate in-medium effects via the dielectric formalism, are indicated by dot-dashed blue curves. Notably, our prototype detector already provides the strongest constraints to date on the electronic interactions of sub-MeV (sub-eV) DM via scattering (absorption) processes, with an exposure of only $4.3\,\mathrm{ng} \times 180\,\mathrm{h}$ or equivalently $8.8 \times 10^{-14}\,\mathrm{kg\,yr}$. We also show projections for future SNSPD experiments with larger exposures in NbN and Al detectors. All bounds and projections are given at 95% confidence level (C.L.) for one-sided Poisson statistics and computed using the Lindhard model for the dielectric function [468], which agrees well with available measurements at zero momentum transfer.

Scattering results are shown in terms of a reference scattering cross section $\bar{\sigma}_e = \frac{1}{\pi}\mu_{e\chi}^2 g_e^2 g_\chi^2 \left[(\alpha_{\mathrm{EM}} m_e)^2 + m_\phi^2\right]^{-2}$, where $\mu_{e\chi}$ is the reduced mass of the DM–electron system; g_e and g_χ are the couplings of the mediator to the electron and DM, respectively; and $\alpha_{\mathrm{EM}} \approx 1/137$ is the fine structure constant. Absorption results are shown in terms of the size of the kinetic mixing κ of a dark photon—essentially its coupling to the electromagnetic current. We take the Fermi energy E_{F} to be 7 eV in both WSi and NbN, and we take the densities to be $9.3\,\mathrm{g/cm^3}$ and $8.4\,\mathrm{g/cm^3}$, respectively. The Fermi energy and density of Al are taken to be 11.7 eV and $2.7\,\mathrm{g/cm^3}$, respectively. We assume a local DM density of $0.3\,\mathrm{GeV/cm^3}$ with velocities distributed according to the Standard Halo Model, *i.e.,* with probability density $f_\chi(\mathbf{v}) \propto \Theta(v_{\mathrm{esc}} - |\mathbf{v}|) \exp\left[-(\mathbf{v} + \mathbf{v}_{\mathrm{E}})^2/v_0^2\right]$. We take $v_0 = 220\,\mathrm{km/s}$, $v_{\mathrm{E}} = 232\,\mathrm{km/s}$, and $v_{\mathrm{esc}} = 550\,\mathrm{km/s}$.

The impressive reach for scattering and absorption at the smallest masses is

due to the low device threshold of 0.73 eV, assisted by its low dark count rate. Future realizations of this experiment may be able to achieve substantially lower thresholds, sensitive to much lower masses. The projections for the reach of future NbN detectors assume thresholds of 248 and 124 meV, which would extend the experimental reach to DM masses of order 50–100 keV. Indeed, sensitivity at the 10 µm-wavelength scale—corresponding to a 124 meV threshold—has already been demonstrated in SNSPDs [509]. We also show the projected reach for a superconducting Al detector with a 10 meV threshold. Such a detector would be capable of detecting DM with mass of order ∼keV, below which structure formation considerations rule out fermionic DM [364, 489, 490].

Solid curves are computed neglecting thin-layer effects, *i.e.*, treating the detector as a bulk volume. Dashed and dotted curves show the projections including these effects: dashed curves neglect dissipation in the other layers, following Ref. [520], while dotted lines incorporate this dissipation in the most conservative form. (See Appendix E for details.) Geometric effects do not significantly affect the reach of the constraints for the current experimental configuration, but these effects are an important consideration for future experimental design: thin-layer effects were not exploited in the original design of the prototype, and have arisen incidentally from the necessarily low-dimensional structure of SNSPDs. Sensitivity of the WSi detector layer to deposits in other layers of the device may allow for enhanced reach even at high DM masses, effectively increasing the detector volume. Such sensitivity may be possible for deposits far above threshold, and could be quantified experimentally. Deliberate optimization of the target geometry may enable even more significant enhancements, particularly in the absorption rate.

The geometric effects included in this chapter are estimated in a simplified

framework. We do not quantify the geometric effects on the absorption rate here, and in the case of scattering, additional corrections may arise from the lower layers of the geometry in Fig. 10.2 or from local-field corrections [530, 531]. The accurate impact of the geometry of the device on the DM interaction rate can be quantified experimentally in a robust manner, and is expected to further improve the reach.

10.4 Discussion

We have reported on a new search for DM–electron scattering and absorption in a prototype SNSPD detector. Our results place the strongest terrestrial constraints to date on DM–electron interactions for sub-MeV (sub-eV) masses for scattering (absorption) processes. This is the first time that superconducting detectors have been used to probe unconstrained parameter space for DM scattering, a crucial milestone in the program of light DM searches that heralds significant collaboration between the DM and quantum-sensing communities. The constraints presented in this chapter are computed using the dielectric function formalism, accounting for the many-body physics of the detector material, and we have also accounted for geometric effects that can significantly enhance the predicted DM interaction rate.

Our small-scale prototype is able to exceed previous experimental constraints thanks to the remarkably low 0.73 eV threshold of the SNSPD detector, along with its extremely low dark count rate. Future iterations of this experiment promise to reach even lower thresholds with even lower dark count rates. At present, we place constraints on DM interactions assuming that the dark counts are due to backgrounds. In the future,

experimental improvements will allow the use of rate modulation [532, 533] and possibly even spectroscopic measurements [516] to differentiate between backgrounds and a DM signal. The SNSPD platform is being heavily developed for numerous applications in quantum sensing and precision metrology, and given the rapid pace of development, Figs. 10.4 to 10.6 can be treated as a realistic indication of the reach of future experiments. The Al projections, with their 10 meV thresholds, represent an ambitious target: achieving such thresholds will require considerable technological development, but there is no fundamental obstacle to constructing such a device.

An additional important challenge is to scale the prototype device to a large-scale experiment. Thus far, SNSPD devices are small: our nanogram-scale prototype is typical. Sensitivity to cross sections as small as those probed by experiments at higher DM masses will require significantly larger detectors at the gram scale and beyond. While the electron lithography techniques used to fabricate our prototype do not scale easily to larger devices, it is possible that optical lithography or other technologies would enable the production of a larger detector.

Finally, future experiments will be in a position to leverage geometric enhancements to the interaction rate. Our prototype detector was designed to demonstrate the capabilities of SNSPDs for DM detection with existing technology and fabrication techniques, and such geometric enhancements were not a design consideration. However, the theoretical methods introduced by Refs. [488, 520, 521] make it possible to accurately compute these geometric effects when designing future detectors. The phenomenology of thin layers and interfaces has been studied thoroughly in the condensed matter literature, and this should allow for the fabrication of designer materials or heterostructures

with highly customized dielectric responses. Such materials could feature even larger geometric enhancements to the DM interaction rate, allowing near-future experiments to delve deep into uncharted parameter space.

Closing remarks

The bright future of dark matter science

I began this thesis by laying out the perilous position of dark matter phenomenology: the field has spent decades preparing for a future that may never be realized. The viability of the WIMP paradigm continues to dwindle, and we must now confront the vast space of alternative models. If nature has been unkind, it is entirely possible that the identity of the DM species will be experimentally inaccessible to us for centuries. Amidst this context, I hope to close this thesis by justifying my sense of tremendous optimism for the future study of DM, cosmology, and associated new physics. Notice that I have reverted to the first-person singular: these remarks are strictly my opinion, and my poor predictions of the future should not tarnish my collaborators.

First, to set the record straight, any pessimism over the future of particle physics is extremely premature. After all, the crisis that we face now is less than ten years old. There are numerous reasonable scenarios in which significant discoveries really do lie around the corner. I point to history: physics has been apparently stagnant for long periods in the past, but nonetheless, we have rarely gone more than a few decades without revolutionary progress.

However, in taking lessons from history, we should be attentive to the typical

modes of progress after impasse. One lesson is widely acknowledged both in and out of the field, almost to the point of cliché: that scientific progress is not smooth, but follows a series of punctuated equilibria, with tensions slowly becoming crises that explode into Kuhnian revolutions. This has become such a strong expectation for the resolution of long-standing scientific mysteries that DM phenomenology has become a frequent target of armchair philosophers of science. If I had a nickel for every time I explained to a skeptical layperson that particle DM is still the best hypothesis, I would have earned a sum comparable to the UC Santa Cruz stipend after the average student's rent.

A less trumpeted conclusion from recent history is that the progress of science tracks the development of *tools* more closely than the development of ideas. Most scientific revolutions can be traced to some new discovery, often a completely unexpected finding, that was enabled by new instrumentation rather than brilliant theory. For example, thermodynamics arose as a science in parallel with the development of practical heat engines in the 18th and 19th centuries. Relativity and quantum mechanics were each formulated within years of the key experimental inputs. The Standard Model of particle physics was incrementally constructed as colliders reached higher and higher scales. Even Newton's development of classical mechanics quickly followed and built on advances by Kepler and Galileo, whose work was in turn enabled by the newly-invented telescope.

This is understandably an uncomfortable notion: it suggests that there are severe limits on what we can accomplish with sheer cleverness in the absence of experimental methods. More subtly, it stands in opposition to the "great man" theory of scientific history, since it asserts that the accomplishments of these great men had more

to do with their time and place than with their unique individual characters. To be clear, there were certainly rare geniuses who played instrumental roles in scientific history—I would not dispute that. But it should be clear that if any of these rare geniuses had never lived, scientific progress would not have been long delayed. (For instance, if anyone would claim that Newton was irreplaceable in scientific history, Messrs. Hooke and Leibniz would famously beg to differ.)

What does this have to do with the future of DM science? To me, it suggests that the most pragmatic path forward is to bring new tools to bear on the problem, and to rapidly explore wide swathes of parameter space, looking for a *surprise* of any kind. Formal theory and phenomenology for its own sake is extremely important, but we should recognize that our field has vast reservoirs of creativity that are more than adequate to the task of positing explanations for the phenomena that we already observe. Indeed, this underlies what some consider to be a distressing overproduction of theoretical results in particle physics. The challenge for the future of DM, and particle physics more generally, is to connect particle physics model space with a broad new set of experimental and observational tools.

This is a fundamentally different exercise from the bread and butter of the last four decades. In the past, the field had a very strong prior for the nature of DM, and a very clear idea of what would constitute the next step in its characterization: the detection of DM particles in an laboratory experiment, perhaps, or a clear collider signature. There was a straight and narrow path towards writing down the Lagrangian of the dark sector. Even at my most optimistic, I admit that we may still be many years away from accomplishing that feat. But such a detailed characterization of DM

microphysics does not need to be our immediate goal. Generic constraints are still extremely powerful, and ultimately shape our concept of DM.

However, such a broad goal presents significant challenges for the design and implementation of experiments. What can we use to build our prior for "reasonable" DM models? After all, broadening the search does not mean discarding all scruples. Simple, concretely realizable models are still the best motivated. The laziness of nature is our best guiding principle, even if scientific history does not adhere to it as a rule. But parsimony does not substantially distinguish many classes of "miraculous" models that lie across the viable DM mass range, and that is precisely why the next generation of experiments must search for a broad variety of representative models across the spectrum.

(Here it is worth remembering that parsimony provides one of the strongest pieces of motivation to search for DM in the first place. Consider that we already know a tremendous amount about the DM species from astrophysics and cosmology. Why bother pursuing its mass, spin, and couplings, other than for academic reasons? The most direct answer is that we would be extremely surprised to find that the nature of DM has nothing to do with any of the other open problems in particle physics and cosmology. That, too, is parsimony in action.)

With these grand principles in mind, what I have attempted to do in this thesis is to advance the connective tissue between new tools and new physics. I have argued in the foregoing chapters that new astronomical observables, new gravitational wave probes, and new quantum sensing capabilities open up powerful opportunities to probe DM models across the spectrum. Ultimately, I hope that one of these new observables

will find something surprising—perhaps something not even connected directly to DM, but nonetheless something that spurs a new set of questions.

Looking at recent history, I would be astonished if no grand surprise appeared on the timescale of my career—and that is what underlies my optimism. I do not know if I will live to know the mass, spin, and couplings of the DM particle. But I do expect to see new positive qualitative statements about the nature of DM, and I do expect to see the field reorganize around new and promising lines of inquiry.

And now, I must emphasize that one such grand surprise might be the revival of the SUSY WIMP. Despite general pessimism in the field, the parameter space is extremely broad, extremely complicated, and far from exhausted. SUSY remains an remarkably well-motivated paradigm, both in general and at the weak scale specifically. My personal focus on broadening the search for DM is not blind to this possibility. Indeed, if WIMPs are discovered, that will be the *start* of an exciting era, and by no means the end of a broad search. The WIMP theory space is sufficiently complicated that broad tools will still serve an important role even if these waning ideas turn out to be substantially correct. Such a development would put us back in the shoes of particle physicists from the 1970s and 1980s, in exciting times when our tools turned out to be so much more capable than we had a right to hope.

Finally, I stress that the tools discussed in this thesis are only a small subset of the new tools that may be available to us in the near future. Since methods development is driven by adancements in parallel fields, it is interesting to consider the DM applications of areas that are advancing rapidly today. Notably, the coming decades should see extraordinary advances in both quantum information science and machine

learning, and these are already quietly revolutionizing many areas of physics. Methods from quantum information science have already made their way into direct detection techniques, and machine learning undergirds many analysis methods in collider experiments and astronomical surveys. It is a sure bet that we have yet to realize the full power of these and other nascent tools in the DM community.

So, despite the viability of nightmare scenarios, I remain extremely optimistic. DM physics is undergoing a kind of phase transition: in the coming years and decades, a vast range of uncharted territory will be probed by new kinds of experiments, and new methodological ideas will transform our search. I cannot predict what these new tools will discover, and I hardly dare to hope that they will serve us the DM Lagrangian on a platter. But I do hope and believe with reasonable confidence that interesting surprises are not far down the road.

Bibliography

[1] F. Kahlhoefer, Int. J. Mod. Phys. **A32**, 1730006 (2017), arXiv:1702.02430 [hep-ph] .

[2] R. Agnese *et al.* (SuperCDMS), Phys. Rev. Lett. **120**, 061802 (2018), arXiv:1708.08869 [hep-ex] .

[3] E. Aprile *et al.* (XENON), Phys. Rev. Lett. **121**, 111302 (2018), arXiv:1805.12562 [astro-ph.CO] .

[4] X. Cui *et al.* (PandaX-II), Phys. Rev. Lett. **119**, 181302 (2017), arXiv:1708.06917 [astro-ph.CO] .

[5] G. Arcadi, M. Dutra, P. Ghosh, M. Lindner, Y. Mambrini, M. Pierre, S. Profumo, and F. S. Queiroz, Eur. Phys. J. **C78**, 203 (2018), arXiv:1703.07364 [hep-ph] .

[6] Y. B. Zel'dovich and I. D. Novikov, Soviet Ast. **10**, 602 (1967).

[7] S. Hawking, Monthly Notices of the Royal Astronomical Society **152**, 75 (1971), https://academic.oup.com/mnras/article-pdf/152/1/75/9360899/mnras152-0075.pdf .

[8] A. de Lavallaz and M. Fairbairn, Phys. Rev. D **81**, 123521 (2010), arXiv:1004.0629 [astro-ph.GA] .

[9] S. Shandera, D. Jeong, and H. S. G. Gebhardt, Phys. Rev. Lett. **120**, 241102 (2018), arXiv:1802.08206 [astro-ph.CO] .

[10] R. A. Allsman *et al.* (Macho), Astrophys. J. Lett. **550**, L169 (2001), arXiv:astro-ph/0011506 .

[11] S. Bird, I. Cholis, J. B. Muñoz, Y. Ali-Haïmoud, M. Kamionkowski, E. D. Kovetz, A. Raccanelli, and A. G. Riess, Phys. Rev. Lett. **116**, 201301 (2016), arXiv:1603.00464 [astro-ph.CO] .

[12] M. Sasaki, T. Suyama, T. Tanaka, and S. Yokoyama, Phys. Rev. Lett. **117**, 061101 (2016), [Erratum: Phys.Rev.Lett. 121, 059901 (2018)], arXiv:1603.08338 [astro-ph.CO] .

[13] H. Niikura *et al.*, Nature Astron. **3**, 524 (2019), arXiv:1701.02151 [astro-ph.CO] .

[14] S. Sugiyama, T. Kurita, and M. Takada, (2019), arXiv:1905.06066 [astro-ph.CO] .

[15] N. Smyth, S. Profumo, S. English, T. Jeltema, K. McKinnon, and P. Guhathakurta, Phys. Rev. D **101**, 063005 (2020), arXiv:1910.01285 [astro-ph.CO] .

[16] B. J. Carr, *Proceedings, 271st WE-Heraeus Seminar on Aspects of Quantum Gravity: From Theory to Experiment Search: Bad Honnef, Germany, February 25-March 1, 2002*, Lect. Notes Phys. **631**, 301 (2003), [,301(2003)], arXiv:astro-ph/0310838 [astro-ph] .

[17] M. Y. Khlopov, Res. Astron. Astrophys. **10**, 495 (2010), arXiv:0801.0116 [astro-ph] .

[18] X. Calmet, B. Carr, and E. Winstanley, *Quantum Black Holes*, SpringerBriefs in Physics (Springer, Berlin, 2014).

[19] J. García-Bellido, B. Carr, and S. Clesse, Universe **8**, 12 (2021), arXiv:1904.11482 [astro-ph.CO] .

[20] J. D. Barrow, E. J. Copeland, E. W. Kolb, and A. R. Liddle, Phys. Rev. D **43**, 984 (1991).

[21] T. Fujita, M. Kawasaki, K. Harigaya, and R. Matsuda, Phys. Rev. D **89**, 103501 (2014), arXiv:1401.1909 [astro-ph.CO] .

[22] A. Hook, Phys. Rev. D **90**, 083535 (2014), arXiv:1404.0113 [hep-ph] .

[23] A. Boudon, B. Bose, H. Huang, and L. Lombriser, Phys. Rev. D **103**, 083504 (2021), arXiv:2010.14426 [astro-ph.CO] .

[24] N. Smyth, L. Santos-Olmsted, and S. Profumo, JCAP **03** (03), 013, arXiv:2110.14660 [hep-ph] .

[25] L. Morrison, S. Profumo, and Y. Yu, JCAP **05**, 005, arXiv:1812.10606 [astro-ph.CO] .

[26] A. Coogan, L. Morrison, and S. Profumo, Phys. Rev. Lett. **126**, 171101 (2021), arXiv:2010.04797 [astro-ph.CO] .

[27] W. DeRocco and P. W. Graham, Phys. Rev. Lett. **123**, 251102 (2019), arXiv:1906.07740 [astro-ph.CO] .

[28] B. J. Carr, K. Kohri, Y. Sendouda, and J. Yokoyama, Phys. Rev. D **81**, 104019 (2010), arXiv:0912.5297 [astro-ph.CO] .

[29] B. Carr, F. Kuhnel, and M. Sandstad, Phys. Rev. D **94**, 083504 (2016), arXiv:1607.06077 [astro-ph.CO] .

[30] B. V. Lehmann, S. Profumo, and J. Yant, JCAP **1804** (04), 007, arXiv:1801.00808 [astro-ph.CO] .

[31] B. Carr, K. Kohri, Y. Sendouda, and J. Yokoyama, (2020), arXiv:2002.12778 [astro-ph.CO] .

[32] A. M. Green and B. J. Kavanagh, J. Phys. G **48**, 043001 (2021), arXiv:2007.10722 [astro-ph.CO] .

[33] B. Carr and F. Kuhnel, (2020), arXiv:2006.02838 [astro-ph.CO] .

[34] K. Griest, A. M. Cieplak, and M. J. Lehner, Astrophys. J. **786**, 158 (2014), arXiv:1307.5798 [astro-ph.CO] .

[35] P. Tisserand *et al.* (EROS-2), Astron. Astrophys. **469**, 387 (2007), arXiv:astro-ph/0607207 .

[36] Y. Ali-Haïmoud and M. Kamionkowski, Phys. Rev. D **95**, 043534 (2017), arXiv:1612.05644 [astro-ph.CO] .

[37] M. A. Monroy-Rodríguez and C. Allen, Astrophys. J. **790**, 159 (2014), arXiv:1406.5169 [astro-ph.GA] .

[38] F. Kühnel and K. Freese, Phys. Rev. D **95**, 083508 (2017), arXiv:1701.07223 [astro-ph.CO] .

[39] B. Carr, M. Raidal, T. Tenkanen, V. Vaskonen, and H. Veermäe, Phys. Rev. D **96**, 023514 (2017), arXiv:1705.05567 [astro-ph.CO] .

[40] N. Bellomo, J. L. Bernal, A. Raccanelli, and L. Verde, JCAP **01**, 004, arXiv:1709.07467 [astro-ph.CO] .

[41] K. Inomata, M. Kawasaki, K. Mukaida, Y. Tada, and T. T. Yanagida, Phys. Rev. D **96**, 043504 (2017), arXiv:1701.02544 .

[42] S. Clesse and J. García-Bellido, Phys. Rev. D **92**, 023524 (2015), arXiv:1501.07565 .

[43] P. Wolfe, Mathematical Programming **11**, 128 (1976).

[44] A. Barnacka, J. F. Glicenstein, and R. Moderski, Phys. Rev. D **86**, 043001 (2012), arXiv:1204.2056 [astro-ph.CO] .

[45] S. M. Koushiappas and A. Loeb, Phys. Rev. Lett. **119**, 041102 (2017), arXiv:1704.01668 [astro-ph.GA] .

[46] T. D. Brandt, Astrophys. J. Lett. **824**, L31 (2016), arXiv:1605.03665 [astro-ph.GA] .

[47] P. W. Graham, S. Rajendran, and J. Varela, Phys. Rev. D **92**, 063007 (2015), arXiv:1505.04444 [hep-ph] .

[48] F. Capela, M. Pshirkov, and P. Tinyakov, Phys. Rev. D **87**, 123524 (2013), arXiv:1301.4984 [astro-ph.CO] .

[49] M. Zumalacarregui and U. Seljak, Phys. Rev. Lett. **121**, 141101 (2018), arXiv:1712.02240 [astro-ph.CO] .

[50] J. Garcia-Bellido, S. Clesse, and P. Fleury, Phys. Dark Univ. **20**, 95 (2018), arXiv:1712.06574 [astro-ph.CO] .

[51] R. Saito and J. Yokoyama, Int. J. Mod. Phys. Conf. Ser. **01**, 126 (2011).

[52] K. Ioka, T. Tanaka, and T. Nakamura, Phys. Rev. D **60**, 083512 (1999), arXiv:astro-ph/9809395 .

[53] S. Clesse and J. García-Bellido, Phys. Dark Univ. **18**, 105 (2017), arXiv:1610.08479 [astro-ph.CO] .

[54] Z.-C. Chen and Q.-G. Huang, Astrophys. J. **864**, 61 (2018), arXiv:1801.10327 [astro-ph.CO] .

[55] A. Barnacka, *Detection techniques for the H.E.S.S. II telescope, data modeling of gravitational lensing and emission of blazars in HE-VHE astronomy*, Other thesis (2013), arXiv:1307.4050 [astro-ph.HE] .

[56] C. J. Moore, R. H. Cole, and C. P. L. Berry, Classical and Quantum Gravity **32**, 015014 (2015).

[57] S. Udry and N. C. Santos, ARA&A **45**, 397 (2007).

[58] J. Schneider, C. Dedieu, P. Le Sidaner, R. Savalle, and I. Zolotukhin, Astron. Astrophys. **532**, A79 (2011), arXiv:1106.0586 [astro-ph.EP] .

[59] J. T. Wright, O. Fakhouri, G. W. Marcy, E. Han, Y. Feng, J. A. Johnson, A. W. Howard, J. A. Valenti, J. Anderson, and N. Piskunov, Publ. Astron. Soc. Pac. **123**, 412 (2011), arXiv:1012.5676 [astro-ph.SR] .

[60] A. Cassan *et al.*, Nature **481**, 167 (2012), arXiv:1202.0903 [astro-ph.EP] .

[61] R. Akeson *et al.*, Publ. Astron. Soc. Pac. **125**, 989 (2013), arXiv:1307.2944 [astro-ph.IM] .

[62] S. E. Thompson, J. L. Coughlin, K. Hoffman, F. Mullally, J. L. Christiansen, C. J. Burke, S. Bryson, N. Batalha, M. R. Haas, J. Catanzarite, J. F. Rowe, G. Barentsen, D. A. Caldwell, B. D. Clarke, J. M. Jenkins, J. Li, D. W. Latham, J. J. Lissauer, S. Mathur, R. L. Morris, S. E. Seader, J. C. Smith, T. C. Klaus, J. D. Twicken, J. E. Van Cleve, B. Wohler, R. Akeson, D. R. Ciardi, W. D. Cochran, C. E. Henze, S. B. Howell, D. Huber, A. Prša, S. V. Ramírez, T. D. Morton, T. Barclay, J. R. Campbell, W. J. Chaplin, D. Charbonneau, J. Christensen-Dalsgaard, J. L. Dotson, L. Doyle, E. W. Dunham, A. K. Dupree, E. B. Ford, J. C. Geary, F. R. Girouard, H. Isaacson, H. Kjeldsen, E. V. Quintana, D. Ragozzine,

M. Shabram, A. Shporer, V. Silva Aguirre, J. H. Steffen, M. Still, P. Tenenbaum, W. F. Welsh, A. Wolfgang, K. A. Zamudio, D. G. Koch, and W. J. Borucki, ApJS **235**, 38 (2018), arXiv:1710.06758 [astro-ph.EP] .

[63] R. K. Leane and J. Smirnov, Phys. Rev. Lett. **126**, 161101 (2021), arXiv:2010.00015 [hep-ph] .

[64] Y.-D. Tsai, Y. Wu, S. Vagnozzi, and L. Visinelli, (2021), arXiv:2107.04038 [hep-ph] .

[65] L. Wyrzykowski *et al.*, Mon. Not. Roy. Astron. Soc. **416**, 2949 (2011), arXiv:1106.2925 [astro-ph.GA] .

[66] H. Niikura, M. Takada, S. Yokoyama, T. Sumi, and S. Masaki, Phys. Rev. D **99**, 083503 (2019), arXiv:1901.07120 [astro-ph.CO] .

[67] B. Abbott *et al.* (LIGO Scientific), Phys. Rev. **D72**, 082002 (2005), arXiv:gr-qc/0505042 [gr-qc] .

[68] R. Magee, A.-S. Deutsch, P. McClincy, C. Hanna, C. Horst, D. Meacher, C. Messick, S. Shandera, and M. Wade, Phys. Rev. **D98**, 103024 (2018), arXiv:1808.04772 [astro-ph.IM] .

[69] B. Abbott *et al.* (LIGO Scientific, Virgo), Phys. Rev. Lett. **123**, 161102 (2019), arXiv:1904.08976 [astro-ph.CO] .

[70] H. Davoudiasl and P. P. Giardino, Phys. Lett. B **768**, 198 (2017), arXiv:1609.00907 [gr-qc] .

[71] Y. Génolini, P. Serpico, and P. Tinyakov, Phys. Rev. D **102**, 083004 (2020), arXiv:2006.16975 [astro-ph.HE] .

[72] F. Özel and P. Freire, Ann. Rev. Astron. Astrophys. **54**, 401 (2016), arXiv:1603.02698 [astro-ph.HE] .

[73] K. W. Smith and I. A. Bonnell, Mon. Not. Roy. Astron. Soc. **322**, L1 (2001), arXiv:astro-ph/0101085 .

[74] J. R. Hurley and M. M. Shara, Astrophys. J. **565**, 1251 (2002), arXiv:astro-ph/0108350 .

[75] H. B. Perets and M. Kouwenhoven, Astrophys. J. **750**, 83 (2012), arXiv:1202.2362 [astro-ph.EP] .

[76] L. Wang, M. B. N. Kouwenhoven, X. Zheng, R. P. Church, and M. B. Davies, MNRAS **449**, 3543 (2015), arXiv:1503.03077 [astro-ph.EP] .

[77] T. Barclay, E. V. Quintana, S. N. Raymond, and M. T. Penny, ApJ **841**, 86 (2017), arXiv:1704.08749 [astro-ph.EP] .

[78] M. J. Valtonen, The Observatory **103**, 1 (1983).

[79] M. V. Torbett, AJ **92**, 171 (1986).

[80] C. R. Stagg and M. E. Bailey, MNRAS **241**, 507 (1989).

[81] H. J. Melosh, Astrobiology **3**, 207 (2003).

[82] N. Goulinski and E. N. Ribak, MNRAS **473**, 1589 (2018), arXiv:1705.10332 [astro-ph.EP] .

[83] M. Lingam and A. Loeb, AJ **156**, 193 (2018), arXiv:1801.10254 [astro-ph.EP] .

[84] E. Grishin, H. B. Perets, and Y. Avni, MNRAS **487**, 3324 (2019), arXiv:1804.09716 [astro-ph.EP] .

[85] D. C. Heggie, MNRAS **173**, 729 (1975).

[86] H. F. Levison and M. J. Duncan, Icarus **108**, 18 (1994).

[87] L. Dones, B. Gladman, H. J. Melosh, W. B. Tonks, H. F. Levison, and M. Duncan, Icarus **142**, 509 (1999).

[88] A. Gould, Astrophys. J. **321**, 571 (1987).

[89] A. Gould, Astrophys. J. **328**, 919 (1988).

[90] J. Lundberg and J. Edsjö, Phys. Rev. D **69**, 123505 (2004), arXiv:astro-ph/0401113 .

[91] X. Xu and E. Siegel, (2008), arXiv:0806.3767 [astro-ph] .

[92] A. H. Peter, Phys. Rev. D **79**, 103533 (2009), arXiv:0902.1348 [astro-ph.HE] .

[93] J. Edsjö and A. H. Peter, (2010), arXiv:1004.5258 [astro-ph.EP] .

[94] M. Valtonen and H. Karttunen, *The Three-Body Problem* (Cambridge University Press, 2005).

[95] E. J. Öpik, Proc. R. Irish Acad. Sect. A **54**, 165 (1951).

[96] E. J. Öpik, AJ **66**, 381 (1961).

[97] G. W. Wetherill, J. Geophys. Res. **72**, 2429 (1967).

[98] D. J. Kessler, Icarus **48**, 39 (1981).

[99] J. R. Arnold, ApJ **141**, 1536 (1965).

[100] H. Rein, D. M. Hernandez, D. Tamayo, G. Brown, E. Eckels, E. Holmes, M. Lau, R. Leblanc, and A. Silburt, MNRAS **485**, 5490 (2019), arXiv:1903.04972 [astro-ph.EP] .

[101] H. Rein and S.-F. Liu, REBOUND: Multi-purpose N-body code for collisional dynamics (2011), ascl:1110.016 .

[102] H. Rein and S.-F. Liu, Astron. Astrophys. **537**, A128 (2012), arXiv:1110.4876 [astro-ph.EP] .

[103] M. J. Baker and A. Thamm, SciPost Phys. **12**, 150 (2022), arXiv:2105.10506 [hep-ph] .

[104] J. B. Pollack, J. A. Burns, and M. E. Tauber, Icarus **37**, 587 (1979).

[105] E. C. Ostriker, Astrophys. J. **513**, 252 (1999), arXiv:astro-ph/9810324 .

[106] E. Grishin and H. B. Perets, ApJ **811**, 54 (2015), arXiv:1503.02668 [astro-ph.EP] .

[107] S. R. Pottasch, A&A **89**, 336 (1980).

[108] S. K. Gorny, G. Stasińska, and R. Tylenda, A&A **318**, 256 (1997).

