

UC Riverside

UC Riverside Electronic Theses and Dissertations

Title

Model-Driven Cosmology With Bayesian Machine Learning and Population Inference

Permalink

<https://escholarship.org/uc/item/55x0s0dr>

Author

Ho, Ming-Feng

Publication Date

2024

Supplemental Material

<https://escholarship.org/uc/item/55x0s0dr#supplemental>

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
RIVERSIDE

Model-Driven Cosmology With *Bayesian* Machine Learning and Population
Inference

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

Doctor of Philosophy

in

Physics

by

Ming-Feng Ho

September 2024

Dissertation Committee:

Dr. Simeon Bird, Chairperson

Dr. Hai-bo Yu

Dr. Anson D'Aloisio

Copyright by
Ming-Feng Ho
2024

The Dissertation of Ming-Feng Ho is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

First and foremost, I would like to thank my advisor, Simeon Bird. Simeon gave me the freedom to focus solely on Bayesian statistics and machine learning throughout my graduate career. This allowed me to develop a variety of mathematical and computational skills that have become my greatest assets. Simeon also supported me both personally and academically. He is an excellent performer, and I learned a great deal from him on how to translate complex physics problems into humorous, everyday life examples.

Next, I am grateful to my family for their unconditional support. My parents have always been there for me, and my brother has been a great friend, offering valuable advice on how to improve my academic journey. My extended family, including my grandparents and aunts, have also been incredibly supportive. Also, my cousins have been my lifelong friends and have provided me supports. I am especially thankful to my partner, who has provided me with so many mental supports. I am grateful for her presence in my life.

Also, I would like to thank my old friends in the PG study group in National Taiwan University: Li-Hong, Po-Ya, Sheng-Chang, Tai-Yin, and Yu-Chung. They have supported me since the beginning of both my academic and non-academic journeys. Li-Hong has been a considerate friend, providing me with numerous intriguing ideas regarding cybersecurity. Po-Ya has been an excellent social leader for the group, helping to bring everyone together. Sheng-Chang has exemplified what a serious theoretical physicist looks like in the modern era. Tai-Yin has shared his insights on the fascinating aspects of machine learning. Yu-Chung has been a sincere friend, showcasing his passion for physics and research.

Next, I would like to thank my committee members, Hai-Bo Yu, Anson D'Aloisio, and George Becker and Christian Shelton (for PhD candidacy exam), who have been extremely helpful and supportive. I am also grateful to my friends in Bird's group: Phoebe has been very helpful in teaching me how to be a good research mentor to undergrads/high-school students. Bryan has thoughtfully helped me understand the astronomy culture in California. Martin has been invaluable in helping me become a better collaborator. Reza has been very supportive and guided me through graduate school as both close friend and colleague. Mahdi (Sum) has been a great comrade throughout our graduate studies. Hurum has been a good officemate to chat with, and Yanhui has been an excellent colleague with excellent programming skills. I am also grateful to my collaborator, Scott, who has been very supportive and helpful in the gravitational wave population project.

In addition, I would like to thank my Graduate Student Mentorship Program group: Krista, Mahdi (Sum), and Yongda, who have been very supportive and helpful throughout the COVID pandemic. Without their support, I wouldn't have survived the isolation of quarantine. I would also like to thank my friends in the union, Pao and Jonathan, who provided me with a space to discuss ways to improve the graduate school experience.

I would like to thank my housemates and colleagues, Wei-Xiang, Chia-Feng, and Wei-Cheng, who have been very supportive throughout my graduate career. Wei-Xiang has always offered interesting perspectives on research and life. Chia-Feng has been a great friend with excellent intuition on physics problems. Wei-Cheng has been a good old friend, always there to support me.

Last but not least, I am also grateful to my friends in the astro group: Marie, who always provided support and life advice; Jose, a good international friend who supported me; Jessica and Nakul, who took me on various hiking adventures; Garrett, who showed me what a good social leader looks like in the astro program; and Bayu, who demonstrated how to enjoy life as a hard-working PhD student. I am thankful to my friends in the physics department: Mehrdad, an excellent friend and passionate physicist; Pak Kau, a close friend who helped me survive graduate school; Mingda and Xilin, good Mandarin-speaking friends in my cohort; and Shirash, who showed me how to relax and enjoy life as a graduate student.

To my parents for all the support.

ABSTRACT OF THE DISSERTATION

Model-Driven Cosmology With *Bayesian* Machine Learning and Population Inference

by

Ming-Feng Ho

Doctor of Philosophy, Graduate Program in Physics

University of California, Riverside, September 2024

Dr. Simeon Bird, Chairperson

This thesis presents new directions for cosmological data analysis using Bayesian techniques and machine learning.

First, I introduce a novel machine learning spectroscopic analysis technique to detect absorption systems in the Lyman- α forest. Using Gaussian processes, I build data-driven models for the quasar emission and apply Bayesian model selection to classify the damped Lyman alpha absorbers (DLAs), which are high column density absorption systems found in quasar spectra. The Gaussian process DLA finder (GP-DLA) is applied to the Sloan Digital Sky Survey (SDSS) quasar spectra and is now adopted by the Dark Energy Spectroscopic Instrument (DESI) collaboration. This GP-DLA technique allows us to construct probabilistic catalogs of damped Lyman- α absorbers, offering a new approach to studying the intergalactic medium and cosmology at $z = 2 - 5$.

Next, I present a new method to infer cosmological parameters using Bayesian surrogate modeling with multi-fidelity emulators. Multi-fidelity emulators are a type of surrogate model that use information from multiple levels of fidelity to improve the accu-

racy of the surrogate model. This approach accelerates both the analysis of cosmological simulations and the inference of cosmological parameters, providing a probabilistic method to quantify and correct the resolution in cosmological simulations. Multi-fidelity emulators make it possible to perform fast and accurate parameter inference on large-scale structure data, such as the matter power spectrum, using computationally expensive simulations in high-dimensional parameter spaces.

Finally, I discuss population inference of gravitational wave (GW) data using a mixture model approach. The population statistics of GW events can provide insights into the formation and evolution of binary black holes (BBHs). I present a data-driven method to infer the mixing fraction between BH populations, along with a Bayesian hierarchical approach to correct for selection effects. The results of the mixing fraction analysis suggest that the population of $35M_{\odot}$ BHs is likely separate from the rest of the population, indicating that current formation channels for this mass bump need to be revised to include explanations for the separation of these massive $35M_{\odot}$ binaries.

Contents





List of Figures	xiv
List of Tables	xxix
1 Introduction	1
1.1 Spectroscopic Inference using Gaussian Processes	5
1.1.1 Data: Quasar Spectra	6
1.1.2 Method: Dataspace Gaussian Process Inference	11
1.1.3 Application: Damped Lyman alpha Absorbers	13
1.2 Multi-Fidelity Emulators for Cosmological Simulations	20
1.2.1 Data: Emulation	25
1.2.2 Method: Bayesian Multi-fidelity Emulation	33
1.2.3 Application: Cosmic Multi-Fidelity Emulators	37
1.3 Population Inference: A Short Note	39
2 Detecting Multiple DLAs per Spectrum in SDSS DR12 with Gaussian Processes	44
2.1 Abstract	44
2.2 Introduction	45
2.3 Notation	47
2.4 Bayesian Model Selection	48
2.5 Gaussian Processes	49
2.5.1 Definition and prior distribution	50
2.5.2 Observation model	51
2.6 Learning A GP Prior from QSO Spectra	52
2.6.1 Data	53
2.6.2 Modelling Decisions	54
2.6.3 Redshift-Dependent Mean Flux Vector	56
2.6.4 Learning the flux covariance	60
2.6.5 Model evidence	64
2.7 A GP Model for QSO Sightlines with Multiple DLAs	64
2.7.1 Absorption function	64



2.7.2	Model Evidence: DLA(1)	67
2.7.3	Model evidence: Occam's Razor Effect for DLA(k)	68
2.7.4	Additional penalty for DLAs and sub-DLAs	70
2.7.5	Parameter prior	72
2.7.6	Sub-DLA parameter prior	74
2.8	Model Priors	75
2.8.1	Sub-DLA model prior	76
2.9	Catalogue	78
2.9.1	Running Time	80
2.10	Example spectra	80
2.11	Analysis of the results	87
2.11.1	ROC analysis	87
2.11.2	CDDF analysis	91
2.11.3	Statistical properties of DLAs	95
2.11.4	Comparison to Garnett's Catalogue	102
2.11.5	Comparison to Parks Catalogue	103
2.12	Conclusion	108
3	Damped Lyman-alpha Absorbers from Sloan Digital Sky Survey DR16Q with Gaussian processes	110
3.1	Abstract	110
3.2	Introduction	111
3.3	Methods	114
3.3.1	Data	115
3.3.2	Gaussian process model	118
3.3.3	Example spectra	127
3.3.4	Selection on the strength of Occam's razor	129
3.3.5	Summary of the modifications	131
3.4	Results	133
3.4.1	Column density distribution function	133
3.4.2	Redshift evolution of DLAs	137
3.5	Checks for systematics	143
3.5.1	Effect of signal to noise ratios	143
3.5.2	Effect of quasar redshifts	147
3.5.3	Additional noise test	148
3.6	Results with DLAs in the Lyman β region	149
3.7	Comparison to the CNN model	151
3.8	Conclusion	154
4	Multi-fidelity Emulation for the Matter Power Spectrum using Gaussian Processes	158
4.1	Abstract	158
4.2	Introduction	159
4.3	Simulations	165
4.3.1	Latin hypercube sampling	167

4.3.2	Preprocessing of the simulated power spectrum	168
4.4	Single-fidelity emulators	171
4.4.1	Cosmological emulators	172
4.5	Multi-fidelity emulator	174
4.5.1	General assumptions	174
4.5.2	Linear multi-fidelity emulator (AR1)	176
4.5.3	Non-linear multi-fidelity emulator (NARGP)	181
4.6	Sampling Strategy for High-Fidelity Simulations	184
4.6.1	Nested training sets	184
4.6.2	Optimizing the loss of low-fidelity simulations	186
4.7	Results	190
4.7.1	Comparison of Linear and Non-Linear Emulators	191
4.7.2	Comparison to single-fidelity emulators	195
4.7.3	Varying the number of training simulations	200
4.7.4	Effect of other emulation parameters	202
4.8	Runtime	207
4.9	Conclusions	208
5	MF-Box: Multi-fidelity and multi-scale emulation for the matter power spectrum	212
5.1	Abstract	212
5.2	Introduction	213
5.3	Simulations	219
5.4	Emulation	226
5.4.1	Gaussian process emulator	227
5.4.2	Multi-Fidelity Emulation	228
5.5	Sampling strategy for high-fidelity simulations	234
5.5.1	Sliced Latin hypercube design (SLHD)	234
5.5.2	Selecting the optimal slice	236
5.6	Computational budget estimation	237
5.6.1	Error bounds for Gaussian process emulators	238
5.6.2	Optimal number of simulations per node	243
5.6.3	Empirical estimate of the error function	245
5.7	Results	252
5.7.1	MF-Box Accuracy (256 + 100 Mpc/h)	252
5.7.2	Emulation with various box sizes	257
5.7.3	Runtime comparison	260
5.8	Conclusions	262
6	Investigating the mixing between two black hole populations in LIGO-Virgo-KAGRA GWTC-3	265
6.1	Abstract	265
6.2	Introduction	266
6.3	Population Model	270
6.3.1	Population model: Subpopulations	270

6.3.2	Visualizations of the population model	275
6.3.3	Average mass spectrum	278
6.4	Population Inference	282
6.4.1	Inference Framework	284
6.4.2	Model averaging	290
6.5	Results	291
6.5.1	Fiducial model	291
6.5.2	Model averaging	295
6.6	Discussion	298
6.6.1	High Proportion of Gaussian-Gaussian BBHs.	298
6.6.2	Limitation on the Interpretation of the Mixing Fraction Posterior.	300
6.7	Conclusion	301
7	Conclusions	304
	Bibliography	308
.1	Sample posteriors for MDLA2	349
.2	Tables for CDDF, dN/dX and OmegaDLA	354
.3	Tables of the measurements for DLAs in SDSS DR16	354
.4	Likelihood Function of Average Mass Spectrum Fitting	354
.5	Fiducial Inference with a Different Power Spectral Density	357

List of Figures

1.1	A quasar spectrum with a DLA. This spectrum is from the SDSS DR16 quasar catalog, and it shows the Lyman- α forest and the Damped Lyman- α absorbers. The x-axis is the rest-frame wavelength, and the y-axis is the normalized flux. The red line is the continuum fit to the quasar emission spectrum, and the blue line is the observed quasar spectrum. The orange line is the Voigt profile fit to the DLA absorption system. The jupyter notebook tutorial in  https://github.com/jibanCat/gpy_dla_detection/blob/master/notebooks/Tutorial%3A%20Automate%20Lyman%20alpha%20Absorption%20Detection.ipynb provides a step-by-step guide on understanding the quasar spectrum and the DLA detection.	7
1.2	The Lyman- α forest, simulated from the neutral hydrogen intergalactic medium from a hydrodynamical simulation. Upper panel: the colors represent the temperature of the gas, and the x-axis is the co-moving box size. The shining arrow represents drawing a quasar sightline through the box. Bottom panel: the shaded blue color represent the absorptions due to the neutral hydrogen in the gas in the sightline. The y-axis is the flux (1: no absorption, 0: fully absorbed), and the x-axis is the comoving distance. YouTube link:  https://youtu.be/xBZLH14Qzyo?si=Jg08ARJju85Ljt5T	10
1.3	A Damped Lyman- α absorber found in a hydrodynamical simulation. The y-axis is the normalized flux (1 means un-absorbed flux and 0 means fully absorbed flux), and the x-axis is the co-moving box size. In contrast to Figure 1.2, the lyman alpha forest is washed out in the yellow region, and the damping wings bias the flux in the blue region.	14
1.4	How GP-DLA works. The GP-DLA finder uses two models to calculate the probability of the data having at least one DLA. A tutorial of GP-DLA can be found in  https://github.com/jibanCat/gpy_dla_detection	17
1.5	DLAs found in the SDSS DR16 quasar spectra, and the corresponding column density distribution function.	18
1.6	A demonstration of the usage of GP redshift estimator from Ref. [1]. This animation is on YouTube:  https://youtu.be/NhUycNaHBzM?si=cRSZhtTKWFJ6mlir	19

1.7	The Ly α 1D (P1D) and 3D (P3D) power spectra. P1D is measured from the auto-correlation of the flux in a single quasar spectrum, and P3D is measured from the cross-correlation of the flux in different quasar spectra. The spikes in the diagram show the Lyman alpha forest, and the black arrows show the quasar sightlines.	21
1.8	A schematic representation from Ref [2], showing the current and future probes of the structure formation, across different redshifts z and scales k . Lyman alpha forest probes the structure formation at $2 \leq z \leq 5$ and $k \sim 0.1 - 10 \text{ Mpc}^{-1}$	21
1.9	Bayesian surrogate modeling of the Forrester function. This is a demonstration of the Gaussian process emulator in combination of Bayesian optimization on the Forrester function. The red dots are the simulations we have run (the red stars being the last simulation we run), and the purple curve is the true simulation function. The emulator prediction is shown as the yellow curve, with the shaded areas showing the (1,2,3)- σ confidence intervals. The lower panel shows the Expected Improvement (EI) function (i.e., the acquisition function), which is used to find the next best point to evaluate the function in the Bayesian optimization. A higher EI value means the point is more likely to better improve the surrogate model fitting. Video tutorial can be found in  https://youtu.be/6JmuqVhSq5Y?si=5TFzIbepU6iCXRno	27
1.10	Bayesian optimization on the Forrester function. This is a demonstration of the Gaussian process emulator in combination of Bayesian optimization on the Forrester function. With a few iterations of Bayesian optimization, the parameters are adaptively allocated around $x \sim 7$ to better capture the sharp drop and rise of the function, and the parameter space of $x < 0.6$ is mostly evenly sampled. Video tutorial can be found in https://youtu.be/6JmuqVhSq5Y?si=5TFzIbepU6iCXRno	32
1.11	An example of the multi-fidelity emulator on the simple function. Red dots as the high-fidelity data, blue dots as the low-fidelity data. Red dashed curve is the high-fidelity true function and blue curve is the low-fidelity true function. Left panel shows the emulator prediction using only the high-fidelity data (purple curve), which is not accurate at $x > 0.5$. Right panel shows the emulator prediction using only the low-fidelity data (blue curve), which is biased, and the multi-fidelity emulator prediction (yellow curve) is more accurate. YouTube video tutorial can be found in  https://youtu.be/tQIytDnW0zk?si=TaKfDkJnhE48yHvK	35

1.12	An example of the non-linear multi-fidelity emulator. Red dots as the high-fidelity data, blue dots as the low-fidelity data. Red dashed curve is the high-fidelity true function and blue curve is the low-fidelity true function. Left panel shows the emulator prediction using only the high-fidelity data (purple curve), which has a wrong frequency. Right panel shows the emulator prediction using only the low-fidelity data (blue curve), which provides prior knowledge on the frequency of the high-fidelity function, and the multi-fidelity emulator prediction (yellow curve) is more accurate with a reasonable uncertainty quantification. YouTube video tutorial can be found in the same video link in the caption of Figure 1.11.	36
1.13	Example of the multi-fidelity emulator on the cosmic power spectrum. The left panel shows the N -body simulations in the density fields, including two low-fidelity nodes (L1, 128^3 in $256 \text{ Mpc}/h$ and L2, 128^3 in $100 \text{ Mpc}/h$) and one high-fidelity node (HF, 512^3 in $256 \text{ Mpc}/h$). The right panel shows the multi-fidelity emulator prediction (red curve) on the power spectrum, and the emulator relative errors are shown in the bottom panels for $z = 2$ and $z = 0$	39
2.1	The effect of the shift to the GP mean vector from the Lyman- α forest effective optical depth model ($\boldsymbol{\mu} \circ \exp(-\tau_0(1+z)^\beta)$). The dotted red curve shows the mean emission model before application of the forest suppression. The solid red curve is the mean model including the forest suppression.	62
2.2	The difference between the original pixel-wise noise variance $\boldsymbol{\omega}$ [3] and the re-trained $\boldsymbol{\omega}$ from Eq. 2.28. The re-trained $\boldsymbol{\omega}$ decreases because the fit no longer needs to account for the mean forest absorption.	62
2.3	The trained covariance matrix \mathbf{M} , which is almost the same as the covariance from [3]. Note that we normalize the diagonal elements to be unity, so this is more like a correlation matrix than a covariance matrix. The values in the matrix are ranging from 0 to 1, representing the correlation between λ and λ' in the QSO emission.	63
2.4	An example of finding DLAs using [3]’s model. Here we use the single-DLA per spectrum version of Garnett’s model. Upper: sample likelihoods $p(\mathbf{y} \theta, \mathcal{M}_{\text{DLA}})$ in the parameter space $\theta = (z_{\text{DLA}}, \log_{10} N_{\text{HI}})$. Red dots show the DLAs predicted by [4], and the blue squares show the maximum a posteriori (MAP) prediction of the [3]. Bottom: the observed spectrum (blue), the null model GP prior (orange), and the DLA model GP prior (Red). So that the upper and bottom panels have the same x-axis, we rescale the observed wavelength to absorber redshift.	81

- 2.5 The same spectrum as Figure 2.4, but using the multi-DLA model reported in this paper. **Upper:** sample likelihoods $p(\mathbf{y} | \theta, \mathcal{M}_{\text{DLA}})$ in the parameter space of the $\mathcal{M}_{\text{DLA}(1)}$, with $\theta = (z_{\text{DLA}}, \log_{10} N_{\text{HI}})$. **Bottom:** the observed spectrum (blue), the null model GP prior before the suppression of effective optical depth (orange), and the multi-DLA GP prior (Red). The orange curve is slightly higher than the one in Figure 2.4 because we try to model the mean spectrum before the forest. However, the DLA quasar model (red curve) matches the level of the observed mean flux better than Figure 2.4 due to the inclusion of a term for the effective optical depth of the Lyman- α forest. 82
- 2.6 **Blue:** the normalised observed flux. The spectral ID represents `spec-plate-mjd-fiber_id`. **Yellow:** Parks’ predictions on top of our null model. Our model predicts only one DLA while the CNN model in [4] predicts two DLAs. One of the DLAs predicted by [4] is coincident with the Ly γ absorption from our predicted DLA. `z_dla` corresponds to the DLA redshifts reported in Parks’ catalogue, and `lognhi` corresponds to the column density estimations of Parks’ catalogue. `p_dla` is the `dla_confidence` reported in Parks. **Red:** Our current model with the highest model posterior and the MAPs of column densities. In this spectrum, we show that it is crucial to include Ly β and Ly γ absorption from the DLA in the DLA profile. It not only helps to localize the DLA, but it also predicts N_{HI} more accurately using information from the Ly β region. The blue line shows the observed flux, the red curve is our multi-DLA GP prior, and the orange curve shows the predicted DLAs from [4] subtracted from our mean model. 84
- 2.7 A spectrum in which we detect two DLAs. **Blue:** Normalised flux. **Red:** GP mean model with two intervening DLAs. **Yellow:** The predictions from Parks’ catalogue. **Pink:** The MAP prediction of [3] on top of the GP mean model without mean flux suppression. The model posterior from [3] is listed in the legend (1) with the MAP value of $\log_{10} N_{\text{HI}}$. The column density estimate for the DLA near $\lambda_{\text{rest}} = 1025\text{\AA}$ has large uncertainty (see Figure 2.8). It is thus possible that this DLA could be a sub-DLA, as preferred by [4]. . . 84
- 2.8 The log sample likelihoods for the DLA model of the spectrum shown in Figure 2.7, normalised to range from $-\infty$ to 0. The DLA at $z_{\text{DLA}} \sim 2.52$ could be a sub-DLA (as preferred by [4]), as the $\log_{10} N_{\text{HI}}$ estimate is uncertain. However, we found that the 2-DLA model posterior $\log p(\mathcal{M}_{\text{DLA}(2)} | \mathbf{y}, \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\text{QSO}}) = -638$ is still higher than the model posterior from combining 1-DLA and 1-sub-DLA, which is $\log p(\mathcal{M}_{\text{DLA}(1)} + \mathcal{M}_{\text{sub}} | \mathbf{y}, \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\text{QSO}}) = -691.47$ 85
- 2.9 A noisy spectrum at $z_{\text{QSO}} = 2.378$ fitted with a large DLA by [3]. **Red:** The model presented in this paper predicts no DLA detection in this spectrum. **Pink:** The MAP prediction of [3] on top the GP mean model without the mean-flux suppression. **Gold:** The prediction of [4] subtracted from our mean model. Note that [4] also indicates a detection of a DLA at $z_{\text{DLA}} = 2.53$, but outside the range of this spectrum. 86

2.10	Top: The sample likelihoods of the spectrum shown in Figure 2.9. The colour bar indicates the normalised log likelihoods ranging from $-\infty$ to 0. Bottom: The orange curve indicates the GP mean model before mean-flux suppression, the red curve represents the mean model after suppression, and the blue line is the normalised flux of this spectrum. The x-axis of this spectrum is rescaled to be the same as the z_{DLA} presented in the upper panel.	86
2.11	The ROC plot made by ranking the sightlines in BOSS DR9 samples using the log posterior odds of containing at least one DLA. Ground truths are from the DR9 concordance catalogue. The orange curve shows the ROC plot of our current multi-DLA model, and the blue curve is derived from [3]. In this plot we consider only the model containing at least one DLA $p(\{\mathcal{M}_{\text{DLA}}\} \mathcal{D})$, rather than the multiple DLAs models, as the concordance catalogue contains only one DLA per spectrum.	89
2.12	The ROC plot for sightlines with one and two DLA detections, by using the catalogue of [4] (with <code>dla_confidence</code> > 0.98) as ground truth.	90
2.13	The MAP estimates of the DLA parameters $\theta = (z_{\text{DLA}}, \log_{10} N_{\text{HI}})$ for DLAs detected by our model in spectra observed by SDSS DR9, compared to the values reported in the concordance catalogue. The straight line indicates a perfect fit. Note that the concordance $\log_{10} N_{\text{HI}}$ values are not ground truth, so the scatter in column density predictions was expected.	92
2.14	The CDDF based on the posterior densities for at least one DLA (blue, ‘GP’). The DLAs are derived from SDSS DR12 spectra using the method presented in this paper. We integrate all spectral lengths with $z < 5$. We also plot the CDDF of [5] (N12; black) as a comparison. The error bars represent the 68% confidence limits, while the grey filled band represents the 95% confidence limits. Note that our CDDF completely overlaps with those of N12 for column densities in the range $10^{21} \text{ cm}^{-2} < N_{\text{HI}} < 10^{22} \text{ cm}^{-2}$	98
2.15	The line density of DLAs as a function of redshift from our DR12 multi-DLA catalogue (blue, ‘GP’). We also plot the results of [5] (N12; black) and [6] (PW09; grey). Note that statistical error was not computed in [5].	99
2.16	The total HI density in DLAs, Ω_{DLA} , from our DR12 multi-DLA catalogue as a function of redshift (blue, ‘GP’), compared to the results of [5] (N12; black), [6] (PW09; grey) and [7] (C15; red).	100
2.17	The redshift evolution (or non-evolution) of the CDDF. Labels show the absorber redshift ranges used to plot the CDDFs. In column density and redshift ranges with no detection at 68% confidence, a down-pointing arrow is shown indicating the 68% upper limit.	101
2.18	The difference between the MAP estimates of the DLA parameters $\theta = (z_{\text{DLA}}, \log_{10} N_{\text{HI}})$, against the predictions of [4]. We consider spectra which both catalogues agree contain one DLA.	104
2.19	The column density distribution function from [4], showing that the CNN algorithm substantially underestimates the number of DLAs in the high- N_{HI} regime.	106

2.20	dN/dX from [4]. The dN/dX agrees well with other surveys, but there is a moderate deficit of DLAs at high redshifts.	107
3.1	Our GP mean function using a precision weighted average of the rest-frame wavelengths. We extended our model compared [8] (light blue), both bluewards past the Lyman break at 912Å and redwards past the SiV emission line.	120
3.2	The correlation matrix learned from data, which is the covariance matrix \mathbf{K} normalised by the diagonal elements. Note that the correlation in the plot is pixel-by-pixel, and the matrix dimension is 2281×2281 . Different emission lines and the Lyman break are visible in the plot.	123
3.3	An example of a spectrum with distinct DLA features. (Top): The normalised observed spectrum in rest-frame wavelengths (blue) with the GP model (red) and the detection from the CNN model reported in DR16Q (orange). The title shows a series of column values in SDSS DR16Q catalogue, including SDSS identifier, best available redshift, PCA redshift, SDSS pipeline redshift, redshift from visual inspection, source for the best available redshift, and object classification from visual inspection (0: not inspected; 1: star; 3: quasar; 4: galaxy; 30: BAL quasar; 50: Blazar(?)). Shaded area (grey) shows the sampling range of z_{DLA} , which is from $\text{Ly}\beta + 3000 \text{ km s}^{-1}$ to $z_{\text{QSO}} - 3000 \text{ km s}^{-1}$. The legend shows the spectrum is from spec-6880-56543-478 (spec-plate-mjd-fiber-id). The CNN model (orange) detected two DLAs, with redshifts of $z_{\text{DLA}} = 2.91, 3.05$ and column densities of $\log_{10} N_{\text{HI}} = 20.5, 20.8$, at DLA confidence = 1 for each DLA. Our GP model (red) also detected two DLAs with the model posterior $p(\mathcal{M}_{\text{DLA}(2)} \mathcal{D}) = 1$ and column densities $\log_{10} N_{\text{HI}} = 20.6, 20.8$. (Middle): The sample likelihoods of detecting DLAs in the parameter space, $\theta \in (z_{\text{DLA}}, \log_{10} N_{\text{HI}})$. Colour bar shows the normalised log likelihoods, $\log p(\mathbf{y} z_{\text{DLA}}, \log_{10} N_{\text{HI}}, \tau_{0,\text{MF}}, \beta_{\text{MF}}, z_{\text{QSO}}, \mathcal{M}_{\text{DLA}})$, with the maximum log likelihood to be zero. We also show the maximum a posteriori estimates of DLAs in the blue squares. The posterior distribution sharply peaks at the parameter space, indicating the detection of these DLAs in high confidence. (Bottom): The observed flux (blue) as a function of absorber redshifts with the GP model (red) and the GP model before the meanflux suppression (yellow). The position on the x-axis directly corresponds to the x-axis in the middle plot.	128

3.4	An example of a noisy spectrum with an uncertain meanflux. The normalised observed spectrum in rest-frame wavelengths (blue) with the GP model (red) and the detection from the CNN model reported in DR16Q (orange). We also plot the result without marginalising the uncertainty of meanflux prior (cyan). Shaded area (grey) shows the sampling range of z_{DLA} , which is from $\text{Ly}\beta + 3\,000\text{ km s}^{-1}$ to $z_{\text{QSO}} - 3\,000\text{ km s}^{-1}$. Our proposed model (red) indicates no DLA in the spectrum, with the null model posterior $p(\mathcal{M}_{\text{-DLA}} \mathcal{D}) = 0.998$. On the other hand, if our model ignores the uncertainty of $(\beta_{\text{MF}}, \tau_{0,\text{MF}})$, it would falsely detect a DLA with $p(\mathcal{M}_{\text{DLA}} \mathcal{D}) = 0.916$ with $\log_{10} N_{\text{HI}} = 22.9$ (cyan). When marginalising over the uncertainty in effective optical depth, our proposed model (red) avoids detecting a false-positive large DLA.	129
3.5	The CDDF, integrated over all $z < 5$ spectral path, derived from SDSS DR16Q spectra with our proposed Gaussian process models (GP; blue). The CDDF measurements from [8] (GP: Ho20; orange) are plotted as a comparison. Error bars show the 68% confidence limits, while grey areas show the 95% confidence limits. Black dots are from [5] (N12).	134
3.6	The CDDFs with different Occam’s razor strengths, which discussed in Section 3.3.4. Occam’s upper (orange) represents $N = 30\,000$ while Occam’s lower (green) represents $N = 30$. We present our main result (GP; blue) with an optimal strength $N = 1\,000$, which we selected from visually inspecting a subset of the dataset. Note that the difference between different Occam’s strengths is well within 95% confidence limits. Black dots are from [5] (N12).	136
3.7	The CDDF derived from DLAs in a variety of redshift bins. Labels show the redshift bins in used. We show 68% confidence limits in error bars and 95% confidence limits in grey areas. If the bin is consistent with no detection at 68% limits, we show a down-pointing arrow indicating the 68% confidence upper limit.	138
3.8	The incident rate of DLAs as a function of redshift, integrated over $\log_{10} N_{\text{HI}} > 20.3$ spectra from our catalogue (GP; blue). We also plot the line densities from [5] (N12; black) and [6] (PW09; pink), and [8] (GP: Ho20; orange) as comparisons.	140
3.9	The total HI density in DLAs, integrated over DLAs with $\log_{10} N_{\text{HI}} > 20.3$ in our catalogue (GP; blue). For comparison, we plot the measurements from [9] (Berg19; green line and shaded area), [5] (N12; black), [7] (C15; red), and [8] (GP: Ho20; orange).	141
3.10	The line density (left) and Ω_{DLA} (right) in DLAs as a function of redshifts with different Occam’s razor strengths. Occam’s upper (orange) represents $N = 30\,000$ while Occam’s lower (green) represents $N = 30$. The main result (GP; blue) is computed with $N = 1\,000$	142

3.11	The CDDF of DLAs for a subset of samples with different minimal SNRs. SNR > 2 (orange) excludes 20% of the noisiest spectra, and SNR > 4 (green) excludes 54% of the spectra. 68% confidence limits are drawn as error bars, while 95% confidence limits are shown as a grey filled band.	144
3.12	The line density (left) and total N_{HI} mass (right) in DLAs as a function of absorber redshift from subsets of samples with different minimal SNRs. SNR > 2 (orange) excludes 20% of the noisiest spectra, and SNR > 4 (green) excludes 54% of the spectra.	145
3.13	The redshift evolution of the incident rate of DLAs, cutting with different quasar redshift intervals. Any correlation between the absorber properties and the background quasars redshifts might indicate systematics.	146
3.14	The redshift evolution of the incident rate of DLAs, cutting with different quasar redshift intervals. Unlike Figure 3.13, we remove the putative absorbers near the Lyman- α emission line with $ z_{\text{QSO}} - z_{\text{DLA}} < 30\,000 \text{ km s}^{-1}$	147
3.15	Comparing the CDDFs between the sampling range from Ly β -Ly α (Blue; GP) and Ly ∞ -Ly β (Red; GP: ly ∞ -lyb). The error bars are 68% confidence limits, and the shaded areas are 95% confidence limits. [8] (GP: Ho20; Orange) also used a sampling range from Ly β -Ly α	150
3.16	(Left) The comparison of dN/dX with different sampling ranges, Ly β -Ly α (blue) and Ly ∞ -Ly β (red). Other plot settings are the same as Figure 3.8. (Right) The comparison of Ω_{DLA} with difference sampling ranges, Ly β -Ly α (blue) and Ly ∞ -Ly β (red). Other plot settings are the same as Figure 3.9.	151
3.17	(Left) The CDDF of the DLAs detected by the CNN model presented in [4]. The z_{DLA} and $\log_{10} N_{\text{HI}}$ values are taken from the SDSS DR16Q catalogue in column Z_DLA and NHI_DLA. We require the confidence of DLAs to be larger than 0.98 and set the search range of the CNN DLAs to be the same as our search range, which is Lyman- β +3 000 km s $^{-1}$ to $z_{\text{QSO}} - 3\,000 \text{ km s}^{-1}$. (Right) The line density of the DLAs detected by the CNN model. All three measurements, GP, PW09, and CNN are consistent on the line density.	152
3.18	The 2D histograms for z_{DLA} (left) and $\log_{10} N_{\text{HI}}$ (right) estimated by the GP code and the CNN. We use the maximum a posteriori (MAP) estimate for parameter estimation for the GP code. The colourbars indicate the number of DLAs within the bin. The blue line is a straight line that shows the diagonal line of the 2D histogram.	153
3.19	Examples showing (top) the case of a sub-DLA overlapping a DLA and (bottom) the case of a DLA near to another DLA. The red line indicates the GP code predictions, and we describe the $\log_{10} N_{\text{HI}}$ in the legend. We intervene the DLAs from the CNN model in the DR16Q catalogue onto our null model in the orange line. Both spectra have high enough SNR: the upper one has SNR = 3.45 while the bottom one has SNR = 7.52. The damping wings and the Lyman- β absorption lines of the DLAs are visible in the plots.	155

4.1	The matter power spectrum output by MP-GADGET at different mass resolutions. The vertical dash lines indicate the mean particle spacing k_{spacing} for a given mass resolution. (Blue) : The matter power spectrum from a dark-matter only MP-GADGET simulation with 64^3 particles. (Orange) : The matter power spectrum from MP-GADGET with 128^3 particles. (Green) : The matter power spectrum from MP-GADGET with 256^3 particles. (Red) : The matter power spectrum from MP-GADGET with 512^3 particles. (Purple) : Linear theory power spectrum. The cosmology parameters are $h = 0.675, \Omega_0 = 0.278, \Omega_b = 0.0474, A_s = 1.695 \times 10^{-1}, n_s = 9.405 \times 10^{-1}$. The dotted line shows the relative error of HR (512^3 simulations) compared with EuclidEmulator2 [10], averaged over four different cosmologies.	169
4.2	The learned scale factor between fidelities in the linear multi-fidelity model, ρ , as a function of k . This scale factor is learned from 50 low-fidelity simulations and 3 high-fidelity simulations.	180
4.3	Two 2-D cross-sections of the 5-D samples of input parameters. The input parameters are designed with a nested structure, $\mathbf{x}_1 \subseteq \mathbf{x}_2$, between HR and LR. (Blue) : \mathbf{x}_1 , 50 sampling points in LR. (Orange) : \mathbf{x}_2 , 3 sampling points in HR. The selection of these 3 points is chosen by the procedure described in Section 4.6.2, which minimizes the LR error in the low-fidelity only emulator. (Green) : 10 points from the HR testing set, which is a different Latin hypercube than \mathbf{x}_1	185
4.4	Training (left) and testing (right) data for the multi-fidelity emulator. (Left) : 50 low-fidelity training simulations (blue) and 3 high-fidelity simulations (orange) used in a 50LR-3HR emulator. A HR is a 512^3 simulation and a LR is a 128^3 simulation. Both HR and LR are in a box with 256 Mpc/h per side. The 50 low-fidelity training simulations are drawn from a 5D Latin hypercube, $(h, \Omega_0, \Omega_b, A_s, n_s)$. The 3 high-fidelity simulations are a subset of the low-fidelity simulation hypercube. (Right) : 10 high-fidelity test simulations (green dashed) and 3 high-fidelity training simulations (orange).	185
4.5	Emulator mean squared errors evaluated from 64^3 emulators and 256^3 emulators. We compute all subsets of 3 samples from a 50 samples Latin hypercube, $\binom{50}{3} = 19\,600$ subsets in total. Colorbar is in log scale. The blue dashed line represents a perfect linear relationship.	189
4.6	Predicted divided by exact power spectrum from a 50 LR-3 HR emulator using a linear multi-fidelity method (AR1). Different colours correspond to 10 test simulations spanning a 5-D Latin hypercube. The shaded area indicates the worst-case $1 - \sigma$ emulator uncertainty. There is one test simulation driving the larger error compared to the non-linear one in Figure 4.7.	191
4.7	Predicted divided by exact power spectrum from a 50 LR-3 HR emulator using a non-linear multi-fidelity method (NARGP). Different colours correspond to 10 test simulations spanning a 5-D Latin hypercube. The shaded area indicates the worst-case $1 - \sigma$ emulator uncertainty. Note that the y-scale in this plot is the same as Figure 4.6.	192

4.8	Relative emulator errors from a 50 LR-3 HR emulator using linear multi-fidelity (blue) and non-linear multi-fidelity (orange). Solid lines represent the average error from test simulations, $\frac{1}{10} \sum_{i=1}^{10} \left \frac{P_{\text{pred},i}}{P_{\text{true}}} - 1 \right $. Shaded areas show the maximum and minimum test errors.	194
4.9	Non-linear multi-fidelity emulator (blue) with 50 LR and 3 HR simulations, compared to single-fidelity emulators with 3 HR (orange) and with 11 HR (green). Shaded area indicates the maximum and minimum emulation errors. The computational cost for a 50 LR-3 HR emulator $\simeq 9\,000$ core hours while the single-fidelity emulator with 11 HR requires $\simeq 25\,000$ core hours. However, a 50 LR-3 HR emulator still outperforms an 11 HR emulator.	195
4.10	Relative emulator errors between a 50 low-fidelity emulator and a non-linear 50 LR-3 HR emulator. Errors are evaluated on 10 HR simulations. Shaded area indicates the maximum and minimum errors. Note that the y-axis is in \log_{10} scale.	197
4.11	Core hours for running the training simulations versus emulation errors for high-fidelity only emulators (orange) and low-fidelity only emulators (blue), linear multi-fidelity emulators (AR1) with 2 HR (green), and non-linear multi-fidelity emulators (NARGP) with 3 HR (purple). The numbers in the labels indicate the number of training simulations used in the emulator. For multi-fidelity emulators, X - Y , X is the number of low-resolution and Y is the number of high-resolution training simulations. The dots show the average errors. The upper shaded areas show the maximum emulator errors among 10 test simulations. The LR samples beyond 100 are drawn from a separate Latin hypercube with 400 samples. For LF-only emulators, we only calculate the relative errors for $k \leq 3$	198
4.12	Relative emulator error of non-linear N LR-3 HR emulator colour coded with different number of LR training simulations, with $N \in \{10, 20, 30, 40, 50\}$. The same as Figure 4.8, solid lines represent the average error from test simulations, $\frac{1}{10} \sum_{i=1}^{10} \left \frac{P_{\text{pred},i}}{P_{\text{true}}} - 1 \right $, and shaded areas show the maximum and minimum test errors.	200
4.13	Relative emulator errors from non-linear 50 LR- N HR emulator with $N = 3$ (blue), $N = 5$ (orange), $N = 7$ (green), and $N = 9$ (red) HR training simulations. Solid lines are the average test errors. Shaded areas show the maximum and minimum test errors.	201
4.14	Relative emulator errors for 50 LR-3 HR emulator emulators using different qualities of LR simulations. (Blue) : using 128^3 simulations as low-fidelity training simulations. (Orange) : using 64^3 simulations as LR, which are $\simeq 8$ times cheaper than 128^3 simulations. (Green) : using 256^3 simulations as LR, which are $\simeq 8$ times most expensive than 128^3 simulations. Shaded area shows the maximum and minimum errors among ten test simulations.	203
4.15	Relative emulator errors for a non-linear emulator at different redshifts, $z \in \{0, 1, 2\}$. Note the y-axis is in \log_{10} scale. The larger error in the $z = 2$ emulator at $k > 2 h\text{Mpc}^{-1}$ may be due to a transient near the mean-particle spacing in the LR simulations, see Figure 4.16.	205