[109] M. A. Abramowicz, J. K. Becker, P. L. Biermann, A. Garzilli, F. Johansson, and L. Qian, Astrophys. J. **705**, 659 (2009), arXiv:0810.3140 [astro-ph] .

[110] S. Nussinov, L.-T. Wang, and I. Yavin, JCAP **08**, 037, arXiv:0905.1333 [hep-ph] .

[111] G. Jungman and M. Kamionkowski, Phys. Rev. D **51**, 328 (1995), arXiv:hep-ph/9407351 .

[112] O. Y. Gnedin, A. V. Kravtsov, A. A. Klypin, and D. Nagai, Astrophys. J. **616**, 16 (2004), arXiv:astro-ph/0406247 .

[113] E. V. Derishev and A. A. Belyanin, A&A **343**, 1 (1999).

[114] F. Capela, M. Pshirkov, and P. Tinyakov, Phys. Rev. D **87**, 023507 (2013), arXiv:1209.6021 [astro-ph.CO] .

[115] F. Capela, M. Pshirkov, and P. Tinyakov, Phys. Rev. D **90**, 083507 (2014), arXiv:1403.7098 [astro-ph.CO] .

[116] M. Heyer and T. Dame, Annual Review of Astronomy and Astrophysics **53**, 583 (2015), https://doi.org/10.1146/annurev-astro-082214-122324 .

[117] J. M. Kirk, D. Ward-Thompson, and P. Andre, Mon. Not. Roy. Astron. Soc. **360**, 1506 (2005), arXiv:astro-ph/0505190 .

[118] J. A. Dror, H. Ramani, T. Trickle, and K. M. Zurek, Phys. Rev. D **100**, 023003 (2019), arXiv:1901.04490 [astro-ph.CO] .

[119] S. E. Thorsett and J. A. Phillips, ApJ **387**, L69 (1992).

[120] R. G. Martin, M. Livio, and D. Palaniswamy, ApJ **832**, 122 (2016), arXiv:1609.06409 [astro-ph.EP] .

[121] S. Archambault (VERITAS), *Proceedings, 35th International Cosmic Ray Conference (ICRC 2017): Bexco, Busan, Korea, July 12-20, 2017*, PoS **ICRC2017**, 691 (2018), [35,691(2017)], arXiv:1709.00307 [astro-ph.HE] .

[122] A. C. Johnson, *Searching for New Physics with the Fermi Large Area Telescope*, Ph.D. thesis, UC, Santa Cruz (2019-03).

[123] P. Chen and R. J. Adler, *Sources and detection of dark matter and dark energy in the universe. Proceedings, 5th International Symposium, Dark Matter 2002, Marina del Rey, USA, February 20-22, 2002*, Nucl. Phys. Proc. Suppl. **124**, 103 (2003), [,103(2002)], arXiv:gr-qc/0205106 [gr-qc] .

[124] S. Alexeyev, A. Barrau, G. Boudoul, O. Khovanskaya, and M. Sazhin, Class. Quant. Grav. **19**, 4431 (2002), arXiv:gr-qc/0201069 [gr-qc] .

[125] P. Chen, *Proceedings, 6th UCLA Symposium on Sources and Detection of Dark Matter and Dark Energy in the Universe: Marina del Rey, CA, USA, February 18-20, 2004*, New Astron. Rev. **49**, 233 (2005), arXiv:astro-ph/0406514 [astro-ph] .

[126] K. Nozari and S. H. Mehdipour, Mod. Phys. Lett. **A20**, 2937 (2005), arXiv:0809.3144 [gr-qc] .

[127] J. H. MacGibbon, Nature **329**, 308 (1987).

[128] J. D. Barrow, E. J. Copeland, and A. R. Liddle, Phys. Rev. **D46**, 645 (1992).

[129] B. J. Carr, J. H. Gilbert, and J. E. Lidsey, Phys. Rev. **D50**, 4853 (1994), arXiv:astro-ph/9405027 [astro-ph] .

[130] A. De Rujula, S. L. Glashow, and U. Sarid, Nucl. Phys. **B333**, 173 (1990).

[131] L. Chuzhoy and E. W. Kolb, JCAP **0907**, 014, arXiv:0809.0436 [astro-ph] .

[132] S. D. McDermott, H.-B. Yu, and K. M. Zurek, Phys. Rev. **D83**, 063509 (2011), arXiv:1011.2907 [hep-ph] .

[133] S. Dimopoulos, D. Eichler, R. Esmailzadeh, and G. D. Starkman, Phys. Rev. **D41**, 2388 (1990).

[134] A. D. Dolgov, S. L. Dubovsky, G. I. Rubtsov, and I. I. Tkachev, Phys. Rev. **D88**, 117701 (2013), arXiv:1310.2376 [hep-ph] .

[135] S. W. Hawking, Nature **248**, 30 (1974).

[136] W. Israel, Phys. Rev. **164**, 1776 (1967).

[137] D. N. Page, Phys. Rev. D **14**, 3260 (1976).

[138] S. Chandrasekhar, Astrophys. J. **74**, 81 (1931).

[139] G. W. Gibbons, Commun. math. Phys **44**, 245 (1975).

[140] D. N. Page, Phys. Rev. **D16**, 2402 (1977).

[141] A. Paul and B. R. Majhi, Int. J. Mod. Phys. **A32**, 1750088 (2017), arXiv:1601.07310 [gr-qc] .

[142] S. W. Hawking, Commun. Math. Phys. **43**, 199 (1975), [Erratum: Commun.Math.Phys. 46, 206 (1976)].

[143] J. S. Schwinger, Phys. Rev. **82**, 664 (1951).

[144] S. A. Teukolsky and W. H. Press, Astrophys. J. **193**, 443 (1974).

[145] D. N. Page, Phys. Rev. D **13**, 198 (1976).

[146] C. Vafa, (2005), arXiv:hep-th/0509212 [hep-th] .

[147] N. Arkani-Hamed, L. Motl, A. Nicolis, and C. Vafa, JHEP **06**, 060, arXiv:hep-th/0601001 [hep-th] .

[148] E. C. Vagenas, Phys. Lett. **B503**, 399 (2001), arXiv:hep-th/0012134 [hep-th] .

[149] I. B. Khriplovich, Phys. Atom. Nucl. **65**, 1259 (2002), [Yad. Fiz.65,1292(2002)].

[150] C.-M. Chen, J.-R. Sun, F.-Y. Tang, and P.-Y. Tsai, Class. Quant. Grav. **32**, 195003 (2015), arXiv:1412.6876 [hep-th] .

[151] F. Belgiorno and M. Martellini, Int. J. Mod. Phys. D **13**, 739 (2004), arXiv:gr-qc/0210026 .

[152] J. Preskill, P. Schwarz, A. D. Shapere, S. Trivedi, and F. Wilczek, Mod. Phys. Lett. **A6**, 2353 (1991).

[153] W. A. Hiscock and L. D. Weems, Physical Review D **41** (1990).

[154] B. J. Carr, J. Mureika, and P. Nicolini, JHEP **07**, 052, arXiv:1504.07637 [gr-qc] .

[155] Y. Bai and N. Orlofsky, (2019), arXiv:1906.04858 [hep-ph] .

[156] P. Madau, F. Haardt, and M. J. Rees, Astrophys. J. **514**, 648 (1999), arXiv:astro-ph/9809058 [astro-ph] .

[157] W. J. Boardman, The Astrophysical Journal Supplement Series **9**, 185 (1964).

[158] W. J. Karzas and R. Latter, Astrophysical Journal Supplement **21**, 167 (1960).

[159] J. Lindhard, M. Scharff, and H. Schiott, Det Kongelige Danske Videnskabernes Selskab **33**, 42 (1963).

[160] M. Tanabashi *et al.* (Particle Data Group), Phys. Rev. D **98**, 030001 (2018).

[161] V. I. Tretyak, Astropart. Phys. **33**, 40 (2010), arXiv:0911.3041 [nucl-ex] .

[162] W. Mu and X. Ji, Astropart. Phys. **62**, 108 (2015), arXiv:1310.2094 [physics.ins-det] .

[163] J. F. Ziegler, M. D. Ziegler, and J. P. Biersack, Nuclear Instruments and Methods in Physics Research B **268**, 1818 (2010).

[164] S. Hawking, MNRAS **152**, 75 (1971).

[165] C. Amole *et al.* (PICO), Phys. Rev. **D93**, 061101 (2016), arXiv:1601.03729 [astro-ph.CO] .

[166] C. Amole *et al.* (PICO), Phys. Rev. Lett. **118**, 251301 (2017), arXiv:1702.07666 [astro-ph.CO] .

[167] E. Vazquez-Jauregui (2017), 15th International Conference on Topics in Astroparticle and Underground Physics.

[168] R. U. Abbasi *et al.* (HiRes), Astrophys. J. **622**, 910 (2005), arXiv:astro-ph/0407622 [astro-ph] .

[169] Pierre Sokolsky, *Introduction to Ultrahigh Energy Cosmic Ray Physics*, Frontiers in Physics (Addison-Wesley, 1989).

[170] R. Abbasi *et al.* (IceCube), Astropart. Phys. **35**, 615 (2012), arXiv:1109.6096 [astro-ph.IM] .

[171] Y. Fukuda *et al.* (Super-Kamiokande), *Advanced computing and analysis techniques in physics research. Proceedings, 8th International Workshop, ACAT 2002, Moscow, Russia, June 24-28, 2002*, Nucl. Instrum. Meth. **A501**, 418 (2003).

[172] T. C. Weekes *et al.*, Astropart. Phys. **17**, 221 (2002), arXiv:astro-ph/0108478 [astro-ph] .

[173] T. DeYoung (HAWC), *Proceedings, 3rd Roma International Conference on Astro-Particle Physics (RICAP 11): Rome, Italy, May 25-27, 2011*, Nucl. Instrum. Meth. **A692**, 72 (2012).

[174] A. Garfagnini, *Proceedings, 33rd International Symposium on Physics in Collision (PIC 2013): Beijing, China, September 3-7, 2013*, International Journal of Modern Physics: Conference Series **31**, 1460286 (2014).

[175] M. Ambrosio *et al.* (MACRO), Eur. Phys. J. **C25**, 511 (2002), arXiv:hep-ex/0207020 [hep-ex] .

[176] C. Rubbia, (1977).

[177] C. Rubbia *et al.*, JINST **6**, P07011, arXiv:1106.0975 [hep-ex] .

[178] C. Anderson *et al.*, JINST **7**, P10019, arXiv:1205.6747 [physics.ins-det] .

[179] F. Cavanna, M. Kordosky, J. Raaf, and B. Rebel (LArIAT), (2014), arXiv:1406.5560 [physics.ins-det] .

[180] M. Antonello *et al.* (MicroBooNE, LAr1-ND, ICARUS-WA104), (2015), arXiv:1503.01520 [physics.ins-det] .

[181] B. Abi *et al.* (DUNE), (2017), arXiv:1706.07081 [physics.ins-det] .

[182] R. Acciarri *et al.* (DUNE), (2016), arXiv:1601.02984 [physics.ins-det] .

[183] R. Acciarri *et al.* (MicroBooNE), JINST **12** (02), P02017, arXiv:1612.05824 [physics.ins-det] .

[184] S. Baum, A. K. Drukier, K. Freese, M. Górski, and P. Stengel, (2018), arXiv:1806.05991 [astro-ph.CO] .

[185] A. K. Drukier, S. Baum, K. Freese, M. Górski, and P. Stengel, Phys. Rev. **D99**, 043014 (2019), arXiv:1811.06844 [astro-ph.CO] .

[186] T. D. P. Edwards, B. J. Kavanagh, C. Weniger, S. Baum, A. K. Drukier, K. Freese, M. Górski, and P. Stengel, Phys. Rev. **D99**, 043541 (2019), arXiv:1811.10549 [hep-ph] .

[187] D. Ghosh and S. Chatterjea, EPL **12**, 25 (1990).

[188] S. Werbowy and B. Pranszke, Physical Review A **93**, 22713 (2016).

[189] N. Cue, N. V. De Castro-Faria, M. J. Gaillard, J. C. Poizat, J. Remillieux, D. S. Gemmell, and I. Plesser, Physical Review Letters **45**, 613 (1980).

[190] Y. Susuki, M. Fritz, K. Kimura, M. Mannami, N. Sakamoto, H. Ogawa, I. Katayama, T. Noro, and H. Ikegami, Physical Review A **50**, 3533 (1994).

[191] R. Garcia-Molina and M. D. Barriga-Carrasco, Physical Review A - Atomic, Molecular, and Optical Physics **68**, 4 (2003).

[192] W. D. Wilson, L. G. Haggmark, and J. P. Biersack, Phys. Rev. B **15**, 2458 (1977).

[193] W. Brandt and M. Kitagawa, Physical Review B **25**, 10.1103/PhysRevB.25.5631 (1982).

[194] A. Arnau, M. Pealba, P. M. Echenique, F. Flores, and R. H. Ritchie, Physical Review Letters **65**, 1024 (1990).

[195] P. L. Grande and G. Schiwietz, Physical Review A **47**, 1119 (1993).

[196] P. Sigmund, Physical Review A - Atomic, Molecular, and Optical Physics **56**, 3781 (1997).

[197] F. Gobet, S. Eden, B. Coupier, J. Tabet, B. Farizon, M. Farizon, M. J. Gaillard, S. Ouaskit, M. Carré, and T. D. Märk, Chemical Physics Letters **421**, 68 (2006).

[198] S. D. Drell, N. M. Kroll, M. T. Mueller, S. J. Parke, and M. H. Ruderman, Phys. Rev. Lett. **50**, 644 (1983).

[199] M. Antonello *et al.*, JINST **10** (12), P12004, arXiv:1504.01556 [physics.ins-det] .

[200] D. Gastler, E. Kearns, A. Hime, L. C. Stonehill, S. Seibert, J. Klein, W. H. Lippincott, D. N. McKinsey, and J. A. Nikkel, Phys. Rev. **C85**, 065811 (2012), arXiv:1004.0373 [physics.ins-det] .

[201] A. Bondar, A. Buzulutskov, A. Dolgov, E. Grishnyaev, S. Polosatkin, L. Shekhtman, E. Shemyakina, and A. Sokolov, EPL **108**, 12001 (2014), arXiv:1407.7348 [physics.ins-det] .

[202] B. P. Abbott *et al.* (LIGO Scientific, Virgo), Phys. Rev. Lett. **116**, 061102 (2016), arXiv:1602.03837 [gr-qc] .

[203] B. P. Abbott *et al.* (LIGO Scientific, Virgo), Phys. Rev. Lett. **116**, 241103 (2016), arXiv:1606.04855 [gr-qc] .

[204] B. P. Abbott *et al.* (LIGO Scientific, VIRGO), Phys. Rev. Lett. **118**, 221101 (2017), [Erratum: Phys.Rev.Lett. 121, 129901 (2018)], arXiv:1706.01812 [gr-qc] .

[205] B. Abbott *et al.* (LIGO Scientific, Virgo), Phys. Rev. Lett. **119**, 141101 (2017), arXiv:1709.09660 [gr-qc] .

[206] B. P. Abbott *et al.* (LIGO Scientific, Virgo), Astrophys. J. **851**, L35 (2017), arXiv:1711.05578 [astro-ph.HE] .

[207] B. Abbott *et al.* (LIGO Scientific, Virgo), Phys. Rev. X **9**, 031040 (2019), arXiv:1811.12907 [astro-ph.HE] .

[208] B. Abbott *et al.* (LIGO Scientific, Virgo), Astrophys. J. Lett. **892**, L3 (2020), arXiv:2001.01761 [astro-ph.HE] .

[209] R. Abbott *et al.* (LIGO Scientific, Virgo), Phys. Rev. D **102**, 043015 (2020), arXiv:2004.08342 [astro-ph.HE] .

[210] B. Abbott *et al.* (LIGO Scientific, Virgo), Phys. Rev. X **6**, 041015 (2016), [Erratum: Phys.Rev.X 8, 039903 (2018)], arXiv:1606.04856 [gr-qc] .

[211] B. Abbott *et al.* (LIGO Scientific, Virgo), Phys. Rev. Lett. **116**, 221101 (2016), [Erratum: Phys.Rev.Lett. 121, 129902 (2018)], arXiv:1602.03841 [gr-qc] .

[212] B. Abbott *et al.* (LIGO Scientific, Virgo, Fermi GBM, INTEGRAL, IceCube, AstroSat Cadmium Zinc Telluride Imager Team, IPN, Insight-Hxmt, ANTARES, Swift, AGILE Team, 1M2H Team, Dark Energy Camera GW-EM, DES, DLT40, GRAWITA, Fermi-LAT, ATCA, ASKAP, Las Cumbres Observatory Group, OzGrav, DWF (Deeper Wider Faster Program), AST3, CAASTRO, VINROUGE, MASTER, J-GEM, GROWTH, JAGWAR, CaltechNRAO, TTU-NRAO, NuSTAR, Pan-STARRS, MAXI Team, TZAC Consortium, KU, Nordic Optical Telescope, ePESSTO, GROND, Texas Tech University, SALT Group, TOROS, BOOTES, MWA, CALET, IKI-GW Follow-up, H.E.S.S., LOFAR, LWA, HAWC, Pierre Auger, ALMA, Euro VLBI Team, Pi of Sky, Chandra Team at McGill

University, DFN, ATLAS Telescopes, High Time Resolution Universe Survey, RIMAS, RATIR, SKA South Africa/MeerKAT), Astrophys. J. Lett. **848**, L12 (2017), arXiv:1710.05833 [astro-ph.HE] .

[213] B. Abbott *et al.* (LIGO Scientific, Virgo, Fermi-GBM, INTEGRAL), Astrophys. J. Lett. **848**, L13 (2017), arXiv:1710.05834 [astro-ph.HE] .

[214] I. Cholis, E. D. Kovetz, Y. Ali-Haïmoud, S. Bird, M. Kamionkowski, J. B. Muñoz, and A. Raccanelli, Phys. Rev. D **94**, 084013 (2016), arXiv:1606.07437 [astro-ph.HE] .

[215] N. Fernandez and S. Profumo, JCAP **08**, 022, arXiv:1905.13019 [astro-ph.HE] .

[216] V. De Luca, G. Franciolini, P. Pani, and A. Riotto, JCAP **04**, 052, arXiv:2003.02778 [astro-ph.CO] .

[217] V. De Luca, G. Franciolini, P. Pani, and A. Riotto, JCAP **06**, 044, arXiv:2005.05641 [astro-ph.CO] .

[218] A. Dolgov, A. Kuranov, N. Mitichkin, S. Porey, K. Postnov, O. Sazhina, and I. Simkin, (2020), arXiv:2005.00892 [astro-ph.CO] .

[219] K. Belczynski, T. Bulik, C. L. Fryer, A. Ruiter, J. S. Vink, and J. R. Hurley, Astrophys. J. **714**, 1217 (2010), arXiv:0904.2784 [astro-ph.SR] .

[220] K. Belczynski *et al.*, Astron. Astrophys. **636**, A104 (2020), arXiv:1706.07053 [astro-ph.HE] .

[221] D. Gerosa and E. Berti, Phys. Rev. D **95**, 124046 (2017), arXiv:1703.06223 [gr-qc] .

[222] B. Farr, D. E. Holz, and W. M. Farr, Astrophys. J. Lett. **854**, L9 (2018), arXiv:1709.07896 [astro-ph.HE] .

[223] G. F. Chapline, Nature **253**, 251 (1975).

[224] R. Abbott *et al.* (LIGO Scientific, Virgo), Astrophys. J. **896**, L44 (2020), arXiv:2006.12611 [astro-ph.HE] .

[225] C. D. Bailyn, R. K. Jain, P. Coppi, and J. A. Orosz, Astrophys. J. **499**, 367 (1998), arXiv:astro-ph/9708032 .

[226] F. Ozel, D. Psaltis, R. Narayan, and J. E. McClintock, Astrophys. J. **725**, 1918 (2010), arXiv:1006.2834 [astro-ph.GA] .

[227] F. Ozel, D. Psaltis, R. Narayan, and A. S. Villarreal, Astrophys. J. **757**, 55 (2012), arXiv:1201.1006 [astro-ph.HE] .

[228] W. M. Farr, N. Sravan, A. Cantrell, L. Kreidberg, C. D. Bailyn, I. Mandel, and V. Kalogera, Astrophys. J. **741**, 103 (2011), arXiv:1011.1459 [astro-ph.GA] .

[229] B. Abbott *et al.* (LIGO Scientific, Virgo), Phys. Rev. Lett. **119**, 161101 (2017), arXiv:1710.05832 [gr-qc] .

[230] M. Soares-Santos *et al.* (DES, Dark Energy Camera GW-EM), Astrophys. J. Lett. **848**, L16 (2017), arXiv:1710.05459 [astro-ph.HE] .

[231] M. W. Coughlin and T. Dietrich, Phys. Rev. D **100**, 043011 (2019), arXiv:1901.06052 [astro-ph.HE] .

[232] V. Vaskonen and H. Veermäe, Phys. Rev. D **101**, 043015 (2020), arXiv:1908.09752 [astro-ph.CO] .

[233] B. Carr and F. Kuhnel, Phys. Rev. D **99**, 103535 (2019), arXiv:1811.06532 [astro-ph.CO] .

[234] Z.-C. Chen and Q.-G. Huang, (2019), arXiv:1904.02396 [astro-ph.CO] .

[235] V. Mandic, S. Bird, and I. Cholis, Phys. Rev. Lett. **117**, 201102 (2016), arXiv:1608.06699 [astro-ph.CO] .

[236] M. Raidal, V. Vaskonen, and H. Veermäe, JCAP **1709**, 037, arXiv:1707.01480 [astro-ph.CO] .

[237] J. Hassall, The old nursery stories and rhymes (Blackie & Son, 1904) Chap. The Story of the Three Bears.

[238] M. Raidal, C. Spethmann, V. Vaskonen, and H. Veermäe, JCAP **02**, 018, arXiv:1812.01930 [astro-ph.CO] .

[239] K. Jedamzik, (2020), arXiv:2006.11172 [astro-ph.CO] .

[240] K. Inomata, M. Kawasaki, K. Mukaida, and T. T. Yanagida, Phys. Rev. D **97**, 043514 (2018), arXiv:1711.06129 [astro-ph.CO] .

[241] A. Katz, J. Kopp, S. Sibiryakov, and W. Xue, JCAP **12**, 005, arXiv:1807.11495 [astro-ph.CO] .

[242] P. Montero-Camacho, X. Fang, G. Vasquez, M. Silva, and C. M. Hirata, JCAP **08**, 031, arXiv:1906.05950 [astro-ph.CO] .

[243] B. Abbott *et al.* (LIGO Scientific, VIRGO), Nature **460**, 990 (2009), arXiv:0910.5772 [astro-ph.CO] .

[244] B. P. Abbott *et al.* (LIGO Scientific, Virgo), Phys. Rev. Lett. **118**, 121101 (2017), [Erratum: Phys.Rev.Lett. 119, 029901 (2017)], arXiv:1612.02029 [gr-qc] .

[245] S. Wang, T. Terada, and K. Kohri, Phys. Rev. **D99**, 103531 (2019), arXiv:1903.05924 [astro-ph.CO] .

[246] T. Regimbau, Res. Astron. Astrophys. **11**, 369 (2011), arXiv:1101.2762 [astro-ph.CO] .

[247] P. A. Rosado, Phys. Rev. D **84**, 084004 (2011), arXiv:1106.5795 [gr-qc] .

[248] X.-J. Zhu, E. Howell, T. Regimbau, D. Blair, and Z.-H. Zhu, Astrophys. J. **739**, 86 (2011), arXiv:1104.3565 [gr-qc] .

[249] S. Wang, Y.-F. Wang, Q.-G. Huang, and T. G. F. Li, Phys. Rev. Lett. **120**, 191102 (2018), arXiv:1610.08725 [astro-ph.CO] .

[250] S. Kirkpatrick, C. Gelatt, and M. Vecchi, Science **220**, 671 (1983).

[251] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).

[252] W. Hastings, Biometrika **57**, 97 (1970).

[253] P. Tisserand *et al.* (EROS-2), Astron. Astrophys. **469**, 387 (2007), arXiv:astro-ph/0607207 .

[254] R. w. Hellings and G. s. Downs, Astrophys. J. Lett. **265**, L39 (1983).

[255] Z. Arzoumanian *et al.* (NANOGrav), Astrophys. J. **859**, 47 (2018), arXiv:1801.02617 [astro-ph.HE] .

[256] L. Lentati *et al.*, Mon. Not. Roy. Astron. Soc. **453**, 2576 (2015), arXiv:1504.03692 [astro-ph.CO] .

[257] R. M. Shannon *et al.*, Science **349**, 1522 (2015), arXiv:1509.07320 [astro-ph.CO] .

[258] Z. Arzoumanian *et al.* (NANOGrav), Astrophys. J. Lett. **905**, L34 (2020), arXiv:2009.04496 [astro-ph.HE] .

[259] R. Nan, D. Li, C. Jin, Q. Wang, L. Zhu, W. Zhu, H. Zhang, Y. Yue, and L. Qian, Int. J. Mod. Phys. D **20**, 989 (2011), arXiv:1105.3794 [astro-ph.IM] .

[260] C. L. Carilli and S. Rawlings, New Astron. Rev. **48**, 979 (2004), arXiv:astro-ph/0409274 .

[261] M. Bonetti, A. Sesana, E. Barausse, and F. Haardt, Mon. Not. Roy. Astron. Soc. **477**, 2599 (2018), arXiv:1709.06095 [astro-ph.GA] .

[262] D. Krause, H. T. Kloor, and E. Fischbach, Phys. Rev. **D49**, 6892 (1994).

[263] S. Alexander, E. McDonough, R. Sims, and N. Yunes, Class. Quant. Grav. **35**, 235012 (2018), arXiv:1808.05286 [gr-qc] .

[264] J. A. Dror, R. Laha, and T. Opferkuch, Phys. Rev. D **102**, 023005 (2020), arXiv:1909.12845 [hep-ph] .

[265] R. A. Hulse and J. H. Taylor, Astrophys. J. Lett. **195**, L51 (1975).

[266] D. Croon, A. E. Nelson, C. Sun, D. G. E. Walker, and Z.-Z. Xianyu, Astrophys. J. **858**, L2 (2018), arXiv:1711.02096 [hep-ph] .

[267] J. Kopp, R. Laha, T. Opferkuch, and W. Shepherd, JHEP **11**, 096, arXiv:1807.02527 [hep-ph] .

[268] M. Fabbrichesi and A. Urbano, JCAP **06**, 007, arXiv:1902.07914 [hep-ph] .

[269] X.-J. Yue and Z. Cao, Class. Quant. Grav. **37**, 245009 (2020).

[270] X.-J. Xu, JHEP **09**, 105, arXiv:2007.01893 [hep-ph] .

[271] G. E. Romero and E. M. Gutiérrez, Universe **6**, 99 (2020), arXiv:2007.09717 [astro-ph.HE] .

[272] C. M. Mingarelli 10.1038/s41550-018-0666-y (2019), arXiv:1901.06785 [gr-qc] .

[273] T. Ryu, R. Perna, Z. Haiman, J. P. Ostriker, and N. C. Stone, Mon. Not. Roy. Astron. Soc. **473**, 3410 (2018), arXiv:1709.06501 .

[274] M. Enoki and M. Nagashima, Prog. Theor. Phys. **117**, 241 (2007), arXiv:astro-ph/0609377 .

[275] A. Sesana, Astrophys. J. **719**, 851 (2010), arXiv:1006.0730 [astro-ph.CO] .

[276] A. Sesana, Mon. Not. Roy. Astron. Soc. **433**, L1 (2013), arXiv:1211.5375 [astro-ph.CO] .

[277] A. Sesana, Class. Quant. Grav. **30**, 224014 (2013), arXiv:1307.2600 [astro-ph.CO] .

[278] C. J. Moore, R. H. Cole, and C. P. L. Berry, Class. Quant. Grav. **32**, 015014 (2015), arXiv:1408.0740 [gr-qc] .

[279] E. S. Phinney, (2001), arXiv:astro-ph/0108028 .

[280] M. Maggiore, *Gravitational Waves. Vol. 1: Theory and Experiments*, Oxford Master Series in Physics (Oxford University Press, 2007).

[281] G. W. Gibbons, Commun. Math. Phys. **44**, 245 (1975).

[282] S. R. Coleman, J. Preskill, and F. Wilczek, Nucl. Phys. B **378**, 175 (1992), arXiv:hep-th/9201059 .

[283] M. Begelman, R. Blandford, and M. Rees, Nature **287**, 307 (1980).

[284] O. Ilbert *et al.*, Astrophys. J. **709**, 644 (2010), arXiv:0903.0102 [astro-ph.CO] .

[285] E. F. Bell, D. H. McIntosh, N. Katz, and M. D. Weinberg, Astrophys. J. Suppl. **149**, 289 (2003), arXiv:astro-ph/0302543 .

[286] N. J. McConnell and C.-P. Ma, Astrophys. J. **764**, 184 (2013), arXiv:1211.2816 [astro-ph.CO] .

[287] C. Lopez-Sanjuan *et al.*, Astron. Astrophys. **548**, A7 (2012), arXiv:1202.4674 [astro-ph.CO] .

[288] M. Milosavljevic and D. Merritt, Astrophys. J. **596**, 860 (2003), arXiv:astro-ph/0212459 .

[289] G. D. Quinlan, New Astron. **1**, 35 (1996), arXiv:astro-ph/9601092 .

[290] F. Sanchez-Salcedo and A. Brandenburg, Mon. Not. Roy. Astron. Soc. **322**, 67 (2001), arXiv:astro-ph/0010003 .

[291] A. Escala, R. B. Larson, P. S. Coppi, and D. Mardones, Astrophys. J. **607**, 765 (2004), arXiv:astro-ph/0310851 .

[292] A. Escala, R. B. Larson, P. S. Coppi, and D. Mardones, Astrophys. J. **630**, 152 (2005), arXiv:astro-ph/0406304 .

[293] P. J. Armitage and P. Natarajan, Astrophys. J. Lett. **567**, L9 (2002), arXiv:astro-ph/0201318 .

[294] L. Mayer, S. Kazantzidis, P. Madau, M. Colpi, T. R. Quinn, and J. Wadsley, Science **316**, 1874 (2007), arXiv:0706.1562 [astro-ph] .

[295] L. Mayer, Class. Quant. Grav. **30**, 244008 (2013), arXiv:1308.0431 [astro-ph.CO] .

[296] V. Ravi, J. Wyithe, R. Shannon, G. Hobbs, and R. Manchester, Mon. Not. Roy. Astron. Soc. **442**, 56 (2014), arXiv:1404.5183 [astro-ph.CO] .

[297] V. Ravi, J. Wyithe, R. Shannon, and G. Hobbs, Mon. Not. Roy. Astron. Soc. **447**, 2772 (2015), arXiv:1406.5297 [astro-ph.CO] .

[298] F. M. Khan, D. Fiacconi, L. Mayer, P. Berczik, and A. Just, Astrophys. J. **828**, 73 (2016), arXiv:1604.00015 [astro-ph.GA] .

[299] S. Burke-Spolaor *et al.*, Astron. Astrophys. Rev. **27**, 5 (2019), arXiv:1811.08826 [astro-ph.HE] .

[300] P. Berczik, D. Merritt, R. Spurzem, and H.-P. Bischof, Astrophys. J. Lett. **642**, L21 (2006), arXiv:astro-ph/0601698 .

[301] S. R. Taylor, J. Simon, and L. Sampson, Phys. Rev. Lett. **118**, 181102 (2017), arXiv:1612.02817 [astro-ph.GA] .