- 4.16 The matter power spectrum at $z = 2$, output by MP-GADGET with different mass resolutions. The vertical dash lines indicated the mean particle spacing k_{spacing} for a given mass resolution. **(Blue)**: The matter power spectrum from dark-matter only MP-GADGET simulation with $N_{\text{ptl,side}} = 64$. **(Orange)**: The matter power spectrum from MP-GADGET with $N_{\text{ptl,side}} = 128$. **(Green)**: The matter power spectrum from MP-GADGET with $N_{\text{ptl,side}} = 256$. **(Red)**: The matter power spectrum from MP-GADGET with $N_{\text{ptl,side}} = 512$. **(Purple)**: Linear theory power spectrum. 206
- 5.1 Illustration of the MF-Box framework and the dark-matter only simulations performed at $z = 0$. MF-Box provides a emulation framework to connect power spectra (denoted as $f(\theta)$, where θ is the input cosmology) from low-fidelity simulations (L1 and L2) to high-fidelity simulations (HF), providing an efficient emulation framework in predicting HF power spectra using only a few HF simulations augmented with many low-fidelity simulations with various volumes. ρ is a learnable multiplicative resolution correction parameter, and δ is a learnable additive resolution correction parameter. Details of the MF-Box model can be found in Section 5.4.2. The particle loads and box sizes for each simulation are listed in Table 5.1. **(a.)** Large-scale structures of each simulation are shown. Simulations L1 and L2 have the same particle load ($N_{\text{ptl,side}} = 128$), but L1 has a smaller box size (100 Mpc/h). As a result, the large scales of L1 resemble those of the high-fidelity (HF) simulation, while L2 lacks the necessary large-scale information to match HF. **(b.)** Zoomed-in view (25.6 Mpc/h) of the small scales from (a.). L1 lacks structures due to the sparsity of particles at this scale, whereas L2 captures more structures by utilizing a smaller box size. As a result, L1 resembles HF at small scales due to its finer mass resolution. 221
- 5.2 Matter power spectra from dark-matter only MP-Gadget simulations with various fidelities, conditioning on the same cosmology. The top panel shows the power spectra from a large-box low-fidelity (L1; blue), a small-box low-fidelity (L2; black), and a large-box high-fidelity simulations (HF; yellow). The numeric values for different fidelities of simulations are tabulated in Table 5.1. The 2nd, 3rd, and bottom panels show the ratios of L1/HF (red) and L2/HF (black) simulations, conditioned on different redshift bins, $z = 3.0, 0.5, 0$. (Bottom panel): We also show the ratio between (L1, L2) and the linear theory power spectrum from CLASS at large scales. The solid lines show the median and shaded areas show the 68% quantiles across 60 different cosmologies. 222

5.3	Experimental design of low- and high-fidelity simulations in this work. The prior volume is chosen to be the same as EuclidEmulator2 [11]. Crosses (black) are the input parameters for the low-fidelity simulations (both L1 and L2). Circles (red and yellow) are the parameters for high-fidelity simulations, which is a subset of the low-fidelity experimental design. We use max-min Sliced Latin Hypercube (SLHD) [12] for the LF design, containing 20 slices with 3 samples in each slice. Red and Yellow circles show two of the slices, which we select to be the input parameters for HF simulations.	224
5.4	MF-Box’s emulation errors, averaged over redshift bins and test simulations, using 60 L1, 60 L2, and 3 HF (see Table 5.1). Here, we show the emulation minimum and maximum errors using different slices from SLHD (blue shaded area), and the best slice found by the grid search method is labeled as yellow.	236
5.5	Relative errors plotted against the number of LF and HF design points in a MF-Box emulator. Here, LF refers to the combined number of L1 and L2 points, where $LF = n_{L1} = n_{L2}$. The plot reveals a trend of decreasing errors as the number of low-fidelity training simulations increases. However, due to the limited number of high-fidelity points compared to LF points, the decreasing trend is relatively modest.	241
5.6	Inferred relative errors for all available MF-Box emulators are displayed. Each subplot corresponds to a fixed number of HF points (as indicated in the title) with varying LF points (on the x-axis). The red curves represent the median predictions (50% posterior). Blue lines indicate the average relative errors obtained from the MF-Box emulators, while the error bars represent the standard deviation of relative errors across 10 simulations in the test set. The shaded area depicts the 25% and 75% confidence interval of the predictions based on the inference results. Overall, the relative errors demonstrate a decreasing trend as the number of LF and HF points increases.	246
5.7	Inferred relative errors as a function of LF points. Shaded area shows the 25% and 75% confidence interval of the prediction from the inference result.	247
5.8	Inferred relative errors as a function of HF points. Shaded area shows the 25% and 75% confidence interval of the prediction from the inference result.	248
5.9	The predicted emulator errors as a function of the budget size, in the unit of the number of LF simulations. The predictions are based on the medians of the parameter posteriors presented in Table 5.3. The plot shows the predicted error functions using different combinations of LF and HF nodes. The red, yellow, blue, and black curves represent the predicted error functions with varying LF nodes and a fixed HF node ($n_{HF} = 3, 4, 5, 6$). In contrast, the purple dashed curve represents the predicted error function with varying HF nodes and a fixed LF node ($n_{LF} = 60$). The green dotted line illustrates the error function corresponding to the optimal budget (Eq 5.21). The vertical gray dotted lines indicate the budget size in terms of the number of HF simulations. The horizontal gray dotted lines denote the predicted errors at the levels of (1%, 0.5%, 0.3%).	250

5.10	Relative errors averaged over $z = [0, 0.2, 0.5, 1, 2, 3]$ for different multi-fidelity models, AR1 (blue), NARGP (red), and MF-Box (yellow). The MF-Box model uses 60 L1 (256 Mpc/h), 60 L2 (100 Mpc/h), and 3 H (256 Mpc/h) simulations for training. Both AR1 and NARGP use 60 L1 and 3 HF for training. The shaded area is the variance among different test simulations.	253
5.11	Relative errors averaged over all k modes (split into large and small scales) for different multi-fidelity models (AR1 (blue), NARGP (red), and MF-Box (yellow)), broken down into different redshift bins. The grey dashed line is the HF-only emulator using 3 H simulations, and the solid grey line is the LF-only emulator using 60 L1 simulations. The shaded area is the variance among different test simulations. MF-Box improves the emulation at small scales at higher redshifts ($z \geq 1$). We do not include the variance of LFEmu (60L1) because the variance is too large.	254
5.12	Relative error as a function of the number of HF training points for different multi-fidelity methods: AR1 (blue), NARGP (red), and MF-Box (yellow). The range of the number of HF points is relatively small, so the error estimate trend is unclear. However, in general, the emulation error decreases with more HF points. (Left) Averaged relative error for $z \in [0, 0.2, 0.5]$. (Right) Averaged relative error for $z \in [1, 2, 3]$	255
5.13	Relative errors for AR1 (blue), NARGP (red), and MF-Box (yellow) as a function of LF points, splitting into two redshift bins. (Left) Averaged error for $z \in [0, 0.2, 0.5]$. (Right) Averaged error for $z \in [1, 2, 3]$	256
5.14	Relative errors of multi-fidelity emulation as a function of L2 boxsize, for AR1 (blue), NARGP (red), and MF-Box (yellow). Note that we use L2 instead of L1 for AR1 and NARGP models.	258
5.15	Relative errors averaged over redshift bins, as a function of k modes. MF-Box with 224 Mpc/h L2 (blue), MF-Box with 160 Mpc/h L2 (yellow), and MF-Box with 100 Mpc/h L2 (red). The gray dashed line is the NARGP model uses 100 Mpc/h L2.	259
5.16	Runtime comparison in node hours. We average the error across redshift bins $z = [0, 0.2, 0.5, 1, 2, 3]$ and average across k bins. AR1 and NARGP perform similarly to MF-Box at $z < 1$. Dashed lines are the predicted error based on the error function Eq 5.13, which we inferred in Section 5.6.	261
6.1	A cartoon illustrates the mixing scenarios used in this work. The size of the circles represents the masses of the black holes, while the color indicates the underlying population. If the power-law and Gaussian bump BHs are mixed, as in the PG model, the resulting two-dimensional probability density of chirp mass and mass ratio (\mathcal{M}, q) will exhibit a distinct morphology, as shown in Figure 6.2.	274
6.2	Map of likelihood density in chirp mass (\mathcal{M}) versus mass ratio (q) space for three subpopulation models, PP, PG, and GG. The shape parameters, $\lambda = (-\alpha, \mu, \sigma) = (-3.66, 31.59, 5.51)$ used to generate the map come from the average mass spectrum of the GWTC-3's Power-law+Peak model as derived in Section 6.3.3.	276

6.3	The one-dimensional marginal distribution of the two-dimensional density shown in Figure 6.2. The chirp mass spectrum, as shown in the upper panel, features three peaks at $\mathcal{M} \sim 8M_{\odot}$, $14M_{\odot}$, and $28M_{\odot}$. The mass ratio spectrum reveals a bump at $q \sim 0.2$ for the PG population, highly equal-mass binaries in the GG population, and a smooth mass ratio distribution for the PP population.	277
6.4	The exploratory models with varying spectral indices, $-\alpha$, in the space of chirp mass and mass ratio, (\mathcal{M}, q) . It ranges from a flat spectral index (left panel) to a sharp spectral index (right panel). In these exploratory plots, each subpopulation model has the same relative abundance, $1/3$	278
6.5	The exploratory models with varying spectral indices, $-\alpha$, on the chirp mass $(p(\mathcal{M}))$ and mass ratio $(p(q))$ marginal distributions. The relative abundance is fixed to $(\psi_{PP}, \psi_{PG}, \psi_{GG}) = (0.92, 0.03, 0.05)$, matching the maximum a posteriori (MAP) of the model averaging results in Section 6.5.2.	279
6.6	The average mass spectrum sampled from GWTC-3 Power-law+Peak model and the best-fit mass spectrum with the fiducial values of our population model. The data points represent the Monte Carlo samples from the Power-law+Peak with best-fit parameters from [13], with the Poisson uncertainty of the Monte Carlo samples. The purple line represents the best-fit power-law function with a Gaussian peak to the sampled average mass spectrum. . . .	281
6.7	Posterior probability for mixing fraction parameters, ψ under different model assumptions (“Fiducial” model in purple and the “Model Averaging” in green). Two extreme hypothetical scenarios are also shown: (1). Red error bars show the case where the Gaussian bump is completely separate from the power-law population. (2). Black error bars show the case where the Gaussian bump and power-law populations are co-located. These hypothetical scenarios are defined using $\lambda_{\text{peak}} = 3.8_{-2.6}^{+5.8}\%$ for the Gaussian bump reported in LVK [13]. The orange dashed lines represent the “Partially Separate/Co-located” scenario, showing a situation in which a portion of the Gaussian bump black holes is co-located with the power-law distribution, while the remainder is separate.	292
6.8	Predicted primary/secondary mass and mass ratio functions from the model averaging inference (in light green) and the fiducial inference (in purple). The solid lines represent the MAP and the shaded areas represent the 95% confidence intervals. The light green lines represent predicted functions sampled from the posterior probability of both ψ and λ . The underlying black dashed lines represent the fiducial model of Power-law+Peak from GWTC-3, with the corresponding fiducial parameters detailed in Table 6.1. The 95% confidence interval for the “Fiducial” inference reflects only the posterior uncertainty in ψ and does not include uncertainty regarding λ . As we fix the minimum mass (m_{min}) and the maximum mass (m_{max}), the shape uncertainty at the low and high mass ends is not incorporated.	294

- 6.9 The model evidences from 1,000 population inferences utilized in the “Model Averaging”, namely, $p(\boldsymbol{\lambda} \mid \{\mathbf{d}_i\}, \{\text{trig}\}, N_{\text{obs}})$, where we treat each set of $\boldsymbol{\lambda}$ as a model. The colors indicate the model evidence for each of the shape parameters, $(-\alpha, \mu, \sigma)$. There is no obvious correlation between σ and μ , but there is some weak correlation between α and μ with correlation $\simeq -0.3$. 296
- 6.10 The chirp mass spectra predicted by the subpopulations in our population model: PP (purple), PG (blue), and GG (green). The shaded regions represent the 95% confidence intervals, and the subpopulations are normalized to have $\psi_{\text{PP}} + \psi_{\text{PG}} + \psi_{\text{GG}} = 1$. The underlying mass spectrum (black) is from the Flexible Mixture model [14], and utilizes fitting results from Ref. [13]. . 297
- .1 Fiducial MCMC chains using various Power Spectral Densities (PSDs) in the computation of the detection probability, p_{det} . Four different PSDs are utilized: the analytical PSD (`AdVMidHighSensitivityP1200087`), the PSD from the GWTC-1 event (`GW150914_095045`), the PSD from the GWTC-2 event (`GW190916_200658`), and the PSD associated with the GWTC-3 event (`GW200322_091133`). There is no evident shift in the posterior with different PSDs. 359

List of Tables

2.1	The confusion matrix for multi-DLAs detections between Garnett with multi-DLAs and Parks. Note we require both the model posteriors in Garnett and DLA confidence in Parks to be larger than 0.98. We also require $\log_{10} N_{\text{HI}} > 20.3$	105
3.1	Mathematical notations and definitions	116
3.2	The confusion matrix for multi-DLAs detections between the GP and the CNN model [4]. Note we require both the model posteriors of our GP model and DLA confidence in Parks to be larger than 0.98. We also require $\log_{10} N_{\text{HI}} > 20.3$. The maximum number of DLAs is fixed to three, and everything larger than three is considered three.	152
4.1	Notations and definitions	166
5.1	Low- and high-fidelity simulation suites used in our study. The definition of low- and high-fidelity nodes is based on a relative scale specific to our approach and is not intended for direct comparison with other matter power spectrum emulators.	220
5.2	Notations and definitions	226
5.3	MCMC analysis of Eq 5.13: $\frac{1}{N} \sum_{i=1}^N \left \frac{f_{\text{HF}}(\theta_i) - m_{f_{\text{HF}}}(\theta_i)}{f_{\text{HF}}(\theta_i)} \right = \Phi(n_{\text{L1}}, n_{\text{L2}}, n_{\text{HF}}) \approx \eta \cdot (\rho_{\text{L1}} \cdot n_{\text{L1}}^{-\frac{\nu_{\text{L1}}}{d}} + \rho_{\text{L2}} \cdot n_{\text{L2}}^{-\frac{\nu_{\text{L2}}}{d}} + n_{\text{HF}}^{-\frac{\nu_{\text{HF}}}{d}})$. The notation $\{\Phi(n_{\text{L1},j}, n_{\text{L2},j}, n_{\text{HF},j})\}_{j=1}^{144}$ means all 144 MF-Box emulator errors used for parameter estimation. The column ‘‘Posterior (50%)’’ reports the medians of the posteriors of the parameters, and ‘‘Posterior (25%, 75%)’’ reports the 25% and 75% quantities of the posterior distributions.	242

6.1	The fiducial shape parameters for our population model, transforming the fiducial values of Power-law+Peak to our average mass spectrum parameterization. The uncertainty in converting the shape parameters from one population model to another can be arbitrarily small, depending on the number of Monte Carlo samples used to construct the KDE for the average mass spectrum. Therefore, we do not include this uncertainty in the table. We do not vary m_{\min} or m_{\max}	283
.1	Average column density distribution function for all DLAs with $2 < z < 5$. The table is generated by using $p(\{\mathcal{M}_{\text{DLA}}\} \mathbf{y}, \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\text{QSO}})$. See also Figure 2.14.	352
.2	Table of dN/dX values from our multi-DLA catalogue for $2 < z < 5$. The table is generated by using $p(\{\mathcal{M}_{\text{DLA}}\} \mathbf{y}, \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\text{QSO}})$. See also Figure 2.15.	353
.3	Table of Ω_{DLA} values. The table is generated by using $p(\{\mathcal{M}_{\text{DLA}}\} \mathbf{y}, \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\text{QSO}})$. See also Figure 2.16.	353
.4	Table of dN/dX values, integrated over all putative absorbers with $N_{\text{HI}} > 10^{20.3}$ in our catalogue.	354
.5	Ω_{DLA} values, integrated over all putative absorbers with $N_{\text{HI}} > 10^{20.3}$ in our catalogue.	355
.6	The column density distribution function integrated over all spectral lengths within $2 < z < 5$	356
.7	The prior for shape parameters in the average mass spectrum fitting.	358

Chapter 1

Introduction

The beginning of the 21st century is the era of Bayesian methods. Bayesian methods have revolutionized the way natural and social scientists analyze data, providing a more accessible way to quantify uncertainty and allowing scientists to develop domain-specific theories to explain their data. Before this revolution, scientists often used classical statistical methods, which typically required numerous assumptions about the data and models, and only well-trained statisticians could understand the results.

The benefits of Bayesian methods are two-fold: they provide a principled way to combine scientific theory with data, and they also propagate uncertainty from the data to the theories. Thus, Bayesian methods free scientists from the burden of understanding complex statistical methods and allow them to focus on the scientific questions they are interested in. This is particularly well-suited for physicists, who often create many untested (sometimes untestable) hypothetical models.

The rise of Bayesian methods coincides with the rise of observational cosmology. When CPUs became fast enough to run Markov Chain Monte Carlo (MCMC) samplers, physicists were able to send satellites into space to observe the cosmic microwave background (CMB) radiation. The COBE (Cosmic Background Explorer) satellite was the first to measure the CMB radiation, revealing that the background temperature of the universe is 2.7K. The success of COBE led to the WMAP (Wilkinson Microwave Anisotropy Probe) satellite, which measured the anisotropy of the CMB radiation, providing insights into the age, composition, and geometry of the universe.

However, we have only one universe and can measure it only once. How can we measure cosmological parameters from this single observation? The answer lies in Bayesian methods. Cosmologists perform Bayesian inference using the well-developed Λ CDM model and run MCMC samplers on the cosmological parameters. We can simulate the universe many times using the Λ CDM model and use these simulations to infer the cosmological parameters. To achieve this, we need robust theoretical models, which physicists are good at creating. This demonstrates the power of Bayesian methods in physics: even when data are scarce, we can still infer parameters using theoretical models. This advantage does not always apply to other fields, where data are abundant but good theoretical models are scarce. Bayesian methods shine in physics for good reasons.

The success of cosmological inference using Bayesian methods has also spread to astrophysics and astronomy. Astronomers have photometric surveys to catalog stars and galaxies. With a limited number of photometric bands (usually 5-10 bands), astronomers infer the redshift of galaxies, star formation rates, and metallicity using knowledge of stellar

population synthesis models. Although this is a challenging problem, and the accuracy of photometric inference is strongly affected by prior knowledge, it opens a window for astronomers to study the universe on a large scale at minimal cost. This illustrates the power of Bayesian methods in astronomy: even when data are noisy and incomplete, we can still infer parameters using theoretical models.

After the dark age of the 20th century, Bayesian statistics have become popular and have survived criticism from frequentists. However, the information explosion and the rise of machine learning and deep learning techniques bring new challenges to the Bayesian paradigm. Data are no longer scarce, and theory models are no longer the only way to explain data. The old strong-model-small-data paradigm is giving way to a new weak-model-big-data paradigm. Instead of spending years developing theoretical models, scientists now use deep learning to replicate what Bayesian methods can do. Deep learning models are extremely flexible and, with enough data, scientists can make predictions without understanding the underlying mechanisms. Moreover, under the umbrella of the AI industry's rise, all deep learning APIs are well-maintained and easy to use even for non-computer scientists. The success of convolutional neural networks, large language models, and reinforcement learning can be directly applied to scientific research with minimal effort. How can Bayesian methods survive in this new era?

I titled my thesis “Model-Driven Cosmology with Bayesian Machine Learning and Population Inference” to emphasize what I believe is the most important aspect of Bayesian methods in the era of big data: being model-driven. In my opinion, Bayesian methods are not just a set of statistical techniques but a set of principles that guide physics theory

research. Their purpose is not to replace theoretical models but to help make these models more accurate, interpretable, and properly connected to data.

In this thesis, I document my research from the past few years and demonstrate how Bayesian methods can be combined with machine learning to solve three different problems in cosmology and astrophysics. By documenting these three methods, I hope this thesis will serve as a reference for future researchers interested in exploring data analysis using Bayesian methods in the era of big data.

This thesis is organized as follows. In this chapter, I will summarize the three Bayesian approaches I took in the past few years, and the motivation behind them. For each method, I will describe the data I used, the method I developed, and the application I applied. Each type of data and application represent a science problem, which I will also describe after the method.

Following the tradition of thesis writing in this field, I append the published papers I have written during my PhD journey in the following chapters. Chapter 2 is the published paper in Ref. [8], Chapter 3 is the published paper in Ref. [15], Chapter 4 is the published paper in Ref. [16], Chapter 5 is the published paper in Ref. [17], and Chapter 6 is the submitted paper to Physical Review D (in review).

These papers provide details of each scientific topic and method I mentioned in this chapter, they also provide a much thorough literature review and discussion on the scientific implications. For interested experts, these papers can be served as a reference for the specific topic. For a general audience, Chapter 1 provides a high-level overview which will be easier to follow.

This section is structured as follows. Section 1.1 will describe the first method I developed during my PhD journey, which is the Gaussian Process Damped Lyman- α Absorber (GP-DLA) method. Section 1.2 will describe the second method I developed during my PhD journey, which is the Multi-fidelity Emulator (MFEimulator) method. Section 1.3 is a short note on the mathematics behind applying Bayesian hierarchical modeling to population inference on a catalog of gravitational wave events, subjecting to the selection bias.