[302] L. Z. Kelley, L. Blecha, and L. Hernquist, Mon. Not. Roy. Astron. Soc. **464**, 3131 (2017), arXiv:1606.01900 [astro-ph.HE] .

[303] P. Ivanov, J. Papaloizou, and A. Polnarev, Mon. Not. Roy. Astron. Soc. **307**, 79 (1999), arXiv:astro-ph/9812198 .

[304] Z. Haiman, B. Kocsis, and K. Menou, Astrophys. J. **700**, 1952 (2009), arXiv:0904.1383 [astro-ph.CO] .

[305] H. Aly, W. Dehnen, C. Nixon, and A. King, Mon. Not. Roy. Astron. Soc. **449**, 65 (2015), arXiv:1501.04623 [astro-ph.HE] .

[306] C. J. Moore, S. R. Taylor, and J. R. Gair, Class. Quant. Grav. **32**, 055004 (2015), arXiv:1406.5199 [astro-ph.IM] .

[307] P. A. Seoane *et al.* (eLISA), (2013), arXiv:1305.5720 [astro-ph.CO] .

[308] P. Amaro-Seoane *et al.* (LISA), (2017), arXiv:1702.00786 [astro-ph.IM] .

[309] V. B. Braginsky, N. S. Kardashev, A. G. Polnarev, and I. D. Novikov, Nuovo Cimento B Serie **105**, 1141 (1990).

[310] C. R. Evans, I. Iben, and L. Smarr, Astrophys. J. **323**, 129 (1987).

[311] P. L. Bender and D. Hils, Class. Quant. Grav. **14**, 1439 (1997).

[312] J. Brod, M. Gorbahn, and E. Stamou, Phys. Rev. **D83**, 034030 (2011), arXiv:1009.0947 [hep-ph] .

[313] A. J. Buras, D. Buttazzo, J. Girrbach-Noe, and R. Knegjens, JHEP **11**, 033, arXiv:1503.02693 [hep-ph] .

[314] S. Adler *et al.* (E787), Phys. Lett. B **537**, 211 (2002), arXiv:hep-ex/0201037 .

[315] V. Anisimovsky *et al.* (E949), Phys. Rev. Lett. **93**, 031801 (2004), arXiv:hep-ex/0403036 .

[316] A. Artamonov *et al.* (E949), Phys. Rev. Lett. **101**, 191802 (2008), arXiv:0808.2459 [hep-ex] .

[317] E. Cortina Gil *et al.* (NA62), Phys. Lett. B **791**, 156 (2019), arXiv:1811.08508 [hep-ex] .

[318] G. Ruggiero (2019), talk at KAON2019.

[319] J. Ahn *et al.* (KOTO), Phys. Rev. Lett. **122**, 021802 (2019), arXiv:1810.09655 [hep-ex] .

[320] S. Shinohara (2019), talk at KAON2019.

[321] T. Kitahara, T. Okui, G. Perez, Y. Soreq, and K. Tobioka, Phys. Rev. Lett. **124**, 071801 (2020), arXiv:1909.11111 [hep-ph] .

[322] Y. Grossman and Y. Nir, Phys. Lett. B **398**, 163 (1997), arXiv:hep-ph/9701313 .

[323] K. Fuyuto, W.-S. Hou, and M. Kohda, Phys. Rev. Lett. **114**, 171802 (2015), arXiv:1412.4397 [hep-ph] .

[324] D. Egana-Ugrinovic, S. Homiller, and P. Meade, Phys. Rev. Lett. **124**, 191801 (2020), arXiv:1911.10203 [hep-ph] .

[325] P. B. Dev, R. N. Mohapatra, and Y. Zhang, Phys. Rev. D **101**, 075014 (2020), arXiv:1911.12334 [hep-ph] .

[326] Y. Jho, S. M. Lee, S. C. Park, Y. Park, and P.-Y. Tseng, JHEP **04**, 086, arXiv:2001.06572 [hep-ph] .

[327] J. Liu, N. McGinnis, C. E. Wagner, and X.-P. Wang, JHEP **04**, 197, arXiv:2001.06522 [hep-ph] .

[328] X.-G. He, X.-D. Ma, J. Tandean, and G. Valencia, JHEP **04**, 057, arXiv:2002.05467 [hep-ph] .

[329] R. Ziegler, J. Zupan, and R. Zwicky, (2020), arXiv:2005.00451 [hep-ph] .

[330] Y. Liao, H.-L. Wang, C.-Y. Yao, and J. Zhang, (2020), arXiv:2005.00753 [hep-ph] .

[331] S. Gori, G. Perez, and K. Tobioka, (2020), arXiv:2005.05170 [hep-ph] .

[332] M. Hostert, K. Kaneta, and M. Pospelov, (2020), arXiv:2005.07102 [hep-ph] .

[333] A. Datta, S. Kamali, and D. Marfatia, (2020), arXiv:2005.08920 [hep-ph] .

[334] M. Pospelov (2019), talk at HC2NP 2019.

[335] S. Gninenko, Phys. Rev. D **91**, 015004 (2015), arXiv:1409.2288 [hep-ph] .

[336] N. Carrasco, P. Dimopoulos, R. Frezzotti, V. Lubicz, G. C. Rossi, S. Simula, and C. Tarantino (ETM), Phys. Rev. D **92**, 034516 (2015), arXiv:1505.06639 [hep-lat] .

[337] B. J. Choi *et al.* (SWME), Phys. Rev. D **93**, 014511 (2016), arXiv:1509.00592 [hep-lat] .

[338] N. Garron, R. J. Hudspith, and A. T. Lytle (RBC/UKQCD), JHEP **11**, 001, arXiv:1609.03334 [hep-lat] .

[339] J. Brod and M. Gorbahn, Phys. Rev. Lett. **108**, 121801 (2012), arXiv:1108.2036 [hep-ph] .

[340] J. Brod, M. Gorbahn, and E. Stamou, (2019), arXiv:1911.06822 [hep-ph] .

[341] M. Tanabashi *et al.* (Particle Data Group), Phys. Rev. D **98**, 030001 (2018).

[342] G. G. Raffelt, *Stars as laboratories for fundamental physics* (1996).

[343] F. Bergsma *et al.* (CHARM), Phys. Lett. B **157**, 458 (1985).

[344] J. Blumlein *et al.*, Z. Phys. C **51**, 341 (1991).

[345] A. Berlin, S. Gori, P. Schuster, and N. Toro, Phys. Rev. D **98**, 035011 (2018), arXiv:1804.00661 [hep-ph] .

[346] C. Aidala *et al.* (SeaQuest), Nucl. Instrum. Meth. A **930**, 49 (2019), arXiv:1706.09990 [physics.ins-det] .

[347] C. Kelso, J. Kumar, P. Sandick, and P. Stengel, Phys. Rev. D **91**, 055028 (2015), arXiv:1411.2634 [hep-ph] .

[348] B. Dutta and L. E. Strigari, Ann. Rev. Nucl. Part. Sci. **69**, 137 (2019), arXiv:1901.08876 [hep-ph] .

[349] M. Blennow, E. Fernandez-Martinez, A. Olivares-Del Campo, S. Pascoli, S. Rosauro-Alcaraz, and A. Titov, Eur. Phys. J. C **79**, 555 (2019), arXiv:1903.00006 [hep-ph] .

[350] L. J. Hall, K. Jedamzik, J. March-Russell, and S. M. West, JHEP **03**, 080, arXiv:0911.1120 [hep-ph] .

[351] F. Elahi, C. Kolda, and J. Unwin, JHEP **03**, 048, arXiv:1410.6157 [hep-ph] .

[352] P. Gondolo and G. Gelmini, Nucl. Phys. **B360**, 145 (1991).

[353] S. Hannestad, Phys. Rev. D **70**, 043506 (2004), arXiv:astro-ph/0403291 .

[354] P. de Salas, M. Lattanzi, G. Mangano, G. Miele, S. Pastor, and O. Pisanti, Phys. Rev. D **92**, 123534 (2015), arXiv:1511.00672 [astro-ph.CO] .

[355] A. Pich, Rept. Prog. Phys. **58**, 563 (1995), arXiv:hep-ph/9502366 .

[356] J. Gasser and H. Leutwyler, Nucl. Phys. B **250**, 465 (1985).

[357] T. Hasegawa, N. Hiroshima, K. Kohri, R. S. Hansen, T. Tram, and S. Hannestad, (2020), arXiv:2003.13302 [hep-ph] .

[358] T. Hasegawa, N. Hiroshima, K. Kohri, R. S. Hansen, T. Tram, and S. Hannestad, JCAP **12**, 012, arXiv:1908.10189 [hep-ph] .

[359] K. Abazajian *et al.*, Bull. Am. Astron. Soc. **51**, 209 (2019), arXiv:1908.01062 [astro-ph.IM] .

[360] M. Viel, J. Lesgourgues, M. G. Haehnelt, S. Matarrese, and A. Riotto, Phys. Rev. **D71**, 063534 (2005), arXiv:astro-ph/0501562 [astro-ph] .

[361] D. S. Akerib *et al.* (LUX), Phys. Rev. Lett. **112**, 091303 (2014), arXiv:1310.8214 [astro-ph.CO] .

[362] T. Marrodán Undagoitia and L. Rauch, J. Phys. G **43**, 013001 (2016), arXiv:1509.08767 [physics.ins-det] .

[363] A. Tan *et al.* (PandaX-II), Phys. Rev. Lett. **117**, 121303 (2016), arXiv:1607.07400 [hep-ex] .

[364] A. Boyarsky, O. Ruchayskiy, and D. Iakubovskyi, JCAP **0903**, 005, arXiv:0808.3902 [hep-ph] .

[365] K. M. Zurek, Phys. Rept. **537**, 91 (2014), arXiv:1308.0338 [hep-ph] .

[366] R. Caputo, T. Linden, J. Tomsick, C. Prescod-Weinstein, M. Meyer, C. Kierans, Z. Wadiasingh, J. P. Harding, and J. Kopp, (2019), arXiv:1903.05845 [astro-ph.HE] .

[367] R. Essig, J. Mardon, and T. Volansky, Phys. Rev. **D85**, 076007 (2012), arXiv:1108.5383 [hep-ph] .

[368] R. Essig, M. Fernandez-Serra, J. Mardon, A. Soto, T. Volansky, and T.-T. Yu, JHEP **05**, 046, arXiv:1509.01598 [hep-ph] .

[369] Y. Hochberg, Y. Zhao, and K. M. Zurek, Phys. Rev. Lett. **116**, 011301 (2016), arXiv:1504.07237 [hep-ph] .

[370] Y. Hochberg, M. Pyle, Y. Zhao, and K. M. Zurek, JHEP **08**, 057, arXiv:1512.04533 [hep-ph] .

[371] Y. Hochberg, T. Lin, and K. M. Zurek, Phys. Rev. D **94**, 015019 (2016), arXiv:1604.06800 [hep-ph] .

[372] S. Derenzo, R. Essig, A. Massari, A. Soto, and T.-T. Yu, Phys. Rev. **D96**, 016026 (2017), arXiv:1607.01009 [hep-ph] .

[373] Y. Hochberg, T. Lin, and K. M. Zurek, Phys. Rev. D **95**, 023013 (2017), arXiv:1608.01994 [hep-ph] .

[374] S. Knapen, T. Lin, and K. M. Zurek, Phys. Rev. D **95**, 056019 (2017), arXiv:1611.06228 [hep-ph] .

[375] Y. Hochberg, Y. Kahn, M. Lisanti, K. M. Zurek, A. G. Grushin, R. Ilan, S. M. Griffin, Z.-F. Liu, S. F. Weber, and J. B. Neaton, Phys. Rev. **D97**, 015004 (2018), arXiv:1708.08929 [hep-ph] .

[376] N. A. Kurinsky, T. C. Yu, Y. Hochberg, and B. Cabrera, Phys. Rev. **D99**, 123005 (2019), arXiv:1901.07569 [hep-ex] .

[377] A. H. Abdelhameed *et al.* (CRESST), (2019), arXiv:1904.00498 [astro-ph.CO] .

[378] M. Battaglieri *et al.*, in *U.S. Cosmic Visions: New Ideas in Dark Matter College Park, MD, USA, March 23-25, 2017* (2017) arXiv:1707.04591 [hep-ph] .

[379] P. W. Graham, D. E. Kaplan, S. Rajendran, and M. T. Walters, Phys. Dark Univ. **1**, 32 (2012), arXiv:1203.2531 [hep-ph] .

[380] R. Essig, T. Volansky, and T.-T. Yu, Phys. Rev. **D96**, 043017 (2017), arXiv:1703.00910 [hep-ph] .

[381] P. Agnes *et al.* (DarkSide), Phys. Rev. Lett. **121**, 111303 (2018), arXiv:1802.06998 [astro-ph.CO] .

[382] R. Agnese *et al.* (SuperCDMS), Phys. Rev. Lett. **121**, 051301 (2018), [erratum: Phys. Rev. Lett.122,no.6,069901(2019)], arXiv:1804.10697 [hep-ex] .

[383] O. Abramoff *et al.* (SENSEI), Phys. Rev. Lett. **122**, 161801 (2019), arXiv:1901.10478 [hep-ex] .

[384] A. Aguilar-Arevalo *et al.* (DAMIC), Phys. Rev. Lett. **123**, 181802 (2019), arXiv:1907.12628 [astro-ph.CO] .

[385] D. Green and S. Rajendran, JHEP **10**, 013, arXiv:1701.08750 [hep-ph] .

[386] E. W. Kolb, M. S. Turner, and T. P. Walker, Phys. Rev. **D34**, 2197 (1986).

[387] P. D. Serpico and G. G. Raffelt, Phys. Rev. **D70**, 043526 (2004), arXiv:astro-ph/0403417 [astro-ph] .

[388] K. Jedamzik and M. Pospelov, New J. Phys. **11**, 105028 (2009), arXiv:0906.2087 [hep-ph] .

[389] C. Boehm, M. J. Dolan, and C. McCabe, JCAP **08**, 041, arXiv:1303.6270 [hep-ph] .

[390] S. Knapen, T. Lin, and K. M. Zurek, Phys. Rev. **D96**, 115021 (2017), arXiv:1709.07882 [hep-ph] .

[391] A. Berlin and N. Blinov, Phys. Rev. Lett. **120**, 021801 (2018), arXiv:1706.07046 [hep-ph] .

[392] G. Krnjaic, JHEP **10**, 136, arXiv:1711.11038 [hep-ph] .

[393] P. F. Depta, M. Hufnagel, K. Schmidt-Hoberg, and S. Wild, JCAP **1904**, 029, arXiv:1901.06944 [hep-ph] .

[394] A. Berlin, N. Blinov, and S. W. Li, Phys. Rev. **D100**, 015038 (2019), arXiv:1904.04256 [hep-ph] .

[395] M. C. Digman, C. V. Cappiello, J. F. Beacom, C. M. Hirata, and A. H. G. Peter, Phys. Rev. **D100**, 063013 (2019), arXiv:1907.10618 [hep-ph] .

[396] K. Bondarenko, A. Boyarsky, T. Bringmann, M. Hufnagel, K. Schmidt-Hoberg, and A. Sokolenko, (2019), arXiv:1909.08632 [hep-ph] .

[397] N. Sabti, J. Alvey, M. Escudero, M. Fairbairn, and D. Blas, (2019), arXiv:1910.01649 [hep-ph] .

[398] J. H. Chang, R. Essig, and A. Reinert, (2019), arXiv:1911.03389 [hep-ph] .

[399] J. Fiaschi, M. Klasen, M. Vargas, C. Weinheimer, and S. Zeinstra, JHEP **11**, 129, arXiv:1908.09882 [hep-ph] .

[400] E. Bertuzzo, C. J. Caniu Barros, and G. Grilli di Cortona, JHEP **09**, 116, arXiv:1707.00725 [hep-ph] .

[401] D. Choudhury and D. Sachdeva, Phys. Rev. **D100**, 035007 (2019), arXiv:1903.06049 [hep-ph] .

[402] A. Caputo, A. Esposito, and A. D. Polosa, in *16th International Conference on Topics in Astroparticle and Underground Physics (TAUP 2019) Toyama, Japan, September 9-13, 2019* (2019) arXiv:1911.07867 [hep-ph] .

[403] G. Arcadi, A. Djouadi, and M. Raidal, (2019), arXiv:1903.03616 [hep-ph] .

[404] J. F. Nieves and P. B. Pal, Am. J. Phys. **72**, 1100 (2004), arXiv:hep-ph/0306087 .

[405] R. H. Cyburt, B. D. Fields, K. A. Olive, and E. Skillman, Astropart. Phys. **23**, 313 (2005), arXiv:astro-ph/0408033 [astro-ph] .

[406] G. Mangano, G. Miele, S. Pastor, and M. Peloso, Phys. Lett. **B534**, 8 (2002), arXiv:astro-ph/0111408 [astro-ph] .

[407] G. Mangano, G. Miele, S. Pastor, T. Pinto, O. Pisanti, and P. D. Serpico, Nucl. Phys. **B729**, 221 (2005), arXiv:hep-ph/0506164 [hep-ph] .

[408] P. A. R. Ade *et al.* (Planck), Astron. Astrophys. **594**, A13 (2016), arXiv:1502.01589 [astro-ph.CO] .

[409] K. N. Abazajian *et al.* (CMB-S4), (2016), arXiv:1610.02743 [astro-ph.CO] .

[410] C. M. Ho and R. J. Scherrer, Phys. Rev. **D87**, 023505 (2013), arXiv:1208.4347 [astro-ph.CO] .

[411] C. Brust, D. E. Kaplan, and M. T. Walters, JHEP **12**, 058, arXiv:1303.5379 [hep-ph] .

[412] K. Enqvist, K. Kainulainen, and V. Semikoz, Nucl. Phys. **B374**, 392 (1992).

[413] M. Escudero, JCAP **1902**, 007, arXiv:1812.05605 [hep-ph] .

[414] C. Boehm, M. J. Dolan, and C. McCabe, JCAP **1212**, 027, arXiv:1207.0497 [astro-ph.CO] .

[415] M. Escudero Abenza, (2020), arXiv:2001.04466 [hep-ph] .

[416] N. Aghanim *et al.* (Planck), (2018), arXiv:1807.06209 [astro-ph.CO] .

[417] S. Reddy, M. Prakash, and J. M. Lattimer, Phys. Rev. D **58**, 013009 (1998), arXiv:astro-ph/9710115 .

[418] J. D. Lewin and P. F. Smith, Astropart. Phys. **6**, 87 (1996).

[419] H. K. Dreiner, H. E. Haber, and S. P. Martin, Phys. Rept. **494**, 1 (2010), arXiv:0812.1594 [hep-ph] .

[420] J. A. Evans, A. Ghalsasi, S. Gori, M. Tammaro, and J. Zupan, JHEP **02**, 151, arXiv:1910.06319 [hep-ph] .

[421] S. K. Lee, M. Lisanti, S. Mishra-Sharma, and B. R. Safdi, Phys. Rev. D **92**, 083517 (2015), arXiv:1508.07361 [hep-ph] .

[422] J. Alexander *et al.* (2016) arXiv:1608.08632 [hep-ph] .

[423] Y. Hochberg, Y. Kahn, M. Lisanti, C. G. Tully, and K. M. Zurek, Phys. Lett. B **772**, 239 (2017), arXiv:1606.08849 [hep-ph] .

[424] B. J. Kavanagh, R. Catena, and C. Kouvaris, JCAP **01**, 012, arXiv:1611.05453 [hep-ph] .

[425] T. Emken, C. Kouvaris, and I. M. Shoemaker, Phys. Rev. D **96**, 015018 (2017), arXiv:1702.07750 [hep-ph] .

[426] T. Emken and C. Kouvaris, JCAP **10**, 031, arXiv:1706.02249 [hep-ph] .

[427] G. Cavoto, F. Luchetta, and A. Polosa, Phys. Lett. B **776**, 338 (2018), arXiv:1706.02487 [hep-ph] .

[428] R. Essig, M. Sholapurkar, and T.-T. Yu, Phys. Rev. D **97**, 095029 (2018), arXiv:1801.10159 [hep-ph] .

[429] T. Emken and C. Kouvaris, Phys. Rev. D **97**, 115047 (2018), arXiv:1802.04764 [hep-ph] .

[430] Y. Ema, F. Sala, and R. Sato, Phys. Rev. Lett. **122**, 181802 (2019), arXiv:1811.00520 [hep-ph] .

[431] R. M. Geilhufe, B. Olsthoorn, A. Ferella, T. Koski, F. Kahlhoefer, J. Conrad, and A. V. Balatsky, Phys. Status Solidi RRL **12**, 1800293 (2018), arXiv:1806.06040 [cond-mat.mtrl-sci] .

[432] D. Baxter, Y. Kahn, and G. Krnjaic, Phys. Rev. D **101**, 076014 (2020), arXiv:1908.00012 [hep-ph] .

[433] R. Essig, J. Pradler, M. Sholapurkar, and T.-T. Yu, Phys. Rev. Lett. **124**, 021801 (2020), arXiv:1908.10881 [hep-ph] .

[434] T. Emken, R. Essig, C. Kouvaris, and M. Sholapurkar, JCAP **09**, 070, arXiv:1905.06348 [hep-ph] .

[435] Y. Hochberg, I. Charaev, S.-W. Nam, V. Verma, M. Colangelo, and K. K. Berggren, Phys. Rev. Lett. **123**, 151802 (2019), arXiv:1903.05101 [hep-ph] .

[436] T. Trickle, Z. Zhang, K. M. Zurek, K. Inzani, and S. Griffin, JHEP **03**, 036, arXiv:1910.08092 [hep-ph] .

[437] S. M. Griffin, K. Inzani, T. Trickle, Z. Zhang, and K. M. Zurek, Phys. Rev. D **101**, 055004 (2020), arXiv:1910.10716 [hep-ph] .

[438] A. Coskuner, A. Mitridate, A. Olivares, and K. M. Zurek, (2019), arXiv:1909.09170 [hep-ph] .

[439] R. M. Geilhufe, F. Kahlhoefer, and M. W. Winkler, Phys. Rev. D **101**, 055005 (2020), arXiv:1910.02091 [hep-ph] .

[440] R. Catena, T. Emken, N. A. Spaldin, and W. Tarantino, Phys. Rev. Res. **2**, 033195 (2020), arXiv:1912.08204 [hep-ph] .

[441] C. Blanco, J. Collar, Y. Kahn, and B. Lillard, Phys. Rev. D **101**, 056001 (2020), arXiv:1912.02822 [hep-ph] .

[442] N. Kurinsky, D. Baxter, Y. Kahn, and G. Krnjaic, Phys. Rev. D **102**, 015017 (2020), arXiv:2002.06937 [hep-ph] .

[443] S. M. Griffin, Y. Hochberg, K. Inzani, N. Kurinsky, T. Lin, and T. C. Yu, (2020), arXiv:2008.08560 [hep-ph] .

[444] A. Radick, A.-M. Taki, and T.-T. Yu, (2020), arXiv:2011.02493 [hep-ph] .

[445] G. B. Gelmini, V. Takhistov, and E. Vitagliano, Phys. Lett. B **809**, 135779 (2020), arXiv:2006.13909 [hep-ph] .

[446] T. Trickle, Z. Zhang, and K. M. Zurek, (2020), arXiv:2009.13534 [hep-ph] .

[447] P. Du, D. Egana-Ugrinovic, R. Essig, and M. Sholapurkar, (2020), arXiv:2011.13939 [hep-ph] .

[448] R. Essig, A. Manalaysay, J. Mardon, P. Sorensen, and T. Volansky, Phys. Rev. Lett. **109**, 021301 (2012), arXiv:1206.2644 [astro-ph.CO] .

[449] J. Tiffenberg, M. Sofo-Haro, A. Drlica-Wagner, R. Essig, Y. Guardincerri, S. Holland, T. Volansky, and T.-T. Yu (SENSEI), Phys. Rev. Lett. **119**, 131802 (2017), arXiv:1706.00028 [physics.ins-det] .

[450] R. Romani *et al.*, Appl. Phys. Lett. **112**, 043501 (2018), arXiv:1710.09335 [physics.ins-det] .

[451] M. Crisler, R. Essig, J. Estrada, G. Fernandez, J. Tiffenberg, M. Sofo haro, T. Volansky, and T.-T. Yu (SENSEI), Phys. Rev. Lett. **121**, 061803 (2018), arXiv:1804.00088 [hep-ex] .

[452] M. Settimo (DAMIC), in *53rd Rencontres de Moriond on Cosmology* (2018) pp. 315–318, arXiv:1805.10001 [astro-ph.IM] .

[453] D. Akerib *et al.* (LUX), Phys. Rev. Lett. **122**, 131301 (2019), arXiv:1811.11241 [astro-ph.CO] .

[454] E. Aprile *et al.* (XENON), Phys. Rev. Lett. **123**, 251801 (2019), arXiv:1907.11485 [hep-ex] .

[455] L. Barak *et al.* (SENSEI), Phys. Rev. Lett. **125**, 171802 (2020), arXiv:2004.11378 [astro-ph.CO] .

[456] Q. Arnaud *et al.* (EDELWEISS), Phys. Rev. Lett. **125**, 141301 (2020), arXiv:2003.01046 [astro-ph.GA] .

[457] D. Amaral *et al.* (SuperCDMS), Phys. Rev. D **102**, 091101 (2020), arXiv:2005.14067 [hep-ex] .

[458] N. W. Ashcroft and N. D. Mermin, *Solid State Physics* (Saunders College Publishing, 1976).

[459] D. Pines, Rev. Mod. Phys. **28**, 184 (1956).

[460] P. Nozieres and D. Pines, Physical Review **113**, 1254 (1959).

[461] H. Raether, *Excitation of plasmons and interband transitions by electrons*, Vol. 88 (Springer, 2006).

[462] W. Schülke, *Electron Dynamics by Inelastic X-Ray Scattering*, Oxford Science Publications (OUP Oxford, 2007).

[463] I. M. Bloch, R. Essig, K. Tobioka, T. Volansky, and T.-T. Yu, JHEP **06**, 087, arXiv:1608.02123 [hep-ph] .

[464] P. Gibbons, S. Schnatterly, J. Ritsko, and J. Fields, Physical Review B **13**, 2451 (1976).

[465] N. Bachar, D. Stricker, S. Muleady, K. Wang, J. Mydosh, Y. Huang, and D. van der Marel, Physical Review B **94**, 235101 (2016).

[466] X. Chu, T. Hambye, and M. H. Tytgat, JCAP **05**, 034, arXiv:1112.0493 [hep-ph] .

[467] C. Dvorkin, T. Lin, and K. Schutz, Phys. Rev. D **99**, 115009 (2019), arXiv:1902.08623 [hep-ph] .

[468] M. Dressel, G. Gruner, and G. Grüner, *Electrodynamics of Solids: Optical Properties of Electrons in Matter* (Cambridge University Press, 2002).

[469] Y. Sun, H. Xu, B. Da, S.-f. Mao, and Z.-j. Ding, Chinese Journal of Chemical Physics **29**, 663 (2016).

[470] J. A. Dror, B. V. Lehmann, P. H. Hiren, and S. Profumo, to appear .

[471] S. Knapen, J. Kozaczuk, and T. Lin, (2020), arXiv:2011.09496 [hep-ph] .

[472] H.-C. Weissker, J. Serrano, S. Huotari, E. Luppi, M. Cazzaniga, F. Bruneval, F. Sottile, G. Monaco, V. Olevano, and L. Reining, Physical Review B **81**, 085104 (2010).

[473] J. P. Walter and M. L. Cohen, Physical Review B **5**, 3101 (1972).

[474] M. K. Kundmann, *Study of semiconductor valence plasmon line shapes via electron energy-loss spectroscopy in the transmission electron microscope*, Tech. Rep. (Lawrence Berkeley Lab., CA (USA), 1988).

[475] T. O. Wehling, A. M. Black-Schaffer, and A. V. Balatsky, Adv. Phys. **63**, 1 (2014), arXiv:1405.5774 [cond-mat.mtrl-sci] .

[476] D. E. Kharzeev, R. D. Pisarski, and H.-U. Yee, Physical review letters **115**, 236402 (2015).

[477] J. Hofmann and S. D. Sarma, Physical Review B **91**, 241108 (2015).

[478] G. S. Jenkins, C. Lane, B. Barbiellini, A. B. Sushkov, R. L. Carey, F. Liu, J. W. Krizan, S. K. Kushwaha, Q. Gibson, T.-R. Chang, and et al., Physical Review B **94**, 10.1103/physrevb.94.085121 (2016).

[479] A. Thakur, R. Sachdeva, and A. Agarwal, Journal of Physics: Condensed Matter **29**, 105701 (2017).

[480] V. Kozii and L. Fu, Physical Review B **98**, 041109 (2018).

[481] G. R. Stewart, Rev. Mod. Phys. **56**, 755 (1984).

[482] P. S. Riseborough, Advances in Physics **49**, 257 (2000).

[483] P. Coleman (2015) arXiv:1509.05769 [cond-mat.str-el] .

[484] A. Millis, M. Lavagna, and P. Lee, Physical Review B **36**, 864 (1987).

[485] C. Bareille, F. Boariu, H. Schwab, P. Lejay, F. Reinert, and A. Santander-Syro, Nature Communications **5**, 1 (2014).

[486] O. Krivanek, N. Dellby, J. Hachtel, J.-C. Idrobo, M. Hotz, B. Plotkin-Swing, N. Bacon, A. Bleloch, G. Corbin, M. Hoffman, C. Meyer, and T. Lovejoy, Ultramicroscopy **203**, 60 (2019), 75th Birthday of Christian Colliex, 85th Birthday of Archie Howie, and 75th Birthday of Hannes Lichte / PICO 2019 - Fifth Conference on Frontiers of Aberration Corrected Electron Microscopy.

[487] C. Blanco, Y. Kahn, B. Lillard, and S. D. McDermott, (2021), arXiv:2103.08601 [hep-ph] .

[488] Y. Hochberg, Y. Kahn, N. Kurinsky, B. V. Lehmann, T. C. Yu, and K. K. Berggren, (2021), arXiv:2101.08263 [hep-ph] .

[489] S. Tremaine and J. E. Gunn, Phys. Rev. Lett. **42**, 407 (1979).

[490] A. Boyarsky, J. Lesgourgues, O. Ruchayskiy, and M. Viel, JCAP **05**, 012, arXiv:0812.0010 [astro-ph] .

[491] J. Bardeen, L. N. Cooper, and J. R. Schrieffer, Phys. Rev. **108**, 1175 (1957).

[492] N. N. Bogolyubov, Sov. Phys. JETP **7**, 41 (1958).

[493] M. Tinkham, *Introduction to Superconductivity*, 2nd ed. (Dover Publications, 2004).

[494] Y. Hochberg, E. D. Kramer, N. Kurinsky, and B. V. Lehmann, to appear (2021).

[495] W. D. Gregory, Phys. Rev. Lett. **20**, 53 (1968).

[496] J.-J. Chang and D. J. Scalapino, Phys. Rev. B **15**, 2651 (1977).

[497] J. Bardeen, G. Rickayzen, and L. Tewordt, Phys. Rev. **113**, 982 (1959).

[498] A. G. Kozorezov, A. F. Volkov, J. K. Wigmore, A. Peacock, A. Poelaert, and R. den Hartog, Phys. Rev. B **61**, 11807 (2000).

[499] Y. N. Ovchinnikov and V. Z. Kresin, Phys. Rev. B **58**, 12416 (1998).

[500] S. Kaplan, C. Chi, D. Langenberg, J. Chang, S. Jafarey, and D. Scalapino, Phys. Rev. B **14**, 4854 (1976).

[501] H. An, M. Pospelov, J. Pradler, and A. Ritz, Phys. Rev. Lett. **120**, 141801 (2018), [Erratum: Phys.Rev.Lett. 121, 259903 (2018)], arXiv:1708.03642 [hep-ph] .

[502] J. B. Dent, B. Dutta, J. L. Newstead, I. M. Shoemaker, and N. T. Arellano, Phys. Rev. D **103**, 095015 (2021), arXiv:2010.09749 [hep-ph] .

[503] H. An, H. Nie, M. Pospelov, J. Pradler, and A. Ritz, (2021), arXiv:2108.10332 [hep-ph] .

[504] R. J. Gaitskell, *Non-equilibrium superconductivity in niobium and its application to particle detection*, Ph.D. thesis, University of Oxford (1993).

[505] I. Esmaeil Zadeh, J. Chang, J. W. N. Los, S. Gyger, A. W. Elshaari, S. Steinhauer, S. N. Dorenbos, and V. Zwiller, Applied Physics Letters **118**, 190502 (2021), https://doi.org/10.1063/5.0045990 .

[506] M. E. Grein, O. Shatrovoy, D. V. Murphy, B. S. Robinson, and D. Boroson, Conference on Lasers and Electro-Optics (CLEO 2014) , SM4J.4 (2014).

[507] Y. P. Korneeva, D. Y. Vodolazov, A. V. Semenov, I. N. Florya, N. Simonov, E. Baeva, A. A. Korneev, G. N. Goltsman, and T. M. Klapwijk, Conference on Lasers and Electro-Optics (CLEO 2014) , SM4J.5 (2014).

[508] C. M. Natarajan, M. M. Härtig, R. E. Warburton, G. S. Buller, R. H. Hadfield, B. Baek, S. W. Nam, S. Miki, M. Fujiwara, M. Sasaki, and Z. Wang, in *Quantum Communication and Quantum Networking*, edited by A. Sergienko, S. Pascazio, and P. Villoresi (Springer Berlin Heidelberg, Berlin, Heidelberg, 2010) pp. 225–232.

[509] V. B. Verma *et al.*, (2020), arXiv:2012.09979 [physics.ins-det] .

[510] V. B. Verma, B. Korzh, A. B. Walter, A. E. Lita, R. M. Briggs, M. Colangelo, Y. Zhai, E. E. Wollman, A. D. Beyer, J. P. Allmaras, H. Vora, D. Zhu, E. Schmidt, A. G. Kozorezov, K. K. Berggren, R. P. Mirin, S. W. Nam, and M. D. Shaw, APL Photonics **6**, 056101 (2021), https://doi.org/10.1063/5.0048049 .

[511] J. Gaffiot (SOX), Nucl. Part. Phys. Proc. **265-266**, 129 (2015).

[512] Q. Chen *et al.*, (2020), arXiv:2011.06699 [physics.app-ph] .

[513] F. Marsili *et al.*, Nano letters **12**, 4799 (2012).

[514] J. Chang, J. W. N. Los, R. Gourgues, S. Steinhauer, S. N. Dorenbos, S. F. Pereira, H. P. Urbach, V. Zwiller, and I. E. Zadeh, (2021), arXiv:2107.06354 [physics.ins-det] .

[515] E. E. Wollman *et al.*, Opt. Express **25**, 26792 (2017), arXiv:1708.04231 [physics.ins-det] .

[516] L. Kong, Q. Zhao, H. Wang, J. Guo, H. Lu, H. Hao, S. Guo, X. Tu, L. Zhang, X. Jia, L. Kang, X. Wu, J. Chen, and P. Wu, Nano Letters **21**, 9625 (2021).

[517] Y. Hochberg, E. D. Kramer, N. Kurinsky, and B. V. Lehmann, (2021), arXiv:2109.04473 [hep-ph] .

[518] A. M. Bhargav, R. K. Rakshit, S. Das, and M. Singh, Advanced Quantum Technologies **4**, 2100008 (2021).

[519] J. Chiles *et al.*, (2021), arXiv:2110.01582 [hep-ex] .

[520] R. Lasenby and A. Prabhu, (2021), arXiv:2110.01587 [hep-ph] .

[521] S. Knapen, J. Kozaczuk, and T. Lin, (2021), arXiv:2101.08275 [hep-ph] .

[522] J. Kischkat, S. Peters, B. Gruska, M. Semtsiv, M. Chashnikova, M. Klinkmüller, O. Fedosenko, S. Machulik, A. Aleksandrova, G. Monastyrskyi, Y. Flores, and W. T. Masselink, Appl. Opt. **51**, 6789 (2012).

[523] H. An, M. Pospelov, J. Pradler, and A. Ritz, Phys. Lett. **B747**, 331 (2015), arXiv:1412.8378 [hep-ph] .

[524] A. Andrianavalomahefa *et al.* (FUNK Experiment), Phys. Rev. D **102**, 042001 (2020), arXiv:2003.13144 [astro-ph.CO] .

[525] H. An, M. Pospelov, and J. Pradler, Phys. Rev. Lett. **111**, 041302 (2013), arXiv:1304.3461 [hep-ph] .

[526] H. An, M. Pospelov, J. Pradler, and A. Ritz, Phys. Rev. D **102**, 115022 (2020), arXiv:2006.13929 [hep-ph] .

[527] D. Nguyen, D. Sarnaaik, K. K. Boddy, E. O. Nadler, and V. Gluscevic, (2021), arXiv:2107.12380 [astro-ph.CO] .

[528] M. A. Buen-Abad, R. Essig, D. McKeen, and Y.-M. Zhong, (2021), arXiv:2107.12377 [astro-ph.CO] .

[529] C. Giovanetti, M. Lisanti, H. Liu, and J. T. Ruderman, (2021), arXiv:2109.03246 [hep-ph] .

[530] S. L. Adler, Phys. Rev. **126**, 413 (1962).

[531] N. Wiser, Phys. Rev. **129**, 62 (1963).

[532] K. Freese, M. Lisanti, and C. Savage, Rev. Mod. Phys. **85**, 1561 (2013), arXiv:1209.3339 [astro-ph.CO] .

[533] S. K. Lee, M. Lisanti, and B. R. Safdi, JCAP **11**, 033, arXiv:1307.5323 [hep-ph] .

[534] M. G. Kitzbichler and S. D. White, Mon. Not. Roy. Astron. Soc. **391**, 1489 (2008), arXiv:0804.1965 [astro-ph] .

[535] A. Sesana, A. Vecchio, and M. Volonteri, Mon. Not. Roy. Astron. Soc. **394**, 2255 (2009), arXiv:0809.3412 [astro-ph] .

[536] T. Masuda *et al.*, PTEP **2016**, 013C03 (2016), arXiv:1509.03386 [physics.ins-det] .

[537] N. Carrasco, P. Lami, V. Lubicz, L. Riggio, S. Simula, and C. Tarantino, Phys. Rev. D **93**, 114512 (2016), arXiv:1602.04113 [hep-lat] .

[538] K. Sato, E. Iwai, K. Shiomi, Y. Sugiyama, M. Togawa, and T. Yamanaka, JPS Conf. Proc. **8**, 024007 (2015).

[539] A. Altland and B. D. Simons, *Condensed matter field theory* (Cambridge University Press, 2010).

[540] H. Banks and M. Mccullough, (2020), arXiv:2009.12399 [hep-ph] .

[541] G. D. Mahan, *Many-particle physics* (Springer Science & Business Media, 2013).

[542] H. Fröhlich, *Phenomenological theory of the energy loss of fast particles in solids* (VEB Deutscher Verlag der Wissenschaften, 1959).

[543] M. E. Peskin and D. V. Schroeder, *An Introduction to quantum field theory* (Addison-Wesley, Reading, USA, 1995).

[544] J. Mydosh and P. Oppeneer, Philosophical Magazine **94**, 3642–3662 (2014).

[545] T. Palstra, A. Menovsky, J. Van den Berg, A. Dirkmaat, P. Kes, G. Nieuwenhuys, and J. Mydosh, Physical Review Letters **55**, 2727 (1985).

[546] J. D. Bjorken and S. D. Drell, (1965).

[547] A. Mitridate, T. Trickle, Z. Zhang, and K. M. Zurek, (2021), arXiv:2106.12586 [hep-ph] .

[548] S. Weinberg, *The Quantum theory of fields. Vol. 1: Foundations* (Cambridge University Press, 2005).

[549] J. Polchinski, in *Theoretical Advanced Study Institute (TASI 92): From Black Holes and Strings to Particles* (1992) arXiv:hep-th/9210046 .

[550] T. Guruswamy, D. J. Goldie, and S. Withington, Superconductor Science and Technology **27**, 055012 (2014).

[551] J. M. Martinis, Saving superconducting quantum processors from qubit decay and correlated errors generated by gamma and cosmic rays (2020), arXiv:2012.06137 [quant-ph] .

[552] J. Rammer, *Quantum Transport Theory*, Frontiers in Physics (Avalon Publishing, 2004).

[553] C. Jacoboni and L. Reggiani, Rev. Mod. Phys. **55**, 645 (1983).

Appendices

Appendix A

Numerical validation of PBH abundance optimization

Given a set of constraints, it is also possible to use numerical methods to find a mass function which maximizes the PBH density. There are significant caveats to such an approach. Most importantly, a maximization algorithm may converge to a local optimum rather than a global optimum. Additionally, computational costs may render numerical approaches impractical unless the functions involved are discretized sparsely. Even so, numerical optimization can be used to validate our analytical results: if the same set of masses is used for discretization, then the numerical result should never reach a greater normalized mass (cf. Eq. (1.19)) than that of our corresponding semi-analytical result. Numerical methods can also be used to check that our semi-analytical optimum is a stationary point of the normalized mass functional.

A.1 Direct validation

We implement these validation steps using a simple Monte Carlo algorithm, as follows: we begin with an initial mass function of the form $\psi_0(M) \propto M^{-1}$, which assigns equal PBH density to each log-spaced mass bin. We then perturb the value of ψ_0 in a random bin k by a value selected from a Gaussian distribution with mean 0 and variance $\sigma^2\psi_0(M_k)^2$, where σ is a parameter of the maximization. We denote the resulting mass function by $\psi_1(M)$. If $\psi_1(M_k) \geq 0$ and $\mathcal{M}[\psi_1] > \mathcal{M}[\psi_0]$, we accept the step, replace ψ_0 by ψ_1, and repeat. For simplicity, we do not accept any steps which reduce the normalized mass. This is not necessary in order to test whether our semi-analytical optimum mass function is a stationary point. We also reject steps which increase the normalized mass by less than 10^{-10} to avoid exceeding the numerical precision of the semi-analytical result.

In order to make the problem numerically tractable, we use only 10^2 log-spaced mass bins. This discretization is different from the one used in Table 1.1, and it does not capture sharp features of the constraints. Consequently, in order to compare the numerical results with semi-analytical results, we regenerate the semi-analytical mass function with the same discretization. Note that this affects both the form of the optimal mass function and the calculated $f_{\rm max,all}$.

We implement the numerical optimization with $\sigma = 10^{-2}$. In what follows, we denote the numerical mass function by $\psi_{\rm N}$, and the semi-analytical optimum by $\psi_{\rm SA}$. The left-hand side of Fig. A.1 shows $f_{\rm PBH}$ for the numerical mass function at each step as a fraction of the semi-analytical $f_{\rm max,all}$. The numerical $f_{\rm PBH}$ converges to $f_{\rm max,all}$

and immediately stabilizes, and in particular, in no step does $f_{\rm PBH}$ exceed $f_{\rm max,all}$.

In principle, $\psi_{\rm N}$ need not converge to $\psi_{\rm SA}$ even given that $f_{\rm PBH}$ converges to $f_{\rm max,all}$, since the mass function with maximal density is not necessarily unique. However, in the top-right panel of Fig. A.1, we show that $\psi_{\rm N}$ tends to $\psi_{\rm SA}$ in the L^2 norm. To compute this distance consistently, we treat the Dirac deltas of $\psi_{\rm SA}$ as constant functions in their respective bins. As an additional test of convergence, we compute the acceptance rate, i.e., the fraction of steps which are accepted, during each window of 10^4 iterations. The acceptance rate vanishes as $\psi_{\rm N}$ approaches $\psi_{\rm SA}$, which further demonstrates that $\psi_{\rm SA}$ is a stationary point of the normalized mass.

The numerical and semi-analytical mass functions are shown in Fig. A.2. In order to compare Dirac deltas with the smooth mass function $\psi_{\rm N}$, the figure shows the integral of the mass function in each bin rather than $\psi_{\rm N}$ and $\psi_{\rm SA}$ themselves. It is clear that in this case, the numerical algorithm converges to the semi-analytical optimum. We have established via analytical arguments that this is not simply a local optimum, but indeed the global maximum of the normalized mass.

A.2 Sensitivity to the constraint prescription

Our analytical work is based on the prescription of [39] for evaluating constraints on extended mass functions. Since other prescriptions have been considered in the literature, it is important to determine the robustness of our results to variations on the constraint prescription. Assessing this analytically is intractable, as it requires the development of independent analytical frameworks for even slight modifications. How-

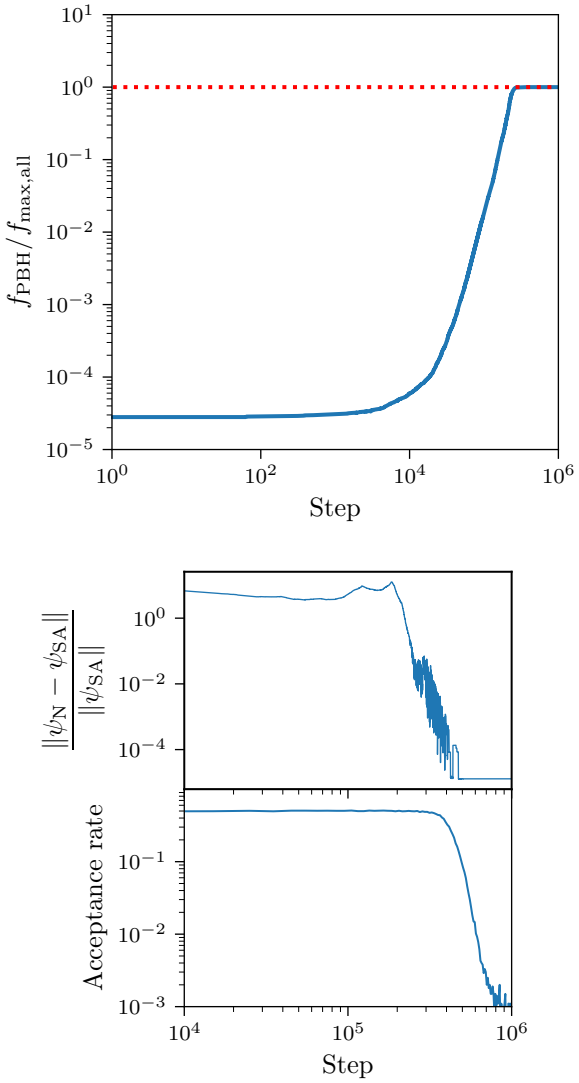

Figure A.1: Left: f_{PBH} attained in each step during numerical maximization, shown as a fraction of the semi-analytical $f_{\mathrm{max,all}}$. The dashed red line indicates $f_{\mathrm{PBH}} = f_{\mathrm{max,all}}$. Right top: L^2 norm of the difference between ψ_{N} (numerical) and ψ_{SA} (semi-analytical) mass functions for each step, shown as a fraction of $\|\psi_{\mathrm{SA}}\|$. In computing the norm, ψ_{SA} is treated as a step function on the mass bins. Right bottom: acceptance rate in bins of 10^4 steps.

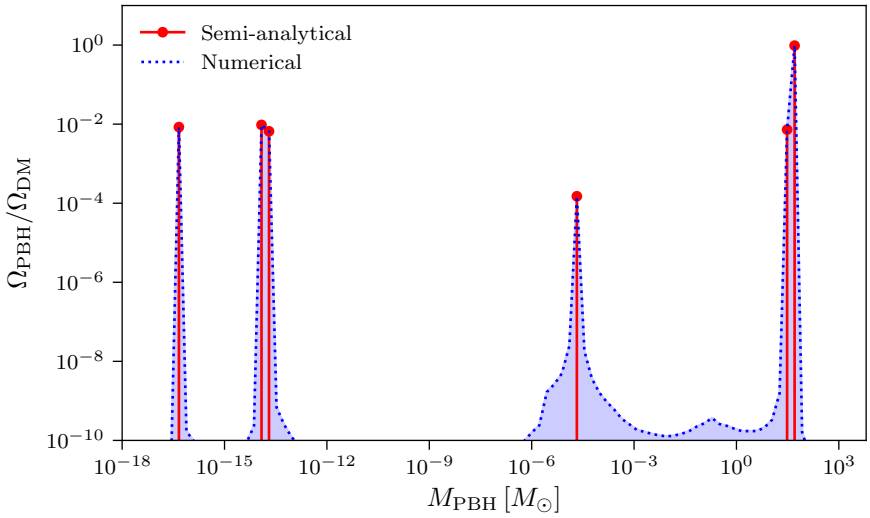

Figure A.2: Blue: numerically-optimized mass function $\psi_{\rm N}$ after 10^6 steps. Red: semi-analytical optimum $\psi_{\rm SA}$. Each curve shows the integral of the mass function in each bin, i.e., the total contribution of that bin to $f_{\rm PBH}$.

ever, numerical methods allow for a comparison of the bounds we obtain analytically with those that would result from any other specified prescription. We thus perform numerical optimization under the prescriptions of [40] and [29], and compare these with our semi-analytical results.

The situation is particularly simple for the constraint prescription of [40]: the major difference is that a mass function is allowed if it is allowed according to each individual constraint, rather than according to their statistical combination. Thus, the normalized mass $\mathcal{M}[\psi]$ is replaced by

$$\widehat{\mathcal{M}}[\psi] \equiv \frac{\int \mathrm{d}M\, \psi(M)}{\max_{j=1,\ldots,N} \mathcal{C}_j[\psi]}. \tag{A.1}$$

It is straightforward to implement numerical optimization with respect to $\widehat{\mathcal{M}}[\psi]$ in place of $\mathcal{M}[\psi]$. For the case shown in Fig. A.2, these two numerical maxima agree to within 1%.

We also implement the constraint procedure of [29], for which the constraints are treated as step functions on a set of mass bins. There is no universal prescription for the size of the bins across all constraints, but they should be chosen small enough that the minimum of each constraint function is not very different from its maximum within any single bin. Altogether, the procedure is as follows:

1. The mass range is divided into bins $I_1, \ldots, I_n$.

2. In each bin, the dominant constraint is identified. If a bin captures the transition between two dominant constraints, the bin is subdivided at the transition point.

3. We evaluate the constraint of Eq. (1.2), considering only the dominant constraint in each bin. We treat ψ as a smooth function, using a more refined set of bins for its numerical representation.

As in Appendix A.1, we take a relatively coarse binning of the mass range for numerical testing purposes. We have evaluated the maxima attained for a range of different constraint bin counts, and we find that our determination of $f_{\rm max,all}$ is robust to changes in binning at the 10% level.

We note as well that prescriptions of this kind have been criticized in the literature for the fact that it is not trivial to determine the range of validity of the individual constraints, and hence to determine the limits of integration for each constraint curve in Eq. (1.17). This is indeed a concern when individual constraints are considered, and it introduces significant potential uncertainty in cases where there are masses for which all constraints are at the edge of their range of validity. In these scenarios—for instance, with the constraints of set **A**—there are effectively "windows" in the constraints that

make any upper bound on the density of PBH quite uncertain. However, in the other cases that we consider, the dominating constraints in a given mass range generally intersect well within their respective ranges of validity. Thus, modifying the limits of integration does not substantially impact our results.

Appendix B

Distribution of SMBH binaries

The prediction of the SGWB spectrum relies on the statistical properties of the sources, which must be extracted from astronomical observations. This enters into the calculation of Eq. (5.2) via the quantity

$$\frac{\mathrm{d}n_{\mathrm{s}}}{\mathrm{d}z\,\mathrm{d}\mathbf{X}} = \frac{\mathrm{d}n_{\mathrm{s}}}{\mathrm{d}z\,\mathrm{d}M_1\,\mathrm{d}M_2}\,, \tag{B.1}$$

where n_{s} is the comoving number density of SMBH binaries. The parameter distributions of such binaries are not measured directly. Instead, observations measure the population statistics of galaxies and the rate at which these galaxies merge. Following Ref. [276], we combine these data with observational relations between galaxies and SMBHs to infer the properties of the SMBH binary population.

This process can be carried out in many different ways, using different astronomical datasets and SMBH–host relations. We now detail the method we use to predict the observed SGWB. As detailed in Ref. [276], the various prescriptions introduce an order-of-magnitude uncertainty in the amplitude of the SGWB. The amplitude is relatively unimportant for the present study since we are most interested in modifications

to the spectral shape. However, note that the spectral shape is in principle sensitive to the redshift distribution of sources, so uncertainties in the relative populations of sources at each redshift propagate to (small) uncertainties in the shape.

We infer the distribution of SMBH binary mergers from the distribution of galaxy mergers, under the assumption that each galaxy hosts an SMBH with mass related to the galaxy mass. Given an SMBH mass M_i, we will denote the host galaxy mass by $M_i^{\rm G}$. To better conform with the astronomy literature, we let M_1 denote the mass of the heavier SMBH, and we use the mass ratio $q \equiv M_2^{\rm G}/M_1^{\rm G} < 1$ instead of $M_2^{\rm G}$ directly. We write the differential merger rate of galaxies per unit comoving volume as

$$\frac{{\rm d}n_{\rm G}}{{\rm d}z\,{\rm d}M_1^{\rm G}\,{\rm d}q} = \frac{\phi(M_1^{\rm G},z)}{M_1^{\rm G}\log 10}\frac{1}{\tau(z,M_1^{\rm G},q)}\frac{{\rm d}\mathscr{F}(z,M_1^{\rm G},q)}{{\rm d}q}\frac{{\rm d}t}{{\rm d}z}\,. \tag{B.2}$$

Here ${\rm d}t/{\rm d}z$ converts the rate with respect to time into a rate with respect to redshift; $\phi(M_{\rm G}) = {\rm d}n_{\rm G}/{\rm d}\log_{10} M_{\rm G}$ is the observed galaxy mass function as reported in the astronomy literature; $\tau(z, M_1^{\rm G}, q)$ is the merger timescale of a given galactic pair; and $\mathscr{F}(z, M_{\rm G}, q)$ is the differential pair fraction, i.e., the fraction of galaxies of mass $M_{\rm G}$ in binaries with mass ratio q at redshift z. We take the pair fraction directly from Ref. [287]. We take the galaxy mass function for $z > 0.2$ from Ref. [284], and we interpolate from $z = 0.2$ to the $z = 0$ mass function of Ref. [285]. We take the merger timescale τ from Eq. (10) of Ref. [534], and following Ref. [276], we include an additional factor of $\frac{1}{2}q^{-0.3}$. We assume that the resulting SMBH binary merges instantaneously, so that Eq. (B.1) and Eq. (B.2) are equivalent up to the change of variable $(M_1, M_2) \mapsto (M_1^{\rm G}, q)$. In other words, we compute the characteristic strain as

$$h_c^2(f) = \int {\rm d}z\,{\rm d}M_1^{\rm G}\,{\rm d}q\,\frac{{\rm d}n_{\rm G}}{{\rm d}z\,{\rm d}M_1^{\rm G}\,{\rm d}q}\frac{f_{\rm s}}{1+z}\left.\frac{{\rm d}E_{\rm GW}}{{\rm d}f_{\rm s}}\right|_{M_1^{\rm G},q}\frac{3H_0^2}{2\pi^2\rho_c f^2}. \tag{B.3}$$

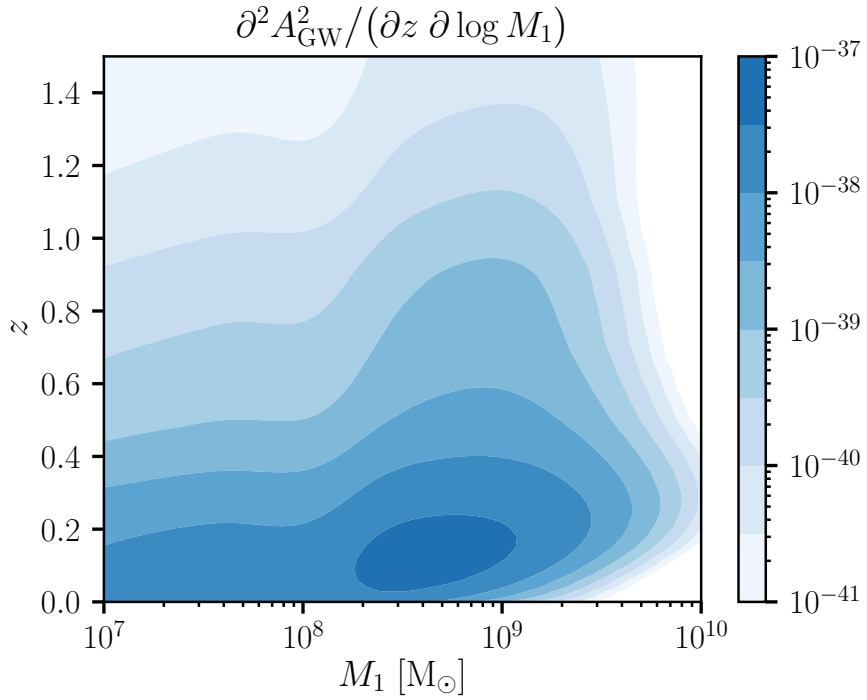

Figure B.1: Differential contribution to the squared amplitude of the SGWB as a function of M_1 and z in the gravity-only case. Here M_1 denotes the more massive component of the binary.

To relate the galaxy mass to the SMBH mass, there are two steps: first, we estimate the stellar mass of the galactic bulge, $M_{\rm bulge}$. Second, we use the scaling relation of Ref. [286] to determine the SMBH mass from the bulge mass. We approximate the bulge fraction $f_{\rm bulge} \equiv M_{\rm bulge}/M_{\rm G}$ as a function of $M_{\rm G}$, following Ref. [276]: the bulge fraction should be ~ 0.9 for galaxy masses above $10^{11}\,\mathrm{M}_\odot$, and ~ 0.25 for galaxy masses below $10^{10}\,\mathrm{M}_\odot$. We smoothly interpolate between these two bulge fractions as $f_{\rm bulge}(M_{\rm G}) = 0.55 - 0.22\tan^{-1}(4 - 1.3\times 10^{-10} M_{\rm G}/\mathrm{M}_\odot)$.

This approach is useful for making a simple estimate with appropriate redshift and mass ratio dependence. However, certain choices must be made to arrive at a single prediction of the SGWB amplitude: Ref. [276] computes many projected strains

based on different observed mass functions, pair fractions, and scaling relations. These estimates span more than a decade in $A_{\rm GW}$, but the values are slightly lower than other estimates, and are used as a "pessimistic" projection by the NANOGrav Collaboration [255]. Therefore, when following the calculation of Ref. [276], we make an optimistic set of choices to bring the normalization of the SGWB into line with the scenario considered to be "moderate" by the NANOGrav Collaboration. In particular, as discussed in that work, we add the uncertainties to the fitted mass function parameters and further add 0.1 dex (i.e., multiply by $10^{0.1}$) to account for systematics. We likewise add the scatter in the SMBH–host scaling relations to the fitted parameter values. We determine the masses of the SMBHs from the properties of the merged galaxy using the "double accretion" prescription of Ref. [535].

Figure B.1 shows the differential contribution to the squared amplitude $A^2_{\rm GW}$ of the SGWB as a function of M_1 and z in the gravity-only case, i.e., the integrand of Eq. (B.3) evaluated at $f = 1\,{\rm yr}^{-1}$. This indicates the relative contribution of different masses and redshifts to the SGWB signal given the observational data and scaling prescriptions used in this chapter. In particular, Fig. B.1 demonstrates that the signal is dominated by binaries with primary masses between $10^8\,{\rm M}_\odot$ and $10^9\,{\rm M}_\odot$ and redshifts $z \lesssim 0.3$.

Appendix C

KOTO simulation

In this appendix, we provide details of our calculation of the quantity R introduced in Eq. (6.23). R is the acceptance of the $K_L \to SP \to \pi^0 PP$ signal relative to the SM $K_L \to \pi^0 \nu\bar{\nu}$ acceptance at KOTO. Our calculation is based on a Monte Carlo simulation following steps similar to the ones described in [321, 332].

The layout of the KOTO beamline and the KOTO detector is described e.g. in [536]. We start by generating K_L momenta, p_{K_L}, and K_L decay vertex locations, z_{K_L}, based on the distribution

$$f(p_{K_L}, z_{K_L}) \propto g(p_{K_L}) \times \exp\left(-\frac{(z_{K_L} - z_{\text{exit}}) m_{K_L}}{\tau_{K_L} p_{K_L}}\right) , \tag{C.1}$$

where $z_{\text{exit}} = 20\,\text{m}$ is the distance of the beam exit from the target and $g(p_{K_L})$ is the measured K_L momentum distribution at the beam exit from [536]. We include a small transverse component of the K_L momentum such that the beam profile at the beam exit is constant within an $8.5\,\text{cm} \times 8.5\,\text{cm}$ square and zero outside [536].

In the case of the SM decay, we generate pion momenta using the $K \to \pi$ form factor from [537]. In the case of the $K_L \to SP \to \pi^0 PP$ decay, we first generate

momenta for S, based on the fixed energy of S in the K_L rest frame, $E_S = (m_{K_L}^2 + m_S^2 - m_P^2)/(2m_{K_L})$. We then decay S with a decay length distribution that is determined by the $S \to \pi^0 P$ and $S \to 3P$ partial widths. The pion momentum is generated based on the known pion energy in the S rest frame, $E_{\pi^0} = (m_S^2 + m_{\pi^0}^2 - m_P^2)/(2m_S)$.

Both in the SM case and the NP case, we let the pion decay promptly into two photons, each with energy $E_\gamma = m_{\pi^0}/2$ in the pion rest frame. We reject events with photons produced less than 2.5 m after the front face of the front barrel (which starts 1.507 m after the beam exit), as they would be rejected by photon veto collar counters. All other photons are propagated to the calorimeter located 6.148 m after the front face of the front barrel [536]. The energy and location of the detected photons in the calorimeter is smeared using the parameters given in [538].

Based on the smeared energy and smeared location of the photons in the calorimeter, the transverse momentum and decay vertex location of the pion is inferred following the procedure described in [536]. If there is more than one solution for the vertex location in the decay volume, we pick the location further away from the calorimeter. We then perform the event selection as in [319], taking into account all cuts but timing and shape related cuts and the trigger related cut on the center of energy deposition. We use the updated signal region in the plane of the inferred pion transverse momentum and the pion decay vertex location from [320].

The results for R in our benchmark scenarios are shown in Fig. 6.3 as function of the S lifetime.

Appendix D

Derivations and details in the dielectric formalism

In this appendix, we provide a number of derivations and further details to support the results in chapter Chapter 8. Appendix D.1 derives our main result for the DM scattering rate in terms of the loss function. Appendix D.2 outlines a number of simple analytic models for dielectric functions in various materials, and compares them to measured data for Al (representative of an ordinary superconductor), Si (a typical semiconductor), and URu_2Si_2 (an example of a heavy-fermion superconductor with meV-scale plasmons). We compare the free-electron gas (FEG) model for Si to the spectrum computed using crystal form factors generated by the publicly-available `QEdark` code [368], and show good qualitative agreement in the range 5–15 eV. Appendix D.3 is devoted to a detailed comparison of our results for superconductors with other results in the literature, justifying our claim of a stronger reach by several orders of magnitude compared to previous estimates, and gives the projected reach for heavy

mediators.