1.1 Spectroscopic Inference using Gaussian Processes

Astronomy has a long history of using various statistical methods to analyze the spectroscopic data. Generally speaking, there are two features in a spectrum: emission lines and absorption lines. Emission lines are the bright spectral lines emitted by the source, i.e., the object we are observing. Absorption lines are the spectral lines absorbed by the objects on the line of sight. The absorption lines are usually caused by the interstellar medium (ISM) or the intergalactic medium (IGM) between the source and the observer. Emission lines inform us about the chemical composition of the source, while absorption lines inform us about the medium on the line of sight.

Quasar spectra, however, do not have well-defined stellar models to predict the emission lines due to the complexity of the active galactic nuclei (AGN) and supermassive black holes. The traditional data analysis in this field is to use low-redshift high-resolution quasar spectra to build a template library, and then use the template library to fit the high-redshift quasar spectra. Astronomers usually apply dimension reduction methods, such as

principal component analysis (PCA), to reduce the dimension of the template library. In this section, I will explain how to interpret the quasar data analysis in a Bayesian way, and describe the template fitting as a probability density function (PDF) through a Gaussian process.

This section describes the Gaussian Process Damped Lyman- α Absorber (GP-DLA) method, which is a new method to classify the absorption systems in the quasar spectra using Gaussian processes. The beauty of this method is that it summarizes and combines the past statistical methods used in the quasar spectroscopic analysis, and provides a single Bayesian way to model both the emission lines and absorption lines under the well-defined probabilistic framework.

1.1.1 Data: Quasar Spectra

Quasars are active galactic nuclei (AGN) powered by supermassive black holes. Astronomers use quasars, the most luminous objects in the universe, as background light sources to trace absorption systems along the line of sight. Here, I will provide a high-level overview of quasar spectroscopic data to give context for the GP-DLA method. The primary goal is to help readers unfamiliar with this field understand how to interpret the data.

Figure 1.1 shows a typical quasar spectrum from Sloan Digital Sky Survey (SDSS). The x-axis is the wavelength (λ), and the y-axis is the flux (f), which counts the number of photons received by the telescope at a given wavelength. A spectrum describes the flux as a continuous function of the wavelength, $f(\lambda)$, however, in practice, we only have a

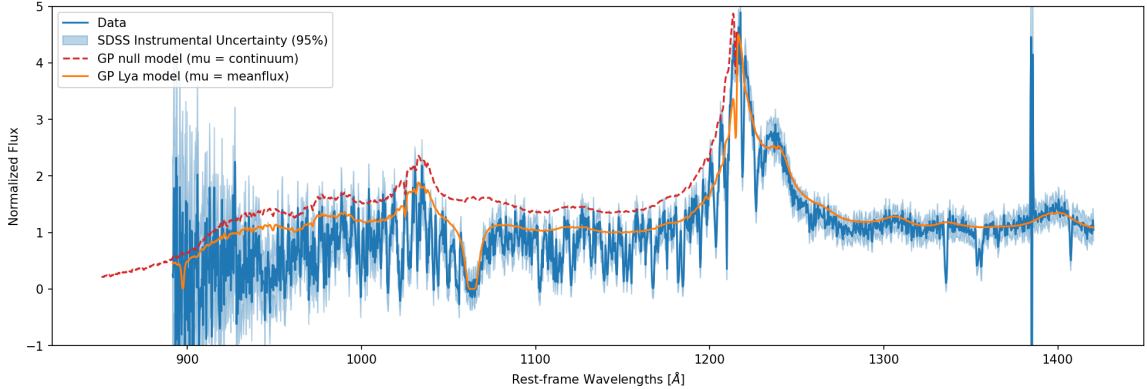


Figure 1.1: A quasar spectrum with a DLA. This spectrum is from the SDSS DR16 quasar catalog, and it shows the Lyman- α forest and the Damped Lyman- α absorbers. The x-axis is the rest-frame wavelength, and the y-axis is the normalized flux. The red line is the continuum fit to the quasar emission spectrum, and the blue line is the observed quasar spectrum. The orange line is the Voigt profile fit to the DLA absorption system. The jupyter notebook tutorial in https://github.com/jibanCat/gpy_dla_detection/blob/master/notebooks/Tutorial%3A%20Automate%20Lyman%20alpha%20Absorption%20Detection.ipynb provides a step-by-step guide on understanding the quasar spectrum and the DLA detection.

discrete number of data points, with the wavelength bins λ and the corresponding flux vector $\mathbf{y} = f(\lambda)$. Observational data always subject to noise, which is usually modeled as a Gaussian noise with the variance $\sigma(\lambda)^2$. To summarize, for each quasar spectrum, we have the following data: $\mathcal{D} = \{\lambda, \mathbf{y}, \sigma(\lambda)\}$.

With the data, now is the matter of how to organize and interpret it. For astronomers, there are a few procedures they usually do with the quasar spectra. These procedures do not change the data itself, but they provide an easier way to interpret it.

- **Normalized Flux:** The flux is usually normalized, so different quasar observations can be compared. One simple way to normalize the flux is to divide the flux by the median flux. Astronomers usually choose a specific absorption-free wavelength range to normalize the flux. Another way to normalize is the continuum-normalized flux,

which is to divide the flux by the continuum model. Continuum-normalized flux is usually used in the absorption line studies when the emission features are not the primary interest.

- **Rest-frame Wavelength:** The observed-frame wavelength, λ_{obs} , is usually converted to the rest-frame wavelength, λ_{rest} , which is the rest-frame of the quasar, with a conversion of cosmic expansion, $\lambda_{\text{obs}} = (1 + z_{\text{QSO}})\lambda_{\text{rest}}$. The benefit of this conversion is that different quasar observations can be compared in the same rest-frame, and they are believed to have similar emission features in the same λ_{rest} .

The above procedures can be thought as normalization procedures for x-axis (λ) and y-axis (f) of the data.

Reading quasar spectrum is not a task we can easily learn from a textbook, it takes time and learning directly from senior practitioners. Here I share my limited knowledge on how to read a quasar spectrum, mostly in the purpose of finding DLAs. There are a few important emission lines to recognize in Figure 1.1. For example, the Lyman- α emission at $\lambda_{\text{rest}} = 1216 \text{ \AA}$ and the Lyman- β emission at $\lambda_{\text{rest}} = 1025 \text{ \AA}$. The above are the hydrogen emission lines. Astronomers usually treat elements heavier than hydrogen and helium as metals. When reading a quasar spectrum, it is useful to find the Lyman- α emission line first. The blueward (left-hand side) of the Lyman- α emission is for hydrogen emissions and the redward (right-hand side) is for metal emissions.¹ Some noteworthy metal lines are SiIV at $\lambda_{\text{rest}} = 1399 \text{ \AA}$, CIV at $\lambda_{\text{rest}} = 1549 \text{ \AA}$, and MgII at $\lambda_{\text{rest}} = 2799 \text{ \AA}$. All of them

¹With some exceptions, such as the OVI emission line at $\lambda_{\text{rest}} = 1035 \text{ \AA}$.

are redward of the Lyman- α emission line. For interested readers, I recommend using the SDSS line table to find more emission lines in the quasar spectrum².

The rest-frame wavelength of the emission lines can be identified from the atomic physics, and usually there are laboratory measurements (on earth) for the wavelengths of the emission lines. The absorption lines, however, are not as easy to identify as the emission lines. Unlike the emission lines, the absorption lines are not from the quasar itself, but from the intervening gas between the quasar and the observer. The absorptions can happen at any location on the line of sight, and usually these gas clouds are not as luminous as the quasar itself. Therefore, the determination of the absorption lines is more challenging than the emission lines because we need to simultaneously know what type of element is in the absorber and the redshift of this absorber. And different types of absorptions from different redshifts can overlap with each other, making it even more challenging.

The rule of thumb for interpreting the absorption lines is that a given type of line only appears at the blueward of the emission line of the same type. For example, the Lyman- α absorption line only appears at the blueward of the Lyman- α emission line at $\lambda_{\text{rest}} = 1216 \text{ \AA}$. Astronomers give a name for the absorptions caused by the Lyman- α line, which is called the Lyman- α forest (see Figure 1.2), falling in the wavelength range of $912 - 1216 \text{ \AA}$, blueward to the Lyman- α emission line. The same rule applies to metal absorptions, such as CIV, SiIV, and MgII. For example, the CIV absorption line only appears at the blueward of the CIV emission line at $\lambda_{\text{rest}} = 1549 \text{ \AA}$. Since the Lyman alpha forest also falls in the wavelength range for the CIV absorptions, it is possible to have a CIV absorption line at the Lyman alpha forest. Usually, high-resolution quasar spectrum

²<https://classic.sdss.org/dr6/algorithms/linestable.php>

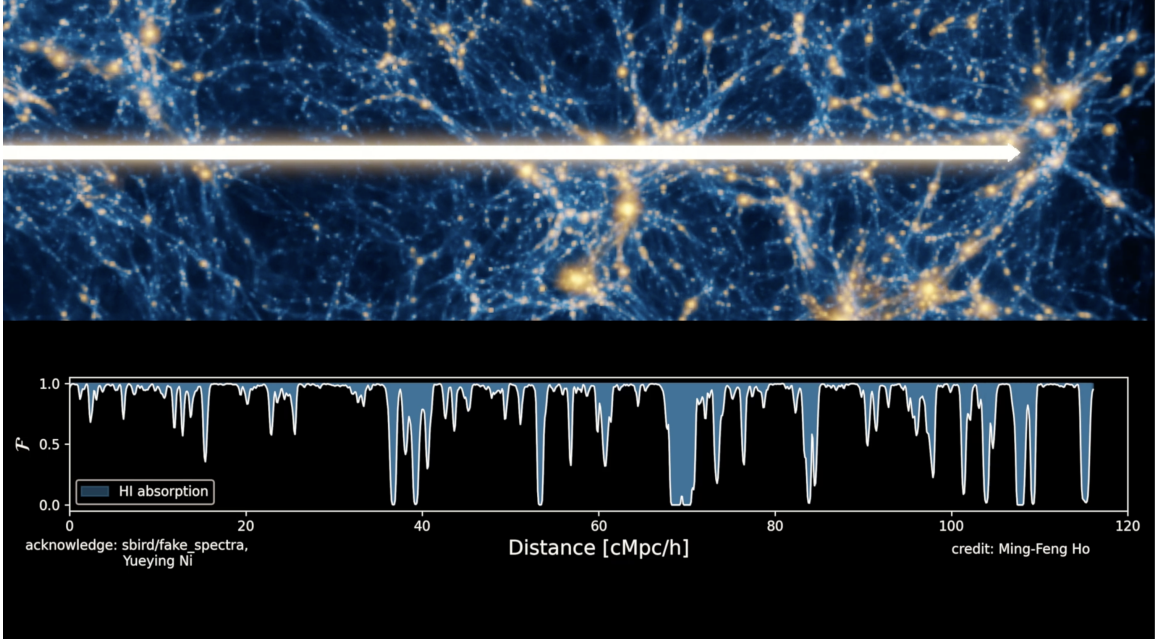


Figure 1.2: The Lyman- α forest, simulated from the neutral hydrogen intergalactic medium from a hydrodynamical simulation. Upper panel: the colors represent the temperature of the gas, and the x-axis is the co-moving box size. The shining arrow represents drawing a quasar sightline through the box. Bottom panel: the shaded blue color represent the absorptions due to the neutral hydrogen in the gas in the sightline. The y-axis is the flux (1: no absorption, 0: fully absorbed), and the x-axis is the comoving distance. YouTube link: <https://youtu.be/xBZLH14Qzzyo?si=Jg08ARJju85Ljt5T>.

is needed to resolve the metals in the Lyman alpha forest. Therefore, at a scale of a cosmological survey such as SDSS and DESI, which heavily rely on low-resolution quasar spectra, it is often very challenging to fully break this degeneracy. A cosmological analysis using Lyman alpha forest therefore usually model metal absorptions as contamination of the forest and marginalize over them as systematics in the Bayesian inference.

A final note I would like to mention is the level of noise in the quasar spectra. Knowing the noise level in different parts of the data is the essence of Bayesian modeling, misinterpreting the noise might lead to interpreting the noise as the signal. Back to Fig-

ure 1.1, one obvious observation is the spectrum at the blueward of the Lyman- α emission line is “noisier” than the redward. I put a quotation for “noiser” because these noisy features are not necessarily noise but mostly from the absorptions of neutral hydrogen. We know that the hydrogen is the most abundant element in the universe, therefore there are more absorption features at $\lambda_{\text{rest}} < 1216 \text{ \AA}$ than $\lambda_{\text{rest}} > 1216 \text{ \AA}$. When these abundant hydrogen lines blending together in the spectrum, they look like noise. Thus, for astronomers interested in the metal absorbers, it is easier to work on the wavelength region redward to the Lyman- α emission line, where the flux is not affected by neutral hydrogen lines.

Finally, for the spectra from Sloan, the bluest end of the optical spectrum, which is around 3800 \AA , is highly affected by the instrumental noise. As seen in Figure 1.1, the flux at the blueward of 3800 \AA , which is $\lambda_{\text{rest}} \sim 908 \text{ \AA}$ for this $z = 3.184$ quasar, is not reliably measured. This bluest part of the spectrum will be redshifted out of the Lyman alpha forest ($912 - 1216 \text{ \AA}$) for quasars at $z > 3.2$, so we need to keep in mind the Lyman alpha forest analysis for $z < 3$ would be affected by this specific instrumental noise.

1.1.2 Method: Dataspace Gaussian Process Inference

After introducing the quasar spectra in the previous subsection, we now narrow our focus to detecting the absorption systems in the quasar spectra. To think this problem in a Bayesian view, the absorption systems are the latent variables in the quasar spectra, and the intrinsic quasar emission function is the background model. The goal is to infer the latent variables from the observed spectrum, based on the emission model and marginalize over the potential systematics in the data, such as other intervening absorbers and the instrumental noise. Schematically, for a given spectroscopic observation, \mathcal{D} , the problem

can be expressed as:

$$P(\text{absorption systems} \mid \mathcal{D}) = \int P(\text{absorption systems} \mid \text{quasar spectrum}, \mathcal{D}) P(\text{quasar spectrum} \mid \mathcal{D}) d\text{quasar spectrum}. \quad (1.1)$$

The $p(\text{quasar spectrum} \mid \mathcal{D})$ is a probability density function (PDF) for the quasar spectrum model, which needs to take into account for both the quasar emission model and data noise model.

The above marginalization is very idealistic, and it is usually not possible to do it in practice as you need to parameterize the quasar systematics, physics of emission, and all possible absorption systems within the search range. The traditional way astronomers do this is to use the template fitting method, ideally this should model the emission physics. The search of absorption systems is then empirically done by chi-squared fitting with Voigt profiles, with human intervention to decide the significance of the absorption systems. All these steps are not Bayesian, and the results are usually not quantified in a probabilistic way.

As a Bayesian, the first thing we might want to quantify is the human intervention on the detection of the absorption systems. In the Bayesian context, this decision making is a model selection problem, where we need to compare the model with the absorption systems and the model without the absorption systems. Suppose \mathcal{M}_+ is the model with the absorption systems, and \mathcal{M}_- is the model without the absorption systems, the probability of the absorption systems given the quasar spectrum (\mathcal{D}) is:

$$P(\mathcal{M}_+ \mid \mathcal{D}) = \frac{P(\mathcal{M}_+)P(\mathcal{D} \mid \mathcal{M}_+)}{P(\mathcal{M}_+)P(\mathcal{D} \mid \mathcal{M}_+) + P(\mathcal{M}_-)P(\mathcal{D} \mid \mathcal{M}_-)}. \quad (1.2)$$

1.1.3 Application: Damped Lyman alpha Absorbers

In observational astronomy, one of the famous absorption lines is the Damped Lyman- α absorbers, or DLAs (see Figure 1.3). DLAs are strong HI absorption lines found in Lyman- α forest. They are optically thick, and their column density is high enough to cause self-shielding, appearing as damping wings on the spectrum. They are interesting objects to study because DLAs trace the neutral hydrogen gas surrounding the small galaxies at $z \sim 3$, and these neutral gases eventually falling into the gravitational well of the galaxies and fuel the galaxy formation activities (see Ref. [18]). Probing DLAs thus traces these small galaxies, which are not luminous enough to be visible through the emission-line observations. Counting and recording the neutral hydrogen in DLAs also tell us about the co-evolution between galaxies and the neutral hydrogen reservoir surrounding them. Observational-wise, the damping wings of DLAs provide accurate measurement of column density compared to depending on lines. This enables better measurement of the metallicity of the gas, which is important for understanding how galaxies deposit or recycle the metals to the surrounding gas. Therefore, DLAs are no-doubt useful tools for studying the galaxy evolution.

On the other hand, cosmology-wise, the damping wings of DLAs introduce a systematic effect in the Lyman- α forest analysis. As shown in Figure 1.3, the damping wings of DLAs are large enough to affect the absorption features not in the same pixel but in a wide range of neighboring pixels. This causes a problem in measuring the neutral hydrogen clustering in the Lyman- α forest, as the damping wings artificially add the scale-dependent absorptions to the forest. In addition, the fully absorbed flux in DLA also wash out the small-scale structures in the Lyman- α forest, which suppresses the small-scale power in the

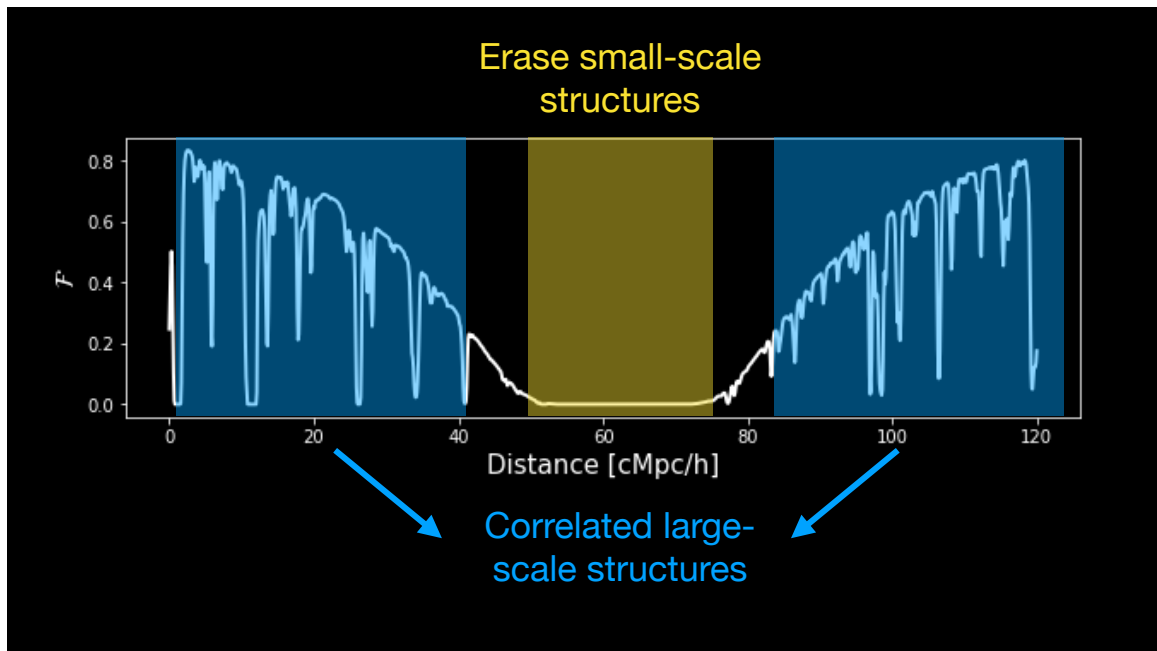


Figure 1.3: A Damped Lyman- α absorber found in a hydrodynamical simulation. The y-axis is the normalized flux (1 means un-absorbed flux and 0 means fully absorbed flux), and the x-axis is the co-moving box size. In contrast to Figure 1.2, the Lyman alpha forest is washed out in the yellow region, and the damping wings bias the flux in the blue region.

flux power spectrum. The impact of DLAs on the Lyman- α 1D power spectrum can be found in Ref. [19].

The current way to deal with DLAs in the Lyman- α forest analysis is to mask out the pixels around the DLAs, therefore, the measurement of two-point correlation function in the Lyman- α forest ($N_{\text{HI}} < 10^{17} \text{ cm}^{-2}$) is not affected by the DLAs. The potential contamination from the error in the DLA masking is usually marginalized as a systematic error in the Bayesian inference. Historically, the search of DLAs is done by human eyes (e.g., DLAs in SDSS DR3, Ref. [20]), which is still a reliable way to find DLAs in some high-resolution surveys when the number of quasar spectra is sparse. However, there are 180,000 quasar spectra in the SDSS-III, and the expected number of quasar spectra in the DESI survey is over millions. It is not only impractical to search DLAs by hands, but also we need a systematic way to access the impact of DLAs on the cosmological analyses. To achieve this, the DLA search needs to be automated and be applied to many mock data to quantify DLA's systematical effects.