D.1 Scattering rate in terms of the loss function

Here we derive Eq. (8.1) and show how the scattering rate for all spin-independent DM-electron interactions is governed by the loss function. Suppose DM couples to electrons through a low-energy Hamiltonian of the form

$$\hat{H}_{\rm int} = \sum_i V(\hat{\mathbf{r}}_\chi - \hat{\mathbf{r}}_i), \tag{D.1}$$

where the sum runs over all electrons in the target. Fourier transforming the potential,

$$V(\hat{\mathbf{r}}_\chi - \hat{\mathbf{r}}_i) = \int \frac{\mathrm{d}^3\mathbf{q}}{(2\pi)^3} e^{i\mathbf{q}\cdot(\hat{\mathbf{r}}_\chi - \hat{\mathbf{r}}_i)} V(\mathbf{q}), \tag{D.2}$$

we can write the interaction Hamiltonian as

$$\hat{H}_{\rm int} = \int \frac{\mathrm{d}^3\mathbf{q}}{(2\pi)^3} e^{i\mathbf{q}\cdot\hat{\mathbf{r}}_\chi} V(\mathbf{q})\hat{\rho}(\mathbf{q}), \tag{D.3}$$

where the momentum-space electron density operator is defined as

$$\hat{\rho}(\mathbf{q}) = \int \mathrm{d}^3\mathbf{x} \sum_i \delta(\mathbf{x} - \hat{\mathbf{r}}_i) e^{-i\mathbf{q}\cdot\mathbf{x}} = \sum_i e^{-i\mathbf{q}\cdot\hat{\mathbf{r}}_i}. \tag{D.4}$$

By Fermi's Golden Rule (equivalently, the Born approximation), we can compute the transition rate $\Gamma(\mathbf{v}_\chi)$ from the ground state $|0\rangle$ for a given incoming DM velocity $\mathbf{v}_\chi$, treating the incoming and outgoing DM as plane waves with energy and momentum $(E_\chi, \mathbf{p}_\chi)$ and $(E'_\chi, \mathbf{p}'_\chi)$ respectively. We take the ground state to have zero energy

without loss of generality. The transition rate is given by [436]

$$\begin{aligned}\Gamma(\mathbf{v}_\chi) &= \sum_f \left|\langle f; \mathbf{p}'_\chi | \hat{H}_{\rm int} | 0; \mathbf{p}_\chi \rangle\right|^2 2\pi\delta(\omega_f + E'_\chi - E_\chi) \\ &= \int \frac{\mathrm{d}^3\mathbf{q}}{(2\pi)^3} |V(\mathbf{q})|^2 \sum_f \left|\langle f|\hat{\rho}(\mathbf{q})|0\rangle\right|^2 2\pi\delta(\omega_f - \omega_\mathbf{q}),\end{aligned} \tag{D.5}$$

where $|f\rangle$ is a final state with energy ω_f and the sum runs over all possible final states of the system, and we recall that

$$\omega_\mathbf{q} = \mathbf{q}\cdot\mathbf{v}_\chi - \frac{q^2}{2m_\chi}. \tag{D.6}$$

Note that the only assumption that was made here was that $\hat{H}_{\rm int}$ is sufficiently weak compared to the unperturbed Hamiltonian $\hat{H}_0$ of the target system; this is the case in Ref. [460] for ordinary electron-electron scattering, so it must be the case for DM-electron scattering where the couplings are much weaker. Note this implies one cannot directly apply our result to regions of parameter space where DM and electrons are strongly coupled, as would be relevant for regions in parameter space where DM may not reach underground detectors due to multiple scattering [424–426, 429, 434].

The insight of Ref. [460] is to relate the density matrix element $|\langle f|\hat{\rho}(\mathbf{q})|0\rangle|^2$ to an experimentally measurable quantity, the dielectric function $\epsilon(\mathbf{q},\omega)$. The dielectric function is *defined* as the linear response of the target to the longitudinal electric field of a test charge. For simplicity and to elucidate the formalism, in this chapter we consider the case of an isotropic material where the dielectric function is a scalar rather than a tensor, and relegate the treatment of the anisotropic case to upcoming work [470]. Since a test charge will also perturb the electron density of the target, it can be shown that this is equivalent to defining the dielectric function as a density-density correlation

function [539]. Of course, the electrons will also couple to ions, and strictly speaking the ion density operator should also appear in the measured loss function. In what follows, we will assume that these contributions are negligible, which is the approximation always made in the condensed matter literature. This assumption makes our formalism independent of the DM coupling to protons or neutrons, and even if ion contributions are significant, the measured loss function will give *exactly* the correct rate for a dark photon mediator.

Because the dielectric function is defined as the *linear* response of the system, the same assumptions are implicit in the setup of [460] (with an electromagnetic probe) as are present in the DM scattering setup: the test charge interactions are weak compared to the internal interactions $\hat{H}_0$. Therefore the Coulomb potential of the test charge may be factored out in Fourier space, separating the (weak) perturbation due to the probe and the (possibly strong) response of the system to such a probe. The result is [460, 539]

$$\mathrm{Im}\left(-\frac{1}{\epsilon(\mathbf{q},\omega)}\right) = \frac{\pi e^2}{q^2}\sum_f \left|\langle f|\hat{\rho}(\mathbf{q})|0\rangle\right|^2 \delta(\omega_f - \omega). \tag{D.7}$$

Here we are using Heaviside-Lorenz conventions for the electron charge e as is common in high-energy physics, which differs from the Gaussian unit definition common in condensed matter physics by a factor of $\sqrt{4\pi}$. (Note also that Eq. (9) of [460] is missing a factor of π.) Plugging Eq. (D.7) into Eq. (D.5), we obtain our main result, Eq. (8.1). For reference, the sum over final states on the right-hand side of Eq. (D.7) is (up to factors of π) conventionally defined in the condensed matter literature as the dynamic structure factor.

Notice that we have made no assumptions whatsoever about the character of

the final state $|f\rangle$. It is an exact eigenstate of the (in general very complicated) many-body condensed matter Hamiltonian, and the only requirement is that it represents some rearrangement of the electrons in the target so that it has a nonzero matrix element with the electron density operator with respect to the ground state. In this sense our treatment is distinct from Ref. [436], which defines a general dynamic structure factor (with slightly different normalization compared to the condensed matter conventions) very similar to the sum in Eq. (D.5), but one which is excitation-specific and requires quantization in terms of single-quasiparticle states in the case of electron scattering. When all many-body states are included, the structure factor defined in Ref. [436] for electron scattering is *identical* to the loss function defined through the complex dielectric function, is directly measurable without the need to compute single-particle wavefunctions, and automatically includes all in-medium effects. (Our formalism is philosophically similar to Ref. [540], which parameterizes non-relativistic potentials using only general principles such as the Källén-Lehmann spectral representation, without relying on the assumption of the perturbative exchange of a single mediator.) On the other hand, the formalism of Ref. [436] is useful when DM couples differently to electrons, protons, and neutrons than the photon, in an energy regime where density perturbations to both electrons and ions are relevant, as might be the case for sub-gap single-phonon excitations.

Finally, we note that other UV Lagrangians considered in Ref. [446] also generate non-relativistic potentials which couple to the electron density, but are often accompanied by other spin- or momentum-dependent operators which may complicate our arguments, so we focus on the case of spin-independent scattering. In particular, if DM

is a Dirac fermion χ which couples to a scalar ϕ of mass m_ϕ through the scalar current $\mathcal{L} \supset g_\chi \phi \bar{\chi}\chi$, or to a vector V_μ of mass m_V through the vector current $\mathcal{L} \supset g_\chi V_\mu \bar{\chi}\gamma^\mu \chi$, and if the mediator couples to electrons in an analogous fashion but with coupling g_e, the resulting potential is the same in both cases [446]:

$$V(\mathbf{q}) = V(q) = \frac{g_\chi g_e}{q^2 + m_{\phi,V}^2}. \tag{D.8}$$

Similar formulas apply when DM is a complex scalar. Note that in contrast with Ref. [446], we leave the DM-electron coupling as its 'bare' value and place all in-medium corrections to this coupling entirely within the loss function. In the case where $g_e \propto e$, as would be the case for a kinetically-mixed dark photon mediator or when the DM is millicharged, the factors of $1/e^2$ cancel in Eq. (8.1) because the DM-induced perturbation to the electron density is exactly proportional to an ordinary electromagnetic probe.

For completeness, we give the expression for the energy spectrum from DM-electron scattering,

$$\frac{\mathrm{d}R}{\mathrm{d}\omega} = \frac{\rho_\chi}{2\pi^2 e^2 \rho_T m_\chi} \int \mathrm{d}q\, q^3 |V(q)|^2 \mathcal{W}(q,\omega) \eta\big(v_{\min}(q,\omega)\big), \tag{D.9}$$

where ρ_T is the mass density of the target, $\eta(v_{\min})$ is the mean inverse DM speed $\int_{v_{\min}} \mathrm{d}^3\mathbf{v}_\chi f(\mathbf{v}_\chi)/v_\chi$, and $v_{\min} = \frac{\omega}{q} + \frac{q}{2m_\chi}$ is the minimum DM speed required to produce an excitation with momentum q and energy ω for DM of mass m_χ. To compare with the literature, we take $f(\mathbf{v}_\chi)$ to be the standard halo model with dispersion $v_0 = 220\,\mathrm{km/s}$, escape velocity $v_{\mathrm{esc}} = 550\,\mathrm{km/s}$, and Earth velocity $v_E = 232\,\mathrm{km/s}$ in the galactic frame. Integrating Eq. (D.9) over ω within the dynamic range of a given experiment gives the total scattering rate.

D.2 Models and measurements of the loss function

In our formalism, the detector response and its influence on the scattering rate are entirely captured by the complex dielectric function $\epsilon(\mathbf{q},\omega)$ via the loss function $\mathcal{W}$ of Eq. (D.7) and Eq. (8.2). In principle, this quantity is directly measurable with electromagnetic probes in a given material. However, most measurements presently available in the literature are made at values of $(\mathbf{q},\omega)$ different than those of interest for the detection of light DM (see Fig. 8.1). Thus, for a first estimate of the scattering rate, we employ analytical approximations to the dielectric function. Important consistency checks can be implemented based on the fact that ϵ^{-1} is defined as a causal correlation function, and thus must have certain analytic properties. In particular, the following two 'sum rule' relations are satisfied exactly by $\mathcal{W}(\mathbf{q},\omega)$ in the limit of an isotropic system [541]:

$$\int_0^\infty \mathrm{d}\omega\, \omega\, \mathcal{W}(\mathbf{q},\omega) = \frac{\pi}{2}\omega_p^2, \tag{D.10}$$

$$\lim_{\mathbf{q}\to 0} \int_0^\infty \mathrm{d}\omega\, \frac{\mathcal{W}(\mathbf{q},\omega)}{\omega} = \frac{\pi}{2}. \tag{D.11}$$

Equation (D.10) is effectively a manifestation of charge conservation, which explains the appearance of the plasma frequency

$$\omega_p^2 = \frac{4\pi\alpha n_e}{m_e}, \tag{D.12}$$

which is proportional to the total electron density n_e in the FEG limit, while Eq. (D.11) follows from causality. Causality also implies that $\mathcal{W}(\mathbf{q},-\omega) = -\mathcal{W}(\mathbf{q},\omega)$ [541], which has important consequences for the projected reach in superconductors, as we will see below.

D.2.1 RPA dielectric function for a homogeneous electron gas

An analytic form for the dielectric function of a homogeneous electron gas can be derived from first principles under the random phase approximation (RPA). Here a word about terminology is in order: screening effects arise from Coulomb interactions between electrons, but in RPA these are embodied in the total scalar potential for the system which is solved for self-consistently [541]. Thus RPA captures only a certain subset of electron interactions without including electron-electron interactions directly in the Hamiltonian; in QFT language, it sums the series of ladder diagrams constructed from the 1-loop vacuum polarization to obtain the resummed photon propagator, but does not include higher-loop diagrams involving additional electron lines. This is the sense in which the electrons are treated as 'free' and ϵ_{RPA} is sometimes referred to as the dielectric function for the free electron gas (FEG). Below we will consider further improvements to this approximation.

The resulting dielectric function at zero temperature is given by Eq. (5.4.21) of Ref. [468] as

$$\epsilon_{\text{RPA}}(\mathbf{q},\omega) = 1 + \frac{3\omega_p^2}{q^2 v_F^2}\left\{\frac{1}{2} + \frac{k_F}{4q}\left(1 - \left(\frac{q}{2k_F} - \frac{\omega+\Gamma_p}{qv_F}\right)^2\right)\text{Log}\left(\frac{\frac{q}{2k_F} - \frac{\omega+i\Gamma_p}{qv_F} + 1}{\frac{q}{2k_F} - \frac{\omega+i\Gamma_p}{qv_F} - 1}\right)\right.$$
$$\left. + \frac{k_F}{4q}\left(1 - \left(\frac{q}{2k_F} + \frac{\omega+\Gamma_p}{qv_F}\right)^2\right)\text{Log}\left(\frac{\frac{q}{2k_F} + \frac{\omega+i\Gamma_p}{qv_F} + 1}{\frac{q}{2k_F} + \frac{\omega+i\Gamma_p}{qv_F} - 1}\right)\right\}. \quad \text{(D.13)}$$

Here Log denotes the principal value of the natural logarithm, k_F and v_F are the Fermi momentum and Fermi velocity respectively, and Γ_p is a free parameter controlling the width of the plasmon which can also be interpreted as a quasiparticle lifetime. The

plasma frequency can also be written in the form

$$\omega_p = \frac{\lambda_{\mathrm{TF}} v_F}{\sqrt{3}} = \frac{v_F}{\sqrt{3}} \left[\frac{e}{\pi} \left(2 E_F m_e^3 \right)^{1/4} \right], \tag{D.14}$$

where λ_{TF} is the Thomas–Fermi screening length. We expect the zero-temperature RPA result to be an excellent approximation for $\omega \gg 2\Delta$, where 2Δ is the superconducting gap. As mentioned in the main text, this approximation ignores possible enhancements to the loss function from scattering off of the condensate at energies near or below the gap, which will be considered in future work [470]. In the literature, Eq. (D.13) is known as the Lindhard dielectric function, though Lindhard's formalism may also be applied to semiconductors as well as metals; in what follows, we will use the terms 'Lindhard,' 'RPA,' and 'FEG' interchangeably to refer to Eq. (D.13).

Observe that the arguments of the logarithms in Eq. (D.13) are in general complex. For some values of q and ω, these arguments lie along the negative real axis in the narrow-width limit $\Gamma_p \to 0$, and the imaginary part of ϵ then depends crucially on the choice of branch. The branch choice is fixed by the causality condition $\mathcal{W}(\mathbf{q}, -\omega) = -\mathcal{W}(\mathbf{q}, \omega)$, which is automatic for positive real values of Γ_p, but the $\Gamma_p \to 0$

limit is non-trivial. The causal result is given by Eq. (5.4.22b) of Ref. [468] as

$$\mathrm{Re}\,\epsilon_{\mathrm{RPA}}(\mathbf{q},\omega) \simeq 1 + \frac{\lambda_{\mathrm{TF}}^2}{q^2}\left(\frac{1}{2} + \frac{k_F}{4q}\left(1-Q_-^2\right)\log\left|\frac{Q_-+1}{Q_--1}\right| + \frac{k_F}{4q}\left(1-Q_+^2\right)\log\left|\frac{Q_++1}{Q_+-1}\right|\right), \tag{D.15}$$

$$\mathrm{Im}\,\epsilon_{\mathrm{RPA}}(\mathbf{q},\omega) \simeq \frac{\pi\omega_p^2}{q^3 v_F^2}\begin{cases} 2\omega/v_F & Q_+ < 1 \\ 3k_F\left(1-Q_-^2\right)/4 & |Q_-| < 1 < Q_+ \\ 0 & |Q_-| > 1. \end{cases} \tag{D.16}$$

where $Q_\pm = \frac{q}{2k_F} \pm \frac{\omega}{qv_F}$. The acausal branch prescription was employed in Ref. [370], which as we will see in Appendix D.3 below, artificially suppresses the scattering rate for low DM masses.

The imaginary part of the Lindhard dielectric function naturally contains the plasmon as a Lorentzian peak at $\omega = \omega_p$ of width Γ_p. For the purposes of light DM detection, kinematics favor energy deposits $\omega \ll \omega_p$. The plasmon has then typically been neglected in the literature in the computation of the scattering rate, *i.e.*, the rate is computed in the limit $\Gamma_p \to 0$. However, for realistic values of Γ_p, the tail of the plasmon peak may significantly contribute to or even dominate the loss function at the relevant values of ω.

For DM–electron scattering in semiconductors, if the deposited energy is $\mathcal{O}(5\,\mathrm{eV})$ or greater, the minimum momentum transfer is $q \gtrsim 5\,\mathrm{keV}$ independent of the DM mass (see Fig. 8.1 in the main text). Since $k_F \simeq 2\pi/a \simeq 5\,\mathrm{keV}$ for typical interatomic spacings a, this means that the behavior of this part of the spectrum will be determined

by the loss function in the region $q \gtrsim k_F$. For these values of q, the DM is probing length scales smaller than the distance between lattice sites, so we might expect that the inhomogeneities due to the lattice become unimportant and the response is similar to a FEG. For $q > 2k_F$ the loss function peaks when $Q_- \approx 0$, corresponding to $\omega = \frac{q^2 v_F}{2k_F} = \frac{q^2}{2m_e}$, which is elastic scattering from free electrons at rest. For a given q, the loss function is nonzero over a range $\Delta\omega \simeq 2qv_F$ around the peak, reflecting the fact that electrons at the Fermi surface have a nonzero velocity. Note however that the loss function vanishes when $|Q_-| > 1$, which can happen for sufficiently small ω at sufficiently large q. This is an artificial feature of the FEG which is not present in semiconductors, where the valence (and core) electron wavefunctions have a tight-binding character with a momentum-space tail that extends to arbitrarily large values. This regime corresponds to $q \gtrsim Z_{\rm eff}/a_0 \simeq 15\,\text{keV}$ where a_0 is the Bohr radius and $Z_{\rm eff} \approx 4$ is the effective nuclear charge felt by the valence electrons in Group 14 elements (carbon, silicon, and germanium). The large-q behavior is especially apparent in some materials like germanium, where the $3d$ shell may become energetically accessible for ω exceeding the binding energy. A corresponding feature is seen in the spectrum in models using tight-binding wavefunctions [421] as well as those using density functional theory (DFT) techniques [368].

D.2.2 Plasmon pole approximation and local field corrections

In the limit that the plasmon dominates, the dielectric function may be derived by modeling the atomic response as a damped harmonic oscillator. This is known as

the Fröhlich model [542], and the result is

$$\epsilon_{\mathrm{F}}(\mathbf{q},\omega) = \epsilon_c + \frac{\omega_p^2}{(\omega_g^2 - \omega^2) - i\omega\Gamma_p}. \tag{D.17}$$

Here ϵ_c denotes the contribution from core electrons, which is assumed to be independent of $\mathbf{q}$ and ω, and ω_g is an average band gap which can be set to zero for metals. The corresponding loss function features a Breit–Wigner-like peak, with the form

$$\mathcal{W}_{\mathrm{F}}(\mathbf{q},\omega) = \frac{\omega_p^2 \omega \Gamma_p}{\epsilon_c^2 \left(\omega_g^2 + \omega_p^2/\epsilon_c^2 - \omega^2\right)^2 + \omega^2\Gamma_p^2}. \tag{D.18}$$

This function satisfies the sum rules of Eqs. (D.10) and (D.11) with $\epsilon_c = 1$ and $\omega_g = 0$. Note that this form of the loss function is linear in ω for $\omega \ll \omega_p$.

The low-energy loss function is also subject to effects which are not included in the Lindhard dielectric function. Ref. [464] (hereafter denoted 'GSRF') fits the plasmon in aluminum including a local-field correction and accounting for the polarizability of atomic cores χ_{core}, resulting in a dielectric function of the form

$$\begin{aligned}\epsilon_{\mathrm{G}}(\mathbf{q},\omega) = 1 + [\omega + i\Gamma_p(\mathbf{q})]\,[\epsilon_{\mathrm{RPA}}(\mathbf{q},\omega) - 1 + 4\pi\chi_{\mathrm{core}}] \div \\ \left[\omega\,(1 - G(\mathbf{q})\,[\epsilon_{\mathrm{RPA}}(\mathbf{q},\omega) - 1]) + i\Gamma_p(\mathbf{q})\,(1 - G(\mathbf{q})\,[\epsilon_{\mathrm{RPA}}(\mathbf{q},0) - 1])\,\frac{\epsilon_{\mathrm{RPA}}(\mathbf{q},\omega) - 1 + 4\pi\chi_{\mathrm{core}}}{\epsilon_{\mathrm{RPA}}(\mathbf{q},0) - 1 + 4\pi\chi_{\mathrm{core}}}\right],\end{aligned} \tag{D.19}$$

where $G(\mathbf{q})$ is known as the exchange parameter and arises in the microscopic theory from 1-loop corrections to the electron-photon vertex [541]. Ref. [464] provides fits to G and Γ_p as functions of $\mathbf{q}$. Complex values of G produce damping, which influences the form of the loss function at small values of ω. However, $\mathrm{Im}\,G(\mathbf{q}) \neq 0$ can lead to unphysical negative values of the loss function at the smallest values of ω, thereby violating the positivity requirements imposed by the sum rules, and moreover G as

computed in various microscopic theories tends to be real [541]. Following Ref. [464], we divide our treatment into two cases, one with complex-valued G ('damped') and one with real-valued G ('undamped').

D.2.3 Dielectric function for Dirac materials

Dirac materials are characterized by electrons with the approximately linear dispersion characteristic of relativistic Dirac fermions, rather than the usual quadratic dispersion expected at a band minimum. In real materials, there are typically two such bands, one below and one above the Fermi energy, with dispersions $E_{\pm}(\mathbf{k}) = \pm\sqrt{v_F^2\mathbf{k}^2 + \Delta^2}$. Here, Δ plays the role of the fermion mass and the Fermi velocity v_F is the analogue of the speed of light; the gap at the Dirac point with $\mathbf{k} = 0$ is 2Δ. The band structure may be anisotropic, with different Fermi velocities along different lattice directions, but for pedagogical purposes we will focus here on isotropic materials; see Refs. [438, 439] for a detailed investigation of anisotropic Dirac materials for DM detection.

In the approximation that only two nondegenerate bands contribute to the Dirac electron spectrum, the dielectric function may be computed using Lindhard's formalism in the Bloch wave basis [468]. At zero temperature, with the valence ($-$) band full and the conduction ($+$) band empty, this reads

$$\epsilon_{\text{Dirac}}(\mathbf{q},\omega) = 1 + \lim_{\eta\to 0}\frac{1}{V}\frac{e^2}{q^2}\int_{\text{BZ}}\frac{V_{\text{uc}}\mathrm{d}^3\mathbf{k}}{(2\pi)^3}\frac{2}{E_+(\mathbf{k}+\mathbf{q}) - E_-(\mathbf{k}) - \omega - i\eta}\left|\langle\mathbf{k}+\mathbf{q};+|e^{i\mathbf{q}\cdot\mathbf{r}}|\mathbf{k};-\rangle\right|^2, \tag{D.20}$$

where $|\mathbf{k};\pm\rangle$ represents a Bloch wavefunction with crystal momentum $\mathbf{k}$ in the band $-$ or $+$, V is the crystal volume, the factor of 2 is for spin degeneracy, and the integral is taken

over the first Brillouin zone (BZ) in the continuum limit using the unit cell volume to regularize the momentum sum, $\sum_{\mathbf{k}} \to \int V_{\text{uc}}\, \mathrm{d}^3\mathbf{k}/(2\pi)^3$. There are some complications with this procedure in the case of anisotropic materials [438], but it yields an accurate estimate for the imaginary part in isotropic materials, which is dominated by the smallest gaps and hence the bands other than the Dirac bands may be neglected. However, as noted in Ref. [438], $\text{Re}[\epsilon(0,0)]$ acts as a background dielectric constant receiving contributions from the entire BZ and thus cannot be reliably calculated analytically. We may therefore estimate the real part as simply $\text{Re}(\epsilon_{\text{Dirac}}) = \kappa \gg 1$ independent of $\mathbf{q}$ and ω over the relevant kinematic range.

To obtain the imaginary part, we may use the identity $\text{Im}(\lim_{\eta\to 0} \frac{1}{x-i\eta}) = \pi\delta(x)$ and perform the integral using spinor wavefunctions with the matrix element given in Ref. [375]. Note that this is precisely analogous to performing the phase space integral over the valence and conduction bands in the single-particle formalism for determining the scattering rate; the dielectric function allows us to express the results of Ref. [375] in a more convenient and generalizable formalism. Equivalently, we may recognize that with the replacements $\Delta \to m_e$ and $v_F \to c$, the imaginary part is identical to that of the 1-loop vacuum polarization in relativistic quantum electrodynamics (QED), which is proportional to the cross section for $\gamma^* \to e^+e^-$ by the optical theorem [543]. The result is

$$\text{Im}\,\epsilon_{\text{Dirac}}(q,\omega) = \frac{e^2}{12\pi v_F}\sqrt{1-\frac{4\Delta^2}{\omega^2-v_F^2q^2}}\left(1+\frac{2\Delta^2}{\omega^2-v_F^2q^2}\right)\Theta(\omega^2-v_F^2q^2-4\Delta^2), \quad \text{(D.21)}$$

where the coefficient $e^2/(12\pi)$ is (up to a factor of π) the familiar 1-loop beta function coefficient of QED. Indeed, the physics of the dielectric function is the same in Dirac

materials as it is in the true QED vacuum; the screening of bare charges due to $\mathrm{Im}(\epsilon)$ at $q \simeq 2m_e$ is known as the Uehling potential.

As long as v_F is not too small, $\mathrm{Im}(\epsilon) \lesssim 1$. (Otherwise perturbation theory would break down, as noted in Ref. [375].) Then if $\kappa \gg 1$, we may approximate $\mathrm{Im}(-1/\epsilon) \approx \mathrm{Im}(\epsilon)/\kappa^2$ and thus

$$\mathcal{W}_{\mathrm{Dirac}}(q,\omega) = \frac{e^2}{12\kappa^2\pi v_F}\sqrt{1-\frac{4\Delta^2}{\omega^2-v_F^2q^2}}\left(1+\frac{2\Delta^2}{\omega^2-v_F^2q^2}\right)\Theta(\omega^2-v_F^2q^2-4\Delta^2)\Theta(\omega_{\mathrm{max}}-\omega) \tag{D.22}$$

Setting $\Delta = 0$ gives Eq. (8.4) in the main text. The last factor may be explained as follows. In real materials, the Dirac band structure does not extend throughout the entire BZ, but deviates from linearity at some point. In Ref. [375] this was expressed as a momentum cutoff Λ, which is required to regularize the real part of $\epsilon_{\mathrm{Dirac}}$. Here, since we are dealing with model functions rather than real materials, we instead impose a cutoff ω_{max} on the depth of the Dirac band, which has typical values of $\omega_{\mathrm{max}} \simeq 0.5\,\mathrm{eV}$ in *e.g.* $ZrTe_5$ [375]. Finally, note that $\mathcal{W}_{\mathrm{Dirac}}$ violates the causality requirement $\mathcal{W}_{\mathrm{Dirac}}(q,-\omega) = -\mathcal{W}_{\mathrm{Dirac}}(q,\omega)$. This indicates that $\mathcal{W}_{\mathrm{Dirac}}$ as computed here does not represent the entire loss function, and in particular (as noted in the main text) it is missing plasmon contributions.

D.2.4 Measurements of the loss function in various materials

Measurements of the loss function in the vicinity of the plasmon peak are available in the literature for certain materials, so it is already possible to fit the Fröhlich model directly to data and to assess the significance of the plasmon tail at $\omega \ll \omega_p$. Figure D.1 (left) shows such a fit to measurements in Al. While the fit is excellent in

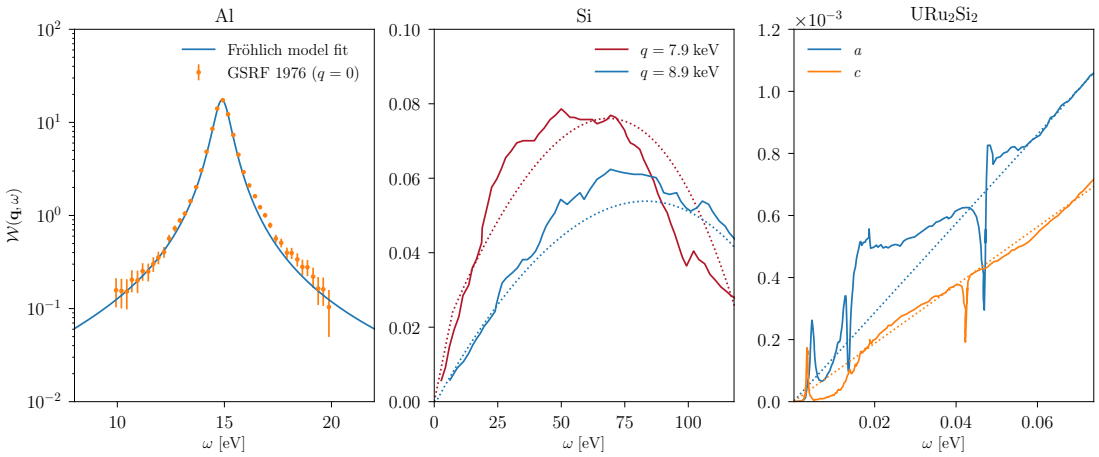

Figure D.1: Measurements of the loss function $\mathcal{W}(\mathbf{q}, \omega) = \mathrm{Im}(-1/\epsilon(\mathbf{q}, \omega))$ in Al, Si, and URu_2Si_2 along with model fits when appropriate. **Left:** loss function for Al at $q = 0$ in the vicinity of the plasmon peak from Ref. [464], fit with the Fröhlich model of Eq. (D.18). The best-fit parameters are $(\omega_p, \Gamma_p) = (14.9\,\mathrm{eV},\ 0.863\,\mathrm{eV})$. Error bars indicate the accuracy with which the data points could be transcribed from Ref. [464]. **Center:** loss function for Si at large momenta $q > 2\pi/a$, measured from X-ray scattering in Ref. [472]. Dashed lines show the Lindhard RPA loss function with $\Gamma_p = 0$, $k_F = m_e v_F$, and $\omega_p = 16.67\,\mathrm{eV}$ [474]. The Fermi velocity is treated as a free parameter and is fixed here to the best-fit value of $v_F = 2.59 \times 10^8\,\mathrm{cm/s} = 8.6 \times 10^{-3}$ in natural units, which is comparable to Fermi velocities of metals with similar densities. **Right:** loss function in URu_2Si_2 at $q = 0$ measured along two different crystal axes a and c at $T = 9\,\mathrm{K}$ [465] (solid), along with a linear fit to both datasets (dashed). If interpreted as the tail of a valence electron plasmon, the slope should be Γ_p/ω_p^2. The fit gives a slope of $\Gamma_p/\omega_p^2 \simeq 14\ (9) \times 10^{-3}\,\mathrm{eV}^{-1}$ along the a (c) axis which implies $\Gamma_p/\omega_p \simeq 0.21\ (0.13)$ for $\omega_p \simeq 15\,\mathrm{eV}$, values which are typical for other metals.

the vicinity of the plasmon peak, the behavior at $\omega \ll \omega_p$ should be viewed only as a benchmark: other physical effects contribute at these energies, notably those encapsulated by the Lindhard dielectric function which incorporates electron screening effects. See Fig. D.4 and Appendix D.3 below for further details.

High-precision measurements of the loss function at nonzero q have also been performed for Si using X-ray scattering [472]. The plasmon is clearly visible at small q, but here we focus on the behavior at large q. Figure D.1 (center) shows the measured loss function along the [100] crystal direction (solid lines), compared to the RPA loss function for the homogeneous electron gas taking $\omega_p = 16.67\,\text{eV}$ for the measured plasmon frequency [474]. While semiconductors and insulators do not, strictly speaking, have a Fermi velocity at zero temperature where there are no free carriers, we may regard v_F as a tuneable parameter which governs the behavior of the loss function at small ω. With $v_F = 8.6 \times 10^{-3}$, on the same scale as v_F for typical metals, the fit is quite good, especially for $\omega < 25\,\text{eV}$. On the other hand, at $q = 10\,\text{keV}$, the RPA loss function vanishes identically for $\omega < 12\,\text{eV}$, which is likely unphysical given that atomic tight-binding wavefunctions have support in this kinematic range. The purpose of this comparison is not to advocate for using this extremely simplified model—indeed, data should be used to compute DM rates whenever possible—but rather to demonstrate how in the absence of data a simple model may provide an accurate estimate for the light-mediator spectrum for $\omega \in [5\,\text{eV}, 10\,\text{eV}]$, where the rate integral is dominated by $q \in [5\,\text{keV}, 10\,\text{keV}]$. Indeed, the success of the RPA model suggests that this part of the spectrum from scattering in any semiconductor or insulator with eV-scale bandgaps is nearly universal, determined only by the valence electron density and an effective Fermi

velocity. This model may be seen as an extension, accounting for screening, of earlier simplified models for scattering in semiconductors using atomic orbitals or tight-binding wavefunctions [379, 421].

To complete our survey of sample loss functions, we show in Fig. D.1 (right) the measured loss function at $q = 0$ for URu_2Si_2 along the a and c crystal axes, measured with Fourier transform infrared spectrometry [465]. URu_2Si_2 has been extensively studied for decades [544] due to its unusual 'hidden order' below 17.5 K, and thus has been synthesized as ultra-pure single crystals. Below $T_c = 1.5\,\mathrm{K}$ it behaves as a conventional superconductor [545]. A number of features are present below 20 meV which may be interpreted as heavy-fermion plasmons, as we discuss in the main text. Based on this interpretation, to perform our rate estimates in the main text, we extrapolate the loss function as independent of q out to $q = q_c \simeq 100\,\mathrm{eV}$. Indeed, this is the standard approximation made in scattering experiments near the plasmon pole [474]. Then, we see from Eq. (D.9) that the spectrum is largely determined by the shape of the zero-momentum loss function $\mathcal{W}(\omega)$, with the inverse mean speed η only serving to enforce the kinematic condition $q > \omega/v_\chi$. All of the approximations we have made may easily be dropped once momentum-resolved data on $\mathcal{W}(\mathbf{q}, \omega)$ within the DM regions shown in Fig. 8.1 is available.

It is also interesting to note that at larger ω, the loss function is linear to an excellent approximation, in the c direction above 20 meV and in the a direction above 50 meV. In Fig. D.1 we show a linear fit to both loss functions with zero offset. In the Fröhlich model Eq. (D.18), the plasmon tail gives a loss function $\mathcal{W}_F(q = 0, \omega) \approx \omega \times (\Gamma_p/\omega_p^2)$ at small ω. The slope of the linear fit is consistent with $\Gamma_p/\omega_p \simeq 0.1 - 0.2$

and $\omega_p \simeq 15\,\text{eV}$, which would be reasonable parameters for the ordinary valence electron plasmon in a generic metal. This data therefore provides some preliminary indication that the linear tail of the plasmon in ordinary superconductors like Al may extend down to the meV scale. We emphasize again that dedicated measurements are needed to confirm this.

D.2.5 Semiconductor spectrum in the free-electron gas approximation

In order to relate the energy loss function to the crystal form factor [368, 436], we compare

$$\Gamma(\mathbf{v}_\chi) = \int \frac{\mathrm{d}^3\mathbf{q}}{(2\pi)^3} |V(q)|^2 \, \mathcal{S}(\mathbf{q},\omega) \tag{D.23}$$

to Eq. (8.1), which gives the relation between the dynamic structure factor $\mathcal{S}(\mathbf{q},\omega)$ defined in Ref. [436] and the loss function via

$$\mathcal{S}(\mathbf{q},\omega) = \frac{2q^2}{e^2}\mathcal{W}(\mathbf{q},\omega). \tag{D.24}$$

On the other hand, the dynamic structure factor in a semiconductor, computed in the basis of single-particle states, can be related to the crystal form factors $|f_{ii'\mathbf{k}\mathbf{k}'\mathbf{G}}|^2$ via [368, 436]

$$\mathcal{S}(\mathbf{q},\omega) = 2\sum_{i,i',\mathbf{G}} \int_{\text{BZ}} \frac{d^3\mathbf{k}}{(2\pi)^3}\frac{d^3\mathbf{k}'}{(2\pi)^3} 2\pi\delta(E_{i'\mathbf{k}'} - E_{i\mathbf{k}} - \omega) 2\pi\delta(|\mathbf{k}' - \mathbf{k} + \mathbf{G}| - q)|f_{ii'\mathbf{k}\mathbf{k}'\mathbf{G}}|^2, \tag{D.25}$$

where the momentum integral is taken over the first BZ, $\mathbf{G}$ runs over all reciprocal

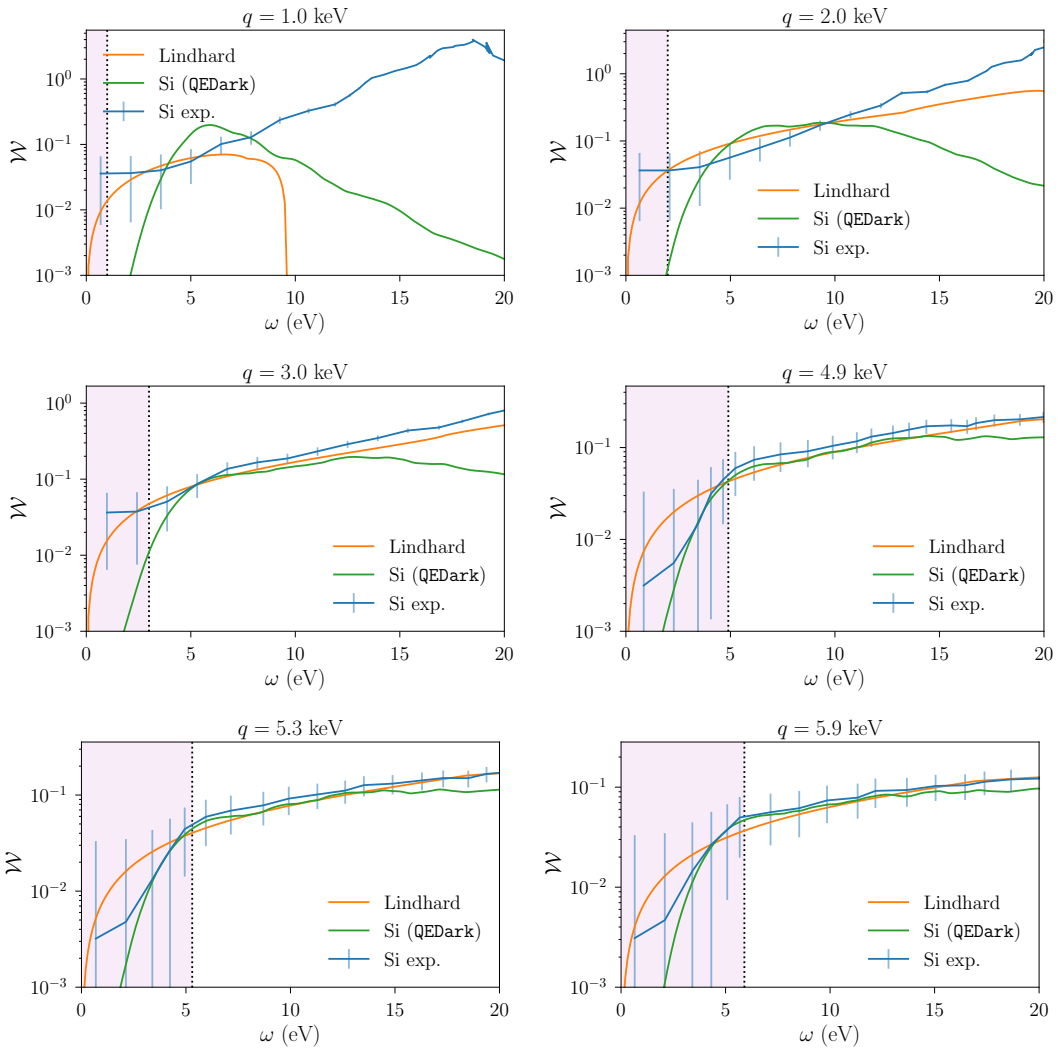

Figure D.2: Loss function comparisons in Si for various q, as a function of ω. Error bars indicate the accuracy with which the data points could be transcribed from Ref. [472]. The shaded purple region represents the kinematically-allowed region for $v_\chi = 10^{-3}$. The measured loss function agrees fairly well with both the loss function computed from the single-particle basis from `QEdark` [368] and the Lindhard FEG approximation in the range 5–10 eV for ω, but there are large differences at both small ω near the gap, and near the plasmon energy $\omega_p \simeq 17\,\text{eV}$ for small q.

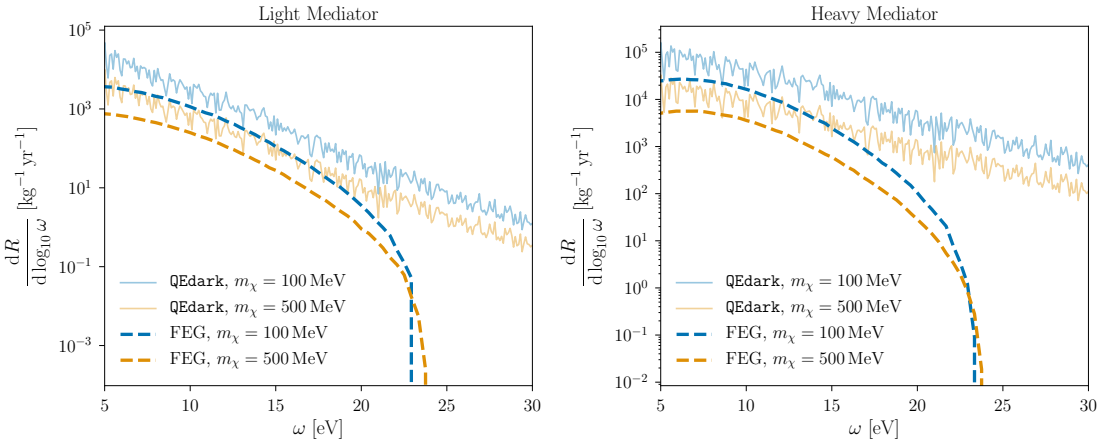

Figure D.3: Recoil spectra in Si at fixed $\overline{\sigma}_e = 10^{-37}\,\mathrm{cm}^2$, for light and heavy mediators (scalar or vector). Solid curves are computed with `QEdark` [368]. Dashed curves are computed from Eq. (8.1) with the Lindhard RPA loss function, Eq. (D.13) with $v_F = 8.6 \times 10^{-3}$, $k_F = m_e v_F$, and $\omega_p = 16.67\,\mathrm{eV}$.

lattice vectors, and i and i' run over all valence and conduction bands, respectively. Thus we can compute an equivalent loss function from `QEdark` [368] crystal form factors by

$$\mathcal{W}(\mathbf{q},\omega) = \frac{e^2}{q^2}\sum_{i,i',\mathbf{G}}\int_{\mathrm{BZ}}\frac{d^3\mathbf{k}}{(2\pi)^3}\frac{d^3\mathbf{k}'}{(2\pi)^3}2\pi\delta(E_{i'\mathbf{k}'} - E_{i\mathbf{k}} - \omega)2\pi\delta(|\mathbf{k}' - \mathbf{k} + \mathbf{G}| - q)|f_{ii'\mathbf{k}\mathbf{k}'\mathbf{G}}|^2 \tag{D.26}$$

Using Eq. (D.26), we can compare the measured loss function to the loss function computed in the single-particle basis by `QEdark`, as well as the Lindhard dielectric function for the FEG with the best-fit v_F in Fig. D.1. The results are shown in Fig. D.2. Note that for a given q, the range of ω which is accessible is $\omega < qv_\chi$, which only comprises a small piece of the total support of $\mathcal{W}(\mathbf{q},\omega)$. Regardless, we see that `QEdark` tends to slightly underpredict the measured loss in the kinematically-allowed region.

Furthermore, `QEdark` accurately reproduces the measured loss in the near-gap region $\omega \in [1\,\mathrm{eV}, 5\,\mathrm{eV}]$ where Lindhard fails to do so, as expected. On the other hand, `QEdark` fails to capture the plasmon which is seen in the measured loss function because the single-particle band structure states do not account for collective effects.

Overall, though, the nearly-linear shape of the measured loss function in the range $\omega \in [5\,\mathrm{eV}, 15\,\mathrm{eV}]$ is reproduced fairly well by the Lindhard model, and matches that of `QEdark`. We therefore expect that the spectral shape (though perhaps not the normalization) will be captured in this energy range by the simple Lindhard model for the loss function. Moreover, since the Lindhard model loss function goes to zero at sufficiently large q for small ω, and since the rate recieves contributions from *all* $q > \omega/v_\chi$, we expect the Lindhard approximation to be best for a light mediator which weights the rate integrand by $|V(q)|^2 \propto 1/q^4$. The results are shown in Fig. D.3. Indeed, the Lindhard FEG model matches the spectrum fairly well for the light mediator, roughly independent of the DM mass as long as the DM kinetic energy is well above the gap. The spectrum for a heavy mediator is a poorer match, especially at large ω where the kinematic mismatch between the FEG and the bound atomic wavefunctions becomes more important. We emphasize once again that these simple arguments are *not* meant to replace a measurement of $\mathcal{W}$ in the relevant kinematic range, which would predict the spectrum unambiguously. However, they do highlight a qualitative understanding of the spectrum in a limited energy range based on simple material properties like the effective v_F, which may be useful for identifying other detector materials suitable for DM-electron scattering. Furthermore, the part of the spectrum where the FEG model performs best corresponds to the 2-electron bin in Si, which is of considerable practical

importance to experiments: the 1-electron bin is typically dominated by backgrounds such as leakage current and Cherenkov radiation [447], while the rates in the bins with 3 or more electrons drop precipitously, at least based on estimates from the single-particle loss function. Integrating the FEG spectra from a threshold of $\omega = 4.7$ eV, corresponding to a $2e^-$ threshold in the model of Ref. [368], we obtain the reach curve shown in Fig. 8.2 in the main text.

D.3 Updated reach projections for superconductors

In Ref. [369], the scattering rate in a superconductor is first computed treating the electrons as free particles, with screening included afterwards in Ref. [370] via a correction to the matrix element. We now show that the result of Ref. [370] at $T = 0$ is exactly reproduced by our Eq. (8.1) when $\epsilon(\mathbf{q}, \omega)$ is taken to be the Lindhard dielectric function in the limit of vanishing plasmon width.

In a relativistic formalism for single-particle scattering, the superconductor scattering rate is given by

$$\Gamma(\mathbf{v}_\chi) = \int \frac{\mathrm{d}^3\mathbf{p}'_\chi}{(2\pi)^3} \frac{\langle |\mathcal{M}|^2 \rangle}{16 E_\chi E'_\chi E_e E'_e} \frac{S(\mathbf{q}, \omega)}{|\epsilon(\mathbf{q}, \omega)|^2}, \tag{D.27}$$

where $\mathbf{q} \equiv \mathbf{p}_\chi - \mathbf{p}'_\chi$ denotes the 3-momentum transfer, $\mathbf{p}'_\chi$ denotes the momentum of the scattered dark matter particle in the final state, and $S(\mathbf{q}, \omega)$ (not to be confused with the dynamic structure factor defined in Eq. (D.24) above) characterizes the available phase space, to be defined shortly. The presence of $|\epsilon|^2$ in the denominator of Eq. (D.27) accounts for screening and was treated in Ref. [370] as an in-medium modification to the dark photon propagator. In the non-relativistic limit, any interaction of the class

considered in Eq. (D.8) gives rise to a matrix element of the form

$$\frac{\langle|\mathcal{M}|^2\rangle}{16E_\chi E'_\chi E_e E'_e} \simeq \left(\frac{g_\chi g_e}{q^2 + m^2_{\phi,V}}\right)^2 = |V(q)|^2\,, \tag{D.28}$$

where $q = |\mathbf{q}|$. Equation (D.27) is trivially transformed to an integral over $\mathbf{q}$, and the rate becomes

$$\Gamma(\mathbf{v}_\chi) = \int \frac{\mathrm{d}^3\mathbf{q}}{(2\pi)^3}\,|V(q)|^2\,\frac{S(\mathbf{q},\omega)}{|\epsilon(\mathbf{q},\omega)|^2}. \tag{D.29}$$

Thus, to agree with Eq. (8.1), it is sufficient to have

$$S(\mathbf{q},\omega) = \frac{2q^2}{e^2}\,\mathrm{Im}\,\epsilon(\mathbf{q},\omega). \tag{D.30}$$

Equation (D.30) holds exactly in the low-temperature limit for the form of S used in Refs. [369, 370], where the superconductor is treated as a free electron gas. In this case, S is given by

$$S(\mathbf{q},\omega) = 2\int \frac{\mathrm{d}^3\mathbf{p}_e}{(2\pi)^3}\frac{\mathrm{d}^3\mathbf{p}'_e}{(2\pi)^3}(2\pi)^4\delta^4(P_\chi + P_e - P'_\chi - P'_e) f_{\mathrm{FD}}(E_e)\left[1 - f_{\mathrm{FD}}(E'_e)\right], \tag{D.31}$$

where f_{FD} is the Fermi–Dirac distribution and the P_i denote 4-momenta. We reserve p_i for the magnitudes of 3-momenta. The integration over $\mathbf{p}'_e$ is readily performed using the 3-momentum delta function. Writing the $\mathbf{p}_e$ integral in spherical coordinates and performing the trivial integral over the azimuthal angle produces

$$S(\mathbf{q},\omega) = 2\int \frac{p_e^2\,\mathrm{d}p_e\,\mathrm{d}(\cos\theta)}{(2\pi)^2}\delta\left(\omega - \frac{q^2 + 2p_e q\cos\theta}{2m_e}\right) f_{\mathrm{FD}}(E_e)\left[1 - f_{\mathrm{FD}}(E'_e)\right], \tag{D.32}$$

where θ denotes the angle between $\mathbf{p}_e$ and $\mathbf{q}$. The remaining delta function can be used to evaluate the integral over $\cos\theta$, but here care must be taken to enforce $|\cos\theta| \leq 1$. With the appropriate Heaviside function, the final integral becomes

$$S(\mathbf{q},\omega) = \int \mathrm{d}p_e\,\frac{m_e p_e}{\pi q}\left[1 - f_{\mathrm{FD}}(E'_e)\right]\Theta\left(1 - \left|\frac{2m_e\omega - q^2}{2p_e q}\right|\right). \tag{D.33}$$

Now the zero-temperature Fermi–Dirac distribution can be inserted and the integral can be performed analytically. The result is

$$S(\mathbf{q},\omega) = \frac{m_e^2}{\pi q}\begin{cases} \omega & 0 < \omega < |E_-| \\ E_F - \frac{(E_q-\omega)^2}{4E_q} & |E_-| < \omega < E_+ \\ 0 & \text{otherwise,} \end{cases} \tag{D.34}$$

where $E_q \equiv q^2/2m_e$ and $E_\pm = E_q \pm qv_F$. The conditions in Eq. (D.34) are equivalent to those in Eq. (D.16), *i.e.*, the imaginary part of the Lindhard dielectric function in the limit that the plasmon is infinitely long-lived. Equation (D.30) follows by direct comparison.

Given this agreement between the single-particle and dielectric-function formalisms, Eq. (8.1) can reproduce prior calculations of the scattering rate in superconductors; essentially, the final-state phase space integral is pre-computed in $\mathrm{Im}(\epsilon)$. However, Eq. (8.1) is more flexible than the traditional calculation in that we are not limited to the narrow-plasmon limit of the Lindhard dielectric function. Any model or measurement of the loss function can be inserted directly in Eq. (8.1).

To evaluate the event rate in a superconducting detector, we take the velocity of the DM in the galactic frame to have a modified Maxwell–Boltzmann distribution,

$$f(\mathbf{v}_\chi) \propto \exp\left(-\mathbf{v}_\chi^2/v_0^2\right)\Theta\big(v_{\rm esc} - |\mathbf{v}_\chi|\big). \tag{D.35}$$

For our reach projections, we take $v_0 = 220\,\mathrm{km/s}$ and $v_{\rm esc} = 550\,\mathrm{km/s}$, and we take Earth to have a velocity $v_E = 232\,\mathrm{km/s}$ in the galactic frame. This matches the conventions of Ref. [375]. In order to facilitate comparison with other results in the literature, we also show some results with $v_E = 0$ and $v_{\rm esc} = 500\,\mathrm{km/s}$, matching the conventions of *e.g.*

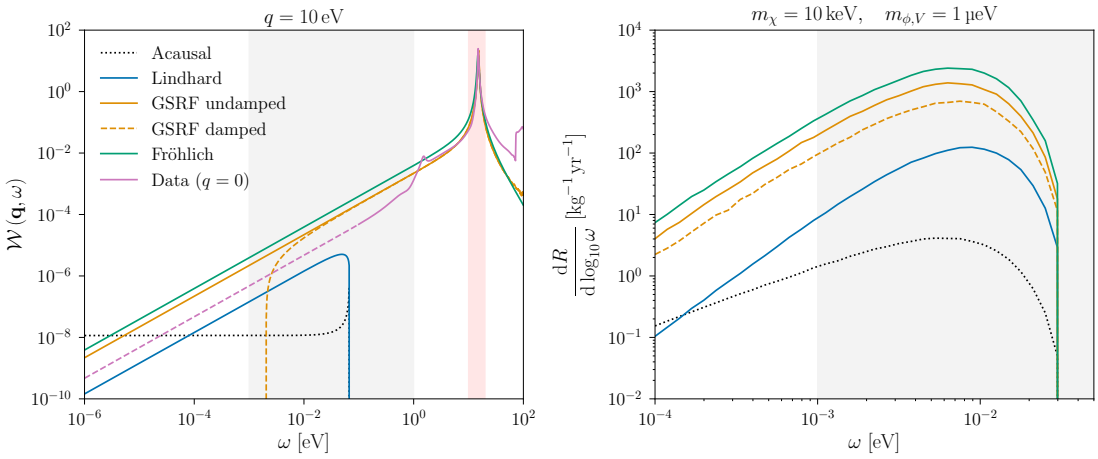

Figure D.4: **Left:** loss function for each of several models for Al, for $q = 10\,\mathrm{eV}$. The curve labeled 'Acausal' shows the loss function used in [370], which involves an unphysical choice of branch cut in the complex logarithm. The Fröhlich model fit is the same as that shown in Fig. D.1, for which measured data are only available within the red band. The Lindhard model is the RPA dielectric function Eq. (D.13) with $\Gamma_p = 0$, and the GSRF models use fit parameters for Eq. (D.19) from Ref. [464], with 'undamped' corresponding to $\mathrm{Im}\, G = 0$ and 'damped' corresponding to $\mathrm{Im}\, G \neq 0$. The damped curve becomes negative at small ω, which is an unphysical consequence of the GSRF model. The curve labeled 'Data' shows the fit to $q = 0$ measurements provided by Ref. [469]. We use dashes to indicate the continuation of the fit beyond the range of measured data. The gray band shows the reference range of 1 meV–1 eV deposits. **Right:** recoil spectra corresponding to each of these loss functions, assuming $(m_\chi,\, m_{\phi,V}) = (10\,\mathrm{keV},\, 1\,\mu\mathrm{eV})$ and $\overline{\sigma}_e = 10^{-39}\,\mathrm{cm}^2$.

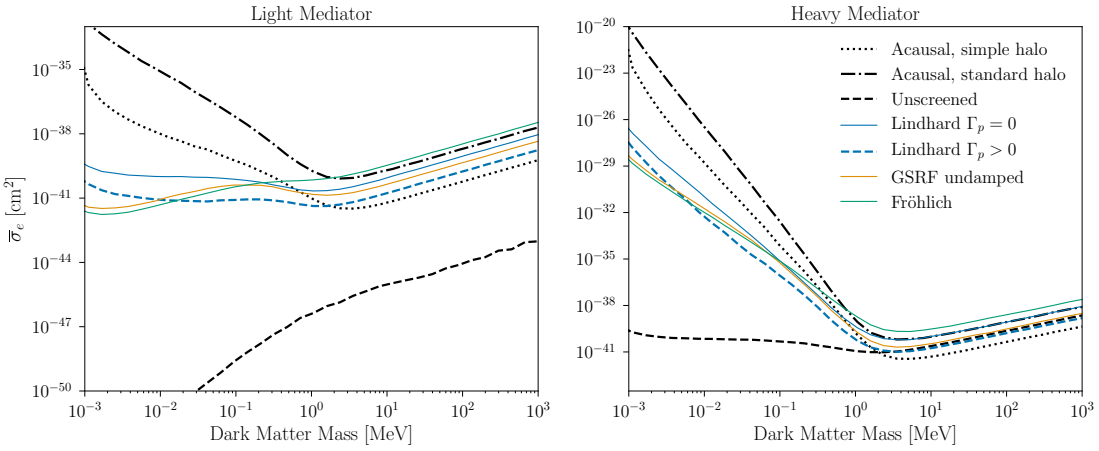

Figure D.5: Projected reach for an aluminum superconductor target for several forms of the loss function for scalar or vector mediators. The dotted curve is computed with the simple halo model ($v_E = 0$, $v_{\rm esc} = 500\,{\rm km/s}$), and all others assume the standard halo model. The dashed line ('unscreened') is computed in the single-particle formalism with no correction for screening, *i.e.*, without the factor of $|\epsilon|^2$ in the denominator of Eq. (D.27); this is unphysical for any spin-independent DM-electron interaction.

Refs. [369, 370]. We refer to this as the 'simple halo' scenario. Finally, for illustrative purposes, we show selected results for a hypothetical halo with $v_0 = v_{\rm esc} = 10^4\,{\rm km/s}$ and $v_E = 0$. In this 'fast DM' scenario, the plasmon peak is kinematically accessible, and this is directly visible as a feature in the recoil spectrum.

The various models for the loss functions in Al are shown in Fig. D.4, together with the corresponding DM recoil spectra for a kg-yr exposure. The undamped GSRF model and the Lindhard model with $\Gamma_p = 0$ correspond to the boundaries of the shaded region in Fig. 8.2. We also show the result obtained by choosing the acausal branch in the Lindhard dielectric function. It is clear from Fig. D.4 that at low energies, a naive extrapolation of the plasmon tail dominates over the Lindhard loss function with its infinitely long-lived plasmon. Moreover, the energy range of interest for light DM

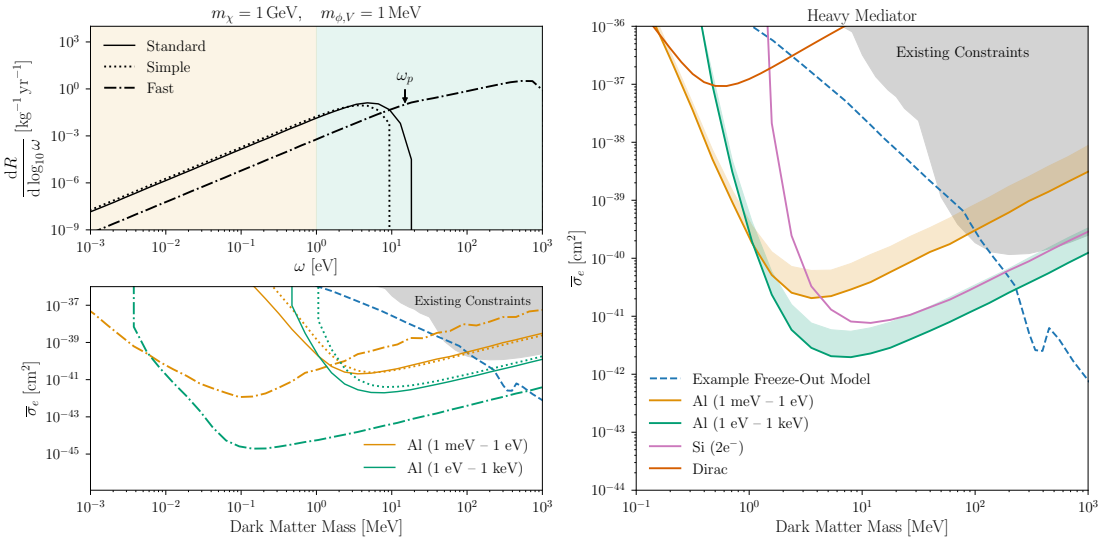

Figure D.6: **Top left:** recoil spectra in an Al superconductor for $(m_\chi, m_\phi) = (1\,\mathrm{GeV}, 1\,\mathrm{MeV})$ and $\overline{\sigma}_e = 10^{-42}\,\mathrm{cm}^2$ assuming the GSRF loss function without damping for several DM velocity distributions. See text for details. The fast halo scenario is unrealistic and is shown for illustrative purposes only: in this case, the plasmon peak is kinematically accessible, and the recoil spectrum exhibits a corresponding kink at $\omega = \omega_p$. The shaded areas indicate two fiducial experimental configurations, one sensitive to deposits 1 meV–1 eV (orange), and the other sensitive to deposits 1 eV–1 keV (green). **Bottom left:** projected reach in an Al superconductor for a 1 kg-yr exposure assuming a heavy mediator. Orange curves show the reach for the low-threshold scenario, and green curves show the reach for the high-threshold scenario. Projections for the standard halo, simple halo, and fast halo scenarios are shown by the solid, dotted, and dot-dashed curves, respectively. **Right:** projected reach for a 1 kg-yr exposure of a Dirac material, Si, and the two Al superconductor configurations assuming a heavy scalar or vector mediator. The parameters of the Dirac material are taken as in Fig. 8.2, with gap $2\Delta = 20\,\mathrm{meV}$, Fermi velocity $v_F = 4 \times 10^{-4}$, background dielectric constant $\kappa = 40$, and Dirac band cutoff $\omega_{\mathrm{max}} = 0.5\,\mathrm{eV}$. The projected reach for Si assumes a two-electron ionization threshold. The projected reach of $\mathrm{URu_2Si_2}$ lies above the top edge of the plot. All curves assume the standard halo model. For the Al target, the shaded regions indicate the range of variation in different models of the loss function. The solid lines are computed using the GSRF loss function without damping, and the top of each shaded band is computed using the Lindhard loss function. An example of the target parameter space for thermal freeze-out through a heavy dark photon mediator [378] is shown in dashed blue.

detection is precisely where the effects of damping in the GSRF loss function become important. While the loss functions given here are valuable benchmarks, the true loss function likely falls somewhere between the Lindhard result and the plasmon tail. This is also suggested by fitting the measurements of Ref. [469], which go down to $\omega = 100\,\text{meV}$ and lie somewhat below the plasmon tail. (See the purple line in Fig. D.4.) To accurately predict the DM scattering rate, it is both essential and feasible to *measure* the loss function in the entire relevant regime of $1\,\text{meV} < \omega < 1\,\text{eV}$.