Our GP-DLA finder (see Sec 2 and Sec 3, or Ref. [21, 15, 8]) is designed to automate the DLA search in large spectroscopic surveys. The GP-DLA finder is a Bayesian model selection method to classify the absorption systems using Gaussian processes. The underlying principle of this GP-DLA finder is the Bayes rule. The probability of finding DLAs in a given quasar spectrum, \mathcal{D} , is:

$$P(\mathcal{M}_{\text{DLA}}|\mathcal{D}) = \frac{P(\mathcal{M}_{\text{DLA}})P(\mathcal{D} | \mathcal{M}_{\text{DLA}})}{P(\mathcal{M}_{\text{DLA}})P(\mathcal{D} | \mathcal{M}_{\text{DLA}}) + P(\mathcal{M}_{-\text{DLA}})P(\mathcal{D} | \mathcal{M}_{-\text{DLA}})}. \quad (1.3)$$

The \mathcal{M}_{DLA} is the model with the DLA, and $\mathcal{M}_{-\text{DLA}}$ is the model without the DLA. $P(\mathcal{D} | \mathcal{M}_{-\text{DLA}})$ is the probability of the data given the model without the DLA, in other words,

this is our model to describe the emission spectrum. GP-DLA uses a Gaussian process to empirically learn the emission model from the data, so,

$$P(\mathcal{D} | \mathcal{M}_{\text{-DLA}}) = \mathcal{GP}(\mu(\lambda), K(\lambda, \lambda')), \quad (1.4)$$

each quasar spectrum can be thought as a realization from this learned Gaussian process. A Gaussian process is a generalization of a multi-variate Gaussian distribution to infinite dimensions, and it is a nature way to model the continuous function such as 1-D spectroscopic data. $\mu(\lambda)$ describes the mean function and $K(\lambda, \lambda')$ describes the covariance function of the Gaussian process. λ denotes that this GP is in the wavelength space.

The DLA model, \mathcal{M}_{DLA} , is a spectrum model with the DLA absorption feature, mathematically, it is a Voigt profile convolved with the learned Gaussian process. Fortunately, applying a multiplication on a Gaussian process is still a Gaussian process, so the DLA model can be thought as a realization from a Gaussian process with slightly different mean and covariance functions,

$$P(\mathcal{D} | \mathcal{M}_{\text{DLA}}) = \mathcal{GP}(\mu(\lambda) \cdot \text{Voigt}(\lambda), \text{Voigt}(\lambda)^\top K(\lambda, \lambda') \text{Voigt}(\lambda')). \quad (1.5)$$

Detailed mathematics can be found in Sec 2. Below, I only provide a hand-waving explanation of how the GP-DLA finder works.

For each incoming quasar spectrum, the GP-DLA finder first uses $\mathcal{M}_{\text{-DLA}}$ to calculate the probability of the data given the learned emission model. Then, the GP-DLA finder uses \mathcal{M}_{DLA} to calculate the probability of the data having at least one DLA. These two probabilities then are compared using Eq. 1.3, and then return a probability of the data having at least one DLA. A deterministic threshold is set to decide whether the quasar

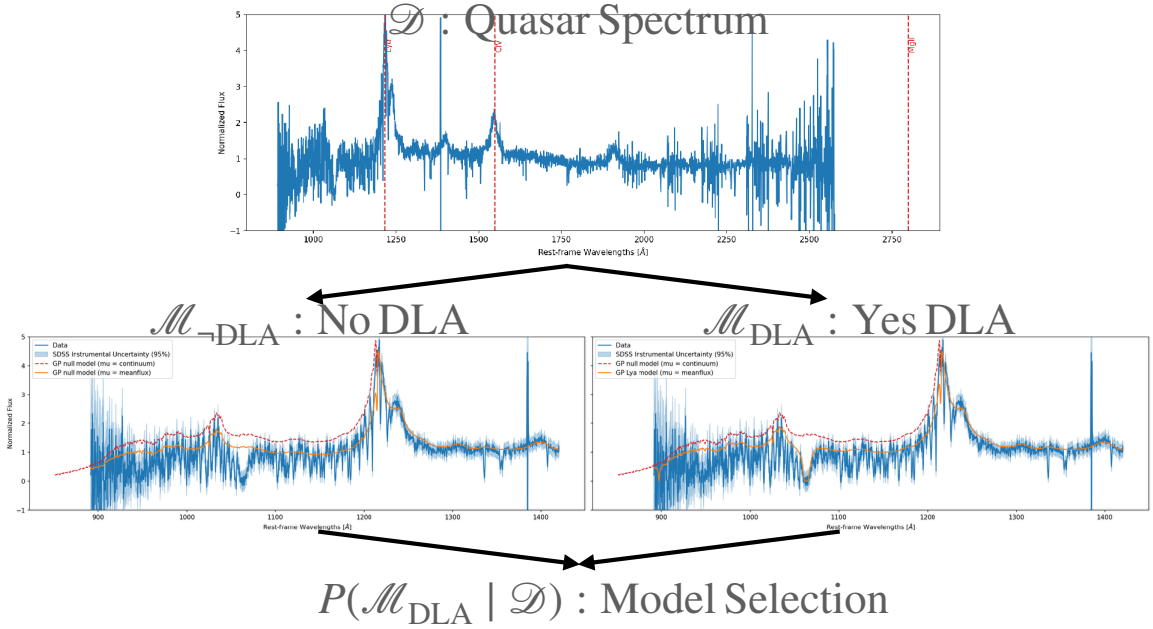


Figure 1.4: How GP-DLA works. The GP-DLA finder uses two models to calculate the probability of the data having at least one DLA. A tutorial of GP-DLA can be found in https://github.com/jibanCat/gpy_dla_detection

spectrum has a DLA or not. A diagram of how the GP-DLA finder works is shown in Figure 1.4.

GP-DLA as a Bayesian method, it allows us to incorporate our prior knowledge to the calculation. For example, we can set a prior on the number of DLAs in the quasar spectrum as a function of redshift, or we can set a prior on the column density of the HI absorptions. Both of these priors could be subjective prior, which can be set by the experts on finding DLAs, or empirical prior, which can be learned from the data or in the literature.

The performance of GP-DLA on low-resolution spectra from large surveys, such as SDSS, has been shown having a high accuracy and a low false positive rate. Sec 2 and Sec 3 show the applications of GP-DLA on the SDSS DR12 and DR16 quasar spectra, respectively. Figure 1.5 shows examples of DLAs found in the SDSS DR16 quasar spectra

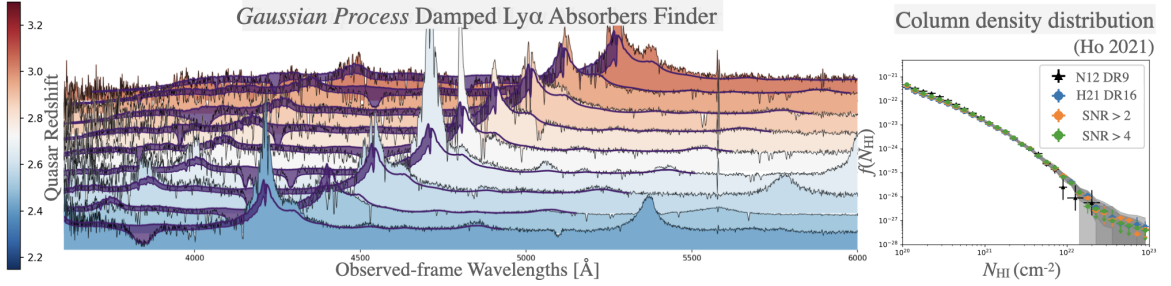


Figure 1.5: DLAs found in the SDSS DR16 quasar spectra, and the corresponding column density distribution function.

using GP-DLA, and the robustness of the constraints on the column density distribution function of DLAs. GP-DLA technique has been adopted in SDSS-IV for the Baryonic Acoustic Oscillation (BAO) analysis [22], and it has been further re-developed and applied to the DESI quasar spectra [23, 24].

The GP-DLA finder is not limited to the DLA search, it can be applied to any absorption systems in the quasar spectra. For example, in Ref. [25], led by Dr. Reza Monadi, a former student in UCR, we applied the same Gaussian process and model selection technique to search for the CIV absorbers in the quasar spectra, one of the common metal absorbers in the quasar spectra. In Ref. [1], led by Leah Fauber, a former computer science student in UCR, we extended the Gaussian process quasar spectrum model to infer the quasar redshift, and it shows a competitive performance compared to the other quasar redshift estimation methods (See Figure 1.6).

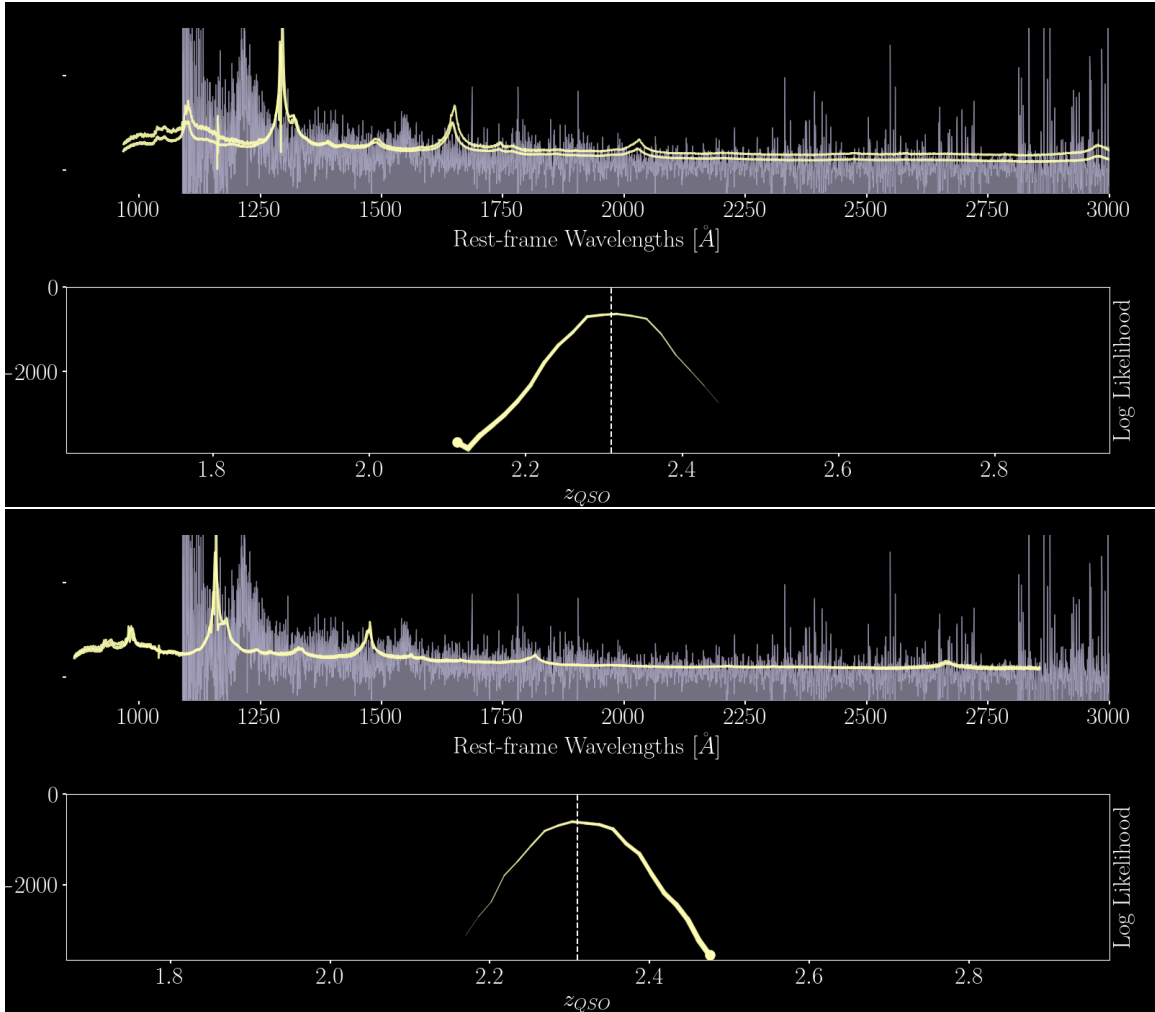



Figure 1.6: A demonstration of the usage of GP redshift estimator from Ref. [1]. This animation is on YouTube:  <https://youtu.be/NhUycNaHBzM?si=cRSZhtTKWFJ6mlir>

1.2 Multi-Fidelity Emulators for Cosmological Simulations

As mentioned in the previous section, the Lyman- α forest is a unique probe to study the large-scale structure of the Universe. Ly α forest probes the structures at the redshift range of $2 < z < 5$, covering a wide range of scales from ~ 100 Mpc to ~ 1 Mpc. On the small scales (~ 1 Mpc), correlating the neutral hydrogen absorptions within a single quasar spectrum (sightline) can be used to measure the power spectrum of the matter density fluctuations. On the large scales (~ 100 Mpc), cross-correlating different quasar sightlines can be used to measure the 3-dimensional power spectrum, constraining the Baryonic Acoustic Oscillations (BAO), which is the imprint of the sound wave in the early universe. These two measurements are complementary to each other. As shown in Figure 1.7, the first one is the Ly α 1-dimensional power spectrum (flux power spectrum, FPS, or P1D), and the second one is the Ly α 3-dimensional power spectrum (P3D).

However, how do we compare the Ly α forest observations to the dark matter's large-scale structure power spectrum? A classic way to predict the large-scale structures is to use the linear perturbation theory, which model the growth of the matter density fluctuations from the early universe to current time. Linear perturbation theory is a good approximation until $k \sim 0.1 h/\text{Mpc}$ (or ~ 10 Mpc; see Ref. [26]), but it is not accurate enough to predict the power spectrum on the non-linear small scales. Ly α P1D probes the scales around $k \sim 1 h/\text{Mpc}$, and around $k \sim 10 h/\text{Mpc}$ when using high-resolution quasar spectra (see Figure 1.8), so the linear theory is apparently not enough. Something more accurate is needed.

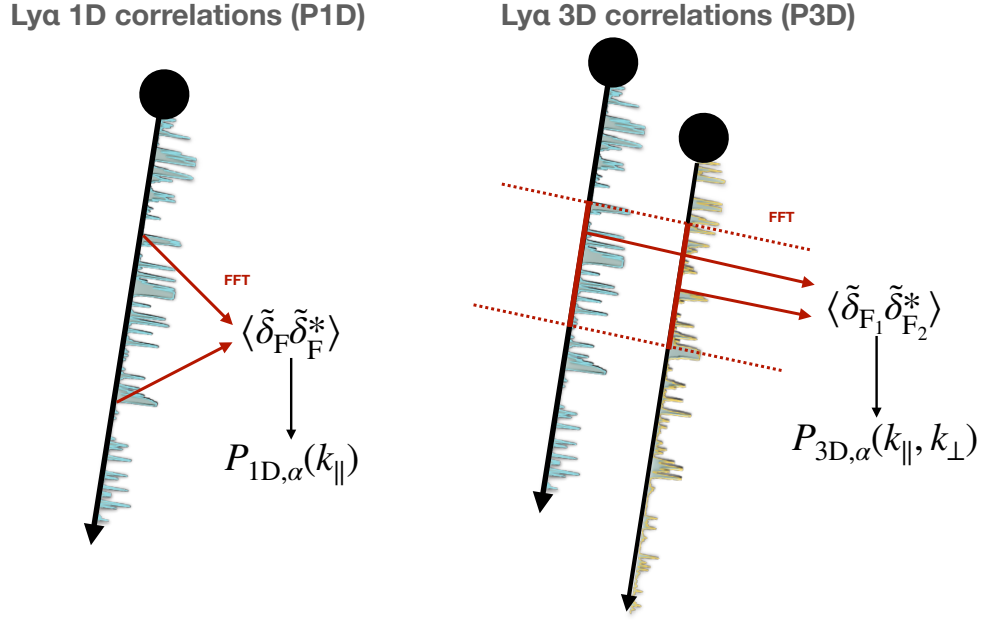


Figure 1.7: The Ly α 1D (P1D) and 3D (P3D) power spectra. P1D is measured from the auto-correlation of the flux in a single quasar spectrum, and P3D is measured from the cross-correlation of the flux in different quasar spectra. The spikes in the diagram show the Lyman alpha forest, and the black arrows show the quasar sightlines.

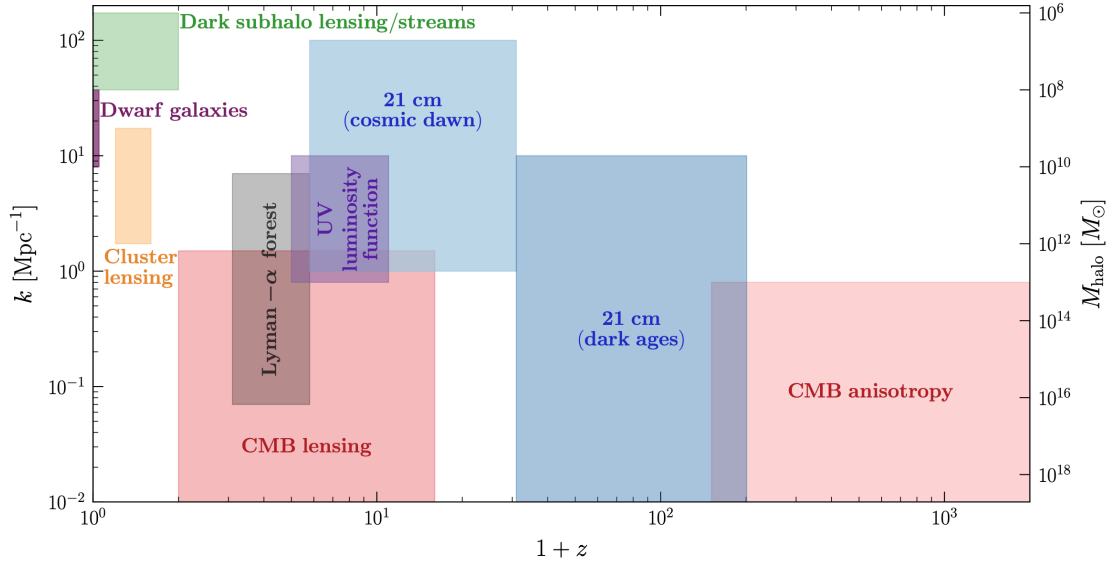


Figure 1.8: A schematic representation from Ref [2], showing the current and future probes of the structure formation, across different redshifts z and scales k . Lyman alpha forest probes the structure formation at $2 \leq z \leq 5$ and $k \sim 0.1 - 10 \text{ Mpc}^{-1}$.

N -body simulations emerged as a powerful tool to predict the structure formation at small scales at lower redshifts. The “ N -body” here means the mutual gravitational interaction between N particles within a simulation box. What N -body simulations do is to solve the non-linear evolution of the matter density fluctuations from the early Universe to the current time. Linear theory is mostly accurate at the early Universe because the non-linear structures still do not have time to form yet. Cosmologists therefore use the linear theory to generate the initial conditions for the N -body simulations at a high enough redshift (e.g., $z = 100$ for most of the simulations in this thesis), and then allow the structures gravitationally evolve to the current time. Pure dark matter N -body simulations are believed to be accurate up to $k \sim 0.5h/\text{Mpc}$ (see Ref. [26]), beyond that, the baryonic effects, such as the feedback from the active galactic nuclei (AGN), need to be included to make the power spectrum prediction accurate.

Unfortunately, N -body simulations are slow. For a simulation box of 1 Gpc^3 with 3000^3 particles, a simulation designed to resolve scales for the current Euclid survey, it takes about 2 000 node hours per simulation on a GPU accelerated supercomputer (see Ref. [11]). In comparison, computing a power spectrum using the linear perturbation theory takes only a few seconds on a laptop. There is a huge computational gap between these two methods, so it means that there are lots of traditional analyses done by the linear theory cannot be directly replaced by the N -body simulations. One of them is the Bayesian inference of the cosmological parameters, which requires re-computing the power spectrum for $> 10^6$ times.

The concept of *emulators* emerged around the 2010s in the cosmology community (see Ref. [27, 28, 29]), primarily due to the surrogate modeling techniques and the devel-

opment of the Gaussian process in the machine learning and statistics community around early 2000s (see Ref. [30, 31]). The idea of the emulator is to build a fast surrogate model to predict the slow simulation results. The underlying principle is simple: take a suite of simulations with different input parameters, and then build a regression model to predict the simulation results as a function of the input parameters. The simulation results here have to be something easy to interpolate in a low-dimensional space, such as the power spectrum, the correlation function, or the bispectrum. The creation of the emulator is not to replace the role of simulations, rather, it is designed to expand the usage of the simulations to a broader range of applications. For example, simulation sensitivity checks (varying 1 input parameter at a time), optimization (finding the best-fit input parameter to the data), model selection (comparing the goodness-of-fit across different simulation codes), and Bayesian inference (quantify the posterior of parameters). This is a powerful tool to bridge the gap between the slow simulations and the fast cosmological data analysis. After around a decade, the emulator approach has become a standard tool in the cosmology community on interpolating the power spectrum from the pure dark matter N -body simulations (until $k \sim 10 h/\text{Mpc}$).

Until this point, we have been talking about the power spectrum from the “pure dark matter” N -body simulations. However, the Ly α forest is a very special probe, to properly predict the Ly α forest, we need to include gas physics in the simulations, which is best done by the hydrodynamical N -body simulations. Each hydrodynamical simulation is more expensive than the pure dark matter simulation. Moreover, unlike the pure dark matter N -body simulations, unfortunately, the astrophysical feedbacks in the hydrodynamical

simulations are not well understood, so different simulation codes do not usually converge to the same answer. These difficulties make the cosmological emulator approach more challenging for the hydrodynamical simulations than the pure dark matter simulations, as the cosmological analysis usually requires to be at a percent level of accuracy.

There are various ways to attack this problem. One way is the baryonification approach, where the pure dark matter N -body simulations are used to post-process and reposition the particles to mimic the baryonic effects according to some physically-motivated phenomenological models (see Ref. [32]). Baryonification method in Ref. [32] is shown to be able to be calibrated to the summary statistics (primarily at the power spectrum and bi-spectrum level) of various hydrodynamical simulations. In a similar vein, Fluctuating Gunn-Peterson Approximation (FGPA) (see Ref. [33, 34]) is a method to post-process the pure dark matter simulations to generate mock Ly α forest spectra, and it is shown to be able to reproduce the Ly α forest produced by the hydrodynamical simulations. The benefit of these post-processing approaches is that it is computationally cheap, as post-processing the pure dark matter simulations is much faster than running a new hydrodynamical simulation.

Another way is to directly emulate the hydrodynamical simulations by running a suite of hydrodynamical simulations with different astrophysical feedback parameters. By varying the feedback parameters, the emulator hopefully can capture all possible outcomes of the feedback effects and help us understand the correlation between the feedback models and the cosmological signals. The benefit of this approach, compared to the baryonification approach, is that we have better physical intuitions on the feedback models because these models are usually designed to match the observations of the galaxy or AGN. Thus, the

feedback parameters, though are considered to be nuisance in the cosmological analysis, are physically meaningful and useful in the galaxy formation studies. The downside of this approach is that it is computationally expensive, as there are way more feedback parameters than the cosmological parameters (see Ref. [35], where 28 subgrid model parameters are used), and each simulation is more expensive than the pure dark matter simulation. Moreover, the issue of the convergence between different hydrodynamical simulation codes is still unsolved. The multi-simulation campaign such as CAMELS [36] is designed to address this issue, by running the same input parameter dimension with different hydrodynamical simulation codes and different feedback models.

In this section, I will briefly recap the concept of emulation and the idea of multi-fidelity emulator. Multi-fidelity emulation is a technique to combine the predictions from different fidelity models to improve the prediction accuracy. Section 1.2.1 will provide a high-level overview of the cosmological emulator. Section 1.2.2 will provide an overview of the multi-fidelity methods. Section 1.2.3 will show the application of the multi-fidelity emulator in the cosmological emulation. Detailed mathematics and the application will be shown in the Chapter 4 and Chapter 5.

1.2.1 Data: Emulation

Why do we need emulators anyway?

In Bayesian modeling, we usually have a statistical model to describe the data. To assess how well the model fits the data, we need to compare the model’s predictions to the actual data. To do this, we can “simulate” “mock” data from the statistical model and

then compare the simulated data to the real data. If the simulated data is close to the real data, then the model is a good fit.

We can vary the parameters of this statistical model, repeat the simulations millions of times, and find the best-fit parameters that can generate the data. This is exactly what Markov Chain Monte Carlo (MCMC) simulations do in Bayesian inference. When people talk about “simulations” in the context of statistics, they usually mean simulations from the statistical model, i.e., something easy to compute, fast to run, and used to better understand the data. While obtaining data can be expensive, simulations are usually straightforward and computationally inexpensive.

However, in cosmology, the “simulations” usually mean the N -body simulations or the hydrodynamical simulations, which are slow to run and expensive to compute. In addition, we only have one realization of the Universe, so we only have one data point. Therefore, in cosmological analysis, not only data are sparse and expensive to obtain, but the simulations are also slow and expensive to run. When both data and simulations are expensive, we need one of them to be cheap to obtain, otherwise, data analysis is not possible.

This is where the emulator comes in. The emulator approach treats the simulations as “data” and fits a flexible statistical model to predict the simulation results as a function of the input parameters. With this fitted statistical model, we can easily generate the “mock simulations” with a cheaper computational cost. This fitted model acts as a surrogate for the simulations, and it is often called the “surrogate model” or the “emulator.”

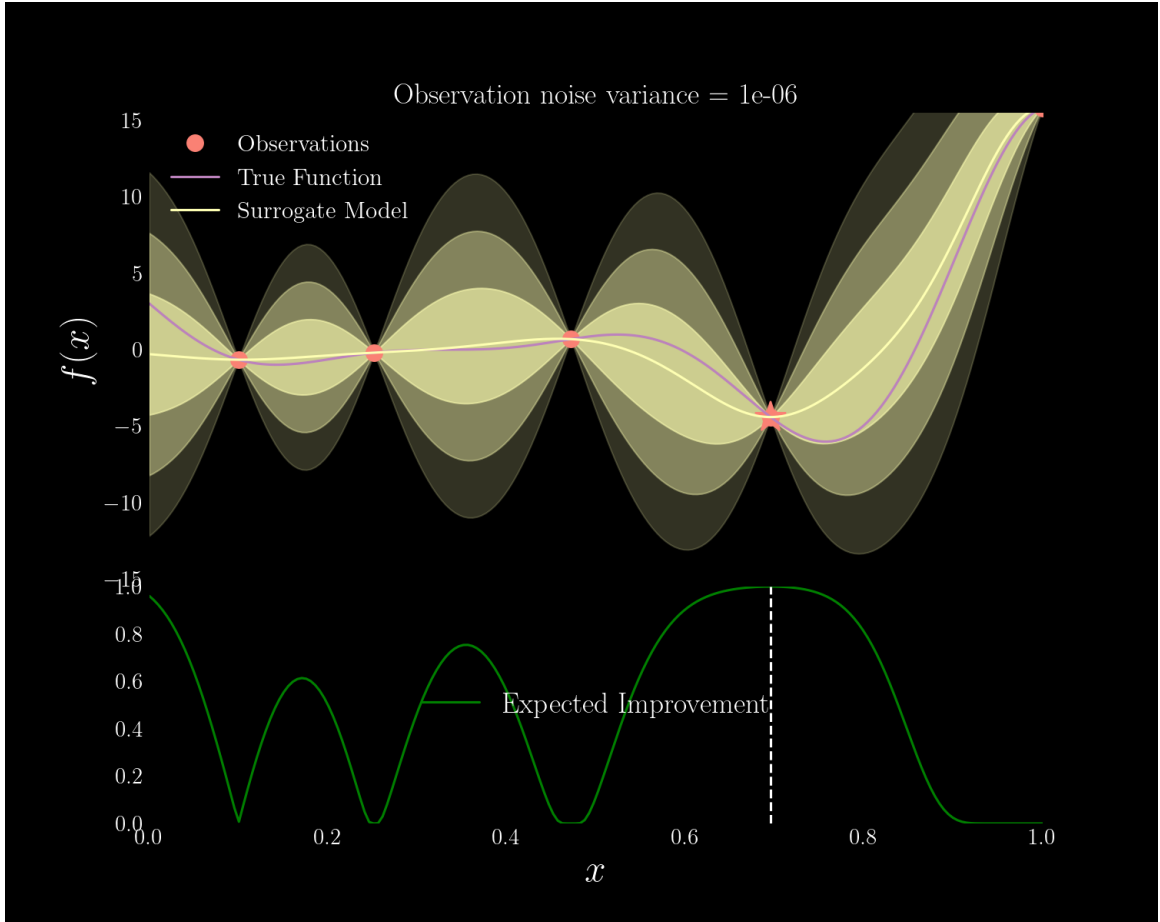


Figure 1.9: Bayesian surrogate modeling of the Forrester function. This is a demonstration of the Gaussian process emulator in combination of Bayesian optimization on the Forrester function. The red dots are the simulations we have run (the red stars being the last simulation we run), and the purple curve is the true simulation function. The emulator prediction is shown as the yellow curve, with the shaded areas showing the (1,2,3)- σ confidence intervals. The lower panel shows the Expected Improvement (EI) function (i.e., the acquisition function), which is used to find the next best point to evaluate the function in the Bayesian optimization. A higher EI value means the point is more likely to better improve the surrogate model fitting. Video tutorial can be found in <https://youtu.be/6JmuqVhSq5Y?si=5TFzIbepU6iCXRno>.

Figure 1.9 shows an example of fitting an emulator on the Forrester function, $f(x) = (6x - 2)^2 \sin(12x - 4)$. This is a classic example to demonstrate the emulator approach. Here, we have a function $f(x)$, which we assume to be the *true simulation function* (purple curve), and we assume $f(x)$ is expensive to compute. Since it is expensive, we can only evaluate the simulation function at a few input x values, which are shown as the red points in Figure 1.9. These red points are the *simulations we have run*, or the “data” we have. The goal of the emulator is to predict the function $f(x)$ at any input x values, given the few data points we already have.

A standard way to build the emulator is to use the Gaussian process (GP) regression. We assume the function $f(x)$ is a realization from a Gaussian process,

$$f(x) \sim \mathcal{GP}(\mu(x), K(x, x')), \quad (1.6)$$

where $\mu(x)$ is the mean function and $K(x, x')$ is the covariance function of the Gaussian process. The usage of Eq. 1.6 is slightly different from the Gaussian process we used in the DLA search. In the DLA search, we directly learn the covariance function from the quasar spectra because there is no simple analytical form for the quasar emission covariance. In the emulator, we usually assume the simulation function, $f(x)$, is smooth and continuous, so we can use a simple covariance function, such as the squared exponential kernel,

$$K(x, x') = \sigma^2 \exp\left(-\frac{(x - x')^2}{2l^2}\right), \quad (1.7)$$

where σ^2 is the variance and l is the length scale of the kernel. The hyperparameters σ^2 and l are usually learned from the data (the red points) by maximizing the marginal likelihood of the Gaussian process. The emulator prediction is shown as the yellow curve in Figure 1.9.

The smoothness of the emulator is controlled by the length scale l of the kernel, where a smaller l means the emulator is more flexible to the data, and a larger l means the emulator is smoother. Given a suite of simulations, $\mathcal{D} = \{\mathbf{x}, f(\mathbf{x})\}$, the emulator prediction at any input x is a Gaussian distribution,

$$p(f | \mathcal{D}) = \mathcal{GP}(f; \mu_{\mathcal{D}}, K_{\mathcal{D}}), \quad (1.8)$$

where

$$\begin{aligned} \mu_{\mathcal{D}}(x^*) &= \mu(x^*) + K(x^*, \mathbf{x})K(\mathbf{x}, \mathbf{x})^{-1}(f(\mathbf{x}) - \mu(\mathbf{x})), \\ K_{\mathcal{D}}(x^*, x^{*'}) &= K(x^*, x^{*'}) - K(x^*, \mathbf{x})K(\mathbf{x}, \mathbf{x})^{-1}K(\mathbf{x}, x^{*'}). \end{aligned} \quad (1.9)$$

Here, the emulator prediction is the posterior mean function $\mu_{\mathcal{D}}(x^*)$ at the new input x^* , and the uncertainty of the prediction is the variance of the Gaussian process at the input x^* , which is the diagonal of the covariance matrix $K_{\mathcal{D}}(x^*, x^*)$.

The smoothness assumption sounds like a strong assumption, but it is usually a good assumption in terms of simulations. For a simulation code, we usually expect the neighboring input parameters to have similar simulation results, so the simulation function is usually smooth and continuous. For example, the power spectrum from the N -body simulations is usually a smooth and continuous function of the cosmological parameters, so the emulator can predict the power spectrum at any cosmological parameters within the range of the simulations.

With the smoothness assumption, a common experimental design for preparing the simulations is the Latin hypercube sampling (LHS). LHS is a space-filling design, which ensures the input parameters are evenly distributed in the input space, even in the high-dimensional space. Detailed explanation of the LHS will be described in Chapter 4 and

Chapter 5. To gain the intuition of the reasoning behind using LHS, let's consider the Forrester function example in Figure 1.9. With the assumption that the underlying true function is smooth, the intuitive way to prepare the simulations is to sample the input parameters evenly in the input space, x . When the input parameters are evenly distributed, the red points in Figure 1.9 are more likely to cover the whole input space. Extending this to the high-dimensional space is the LHS, which ensures the design of simulation suite represent the real variability of the input parameters and better capture the underlying function.

In most of the cosmological emulator applications the response function of the power spectrum is a smooth function of the input cosmology. However, there might be some concerns about how emulator can handle the not-very-smooth functions. For example, if there is a sharp drop or peak in the $f(x)$, the evenly distributed design might not be the best design for the emulator. In this case, emulator requires more data points at the input x values where the sharp drop or peak happens. For example, in Figure 1.9, we might need more data points around $x \sim 7$ to better predict the sharp drop and rise of the function. However, we often do not know where the sharp drop or peak is beforehand, so we need to adaptively sample the input space to better capture the sharp features of the function. This is where the Bayesian optimization comes in.

Bayesian optimization is a sequential design strategy for global optimization of expensive-to-evaluate functions. In the context of the emulator, Bayesian optimization can be used to find the best next input x to evaluate the simulation function, $f(x)$, to better improve the emulator prediction. When we talk about emulator, we cannot avoid

talking about the Bayesian optimization, as the Bayesian optimization and Gaussian process emulation are often used together. However, the Bayesian optimization will not be covered in the later chapters, so I will only provide a hand-waving demonstration below. Interested readers can refer to Ref. [37] for a detailed explanation of the Bayesian optimization.

Figure 1.10 shows an example of the Bayesian optimization on the Forrester function. In comparison to Figure 1.9, the emulator prediction (yellow curve) in Figure 1.10 is more accurate. This is because the Bayesian optimization adaptively allocate more data points around $x \sim 7$ to better capture the sharp drop and rise of the function. The heart of this optimization is the acquisition function, which is a function to balance the “exploration” and “exploitation” of the input space. This is so-called the “multi-armed bandit” problem, or the “explore-exploit” dilemma. We decide between allocating resources to an input x where the emulator is uncertain (exploration) for improving future rewards, or allocating resources to an input x where the emulator is certain (exploitation) for immediate high rewards. In another word, we need to balance between the “exploitation” of a suspected local maximum and the “exploration” of the new input space to find the global maximum.

The above description sounds very familiar because exploration-exploitation is a common problem in everyday life. For example, when we go for lunch, we need to decide whether to go to the same restaurant we always go to (exploitation) or to try a new restaurant (exploration). The formal definition can be found in Ref. [37]. To view the Bayesian optimization in action, I have made a tutorial video on the Forrester function, which can be found in the video link in the caption of Figure 1.10, where the “Expected Improvement” is the acquisition function used in the optimization.

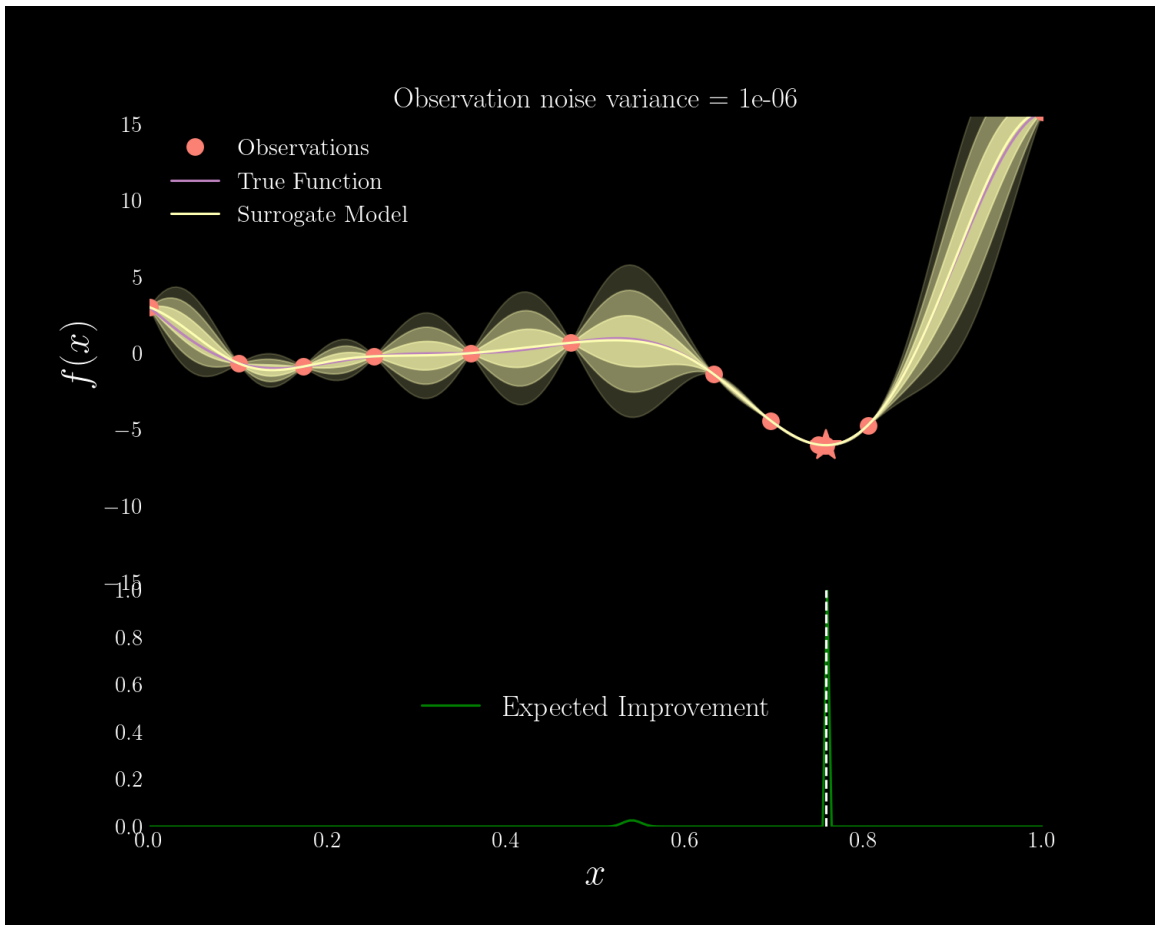


Figure 1.10: Bayesian optimization on the Forrester function. This is a demonstration of the Gaussian process emulator in combination of Bayesian optimization on the Forrester function. With a few iterations of Bayesian optimization, the parameters are adaptively allocated around $x \sim 7$ to better capture the sharp drop and rise of the function, and the parameter space of $x < 0.6$ is mostly evenly sampled. Video tutorial can be found in <https://youtu.be/6JmuqVhSq5Y?si=5TFzIbepU6iCXRno>.

1.2.2 Method: Bayesian Multi-fidelity Emulation

In previous section, we have discussed the situation the simulation function is expensive to compute, and the emulator is used to predict the simulation function at any input parameters, so we can easily generate the “mock simulations” for downstream data analysis. However, what if the target function is not only expensive to compute but almost impossible to compute? In the astrophysics community, this is a common situation. For example, cosmological galaxy formation simulations, such as the IllustrisTNG [38], ASTRID [39], or FLAMINGO [40], these simulations have to resolve the large-scale structure and also resolve the small-scale galaxy physics. First, the simulation volume is always not large enough to match the size of the Universe. Also, the resolution is always not high enough to resolve the small-scale physics such as supernovae or black holes. The simulation is always a compromise between the volume and the resolution. Thus, we know the simulation is always not perfect, and the true simulation function is almost not computable.

Nevertheless, something we know is that the simulation is not completely wrong, and, to some extent, we know the simulation is better when the resolution is higher and the volume is larger.³ This is where the multi-fidelity emulation comes in. Usually, simulation codes have different fidelities, where the high-fidelity simulation (high-resolution) is more accurate but more expensive, and the low-fidelity simulation (low-resolution) is less accurate but cheaper. The goal of the multi-fidelity emulation is to build an emulator for the high-fidelity simulation, a fidelity that is *impossible to compute enough data points to build an*

³This is mostly correct when all the physics in the simulation are from the first principle (e.g., N -body simulations), and the simulation is not tuned to match the observations. When empirical models are used in the simulation to tuned to match the observations at a certain resolution (e.g., subgrid models), the fidelity of the simulation is not necessarily increasing with the resolution.

emulator, by using the data from the low-fidelity simulation, a fidelity that is computable for many data points.