Figure D.5 shows updated reach curves for an aluminum superconductor target alongside the results of Refs. [369, 370]. The reach curves are specified with respect to a reference cross section defined by

$$\overline{\sigma}_e = \frac{16\pi\mu_{e\chi}^2\alpha_e\alpha_\chi}{\left((\alpha_{\text{EM}}m_e)^2 + m_\phi^2\right)^2}, \tag{D.36}$$

where $\mu_{\chi e}$ denotes the reduced mass of the electron–DM system, m_ϕ is the mediator mass, and $\alpha_{e,\chi} = g_{e,\chi}^2/(4\pi)$ in terms of the couplings which define the potential in Eq. (D.8). In Fig. D.5, 'light mediator' means $m_\phi \ll \alpha_{\text{EM}}m_e$ (defined with respect to the ordinary electromagnetic fine-structure constant $\alpha_{\text{EM}} \simeq 1/137$) and 'heavy mediator' means $m_\phi \gg \alpha_{\text{EM}}m_e$. All reach projections are computed in the zero-temperature limit and assume that the detector is sensitive to deposits between 1 meV and 1 eV. We show reach curves for a high-threshold experiment sensitive to deposits 1 eV–1 keV for a heavy mediator in Fig. D.6, along with recoil spectra for selected model points. We also illustrate the appearance of a feature in the recoil spectrum at ω_p in the fast halo model, where the plasmon peak is kinematically accessible. To facilitate comparison with the literature, we show reach curves corresponding to an event rate of $3\,\text{kg}^{-1}\,\text{yr}^{-1}$,

which corresponds roughly to a 95% C.L. constraint.

Figure D.5 in particular underscores the importance of properly treating the material response. For any interaction of the kind we consider in this chapter, screening is significant at low DM mass or for a light mediator. However, the implementation of screening in Ref. [370] overestimated the size of the effect for a vector mediator: at the lowest DM masses, the causal branch choice in the logarithms of Eq. (D.13) yields a rate as much as seven orders of magnitude greater than that produced by the acausal choice. Furthermore, accounting for the non-zero width of the plasmon peak further enhances the rate by an order of magnitude or more. The lingering uncertainty in analytical predictions of the loss function can be easily resolved by directly measuring the loss function in promising target materials.

Appendix E

Geometric enhancement to the DM interaction rate

The rate of Eq. (8.1) is written in a form appropriate for the scattering rate in a bulk volume. However, for thin layers, the dielectric response of the detector is different from that of a bulk sample of material. In particular, the relationship between the scattering rate and the dielectric function is modified: $\mathcal{W}[\epsilon]$ is replaced by a new response function $\mathcal{V}[\epsilon]$. This can significantly influence the DM interaction rate. This thin-layer response function can still be measured experimentally, but in the absence of experimental data, it is also possible to predict $\mathcal{V}[\epsilon]$ given a model for the dielectric function ϵ.

These effects are newly explored in Ref. [520]. Ref. [520] derives a function $R[\epsilon]$ such that $\mathcal{V} = \frac{1}{d}\,\mathrm{Re}(R)$, where d is the thickness of the detector layer (WSi in our prototype), and shows that the scattering rate per unit volume is exactly as given in Eq. (8.1) with the replacement $\mathcal{W} \to \mathcal{V}$. The response function $\mathcal{V}$ is determined

by solving the Poisson equation subject to the appropriate boundary conditions for a perturbing source with charge density $\rho = \rho_0 e^{i(\mathbf{q}\cdot\mathbf{x}-\omega t)}$ and evaluating the time-averaged power deposited in each layer. Schematically, one makes the ansatz $\phi = \psi(z)e^{i(\mathbf{q}\cdot x-\omega t)}$, where z is the coordinate normal to the layers. Then the Poisson equation reduces to an equation for $\psi(z)$, with the form

$$-q^2\psi(z) + 2iq_z\psi'(z) + \psi''(z) = -\rho_0/\epsilon(z). \tag{E.1}$$

After imposing the appropriate boundary conditions and solving for ψ, the thin-layer loss function can be written as

$$\mathcal{V} = \frac{q^2}{d}\,\mathrm{Re}\left[-i\frac{1}{\rho}\int \mathrm{d}z\left(i\psi(z) + \frac{q_z}{q^2}\psi'(z)\right)\right]. \tag{E.2}$$

Note that the integral in Eq. (E.2) is taken over all space, and the integrand has support outside the detector layer.

For a layer of thickness $d \ll q$, the resonance at the plasma frequency is suppressed compared to the bulk loss function. However, the thin-layer loss function exhibits a second resonance at smaller deposits, at $\omega \sim (qd/2)^{1/2}\omega_\mathrm{p}$, in the most important kinematic regime for light DM scattering. Thus, the DM scattering rate per unit volume for a thin layer can be enhanced significantly with respect to a bulk detector. Like the loss function $\mathcal{W}$, the thin-layer response function $\mathcal{V}$ is measurable for a particular target system.

One can make a first estimate of the geometric enhancements to absorption by assuming that the relationship between absorption and scattering is preserved, *i.e.*, that the bulk response function $\mathcal{W}$ in Eq. (10.3) can also be replaced with the thin-layer response function $\mathcal{V}$. An estimate carried out in this manner suggests that the

absorption rate can be enhanced by one or two orders of magnitude in some regimes. However, Eq. (E.2) is derived under the assumption that the momentum transfer q is much larger than the deposited energy ω, which is not the case for absorption. Thus, we do not show thin-layer curves in Fig. 10.6, and leave a quantitative treatment to future work.

In the absence of experimental data, we use the calculation of Ref. [520] to assess the relevance of the detector geometry to the DM scattering rate, considering only the WSi detector layer and the immediately adjacent SiO_2 layers. This calculation requires the dielectric function ϵ to be purely real outside the detector layer, meaning that these layers are dissipationless. We enforce this condition by explicitly taking the real part of ϵ outside the detector layer. This approximation is valuable to highlight a unique effect that takes place when the detector layer is much more strongly dissipative than the other layers: in this case, deposits in those other layers must be conducted to the detector layer before they can dissipate. This means that the detector is sensitive to deposits far from the detector layer, dramatically enhancing the effective volume of the system. This is also the reason for the integral in Eq. (E.2) to be extended over all space. Indeed, in the presence of dissipation in all space, this integral would diverge.

However, in our prototype, dissipation in the other layers is in fact non-negligible. Preliminary experimental results suggest that a deposit in another layer must be above the threshold by a factor of $\mathcal{O}(100)$ in order to reliably trigger the SNSPD, and understanding the effective available detector volume as a function of the deposited energy requires more detailed laboratory characterization. We thus show an additional conservative benchmark (dotted curves in Figs. 10.4 and 10.5) in which the dielectric

function is allowed to be complex everywhere, but only deposits within the WSi detector layer are included, *i.e.* the domain of the integral in Eq. (E.2) is restricted. In addition to the SiO_2 layers, we include the ZEP520A layer, treating it as semi-infinite in extent. This simplistic estimate demonstrates that when ϵ is allowed to be complex everywhere, the scattering rate is enhanced even when deposits outside the detector layer are neglected. Ultimately, direct experimental characterization can eliminate uncertainty in our treatment of geometric effects for both scattering and absorption.

Appendix F

Dark matter interactions in superconductors

In this appendix, we detail our treatment of DM–electron interactions in the language of QP pair production. In particular, we compute $\mathrm{Im}\,\epsilon_{\mathrm{BCS}}$ using the BCS coherence factor $\mathcal{F}_{\mathrm{BCS}}(p_1, p_2)$, and we show that the free electron scattering picture is recovered in the limit of large deposits. Throughout this section, q denotes the 4-momentum transfer $(\omega, \mathbf{q})$.

Consider a DM–electron interaction mediated by a scalar particle φ, with interaction Lagrangian

$$\mathcal{L}_{\mathrm{int}} = g_\chi \phi\, \overline{\chi}\chi + g_e \phi\, \overline{\psi}\psi \,, \tag{F.1}$$

where χ is a spin-1/2 DM fermion and ψ is the electron. Using the projection operator $P = (1 + \gamma^0)/2$, we can project out the so-called "large part" [546] of the electron field ψ_s, where $s = \uparrow, \downarrow$ refers to the two different spin states. This gives, at lowest order, the

interaction Hamiltonian for the electron–mediator interaction:

$$H_{\rm int} \simeq -g_e \int \mathrm{d}^3\mathbf{x}\, \phi(\mathbf{x}) \Big(\psi_\uparrow^\dagger(\mathbf{x})\psi_\uparrow(\mathbf{x}) + \psi_\downarrow^\dagger(\mathbf{x})\psi_\downarrow(\mathbf{x}) \Big) \,. \tag{F.2}$$

The DM–mediator interaction is governed by a Hamiltonian of the same form with the replacement $\psi \to \chi$. Higher-order terms are discussed by Ref. [547]. (Similarly, as shown in that reference, for a light vector mediator A^μ, we will have in the low-energy limit $\mathcal{L}_{\rm int} = g_\chi A_0 \chi^\dagger \chi + g_e A_0 \psi^\dagger \psi$. Higher-order terms and magnetic interactions will be suppressed by factors of v_χ/c in the low-energy limit. In the case of a heavy vector mediator, the A_0 interaction will again dominate, because the currents in the interaction $\mathbf{A}\cdot\mathbf{j}$ will be suppressed by factors of v_χ/c. Thus, light and heavy vector mediators should also be described by interaction Eq. (F.2) in this limit.[1]) Defining the density operator $\rho_e(\mathbf{x}) \equiv \sum_s \psi_s^\dagger(\mathbf{x})\psi_s(\mathbf{x})$, its Fourier transform is

$$\rho_e(\mathbf{q}) = \sum_s \int \frac{\mathrm{d}^3\mathbf{p}}{(2\pi)^3} c^\dagger_{\mathbf{p}-\mathbf{q},s} c_{\mathbf{p},s}, \tag{F.3}$$

where $c_{\mathbf{p},s}$ annihilates an electron with momentum $\mathbf{p}$ and spin s. A similar expression is obtained for ρ_χ. We can then write the interaction Hamiltonian as

$$H_{\rm int} = -\int \mathrm{d}^3\mathbf{x}\, \varphi(\mathbf{x}) \Big(g_e \rho_e(\mathbf{x}) + g_\chi \rho_\chi(\mathbf{x}) \Big) \,. \tag{F.4}$$

[1] One might object that a propagating A_0 should be suppressed in the non-relativistic limit, by virtue of the constraint $\partial_\mu A^\mu = 0$. Indeed, one can compute the time-ordered propagator $\int \mathrm{d}t\, \mathrm{d}^3\mathbf{x}\, \exp(iq^0 t - i\mathbf{q}\cdot\mathbf{x}) \langle 0|T\{A^\mu(\mathbf{x},t), A^\nu(\mathbf{0},0)\}|0\rangle$ in the interaction picture and verify that it is not Lorentz-covariant, and that its 00 component is highly suppressed when $|\mathbf{q}| \ll m_A$. However, the absence of a kinetic term for A_0 in the Lagrangian introduces an additional terms in the Hamiltonian to precisely cancel this suppression. The effective propagator becomes the Lorentz-covariant propagator $-i/(q^2 - m_A^2 + i\varepsilon)(\eta^{\mu\nu} - q^\mu q^\nu/m_A^2)$, which is unsuppressed for $q^0 \ll |\mathbf{q}|$, or when coupling to a conserved current. This is equivalent to the statement that an off-shell vector can be polarized in any direction. See Sec. 6.2 of Ref. [548] for further insight.

At second order in perturbation theory, the S-matrix will therefore contain a term

$$\begin{aligned}\hat{\mathbb{S}}^{(2)} &\supset -g_e g_\chi \int \mathrm{d}^4x\, \mathrm{d}^4x'\, \overline{\chi}(x)\chi(x)\Delta(x-x')\overline{\psi}(x')\psi(x') \\ &= -g_e g_\chi \int \frac{\mathrm{d}^4q}{(2\pi)^4}\, i\frac{\rho_\chi^\dagger(q)\,\rho_e(q)}{q^2-m_\phi^2+i\varepsilon}\,,\end{aligned} \tag{F.5}$$

where

$$\rho_e(q) \equiv \rho_e(\mathbf{q},\omega) = \int \mathrm{d}t\, e^{i\omega t}\rho_e(\mathbf{q},t) = \int \mathrm{d}t\, e^{i\omega t}e^{iH_0 t}\rho_e(\mathbf{q})\, e^{-iH_0 t}\,, \tag{F.6}$$

with H_0 the free Hamiltonian. In the presence of the lattice potential, an effective electron-electron potential is induced through a phonon loop [549]. The energy eigenstates in the presence of this effective potential are now given by the Bogoliubov QPs, with creation/annihilation operators $\gamma^\dagger$, γ respectively [492]. To implement the unitary transformation to the QP basis, we simply replace

$$c_{\mathbf{p}\uparrow} = u_{\mathbf{p}}\gamma_{\mathbf{p}\uparrow} + v_{\mathbf{p}}\gamma^\dagger_{-\mathbf{p},\downarrow}, \qquad c^\dagger_{-\mathbf{p},\downarrow} = -v_{\mathbf{p}}\gamma_{\mathbf{p}\uparrow} + u_{\mathbf{p}}\gamma^\dagger_{-\mathbf{p}\downarrow}, \tag{F.7}$$

where the coefficients $u_{\mathbf{p}}$ and $v_{\mathbf{p}}$ satisfy

$$|u_{\mathbf{p}}|^2 = \frac{1}{2}\left(1+\frac{\mathcal{E}_{\mathbf{p}}}{E_{\mathrm{QP}}(\mathbf{p})}\right), \qquad |v_{\mathbf{p}}|^2 = \frac{1}{2}\left(1-\frac{\mathcal{E}_{\mathbf{p}}}{E_{\mathrm{QP}}(\mathbf{p})}\right)\,, \tag{F.8}$$

with $\mathcal{E}$ and E_{QP} defined as in Eq. (9.1). We can then isolate the term in ρ_e that creates two quasiparticles (breaks a Cooper-pair):

$$\rho_e(\mathbf{q}) \supset \int \frac{\mathrm{d}^3\mathbf{p}}{(2\pi)^3}(u^*_{\mathbf{p}+\mathbf{q}}v_{\mathbf{p}} + u_{\mathbf{p}}v^*_{\mathbf{p}+\mathbf{q}})\gamma^\dagger_{-\mathbf{p}-\mathbf{q}\uparrow}\gamma^\dagger_{\mathbf{p}\downarrow} \tag{F.9}$$

giving, according to Eq. (F.6),

$$\rho_e(\mathbf{q},\omega) = \int \frac{\mathrm{d}^3\mathbf{p}}{(2\pi)^3}(u^*_{\mathbf{p}+\mathbf{q}}v_{\mathbf{p}} + u_{\mathbf{p}}v^*_{\mathbf{p}+\mathbf{q}})\gamma^\dagger_{-\mathbf{p}-\mathbf{q}\uparrow}\gamma^\dagger_{\mathbf{p}\downarrow}(2\pi)\delta\big(\omega - E_{\mathrm{QP}}(\mathbf{p}) - E_{\mathrm{QP}}(\mathbf{p}+\mathbf{q})\big)\,. \tag{F.10}$$

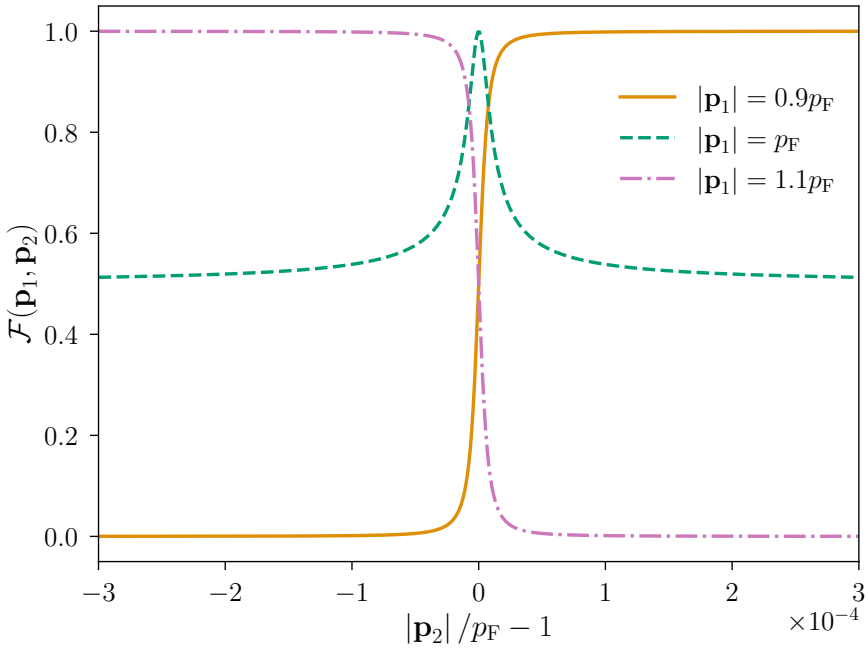

Figure F.1: The BCS coherence factor, $\mathcal{F}_{\mathrm{BCS}}(\mathbf{p}_1, \mathbf{p}_2)$, for several fixed values of $|\mathbf{p}_1|$. Note that $\mathcal{F}_{\mathrm{BCS}}(\mathbf{p}_1, \mathbf{p}_2) = \mathcal{F}_{\mathrm{BCS}}(\mathbf{p}_2, \mathbf{p}_1)$. When both momenta are far from the Fermi surface, the coherence factor reduces to the Pauli blocking factor. Pauli blocking only permits the creation of an electron–hole pair if the electron is above the Fermi surface and the hole is below. Accordingly, if both of $\mathbf{p}_1$ and $\mathbf{p}_2$ are on the same side of the Fermi surface, the coherence factor vanishes rapidly. Otherwise, it quickly approaches 1. That $\mathcal{F}_{\mathrm{BCS}}(p_\mathrm{F}, p_\mathrm{F}) = 1$ is a consequence of the sign in the coherence factor for the interactions considered here. For interactions with the opposite sign, $\mathcal{F}_{\mathrm{BCS}}(p_\mathrm{F}, p_\mathrm{F}) = 0$.

Plugging this into the S-matrix of Eq. (F.5), and using the fact that the scattering rate is given by $\Gamma = \frac{\mathrm{d}}{\mathrm{d}t}\sum_f |\langle f|\hat{\mathbb{S}}|i\rangle|^2$, we find that the lowest order Cooper-pair breaking rate is then simply given by

$$\Gamma(v_\chi) = g_e^2 g_\chi^2 \int \frac{\mathrm{d}^3\mathbf{p}_1}{(2\pi)^3}\frac{\mathrm{d}^3\mathbf{p}_2}{(2\pi)^3}\, 2\pi\delta\Big(\omega_{\mathbf{p}_1+\mathbf{p}_2} - E_{\mathrm{QP}}(\mathbf{p}_1) - E_{\mathrm{QP}}(\mathbf{p}_2)\Big)\frac{\left|u^*_{\mathbf{p}_1}v_{\mathbf{p}_2} + u_{\mathbf{p}_2}v^*_{\mathbf{p}_1}\right|^2}{\left|(\mathbf{p}_1+\mathbf{p}_2)^2 + m_\phi^2 - \omega^2\right|^2}, \tag{F.11}$$

where $\omega_{\mathbf{p}_1+\mathbf{p}_2} = (\mathbf{p}_1+\mathbf{p}_2)\cdot\mathbf{v}_\chi - (\mathbf{p}_1+\mathbf{p}_2)^2/2m_\chi$ is the energy deposited. The quantity

$$\mathcal{F}_{\mathrm{BCS}}(\mathbf{p}_1,\mathbf{p}_2) \equiv \left|u^*_{\mathbf{p}_1}v_{\mathbf{p}_2} + u_{\mathbf{p}_2}v^*_{\mathbf{p}_1}\right|^2 = \frac{1}{2}\left(1 - \frac{\mathcal{E}_{\mathbf{p}_1}\mathcal{E}_{\mathbf{p}_2} - \Delta^2}{E_{\mathrm{QP}}(\mathbf{p}_1)\,E_{\mathrm{QP}}(\mathbf{p}_2)}\right) \tag{F.12}$$

is the BCS *coherence factor*[2] [493]. The $\mathcal{E}_{\mathbf{p}_1}\mathcal{E}_{\mathbf{p}_2}$ term can be dropped if $\mathbf{p}_1$ or $\mathbf{p}_2$ (or both) are integrated over, provided the remainder of the integrand is even in $\mathbf{p}_1$ and $\mathbf{p}_2$. But if *e.g.* $\mathbf{p}_2$ is fixed to be $\mathbf{q}-\mathbf{p}_1$, then this term must be kept when integrating over $\mathbf{p}_1$.

It is straightforward to see that this matches onto the rate for free electron scattering when $\mathbf{p}_1$ and $\mathbf{p}_2$ are away from the gap ($|\mathbf{p}_i^2/2m_\mathrm{e} - E_\mathrm{F}| \gg \Delta$). In this limit, the matrix element for QP pair production becomes that for electron scattering, the coherence factor becomes a Pauli blocking factor, and the two QPs become an electron–hole pair. In particular, because of the functional form of the coherence factor, we always have $p_1 < \sqrt{2m_\mathrm{e}E_\mathrm{F}}$ and $p_2 > \sqrt{2m_\mathrm{e}E_\mathrm{F}}$ (or vice versa), with $\mathcal{F}_{\mathrm{BCS}}(p_1,p_2) \simeq 1$. Then $E(\mathbf{p}_1) \simeq E_\mathrm{F} - \mathbf{p}_1^2/2m_\mathrm{e}$ and $E(\mathbf{p}_2) \simeq \mathbf{p}_2^2/2m_\mathrm{e} - E_\mathrm{F}$, so the energy delta function in the rate of Eq. (F.11) reduces to

$$\delta\left(\omega - E_1 - E_2\right) \longrightarrow \delta\left[\omega - \left(\frac{\mathbf{p}_2^2}{2m_\mathrm{e}} - \frac{\mathbf{p}_1^2}{2m_\mathrm{e}}\right)\right]. \tag{F.13}$$

[2]The signs of the various terms in the coherence factor depend on the type of interaction. See, e.g. [491, 493].

That is, the kinematical constraint reduces to that of ordinary non-relativistic scattering. Meanwhile, the quantity

$$S(\mathbf{q},\omega) = \sum_f |\langle f|\rho_e(\mathbf{q})|0\rangle|^2 \delta(\omega - E_f) \tag{F.14}$$

is known in the literature as the *dynamic structure factor*, in terms of which we have, at lowest order in perturbation theory,

$$\mathrm{Im}\left(\frac{-1}{\epsilon^{(1)}(\mathbf{q},\omega)}\right) = \frac{\pi e^2}{\mathbf{q}^2} S(\mathbf{q},\omega)\ . \tag{F.15}$$

In our case, the structure factor is simply given by

$$S(\mathbf{q},\omega) = \int \frac{\mathrm{d}^3\mathbf{p}_1}{(2\pi)^3}\frac{\mathrm{d}^3\mathbf{p}_2}{(2\pi)^3}\,\mathcal{F}_{\mathrm{BCS}}(\mathbf{p}_1,\mathbf{p}_2)(2\pi)^3\delta^{(3)}(\mathbf{q}-\mathbf{p}_1-\mathbf{p}_2)\delta\Big(\omega - E_{\mathrm{QP}}(\mathbf{p}_1) - E_{\mathrm{QP}}(\mathbf{p}_2)\Big)\ , \tag{F.16}$$

allowing the loss function, $\mathrm{Im}(-1/\epsilon_{\mathrm{BCS}})$, to be expressed in terms of the QP dispersion relation and the BCS coherence factor. When higher order terms are included, we can resum the series (at zero temperature) in the *random phase approximation* (RPA) [460, 541]:

$$\hat{\mathbb{S}}^{(2,\mathrm{RPA})} \supset -g_e g_\chi \int \frac{\mathrm{d}^4 q}{(2\pi)^4}\, i\frac{\rho_\chi^\dagger(q)\,\rho_e(q)}{q^2 - m_\phi^2 + i\varepsilon}\,\frac{1}{\epsilon^{(\mathrm{RPA})}(q)}\ , \tag{F.17}$$

where $\epsilon^{(\mathrm{RPA})}(\mathbf{q},\omega) \equiv 1 + \chi(\mathbf{q},\omega)$, for $1 - \chi(\mathbf{q},\omega) \equiv -1/\epsilon^{(1)}(\mathbf{q},\omega)$. Importantly, we have

$$\mathrm{Im}\,\epsilon^{(\mathrm{RPA})}(\mathbf{q},\omega) = \mathrm{Im}\left(\frac{-1}{\epsilon^{(1)}(\mathbf{q},\omega)}\right) = \mathrm{Im}\,\chi(\mathbf{q},\omega) = \frac{\pi e^2}{\mathbf{q}^2} S(\mathbf{q},\omega)\ . \tag{F.18}$$

Putting everything together, we have

$$\Gamma^{(\mathrm{RPA})}(v_\chi) = g_e^2 g_\chi^2 \int \frac{\mathrm{d}^3\mathbf{p}_1}{(2\pi)^3}\frac{\mathrm{d}^3\mathbf{p}_2}{(2\pi)^3}\,\frac{\left|u^*_{\mathbf{p}_1} v_{\mathbf{p}_2} + u_{\mathbf{p}_2} v^*_{\mathbf{p}_1}\right|^2}{\left|(\mathbf{p}_1+\mathbf{p}_2)^2 + m_\phi^2 - \omega^2\right|^2}\,\frac{1}{|\epsilon^{(\mathrm{RPA})}(\mathbf{p}_1+\mathbf{p}_2,\omega_{\mathbf{p}_1+\mathbf{p}_2})|^2}$$
$$\times\, 2\pi\delta\Big(\omega_{\mathbf{p}_1+\mathbf{p}_2} - E_{\mathrm{QP}}(\mathbf{p}_1) - E_{\mathrm{QP}}(\mathbf{p}_2)\Big)\ , \tag{F.19}$$

and writing the coherence factor in terms of $\epsilon^{(1)}$ gives

$$\begin{aligned}\Gamma^{(\mathrm{RPA})}(v_\chi) &= \int \frac{\mathrm{d}^3\mathbf{q}}{(2\pi)^3} \frac{|V(\mathbf{q})|^2}{|\epsilon^{(\mathrm{RPA})}(\mathbf{q},\omega_\mathbf{q})|^2} \frac{2\mathbf{q}^2}{e^2} \operatorname{Im}\left(\frac{-1}{\epsilon^{(1)}(\mathbf{q},\omega_\mathbf{q})}\right) \\ &= \int \frac{\mathrm{d}^3\mathbf{q}}{(2\pi)^3} \frac{|V(\mathbf{q})|^2}{|\epsilon^{(\mathrm{RPA})}(\mathbf{q},\omega_\mathbf{q})|^2} \frac{2\mathbf{q}^2}{e^2} \operatorname{Im}\left(\epsilon^{(\mathrm{RPA})}(\mathbf{q},\omega_\mathbf{q})\right) \\ &= \int \frac{\mathrm{d}^3\mathbf{q}}{(2\pi)^3} |V(\mathbf{q})|^2 \frac{2\mathbf{q}^2}{e^2} \operatorname{Im}\left(\frac{-1}{\epsilon^{(\mathrm{RPA})}(\mathbf{q},\omega_\mathbf{q})}\right) . \end{aligned} \tag{F.20}$$

To the accuracy that $\epsilon^{(\mathrm{RPA})}$ represents the true dielectric function ϵ, we have derived Eq. (9.2).[3] The full form of $\epsilon^{(\mathrm{RPA})}$ in the BCS vacuum will be discussed in future work [494]. In this chapter, we make the approximation

$$\operatorname{Im}\left(\frac{-1}{\epsilon^{(\mathrm{RPA})}(\mathbf{q},\omega_\mathbf{q})}\right) \simeq \frac{\operatorname{Im}\epsilon^{(\mathrm{RPA})}_{\mathrm{BCS}}(\mathbf{q},\omega_\mathbf{q})}{|\epsilon_{\mathrm{L}}(\mathbf{q},\omega_\mathbf{q})|^2} , \tag{F.21}$$

where the Lindhard function [460, 468] $\epsilon_{\mathrm{L}} \equiv \epsilon^{(\mathrm{RPA})}_{\mathrm{FEG}}$ is the RPA dielectric function for a free electron gas (FEG), and accounts for screening and in-medium effects in a normal metal. Fortunately, by Eq. (F.18), $\operatorname{Im}\epsilon^{(\mathrm{RPA})}_{\mathrm{BCS}}$ can be evaluated from the dynamic structure function without knowing the full form of $\epsilon^{(\mathrm{RPA})}_{\mathrm{BCS}}(\mathbf{q},\omega)$:

$$\begin{aligned}\operatorname{Im}\left(\epsilon^{(\mathrm{RPA})}_{\mathrm{BCS}}(\mathbf{q},\omega)\right) &= \frac{\pi e^2}{\mathbf{q}^2} S(\mathbf{q},\omega) \\ &= \frac{e^2}{2\mathbf{q}^2} \int \frac{\mathrm{d}^3\mathbf{p}_1}{(2\pi)^3} \frac{\mathrm{d}^3\mathbf{p}_2}{(2\pi)^3} \mathcal{F}_{\mathrm{BCS}}(\mathbf{p}_1,\mathbf{p}_2) \times \\ &\qquad (2\pi)^4 \delta^{(3)}(\mathbf{q}-\mathbf{p}_1-\mathbf{p}_2) \delta\Big(\omega - E_{\mathrm{QP}}(\mathbf{p}_1) - E_{\mathrm{QP}}(\mathbf{p}_2)\Big) . \end{aligned} \tag{F.22}$$

This is because the imaginary part depends only on the spectrum of the Hamiltonian and on its relation to the operator ρ_e. It corresponds precisely to the sum over states in the scattering rate.

[3] Note the subtle difference in our approach (following Ref. [460]) from Ref. [521]. We define $S(\mathbf{q},\omega)$ to be given strictly by Eq. (F.14), while $\epsilon^{(\mathrm{RPA})}$ has been resummed in perturbation theory. One may optionally redefine S in terms of a resummed density operator $\rho^{(\mathrm{RPA})}$ following Ref. [460].

Appendix G

Quasiparticle downconversion in superconductors

We simulate QP down-conversion following a similar procedure to the calculations of Refs. [498, 500, 550, 551]. The principal difference is that we track the full momentum vector for scattered particles to retain information about the scattering direction relative to the momentum of the initial scattering event. Here we review the relevant scattering rate calculations, and we derive the emission angles of QPs and phonons in each relaxation process. We compare the final result to the well-established normal metal case described in Refs. [552, 553]. Down-conversion in the limit of high-energy initial QPs is also discussed by Refs. [498, 499].