A formal definition of the multi-fidelity emulation is as follows. Given a suite of simulations from various fidelities, $\mathcal{D} = \{\mathbf{x}, \{f_t(\mathbf{x})\}_{t=1}^T\}$, where t is the fidelity index, the goal of the multi-fidelity emulation is to predict the simulation function at the high fidelity, $f_T(\mathbf{x})$, given the data from the lower fidelities, $f_t(\mathbf{x})$, where $T \geq t$ and $t = 1, 2, \dots, T$. Assuming the simulation function at a fidelity t is a realization from a Gaussian process,

$$f_t(\mathbf{x}) \sim \mathcal{GP}(\mu_t(\mathbf{x}), K_t(\mathbf{x}, \mathbf{x}')). \quad (1.10)$$

The multi-fidelity emulation models the high-fidelity simulation function as a function of the low-fidelity simulation functions,

$$f_t(\mathbf{x}) = g(\mathbf{x}, f_{t-1}(\mathbf{x})), \quad (1.11)$$

where $g(\mathbf{x}, f_{t-1}(\mathbf{x}))$ is also a Gaussian process. Function $g(\mathbf{x}, f_{t-1}(\mathbf{x}))$ can be thought as a fidelity correction function, which corrects the low-fidelity simulation function to predict the high-fidelity simulation function.

One of the most common way to model the fidelity correction function is to use the linear auto-regressive model (Ref. [41]), which separate the correction between the low-fidelity simulation function and the high-fidelity simulation function into two parts: the linear scaling part and the bias part,

$$f_t(\mathbf{x}) = \rho_{t-1} f_{t-1}(\mathbf{x}) + \delta_{t-1}(\mathbf{x}), \quad (1.12)$$

where ρ_{t-1} is the scaling factor and $\delta_{t-1}(\mathbf{x})$ is the bias term. ρ_{t-1} is a constant, and when $\rho_{t-1} = 1$ and $\delta_{t-1}(\mathbf{x}) = 0$, the low-fidelity simulation function is the same as the high-fidelity

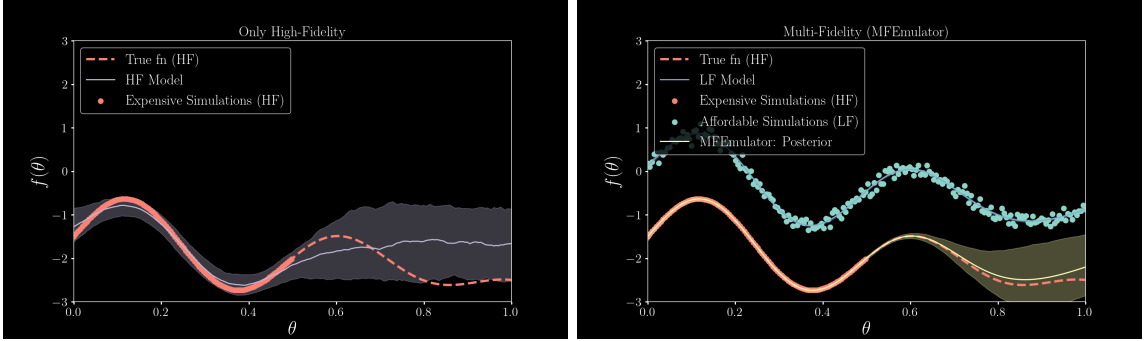


Figure 1.11: An example of the multi-fidelity emulator on the simple function. Red dots as the high-fidelity data, blue dots as the low-fidelity data. Red dashed curve is the high-fidelity true function and blue curve is the low-fidelity true function. Left panel shows the emulator prediction using only the high-fidelity data (purple curve), which is not accurate at $x > 0.5$. Right panel shows the emulator prediction using only the low-fidelity data (blue curve), which is biased, and the multi-fidelity emulator prediction (yellow curve) is more accurate. YouTube video tutorial can be found in <https://youtu.be/tQIytDnW0zk?si=TaKfDkJnhE48yHvK>.

simulation function. $\delta_{t-1}(\mathbf{x})$ is also a Gaussian process. Both ρ_{t-1} and $\delta_{t-1}(\mathbf{x})$ are learned from the data. The linear multi-fidelity model is a simple model, but it works well in the cosmic power spectrum emulation, as shown in Chapter 4 and Ref. [42, 43, 44].

Figure 1.11 shows an example of the linear multi-fidelity emulator (Eq. 1.12) on a simple function, where the high-fidelity data (red dots) is only available at $x < 0.5$. The low-fidelity data (blue dots) are noisy realization of the high-fidelity data with a biased shift to a higher y , but the low-fidelity data is available at all x . If we only use the high-fidelity data, the emulator prediction (purple curve) is not accurate at $x > 0.5$ due to lack of data. If we only use the low-fidelity data, the emulator prediction (blue curve) is biased. The multi-fidelity emulator (yellow curve) combines the high-fidelity data and the low-fidelity data, and it provides a more accurate prediction at all x .

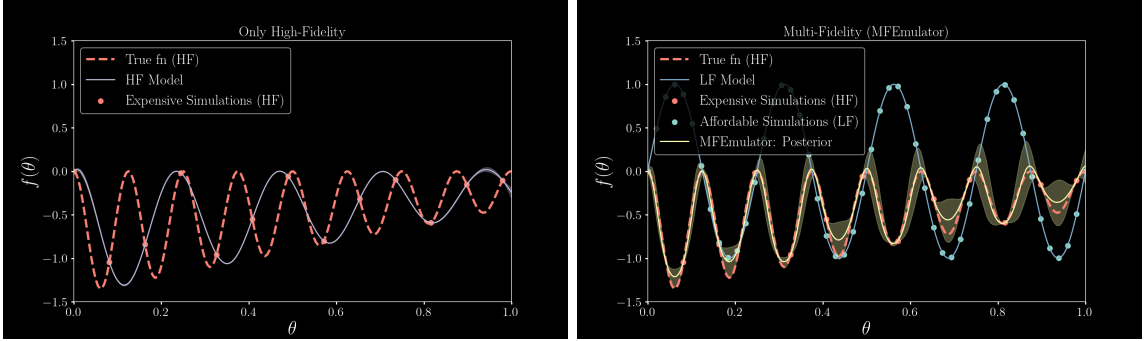



Figure 1.12: An example of the non-linear multi-fidelity emulator. Red dots as the high-fidelity data, blue dots as the low-fidelity data. Red dashed curve is the high-fidelity true function and blue curve is the low-fidelity true function. Left panel shows the emulator prediction using only the high-fidelity data (purple curve), which has a wrong frequency. Right panel shows the emulator prediction using only the low-fidelity data (blue curve), which provides prior knowledge on the frequency of the high-fidelity function, and the multi-fidelity emulator prediction (yellow curve) is more accurate with a reasonable uncertainty quantification. YouTube video tutorial can be found in the same video link in the caption of Figure 1.11.

Another more complex example is shown in Figure 1.12, where the high-fidelity data (red dots) is a sinusoidal function with a decreasing amplitude and the low-fidelity data (blue dots) is also a sinusoidal function but with a constant amplitude. The frequency of the high-fidelity function is 2 times higher than the low-fidelity function. The low-fidelity and high-fidelity are correlated in a non-linear way, so the linear scaling parameter, ρ_{t-1} , is not enough to capture the correlation. If we only use the high-fidelity data, the emulator prediction (purple curve) has a wrong frequency due to the lack of data. However, if we have the low-fidelity data, it provides prior knowledge on the frequency of the high-fidelity function and the emulator prediction (yellow curve) is more accurate.

The YouTube video tutorial of the multi-fidelity emulator on the simple function and the non-linear function can be found in the video link in the caption of Figure 1.11

and Figure 1.12. The jupyter notebook tutorial can be found in my GitHub repository 
https://github.com/jibanCat/nargp_tensorflow.

1.2.3 Application: Cosmic Multi-Fidelity Emulators

One of the most common applications of emulators in cosmology is the cosmic power spectrum emulation (e.g., see Ref. [28, 45, 10, 11], etc). These power spectrum emulators are built based on the pure dark matter N -body simulations. These emulators act as an extension of the linear perturbation theory on predicting the power spectrum at the non-linear scales of the large scale structure. The N -body simulation codes usually have different fidelities. For a given volume, the high-resolution simulation can resolve more small-scale structures, but it is more expensive to run. The resolution of the simulation is usually controlled by the number of particles in the simulation box or the size of the grids.

The power spectrum is a summary statistic describing the density fluctuations in the Universe. The power spectrum is in the Fourier space, thus, it is a function of the wave number, k . A lower k corresponds to lower frequency fluctuations (large scales), and a higher k corresponds to higher frequency fluctuations (small scales). Thus, for power spectrum emulation, low fidelity and high fidelity are highly correlated in a lower k region, but they are less correlated in a higher k region. The exact k value where the correlation breaks is usually determined by the resolution of the low-fidelity simulation.

Suppose we have a suite of simulations from low-fidelity and high-fidelity simulations, $\mathcal{D} = \{\boldsymbol{\theta}, \{P_t(\boldsymbol{\theta})\}_{t=1}^T\}$, where $P_t(\boldsymbol{\theta})$ is the power spectrum from the t -th fidelity simulation. $\boldsymbol{\theta}$ is the input cosmological parameters and $P_t(\boldsymbol{\theta})$ is the power spectrum at the cosmological parameters $\boldsymbol{\theta}$, which is also a function of the wave number, k . Thus, $P_t(\boldsymbol{\theta})$

should be $P_t(\boldsymbol{\theta}, k)$. The goal of the multi-fidelity emulation is to predict the power spectrum at the high-fidelity, $P_T(\boldsymbol{\theta}, k)$, given the data from the low-fidelity simulations, $P_t(\boldsymbol{\theta}, k)$, where $T \geq t$ and $t = 1, 2, \dots, T$,

$$P_t(\boldsymbol{\theta}, k) = \rho_{t-1} \cdot P_{t-1}(\boldsymbol{\theta}, k) + \delta_{t-1}(\boldsymbol{\theta}, k), \quad (1.13)$$

which is the same as Eq. 1.12 but with the wave number, k . Here, ρ_{t-1} could be a function of k depending on the resolution difference between t -th and $t - 1$ -th simulations. When the difference is small, ρ_{t-1} is likely to be a constant.

The modeling form of Eq. 1.13 has a similar form to the halo model (see e.g., Ref. [46, 47, 48]):

$$P(\boldsymbol{\theta}, k) = P_Q(\boldsymbol{\theta}, k) + P_H(\boldsymbol{\theta}, k), \quad (1.14)$$

where $P_Q(\boldsymbol{\theta}, k)$ is the quasi-linear part of the power spectrum, which is usually computed from the linear perturbation theory, and $P_H(\boldsymbol{\theta}, k)$ is the non-linear part of the power spectrum, which is usually computed from the N -body simulations. There is a clear connection between the two-halo term, P_Q , and the low-fidelity simulation ($\rho_{t-1}P_{t-1}(\boldsymbol{\theta}, k)$), and the one-halo term, P_H , and the high-fidelity simulation ($\delta_{t-1}(\boldsymbol{\theta}, k)$). The multi-fidelity modeling in Eq. 1.13 gives a much better prediction than the halo model because the low-fidelity simulations are much better than the linear theory

Figure 1.13 shows an example of the multi-fidelity emulator on the matter power spectrum emulation. It extends the Eq. 1.13 to two low-fidelity simulations (L1 and L2) and one high-fidelity simulation (HF). Due to the scale-separation property of the power spectrum, we can use L1 to correct the HF at the large scales, and use L2 to correct the HF

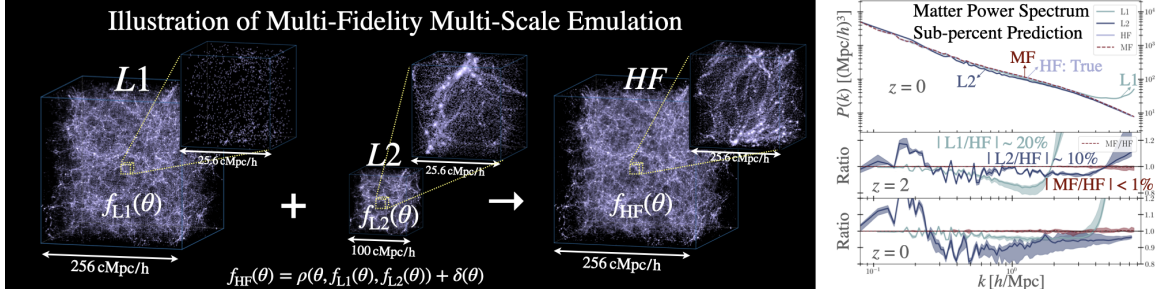


Figure 1.13: Example of the multi-fidelity emulator on the cosmic power spectrum. The left panel shows the N -body simulations in the density fields, including two low-fidelity nodes (L1, 128^3 in $256 \text{ Mpc}/h$ and L2, 128^3 in $100 \text{ Mpc}/h$) and one high-fidelity node (HF, 512^3 in $256 \text{ Mpc}/h$). The right panel shows the multi-fidelity emulator prediction (red curve) on the power spectrum, and the emulator relative errors are shown in the bottom panels for $z = 2$ and $z = 0$.

at the small scales. Here, L1 node is low-resolution simulation in a large volume, and L2 node is high-resolution simulation in a small volume. The HF node is the high-resolution simulation in a large volume. The multi-fidelity emulator (red curve) combines the L1, L2, and HF data, and it provides a more accurate prediction at all k . Chapter 5 will provide a detailed explanation of the multi-fidelity emulator in Figure 1.13.

1.3 Population Inference: A Short Note

In this section, I will briefly explain how to apply Bayesian inference to the population analysis. In astronomical data analysis, there is a difference between the population analysis and the individual source analysis. For example, an astronomer observe a single galaxy, the goal might be to understand the properties of the galaxy, such as the mass, the morphology, the color, etc. This is the individual source analysis. On the other hand, an astronomer analyzes a collection of galaxies, the goal might be to understand galaxy proper-

ties as a distribution, such as the galaxy stellar mass function, the galaxy color distribution, etc. This is the population analysis.

For each individual source, there is measurement uncertainty. For a population of sources, there is bias due to selection effects. The population analysis is challenging because it needs to propagate the measurement uncertainty from each individual source to the population level, subjecting the selection bias. The population inference method is designed to address this issue using Bayesian hierarchical modeling.

Gravitational wave astronomy community has developed a tradition on using Bayesian hierarchical modeling to analyze the population of the gravitational wave events. Chapter 6 describes my work on using Bayesian hierarchical inference to understand the black hole mass spectrum from the gravitational wave events. Below, I will provide a brief overview on how to apply Bayesian hierarchical inference to the population analysis in the context of the gravitational wave events.

Here, we define some notations. In what follows, $\boldsymbol{\theta}$ are the parameters for a single event. For example, $\boldsymbol{\theta}$ could be the masses of the binary black hole, the spins of the binary black hole, the sky location of the binary black hole, etc. \boldsymbol{d} is the data for a single event. The event parameters $\boldsymbol{\theta}$ are usually inferred from the data \boldsymbol{d} using a template bank, which is a set of waveform models to predict the gravitational wave signal. When we have a catalog of events, we have $\{\boldsymbol{\theta}\}$ and $\{\boldsymbol{d}\}$.

The population parameters, $\boldsymbol{\Lambda}$, are the parameters for the population of the events according to our population model. The population models are usually designed to understand the distribution of the event parameters, such as the mass spectrum, the redshift

distribution, the spin distribution, etc. They can be either astrophysical models or phenomenological models. For instance, if we have a population model that the black hole mass spectrum follows a power-law distribution, then the population parameters $\mathbf{\Lambda}$ are the power-law index and the normalization of the power-law distribution.

The distinction between event parameters, $\boldsymbol{\theta}$, and population parameters, $\mathbf{\Lambda}$, is important. We can infer the $\boldsymbol{\theta}$ from a single event \mathbf{d} , but we cannot infer the $\mathbf{\Lambda}$ from a single event. To infer $\mathbf{\Lambda}$, we need to consider the entire catalog of events $\{\mathbf{d}\}$ and the “detection efficiency” due to the sensitivity of instrument.

We need to have the concept of “detection” in the population modeling. The detection is the process of deciding whether an event is detected by the instrument. Detection bias means that our observed catalog of events is not a fair sample of the true underlying distribution. For example, in gravitational wave, the lighter binary black holes are harder to detect than the heavier binary black holes, so the observed catalog of events is biased towards the heavier binary black holes. And in extragalactic astronomy, the fainter galaxies are harder to detect than the brighter galaxies.

To bring the concept of detection to the probability notations, we can introduce the “trigger” term. When the event data \mathbf{d} have been recorded at the detector, it either triggers the detection or not. This trigger is determined according to some deterministic criteria ($\rho(\mathbf{d}) \geq \rho_{\text{threshold}}$),

$$p(\text{trig}|\mathbf{d}) = \begin{cases} 0 & \rho(\mathbf{d}) < \rho_{\text{threshold}} , \\ 1 & \rho(\mathbf{d}) \geq \rho_{\text{threshold}} . \end{cases} \quad (1.15)$$

The criteria is usually determined by the signal-to-noise ratio.

For a given event with event parameters $\boldsymbol{\theta}$, we want to know what is the probability of the event being detected. This requires us to marginalize over the data \mathbf{d} ,

$$p(\text{trig}|\boldsymbol{\theta}) = \int p(\text{trig}|\mathbf{d})p(\mathbf{d}|\boldsymbol{\theta})d\mathbf{d}. \quad (1.16)$$

The probability, $p(\text{trig}|\boldsymbol{\theta})$, is the *detection probability* of the event. The probability, $p(\mathbf{d}|\boldsymbol{\theta})$, is the noise model of the event, in another word, it is the single-event likelihood model. The above marginalization simply states that we marginalize over all possible noisy realizations of the data \mathbf{d} (using the noise model $p(\mathbf{d}|\boldsymbol{\theta})$) to get the detection probability. Eq. 1.16 is usually computed using the Monte Carlo integration.

For a population model, we parameterize the distribution of the event parameters, $\boldsymbol{\theta}$, according to the population parameters $\boldsymbol{\Lambda}$, i.e., $p(\boldsymbol{\theta}|\boldsymbol{\Lambda})$. To acquire the detection efficiency of the population model, we need to marginalize over the event parameters $\boldsymbol{\theta}$ in Eq. 1.16,

$$\begin{aligned} p(\text{trig}|\boldsymbol{\Lambda}) &= \int d\mathbf{d} \int p(\text{trig}|\mathbf{d})p(\mathbf{d}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\Lambda})d\boldsymbol{\theta} \\ &= \int p(\text{trig}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\Lambda})d\boldsymbol{\theta}. \end{aligned} \quad (1.17)$$

The detection efficiency is often defined as $\alpha \equiv p(\text{trig}|\boldsymbol{\Lambda})$ in the literature. The detection efficiency α can be understood as how efficient the population model can produce detectable events. α can be interpreted as the fraction of the expected detectable events (N_{det}) over the total number of events (N_{tot}), $\alpha = N_{\text{det}}/N_{\text{tot}}$. To compute the detection efficiency in Eq. 1.17, we can simulate the values of $\boldsymbol{\theta}$ from the population model according to the prior $p(\boldsymbol{\theta}|\boldsymbol{\Lambda})$, and then compute the detection probability $p(\text{trig}|\boldsymbol{\theta})$ for each $\boldsymbol{\theta}$, and then average over all $\boldsymbol{\theta}$. Thus, one way to interpret the population model, $p(\boldsymbol{\theta}|\boldsymbol{\Lambda})$, is that it is a physics-informed parametric prior over the event parameters, and we want to use a set of events to infer the population parameters $\boldsymbol{\Lambda}$.

Now, suppose we have a set of N_{obs} events, $\{\mathbf{d}_i\}$, and detection information, $\{\text{trig}\}$.

We can infer the posterior probability density of the population parameters $\mathbf{\Lambda}$ using

$$p(\mathbf{\Lambda}|\{\mathbf{d}_i\}, \{\text{trig}\}, N_{\text{obs}}) \propto \frac{p(\mathbf{\Lambda})e^{-\alpha N} N^{N_{\text{obs}}}}{p(\{\mathbf{d}_i\}, \{\text{trig}\}, N_{\text{obs}})} \prod_{i=1}^{N_{\text{obs}}} \mathcal{L}_i^{\text{obs}}, \quad (1.18)$$

where $\mathcal{L}_i^{\text{obs}}$ is the event likelihood of the i -th event, which is independent of the selection effects,

$$\mathcal{L}_i^{\text{obs}} = \int p(\mathbf{d}_i|\boldsymbol{\theta}_i)p(\boldsymbol{\theta}_i|\mathbf{\Lambda})d\boldsymbol{\theta}_i. \quad (1.19)$$

Here, $p(\mathbf{\Lambda})$ is the prior of the population parameters, and $p(\{\mathbf{d}_i\}, \{\text{trig}\}, N_{\text{obs}})$ is the evidence of the data. The important part of the posterior is the detection efficiency α , which is in the term $e^{-\alpha N} N^{N_{\text{obs}}}$. This term comes from assuming the events are generated from an inhomogeneous Poisson process, and it penalizes the population model when the expected number of events (αN) is different from the detected number of events. For example, when detection efficiency α is higher, the expected number of events is higher, and the population model is penalized when the detected number of events is lower than the expected number of events via the term $e^{-\alpha N}$.