G.1 Phonon Scattering at the Fermi Surface

We treat QP–phonon interactions following Ref. [500]. The simplified model in this treatment contains a single acoustic phonon branch with the dispersion relation

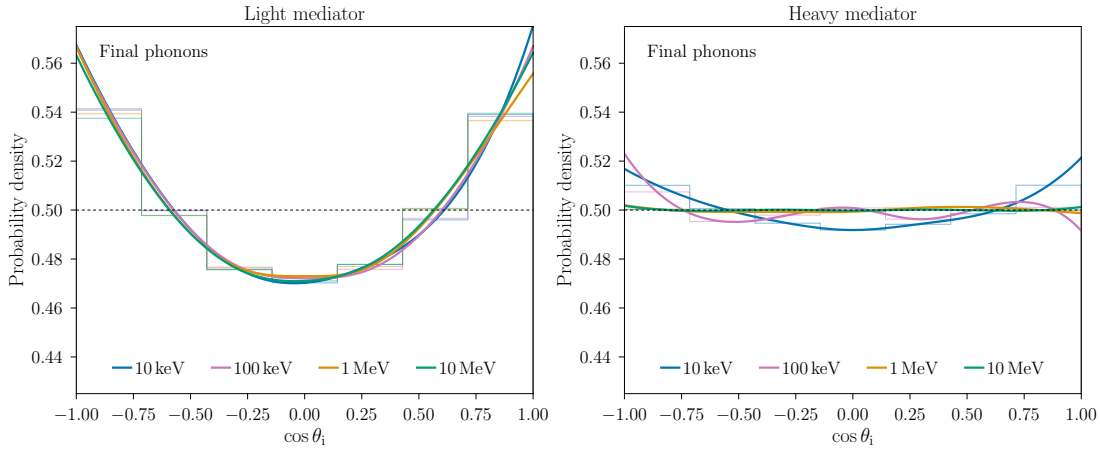

Figure G.1: Angular distributions of phonons produced by DM scattering in Al. The angles shown are defined with respect to the axis of the DM wind. The distribution of DM orientations in the Standard Halo Model is included. The left and right panel show the distributions in the light- and heavy-mediator limits, respectively. In each panel, the colors correspond to different DM masses, and a dashed horizontal line at $\cos\theta_{\mathrm{i}} = \frac{1}{2}$ indicates the isotropic distribution. Thick lines interpolate between histogram values (thin lines) for ease of visualization. In the right panel, the curves include only events with total deposit $\omega < 20\Delta$, for which the effects of down-conversion are less significant.

$\omega_q = qc_s$, where $\mathbf{q}$ is the phonon momentum and c_s is the sound speed in the material. For the dynamics of the problem, scattering is contained to the first Brillouin zone ($q < \frac{2\pi}{a}$), so we do not include an explicit upper limit in momentum in the rate integral. As significant down-conversion already occurs at energies well below the optical phonon modes in most superconductors, we do not explicitly include optical phonon emission in our calculations. This is compatible with the conclusions of past down-conversion codes (see *e.g.* Refs. [498, 550]).

For acoustic phonon emission, we first adopt the result of Ref. [500] for the emission rate of an acoustic phonon of momentum q in the zero-temperature limit:

$$\frac{\mathrm{d}\Gamma}{\mathrm{d}\omega_q}(E_{\mathrm{QP}}) = \frac{2\pi}{Z_0}\alpha^2 F(q)\,\mathrm{Re}\left(\frac{E_{\mathrm{QP}}-\omega_q}{\sqrt{(E_{\mathrm{QP}}-\omega_q)^2-\Delta^2}}\right)\times\left[1-\frac{\Delta^2}{E_{\mathrm{QP}}(E_{\mathrm{QP}}-\omega_q)}\right] \tag{G.1}$$

where Z_0 is the "renormalization parameter" defined in Ref. [500] (Typically $Z_0 \sim 2$). We have also assumed that $\alpha^2(q)F(q) \approx b\,\omega_q^2$, the Debye solution for this quantity [552]. The QP energy E_{QP} is defined in Eq. (9.1).

G.2 Computing the Scattering Angle

Given the general scattering rate of Eq. (G.1), we also need to determine the scattering angle for a phonon of energy $\omega_{\mathbf{q}} = qc_s$ emitted by an QP of energy $E_{\mathrm{QP}}(\mathbf{k})$. We find this angle by first solving for k' from the conservation of energy relation, $E_{\mathrm{QP}}(\mathbf{k}') = E_{\mathrm{QP}}(\mathbf{k}) - qc_s$, where $E_{\mathrm{QP}}(\mathbf{k}) \equiv E_{\mathrm{QP}}(\mathbf{k})$ as defined in Eq. (9.1). Solving, we get

$$k'^2 = 2m_{\mathrm{e}}\left[(E_{\mathrm{QP}}(\mathbf{k})-\omega_{\mathbf{q}})^2-\Delta^2\right]^{1/2}+p_{\mathrm{F}}^2. \tag{G.2}$$

In a normal metal or a semiconductor, the $p_{\rm F}$ dependence cancels. However, for a superconductor with $\Delta > 0$, the $p_{\rm F}$ dependence is retained.

We can now use momentum conservation to solve for scattering angle. Writing $\mathbf{k} = \mathbf{k}' + \mathbf{q}$, we have $k'^2 = k^2 + q^2 - 2kq\cos\theta_q$. Solving explicitly for k in the QP dispersion relation gives

$$k = \left[2m_{\rm e}\left(E_{\rm F} + s\sqrt{E_{\rm QP}^2 - \Delta^2}\right)\right]^{1/2}, \tag{G.3}$$

where $s = \pm 1$. An analogous sign s' appears in the solution for k'. We then find

$$\cos\theta_q = \frac{\omega_q^2 + 4\gamma^2\Delta\left[s\sqrt{(E_{\rm QP}^2 - \Delta^2)} - s'\sqrt{(E_{\rm QP} - \omega_q)^2 - \Delta^2}\right]}{4\gamma\omega_q\left[E_{\rm F}\Delta + s\Delta\sqrt{(E_{\rm QP}^2 - \Delta^2)}\right]^{1/2}}, \qquad \gamma \equiv \sqrt{\frac{\frac{1}{2}m_{\rm e}c_s^2}{\Delta}}. \tag{G.4}$$

We now consider a few limiting cases to understand the angular spread in the phonon spectrum. First, observe that for QPs far from the gap, with $\Delta \ll E_{\rm QP} \ll E_{\rm F}$, the scattering angle is unrestricted. In this limit, we have

$$\cos\theta_q \simeq \frac{1}{4\gamma}\frac{\omega_q}{\Delta}\left(\frac{\Delta}{E_{\rm F}}\right)^{1/2} + s'\gamma\left(\frac{\Delta}{E_{\rm F}}\right)^{1/2} + (s - s')\gamma\frac{E_{\rm QP}}{\omega_q}\left(\frac{\Delta}{E_{\rm F}}\right)^{1/2}. \tag{G.5}$$

For typical materials, γ is $\mathcal{O}(1)$ and $E_{\rm F} \gg \Delta$. Thus, if $s \neq s'$, the last term dominates in the limit of small ω_q, and $|\cos\theta_q| = 1$ is allowed. On the other hand, if $s = s'$, then the last term vanishes. The first term can be made arbitrarily small in the small-ω_q limit, and can even be made to cancel with the second term, which is itself always small. In this case, $\cos\theta_q = 0$ is allowed.

On the other hand, consider the near-gap regime, where $k \sim p_{\rm F}$. Here we can write $E_{\rm QP} = (1 + \delta)\Delta$ with $\delta \ll 1$, and since the final-state QP energy is at least Δ, we must have $\omega_q = a\delta\Delta$ with $0 < a < 1$. Inserting these expressions into Eq. (G.4) and

expanding for small δ gives

$$\cos\theta_q \simeq \frac{s\delta - s'\sqrt{1-a}}{(\delta/2)^{1/2}(a/\gamma)} \left(\frac{\Delta}{E_\mathrm{F}}\right)^{1/2}. \tag{G.6}$$

However, the phonon emission process is kinematically forbidden if $|\cos\theta_q| > 1$, and minimizing Eq. (G.6) over a gives

$$\min_{a\in(0,1)} \cos\theta_q \simeq \gamma \left(\frac{E_\mathrm{F}}{\Delta}\right)^{-3/2} \left(s\frac{E_\mathrm{F}}{\Delta}\sqrt{\frac{2}{\delta}} - 1\right). \tag{G.7}$$

Thus, phonon emission is only kinematically allowed for sufficiently large δ, i.e., for

$$\delta \gtrsim \delta_\mathrm{min} \equiv \frac{2\gamma^2 (E_\mathrm{F}/\Delta)^2}{\left[(E_\mathrm{F}/\Delta)^{3/2} + \gamma\right]^2}. \tag{G.8}$$

This result is self-consistent: in typical materials, γ is $\mathcal{O}(1)$, and $E_\mathrm{F} \gg \Delta$, so $\delta_\mathrm{min} \ll 1$. This gives rise to a condition for phonon emission:

$$E_\mathrm{QP} \gtrsim \Delta + \frac{4c_s^2 E_\mathrm{F}^2 m_\mathrm{e}}{\left(2E_\mathrm{F}^{3/2} + \Delta\sqrt{2m_\mathrm{e}c_s^2}\right)^2}\Delta. \tag{G.9}$$

QPs with energies below this threshold are ballistic: no phonon emission is allowed.

The angular distribution of final-state phonons is peaked oppositely to that of final-state QPs, and is closer to the isotropic distribution. The distribution is shown explicitly in Fig. G.1 for the same cases as in Fig. 9.1.

G.3 Relation to Normal Metal Scattering

It is instructive to compare Eq. (G.1) to the equivalent rate in the normal metal case. In a normal metal, this emission rate becomes [552]

$$\Gamma(E_\mathbf{k}) = 2\pi \int \frac{\mathrm{d}^3\mathbf{k}'}{(2\pi)^3} |\alpha_\mathbf{q}|^2 \, \delta(E_\mathbf{k} - E_{\mathbf{k}'} - \omega_\mathbf{q}) \delta^{(3)}(\mathbf{k} - \mathbf{k}' - \mathbf{q}), \tag{G.10}$$

where $\alpha_{\mathbf{q}}$ is the coupling for electron–phonon scattering. If we assume that scattering is isotropic, we can make a change of variables such that $\int \mathrm{d}\mathbf{p}' \propto \int \mathrm{d}\omega_q \, \mathrm{d}(\cos\theta)$, where $\cos\theta$ is the scattering angle between $\mathbf{k}$ and $\mathbf{k}'$. For scattering near the Fermi energy in the metal, conservation of momentum gives

$$q^2 = k^2 + k'^2 - 2kk'\cos\theta \approx 2p_{\mathrm{F}}^2(1-\cos\theta) \tag{G.11}$$

where $p_{\mathrm{F}} = \sqrt{2mE_{\mathrm{F}}}$ is the Fermi momentum. This implies $\mathrm{d}(\cos\theta) = -q\,\mathrm{d}q/p_{\mathrm{F}}^2$, which in turn allows us to write the differential scattering rate for $\omega_q < E_{\mathrm{QP}}$ as

$$\frac{\mathrm{d}\Gamma_n}{\mathrm{d}\omega_q}(E_{\mathrm{QP}}) = 2\pi\alpha^2 F(q) \; . \tag{G.12}$$

Here the subscript n indicates the normal metal case; N_0 is the normal metal density of states at E_{F}; and $\alpha^2 F(q)$ is the coupling-weighted phonon density of states, given by

$$\alpha^2 F(q) = \frac{N_0}{2p_{\mathrm{F}}^2}\int_0^{q_{\max}} \mathrm{d}q' \, q' |\alpha_{q'}^2| \delta(q'-q) \; . \tag{G.13}$$

For the scaling used earlier this gives the normal result that the acoustic scattering rate goes as E_{QP}^3 when integrated over energy, and the differential spectrum goes as ω_q^2.

When we derive the emission rate for superconductors, there are two important modifications required to get from Eq. (G.10) to the final differential rate in emitted phonon energy. First, we modify the dispersion relation to that of the QPs in the superconductor. In the metal, we had $E_{\mathbf{k}} = \mathcal{E}_{\mathbf{k}} \equiv \mathbf{p}^2/(2m_{\mathrm{e}}) - E_{\mathrm{F}}$ relative to the Fermi surface, but in a superconductor, the gap energy modifies this to

$$E_{\mathrm{QP}}(\mathbf{k}) = \sqrt{\mathcal{E}_{\mathbf{k}} + \Delta^2} \; . \tag{G.14}$$

In the normal metal phonon interaction, the total coupling has a density of states term that is valid in the metal, but not in the superconductor, since there are no states at

$E_{\mathbf{k}} < \Delta$. We can use the fact that the total number of states is the same to find the modified density of states at a given energy, i.e., we have $\mathrm{d}E\, N_s(E) = \mathrm{d}\mathcal{E}\, N_n(\mathcal{E})$, where the subscript s indicates the superconductor case. We thus find that

$$N_s(E) \approx N_0 \frac{\mathrm{d}\mathcal{E}}{\mathrm{d}E} = N_0 \frac{E}{\sqrt{E^2 - \Delta^2}}. \tag{G.15}$$

Thus, to rescale $\alpha^2 F(q)$ for the superconducting case, we have to rescale by the ratio of superconducting to normal states at a given energy. This gives us the first additional factor in the superconducting rate equation. Ref. [493] points out that, in principle, the gap function is complex-valued, hence the need to take only the real part.

The second correction factor is the coherence factor described in the previous section. This is a purely BCS effect that arises from the collective nature of the superconducting states. This factor ensures that the divergence in the density of states does not lead to a divergence in the phonon scattering rate. Taking the type I coherence factor for phonon emission from Ref. [493], we find

$$\mathcal{F}_{\mathrm{BCS}}(\Delta, E_{\mathrm{QP}}, \omega_q) \approx \frac{1}{Z_0}\left(1 - \frac{\Delta^2}{E_{\mathrm{QP}}(E_{\mathrm{QP}} - \omega_q)}\right), \tag{G.16}$$

using the same normalization procedure as Ref. [500]. Combining this with the previous correction factor and multiplying by the normal metal scattering rate produces the scattering rate of Eq. (G.1). This heuristic argument elucidates the origin of Eq. (G.1) in the low-temperature limit with a less formal approach than that in Ref. [500].

Appendix H

Directional reach estimation in superconductors

In this appendix, we detail the methodology used to produce Fig. 9.2 in the main text, and we demonstrate the impact of total-deposit cuts on the directionality of the signal. As in the main text, we assume the Standard Halo Model (SHM), *i.e.*, $f_\chi(v) \propto \Theta(v_{\rm esc} - v) e^{-v^2/v_0^2}$ in the galactic frame, taking $v_0 = 220\,\mathrm{km/s}$, $v_{\rm esc} = 550\,\mathrm{km/s}$, and Earth velocity $v_{\rm E} = 230\,\mathrm{km/s}$ relative to the galactic frame [418].

H.1 Statistical methods

The directional reach in Fig. 9.2 is estimated in two distinct ways. The dashed curves are based on measurement of an asymmetry in the counts of final-state QPs between two bins of equal solid angle, and the dotted curve is based on comparison of the full angular distribution against the null hypothesis.

H.1.1 Two-bin reach

We first discuss the two-bin reach estimate. The premise of this test is that an isotropic background gives rise to an isotropic distribution of final-state quasiparticles, and, in particular, produces statistically-indistinguishable counts in any two bins of equal solid angle. Since the DM wind is not isotropic, it is possible to statistically distinguish the counts in the two bins given a sufficient number of events.

We make the simplifying assumption that the angles of the final-state QPs are independent random variables. This is not strictly the case, since QPs that originate from the same event have some angular correlation. However, given a large number of events, such correlations are extremely sparse. Moreover, by simulating an ensemble of *isotropic* DM scattering events, we have directly checked that such correlations are irrelevant at the number of events needed to establish directionality. Having made this assumption, the assignment of an angular bin to each QP can be treated as a Bernoulli trial. We can then use the binomial test to determine whether to reject the isotropic distribution given a particular set of QPs.

This procedure allows us to determine whether a particular sample of final-state QPs is consistent with an isotropic signal. Next, we must translate this to a minimum number of events needed to establish directionality. To that end, we randomly draw samples of $N_{\mathrm{QP}} = 2, 3, 4, \ldots$ QPs and evaluate the binomial test for each sample, repeating the process many times for each fixed N to obtain a median p-value. We advance N_{QP} until this median p-value drops below the threshold value of 0.05, and we interpret the resulting value of N as the typical number of QPs needed in order to

establish directionality. Finally, this number of QPs must be translated to a number of scattering events. We estimate this number as $N_{\rm e} \equiv N_{\rm QP}/\bar{n}$, where $\bar{n}$ is the average number of final-state QPs produced by a scattering event. This $N_{\rm e}$ is indicated by the dashed curves in Fig. 9.2.

For the heavy mediator case, we introduce an additional step. As we discuss below, the distribution of final-state QPs produced by an event approaches the isotropic distribution as the deposit becomes much larger than the superconducting gap 2Δ. Thus, it is advantageous to restrict attention to events with total deposit below some cut, even at the cost of a reduced event rate below the cut. The dashed curve in the right panel of Fig. 9.2 is a composite of two reach curves obtained with cuts $\omega < 10\Delta$ and $\omega < 50\Delta$. Due to the complicated relationship between the total deposit, the initial directionality, and the effects of down-conversion, each of these cuts preserves overall directionality for a different mass range, and the combination of the two gives directional sensitivity over the entire mass range. Due to the large deposits favored in the heavy-mediator case, a cut on the total deposit is essential to establish directionality for all but the lowest masses.

This simplistic treatment produces a rough upper bound on the number of events needed to detect directionality, and admits a very direct interpretation. Even at the level of a two-bin experimental configuration, more sophisticated statistical treatments may yield slightly stronger results. In particular, it is possible to extract directionality using the Skellam distribution, as in rate modulation experiments [487]. Here one treats the count in each bin as a Poisson random variable, so that the difference in the number of counts between the two bins is a Skellam-distributed random variable. It

is then possible to test whether the two Poisson counts are produced with the same rate. However, our case is simpler than a traditional rate modulation experiment in that the target is isotropic, so the total rate is fixed. We have checked that using the Skellam distribution offers a slight enhancement to the two-bin reach, but does not qualitatively change the result.

H.1.2 Full angular distribution

Figure 9.2 also includes an estimate based on the full angular distribution of the final-state QPs, assuming an experimental configuration with great angular precision. For this estimate, we begin with an ensemble of simulated DM scattering events. Next, a second "null" ensemble of DM scattering events is simulated with an isotropic distribution of DM directions, using the same speed distribution as the Standard Halo Model. For each simulated event, we compute the mean angle of the final-state QPs, $\langle \cos\theta_{\rm i} \rangle$. We thus obtain two sets of samples $\{\langle \cos\theta_{\rm i} \rangle\}_{\rm SHM}$ and $\{\langle \cos\theta_{\rm i} \rangle\}_{\rm null}$ for the two ensembles. We then determine the average number of events needed to reject at 95% C.L. the hypothesis that the "SHM" and "null" samples are drawn from the same distribution, using the two-sample Kolmogorov–Smirnov test. Note that unlike the binomial test of the previous case, the samples being compared in this case are truly independent: since each mean angle corresponds to a single event, and vice versa, all of the values we draw originate from different scattering events, and they are thus independent random variates.

For light mediators, the two procedures give a nearly identical result, and Fig. 9.2 shows only the simpler two-bin result. This is easily understood in light of the

typical number of QPs produced in a scattering event: for a light mediator, a typical final state consists of an $\mathcal{O}(1)$ number of QPs, and deviation of these QPs from the isotropic distribution generally gives rise to a two-bin asymmetry. In fact, for large DM masses, taking the mean angle in each event discards information to the point that the distributional test slightly *underperforms* the two-bin test. For a heavy mediator, on the other hand, a typical final state may consist of $\mathcal{O}(10^4)$ QPs for large DM masses. The angular mean $\langle \cos\theta_{\rm i} \rangle$ will typically be very close to zero, and distinguishing the distribution of these means from the isotropic case becomes a problem of precision measurement. The large number of QPs makes this a realistic possibility. However, we estimate that achieving the high-resolution dotted curve in Fig. 9.2 would require a measurement precision of $\mathcal{O}(10^{-2})$ in $\cos\theta_{\rm i}$.

H.2 Energy dependence of directionality

In this section, we demonstrate the dependence of final-state directionality on the deposited energy. As noted in the main text, larger deposits allow for a larger number of relaxation events during the down-conversion process, which attenuates the correlation between the directions of the final-state excitations and those of the initial excitations produced by DM scattering. This is illustrated for particular realizations in Fig. H.1. In the left panel, the relatively small number of relaxation events and the low energies of the emitted phonons guarantee that the directions of the initial QPs are well-preserved, and no additional QPs are produced. In the right panel, on the other hand, the larger deposit allows for a larger number of relaxation events, with additional QP

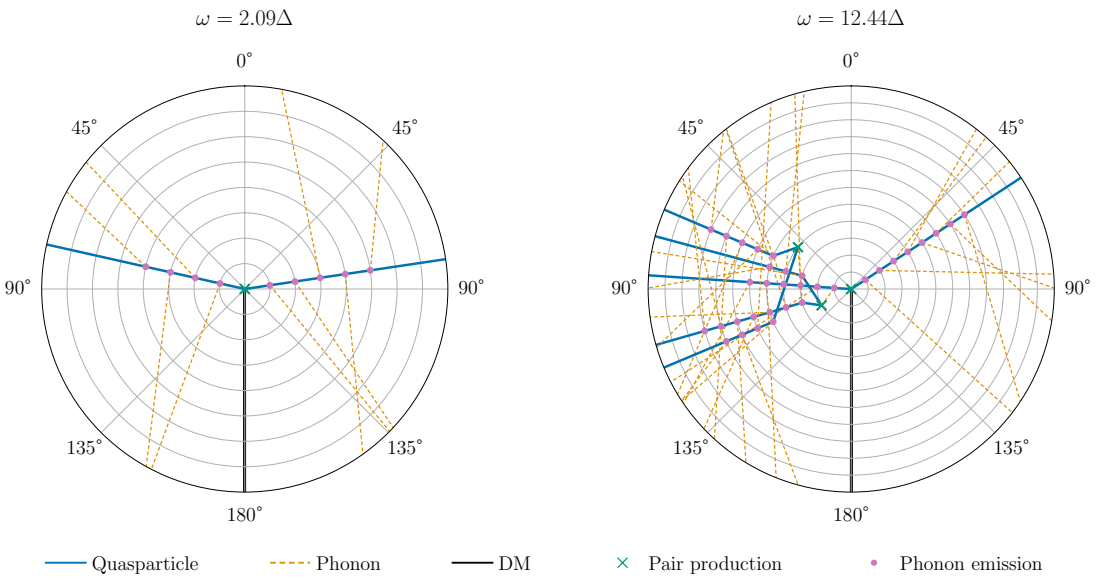

Figure H.1: Simulation of the QP down-conversion process for two initial energy deposits. The black line indicates the direction of the incoming DM. For visualization purposes, each line terminates at an angular coordinate corresponding to its angle from the incoming DM axis. Thus, the directions of the plotted excitations are represented by their endpoints, not by their slopes. *Left:* since $\omega = 2.09\Delta$, neither initial QP can emit a phonon with energy above 2Δ, so no phonons can produce QP pairs. Thus, the only QPs in the final state are those produced in the DM scattering event, with their directions barely altered. *Right:* now $\omega = 12.44\Delta$ and the emission of above-gap phonons is allowed. Thus, a chain of phonon emissions and decays erases much of the initial directionality, although a preference for off-axis final states remains visible.

pair production. While there remains a directional correlation between the final state and the initial QPs, this correlation is partially erased by down-conversion. Directional information is effectively lost for very large deposits.

The impact of down-conversion means that small deposits are favorable for directionality even if the directionality of the initial QPs is smaller in this regime. We can demonstrate this explicitly by evaluating the final-state asymmetry as a function of the deposited energy. To facilitate quantitative discussion of directionality, we introduce a quantitative "two-bin asymmetry" $\mathcal{A}_2$, defined as follows. As discussed above, we divide the final-state QPs into two bins of equal solid angle: the "on-axis" bin, with $|\cos\theta| > \frac{1}{2}$, and the "off-axis" bin, with $|\cos\theta| < \frac{1}{2}$. We denote the counts in these two bins by n_{on} and n_{off}, respectively, and then define

$$\mathcal{A}_2 \equiv \left| 2 \times \frac{n_{\text{on}}}{n_{\text{on}} + n_{\text{off}}} - 1 \right|. \tag{H.1}$$

In particular, the isotropic distribution gives $\mathcal{A}_2 = 0$, and a totally asymmetric distribution gives $\mathcal{A}_2 = 1$.

The left panel of Fig. H.2 shows $\mathcal{A}_2$ for the final-state QPs produced by a single QP injected at fixed energy ω with $\cos\theta = 1$. This serves as a proxy for the preservation of directionality at fixed deposit. Directionality is almost perfectly preserved for very low-energy QPs with $\omega < 3\Delta$. Above this threshold, it becomes possible for the QP to emit an above-gap phonon, with $E_{\text{ph}} > 2\Delta$. Such a phonon subsequently decays to another pair of QPs, which have only weak angular correlation with the original QP. The nature of the 3Δ threshold is also clearly visible in the right panel of Fig. H.2, which shows the fraction of the total energy that resides in the QP system after down-

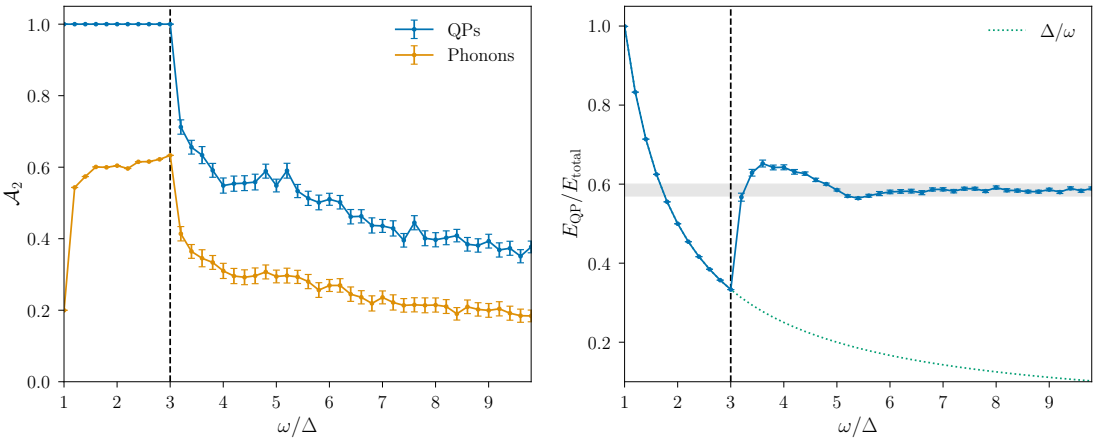

Figure H.2: Directionality of final-state QPs resulting from the down-conversion of a single QP of energy ω oriented with $\cos\theta = 1$. **Left:** asymmetry $\mathcal{A}_2$ of final-state QPs (blue) and phonons (orange). Directionality of the final excitations is quickly erased for $\omega \gg \Delta$, and $\mathcal{A}_2$ approaches zero in the large-ω limit. **Right:** Fraction of the total energy residing in the QP system after down-conversion. The gray band shows the range of asymptotic results 0.57–0.60 obtained in the literature for bulk Al superconductors [498, 550].

conversion. For $E_{\mathrm{QP}} < 3\Delta$, the QP relaxes almost all the way to the gap by emission of sub-gap phonons, which cannot produce any additional QPs. Thus, the final state consists of a single QP with $E_{\mathrm{QP}} \approx \Delta$, and a set of phonons with all the remaining energy from the deposit. The fraction of the initial energy in the QP system is approximately Δ/ω. Upon reaching $\omega > 3\Delta$, emission of above-gap phonons produces additional QPs in the final state, sharply raising the fraction of the total energy in the QP system. At large ω, this fraction asymptotically reaches $\sim$0.6. This is consistent with previous studies of down-conversion in the high-energy limit, which find fractions between 0.57 and 0.60 [498, 550] (gray band in Fig. H.2).

Naively, Fig. H.2 suggests that the smallest deposits will yield the strongest directionality. However, imposing an upper limit on the deposit also influences the

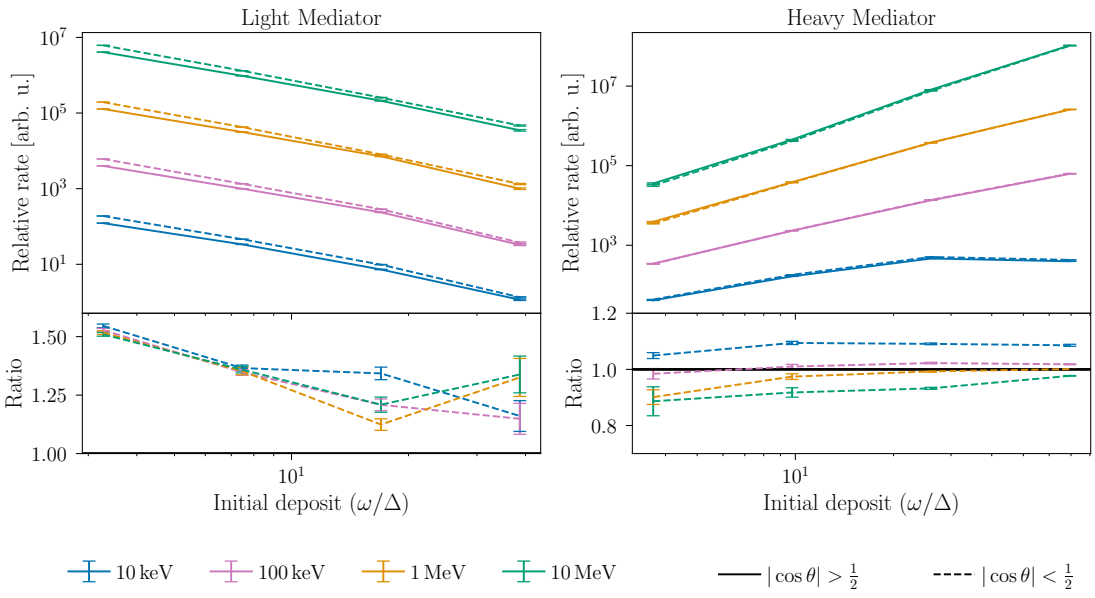

Figure H.3: *Top panels*: total event rates as a function of initial deposit in the on-axis ($|\cos\theta| > \frac{1}{2}$, solid) and off-axis ($|\cos\theta| > \frac{1}{2}$, dashed) bins for several DM masses, for light (*left panel*) and heavy (*right panel*) mediators. Normalization of each pair of curves is arbitrary and fixed for ease of visualization. *Bottom panels:* ratio of off-axis to on-axis QP counts. The left and right panels assume a light and heavy mediator, respectively.

directionality of the initial excitations, and, of course, the rate of events which fall below the cut. To study the total directionality as a function of the deposited energy, in Fig. H.3, we show the spectrum of final-state QPs in each of the two angular bins ("on-axis" and "off-axis") for several DM masses in the light- and heavy-mediator limits. The bottom panels of Fig. H.3 show the ratios of these spectra, i.e., a signed and shifted version of the two-bin asymmetry $\mathcal{A}_2$.

For light mediators, directionality is quickly lost for deposits well above the gap, and the ratio approaches 1. Moreover, the ratio is generally above 1. For heavy mediators, due to the directionality of the initial excitations, the ratio declines less noticeably, but it is near 1 throughout the plot and asymptotically reaches 1. As anticipated in Fig. 9.1, light mediators always enhance the off-axis rate, while heavy mediators can enhance either the off-axis or on-axis rates, depending on the DM mass: since scattering through a heavy mediator can produce a QP distribution peaked either in the forward direction or off-axis, the ratio can be either below or above 1.