Chapter 2

Detecting Multiple DLAs per Spectrum in SDSS DR12 with Gaussian Processes

2.1 Abstract

We present a revised version of our automated technique using Gaussian processes (GPs) to detect Damped Lyman- α absorbers (DLAs) along quasar (QSO) sightlines. The main improvement is to allow our Gaussian process pipeline to detect multiple DLAs along a single sightline. Our DLA detections are regularised by an improved model for the absorption from the Lyman- α forest which improves performance at high redshift. We also introduce a model for unresolved sub-DLAs which reduces mis-classifications of absorbers without detectable damping wings. We compare our results to those of two different large-scale DLA catalogues

and provide a catalogue of the processed results of our Gaussian process pipeline using 158 825 Lyman- α spectra from SDSS data release 12. We present updated estimates for the statistical properties of DLAs, including the column density distribution function (CDDF), line density (dN/dX), and neutral hydrogen density (Ω_{DLA}).

2.2 Introduction

Damped Ly α absorbers (DLAs) are absorption line systems with high neutral hydrogen column densities ($N_{\text{HI}} > 10^{20.3} \text{cm}^{-2}$) discovered in sightlines of quasar spectroscopic observations [49]. The gas which gives rise to DLAs is dense enough to be self-shielded from the ultra-violet background (UVB) [50] yet diffuse enough to have a low star-formation rate [51]. DLAs dominate the neutral-gas content of the Universe after reionisation [52, 5, 53, 7]. Simulations tell us DLAs are connected with galaxies over a wide range of halo masses [54, 55, 18], and at $z \geq 2$ are formed from the accretion of neutral hydrogen gas onto dark matter halos [56, 57]. The abundance of DLAs at different epochs of the universe ($2 < z < 5$) thus becomes a powerful probe to understand the formation history of galaxies [52, 58].

Finding DLAs historically involves a combination of template fitting and visual inspection of spectra by the eyes of trained astronomers [20, 59]. Recent spectroscopic surveys such as the Sloan Digital Sky Survey (SDSS) [60] have taken large amount of quasar spectra ($\sim 500\,000$ in SDSS-IV [61]). Future surveys such as the Dark Energy Spectroscopic Instrument (DESI¹) will acquire more than 1 million quasars, making visual inspection of the spectra impractical. Moreover, the low signal-to-noise ratios of SDSS data makes the

¹<http://desi.lbl.gov>

task of detecting DLAs even harder, and induces noise related detection systematics. Since the release of the SDSS DR14 quasar catalogue [61], visual inspection is no longer performed on all quasar targets. A fully automated and statistically consistent method thus needs to be presented for current and future surveys.

We provide a catalogue of DLAs using SDSS DR12 with 158 825 quasar sightlines. We demonstrate that our pipeline is capable of detecting an arbitrary number of DLAs within each spectroscopic observation, which makes it suitable for future surveys. Furthermore, since our pipeline resides within the framework of Bayesian probability, we have the ability to make probabilistic statements about those observations with low signal-to-noise ratios. This property allows us to make probabilistic estimations of DLA population statistics, even with low-quality noisy data [62].

Other available searches of DLAs in SDSS include: a visual-inspection survey [59], visually guided Voigt-profile fitting [20, 6]; and three automated methods: a template-fitting method [5], an unpublished machine-learning approach using Fisher discriminant analysis [63], and a deep-learning approach using a convolutional neural network [4]. Although these methods have had some success in creating large DLA catalogues, they suffer from hard-to-control systematics due to reliance either on templates or black-box training.

We present a revised version of our previous automated method based on a Bayesian model-selection framework [3]. In our previous model [3], we built a likelihood function for the quasar spectrum, including the continuum and the non-DLA absorption, using Gaussian processes [31]. The SDSS DR9 concordance catalogue was applied to learn the covariance of the Gaussian process model. In this paper, we use the effective optical depth of the

Lyman-series forest to allow the mean model of the likelihood function to be adjustable to the mean flux of the quasar spectrum, which reduces the probability of falsely fitting high-column density absorbers at high redshifts. We also improve our knowledge of low-column density absorbers and build an alternative model for sub-DLAs, which are the HI absorbers with $19.5 < \log_{10} N_{\text{HI}} < 20$. These modifications allow us to extend our previous pipeline to detect an arbitrary number of DLAs within each quasar sightline without overfitting.

Alongside the revised DLA detection pipeline, we present the new estimates of DLA statistical properties at $z > 2$. Since the neutral hydrogen gas in DLAs will eventually accrete onto galactic haloes and fuel the star formation, these population statistics can give an independent constraint on the theory of galaxy formation. Our pipeline relies on a well-defined Bayesian framework and contains a full posterior density on the column density and redshift for a given DLA. We thus can properly propagate the uncertainty in the properties of each DLA spectrum to population statistics of the whole sample. Additionally, we are also able to account for low signal-to-noise ratio samples in our population statistics since the uncertainty will be reflected in the posterior probability. We thus substantially increase the sample size in our measurements by including these noisy observations.

2.3 Notation

We will briefly recap the notation we defined in [3]. Imagine we are observing a QSO with a known redshift z_{QSO} . The underlying true emission function $f(\lambda_{\text{rest}})$ ($f: \mathcal{X} \rightarrow \mathbb{R}$) of the QSO is a mapping relation from rest-frame wavelength to flux. We will always assume the z_{QSO} is known and rescale the observed-frame wavelength λ_{obs} to the rest-frame wavelength

with $\lambda_{\text{rest}} (= \lambda_{\text{obs}} / (1 + z_{\text{QSO}}))$. We will use λ to replace λ_{rest} in the rest of the text because we only work on λ_{rest} .

The quasar spectrum observed is not the intrinsic emission function $f(\lambda)$. Both the instrumental noise and absorption due to the intervening intergalactic medium along the line of sight will affect the observed flux. We thus denote the observed flux as a function $y(\lambda)$.

For a real spectroscopic observation, we measure the function $y(\lambda)$ on a discrete set of samples $\boldsymbol{\lambda}$. We thus denote the observed flux as a vector \boldsymbol{y} , which is defined as $y_i = y(\lambda_i)$ with i representing i^{th} pixel. For a given QSO observation, we use \mathcal{D} to represent a set of discrete observations $(\boldsymbol{\lambda}, \boldsymbol{y})$.

We exclude missing values of the spectroscopic observations in our calculations. These missing values are due to pixel-masking in the spectroscopic observations (e.g., bad columns in the CCD detectors). We will use NaN (‘not a number’) to represent those missing values in the text, and we will always ignore NaNs in the calculations.

2.4 Bayesian Model Selection

The classification approach used in our pipeline depends on Bayesian model selection. Bayesian model selection allows us to compute the probability that a spectroscopic sightline \mathcal{D} contains an arbitrary number of DLAs through evaluating the probabilities of a set of models $\{\mathcal{M}_i\}$, where i is a positive integer. This set of \mathcal{M}_i contains all potential models we want to classify: a model with no DLA and models having between one DLA and k DLAs.

For each \mathcal{M}_i , we want to compute the probability that best explains the data \mathcal{D} given a model \mathcal{M} . To do this, we have to marginalize the model parameters θ and evaluate the model evidence,

$$p(\mathcal{D} | \mathcal{M}) = \int p(\mathcal{D} | \mathcal{M}, \theta) p(\theta | \mathcal{M}) d\theta. \quad (2.1)$$

Given a set of model evidences $p(\mathcal{D} | \mathcal{M}_i)$ and model priors $\Pr(\mathcal{M}_i)$, we are able to evaluate the posterior of a model given data based on Bayes's rule,

$$\Pr(\mathcal{M} | \mathcal{D}) = \frac{p(\mathcal{D} | \mathcal{M}) \Pr(\mathcal{M})}{\sum_i p(\mathcal{D} | \mathcal{M}_i) \Pr(\mathcal{M}_i)}. \quad (2.2)$$

We will select the model from $\{\mathcal{M}_i\}$ with the highest posterior. Readers may think of this method as an application of Bayesian hypothesis testing. Instead of only getting the likelihoods conditioned on models, we get posterior probabilities for each model given data.

Let k be the maximum number of DLAs we will want to detect in a quasar spectrum. For our multi-DLA model selection, we will develop $k+2$ models, which include a null model for no DLA detection ($\mathcal{M}_{\text{-DLA}}$), models for detecting exactly k DLAS ($\mathcal{M}_{\text{DLA}(k)}$), and a model with sub-DLAs (\mathcal{M}_{sub}). With a given spectroscopic sightline \mathcal{D} , we will compute the posterior probability of having exactly k DLAS in data \mathcal{D} , $\Pr(\mathcal{M}_{\text{DLA}(k)} | \mathcal{D})$.

2.5 Gaussian Processes

In this section, we will briefly recap how we use *Gaussian processes* (GPs) to describe the QSO emission function $f(\lambda)$, following [3]. The QSO emission function is a complicated function without a simple form derived from physically motivated parameters. We thus use a nonparametric framework, Gaussian processes, for modelling this physically unknown function $f(\lambda)$. A detailed introduction to GPs may be found in [31].

2.5.1 Definition and prior distribution

We wish to use a Gaussian process to model the QSO emission function $f(\lambda)$. We can treat a Gaussian process as an extension of the joint Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to infinite continuous domains. The difference is that a Gaussian process is a distribution over functions, not just a distribution over a finite number of random variables (although since we are dealing with pixelised variables here the distinction is less important).

A GP is completely specified by its first two central moments, a mean function $\mu(\lambda)$ and a covariance function $K(\lambda, \lambda')$:

$$\begin{aligned}\mu(\lambda) &= \mathbb{E}[f(\lambda) \mid \lambda], \\ K(\lambda, \lambda') &= \mathbb{E}[(f(\lambda) - \mu(\lambda))(f(\lambda') - \mu(\lambda')) \mid \lambda, \lambda'] \\ &= \text{cov}[f(\lambda), f(\lambda') \mid \lambda, \lambda'].\end{aligned}\tag{2.3}$$

The mean vector describes the expected behaviour of the function, and the covariance function specifies the covariance between pairs of random variables. We thus will write the GP as,

$$f(\lambda) \sim \mathcal{GP}(\mu(\lambda), K(\lambda, \lambda')).\tag{2.4}$$

We can write the prior probability distribution of a GP as,

$$p(f) = \mathcal{GP}(f; \mu, K).\tag{2.5}$$

Real spectroscopic observations measure a discrete set of inputs $\boldsymbol{\lambda}$ and the corresponding $f(\boldsymbol{\lambda})$, so we get a multivariate Gaussian distribution

$$p(\mathbf{f}) = \mathcal{N}(f(\boldsymbol{\lambda}); \mu(\boldsymbol{\lambda}), K(\boldsymbol{\lambda}, \boldsymbol{\lambda}')).\tag{2.6}$$

Assuming the dimension of $\boldsymbol{\lambda}$ and \boldsymbol{f} is d , the form of the multivariate Gaussian distribution is written as

$$\mathcal{N}(\boldsymbol{f}; \boldsymbol{\mu}, \mathbf{K}) = \frac{1}{\sqrt{(2\pi)^d \det \mathbf{K}}} \exp\left(-\frac{1}{2}(\boldsymbol{f} - \boldsymbol{\mu})^\top \mathbf{K}^{-1}(\boldsymbol{f} - \boldsymbol{\mu})\right). \quad (2.7)$$

2.5.2 Observation model

We now have a Gaussian process model for a discrete set of wavelengths $\boldsymbol{\lambda}$ and true emission fluxes \boldsymbol{f} . To build the likelihood function for observational data $\mathcal{D} = (\boldsymbol{\lambda}, \boldsymbol{y})$, we have to incorporate the observational noise. Here we assume the observational noise is modelled by an independent Gaussian variable for each wavelength pixel, allowing the noise realisation to differ between pixels but neglecting inter-pixel correlations.

The noise variance for a given λ_i is written as $\nu_i = \sigma(\lambda_i)^2$. $\sigma(\lambda_i)$ is the measurement error from a single observation on a given wavelength point λ . With the above assumptions, we can write down the mechanism of generating observations as:

$$p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{f}, \boldsymbol{\nu}) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{f}, \mathbf{V}), \quad (2.8)$$

where $\mathbf{V} = \text{diag } \boldsymbol{\nu}$, which means we put the vector $\boldsymbol{\nu}$ on the diagonal terms of the diagonal square matrix \mathbf{V} .

Given an observational model $p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{f}, \boldsymbol{\nu})$ and a Gaussian process emission model $p(\boldsymbol{f} \mid \boldsymbol{\lambda})$, the prior distribution for observations \boldsymbol{y} is obtained by marginalizing the latent

function \mathbf{f} :

$$\begin{aligned}
p(\mathbf{y} \mid \boldsymbol{\lambda}, \boldsymbol{\nu}) &= \int p(\mathbf{y} \mid \boldsymbol{\lambda}, \mathbf{f}, \boldsymbol{\nu}) p(\mathbf{f} \mid \boldsymbol{\lambda}) d\mathbf{f} \\
&= \int \mathcal{N}(\mathbf{y}; \mathbf{f}, \mathbf{V}) \mathcal{N}(\mathbf{f}; \boldsymbol{\mu}, \mathbf{K}) d\mathbf{f} \\
&= \mathcal{N}(\mathbf{y}; \boldsymbol{\mu}, \mathbf{K} + \mathbf{V}),
\end{aligned} \tag{2.9}$$

where the Gaussians are closed under the convolution. Our observation model thus becomes a multivariate normal distribution described by a mean model $\boldsymbol{\mu}(\boldsymbol{\lambda})$, covariance structure $K(\lambda, \lambda')$, and the instrumental noise \mathbf{V} . The instrumental noise is derived from SDSS pipeline noise, so it is different from QSO-to-QSO; however, since \mathbf{K} encodes the covariance structure of quasar emissions, \mathbf{K} should be the same for all quasars.

As explained in [3], there is no obvious choice for a prior covariance function \mathbf{K} for modelling the quasar emission function. Most off-the-shelf covariance functions assume some sort of translation invariance, but this is not suitable for spectroscopic observations². However, we understand the quasar emission function will be independent of the presence of a low redshift DLA. We also assume that quasar emission functions are roughly redshift independent in the wavelength range of interest (Lyman limit to Lyman- α), as accretion physics should not strongly vary with cosmological evolution. We thus build our own custom $\boldsymbol{\mu}$ and K for the GP prior to model the quasar spectra.

2.6 Learning A GP Prior from QSO Spectra

In this section, we will recap the prior modelling choices we made in [3] and the modifications we made to reliably detect multiple DLAs in one spectrum. We first build a

²Detailed explanations are in [3] Section 4.2.1.

GP model for QSO emission in the absence of DLAs, the null model $\mathcal{M}_{\text{-DLA}}$. Our model with DLAs (\mathcal{M}_{DLA}) extends this null model. With the model priors and model evidence of all models we are considering, we compute the model posterior with Bayesian model selection.

The GP prior is completely described by the first two moments, the mean and covariance functions, which we derive from data. We must consider the mean flux of quasar emission, the absorption effect due to the Lyman- α forest, and the covariance structure within the Lyman series.

2.6.1 Data

Our training set to learn our GP null model comprises the spectra observed by SDSS BOSS DR9 and labelled as containing (or not) a DLA by [64].³ The DR9 dataset includes 54 468 QSO spectra with $z_{\text{QSO}} > 2.15$. We removed the following quasars from the training set:

- $z_{\text{QSO}} < 2.15$: quasars with redshifts lower than 2.15 have no Lyman- α in the SDSS band.
- BAL: quasars with broad absorption lines as flagged by the SDSS pipeline.
- spectra with less than 200 detected pixels.
- ZWARNING: spectra whose analysis had warnings as flagged by the SDSS redshift estimation. Extremely noisy spectra (the `TOO_MANY_OUTLIERS` flag) were kept.

³However, we use the DR12 pipeline throughout.

2.6.2 Modelling Decisions

Consider a set of quasar observations $\mathcal{D} = (\boldsymbol{\lambda}, \mathbf{y})$; we always shift the observer's frame λ_{obs} to rest-frame λ so that we can set the emissions of Lyman series from different spectra to the same rest-wavelengths. The assumption here is that the z_{QSOS} of quasars are known for all the observed spectra, which is not precisely true for the spectroscopic data we have here. Accurately estimating the redshift of quasars is beyond the scope of this paper, and is tackled elsewhere [65].

The observed magnitude of a quasar varies considerably, based on its luminosity distance and the properties of the black hole. For the observation \mathbf{y} to be described by a GP, it is necessary to normalize all flux measurements by dividing by the median flux observed between 1310 Å and 1325 Å, a wavelength region which is unaffected by the Lyman- α forest.

We model the same wavelength range as in [3]:

$$\lambda \in [911.75\text{\AA}, 1215.75\text{\AA}], \quad (2.10)$$

going from the quasar rest frame Lyman limit to the quasar rest frame Lyman- α . The spacing between pixels is $\Delta\lambda = 0.25\text{\AA}$. Note that we prefer not to include the region past the Lyman limit. This is partly due to the relatively small amount of data in that region and partly because the non-Gaussian Lyman break associated with Lyman limit systems can confuse the model. In particular, it occasionally tries to model a Lyman break with a wide DLA profile with a high column density. We shall see this is especially a problem if the quasar redshift is slightly inaccurate. The code considers the prior probability of a Lyman break at a higher redshift than the putative quasar rest frame to be zero and thus is especially prone to finding other explanations for the large absorption trough.

To model the relationship between flux measurements and the true QSO emission spectrum, we have to add terms corresponding to instrumental noise and weak Lyman- α absorption to the intrinsic correlations within the emission spectrum. Instrumental noise was already added in Eq. 2.9 as a matrix \mathbf{V} .

The remaining part of the modelling is to define the GP covariance structure for quasars across different redshifts. In [3], Lyman- α absorbers were modelled by a single additive noise term, $\mathbf{\Omega}$, accounting for the effect of the forest as extra noise in the emission spectrum. This is not completely physical: it assumes that the Lyman- α forest is just as likely to cause emission as absorption.

Here we rectify this by not only including the Lyman- α perturbation term in our Gaussian process as $\mathbf{\Omega}$, but introducing a redshift dependent mean flux ($\mu(\mathbf{z})$) with a dependence on the absorber redshift ($z(\lambda_{\text{obs}})$). We model the overall mean model with a redshift dependent absorption function and a mean emission vector: $\mu(\mathbf{z}) = a(\mathbf{z}) \circ \boldsymbol{\mu}$. The notation \circ refers to Hadamard product, which is the element-wise product between two vectors or matrices. The covariance matrix is decomposed into $\mathbf{A}_F(\mathbf{K} + \mathbf{\Omega})\mathbf{A}_F$, where $\text{diag}(\mathbf{A}_F) = a(\mathbf{z})$ and \mathbf{A}_F is a diagonal matrix.⁴ The \mathbf{K} matrix describes the covariance between different emission lines in the quasar spectrum, which we will learn from data. The \mathbf{A}_F matrix is applied to \mathbf{K} because we assume that \mathbf{K} is learned before the absorption noise $a(\mathbf{z})$ is applied. See Sec2.6.4 for how we learn the covariance.

⁴ $\mathbf{A}_F^T = \mathbf{A}_F$ because it is diagonal.

Combining all modelling decisions, the model prior for an observed QSO emission is:

$$p(\mathbf{y} \mid \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\text{QSO}}, \mathcal{M}_{\text{-DLA}}) = \mathcal{N}(\mathbf{y}; \boldsymbol{\mu}(z), \mathbf{A}_F(\mathbf{K} + \boldsymbol{\Omega})\mathbf{A}_F + \mathbf{V}). \quad (2.11)$$

The mean emission flux is now redshift- and wavelength-dependent, so the optimisation steps will differ slightly from [3]. We will address the modifications in the following subsections.

2.6.3 Redshift-Dependent Mean Flux Vector

In this paper, instead of using a single mean vector $\boldsymbol{\mu}$ to describe all spectra, we adjust the mean model of the GP to fit the mean flux of each quasar spectrum. For modeling the effect of forest absorption on the flux, we adopt an empirical power law with effective optical depth $\tau_0(1+z)^\beta$ for Ly α forest [66]:

$$a(z) = \exp(-\tau_0(1+z)^\beta), \quad (2.12)$$

where the absorber redshift z is related to the observer's wavelength λ_{obs} as:

$$\begin{aligned} 1+z &= \frac{\lambda_{\text{obs}}}{\lambda_{\text{Ly}\alpha}} \\ &= \frac{\lambda_{\text{obs}}}{1215.7\text{\AA}} \\ &= (1+z_{\text{QSO}}) \frac{\lambda}{1215.7\text{\AA}}, \end{aligned} \quad (2.13)$$

so the absorber redshift $z(\lambda_{\text{obs}}) = z(\lambda, z_{\text{QSO}})$ is a function of the quasar redshift and the wavelength.

In [3], we assumed the absorption from the forest would only play a role in the additive noise term (ω) in our likelihood model $p(\mathbf{y} | \boldsymbol{\lambda}, \boldsymbol{\nu}, \boldsymbol{\omega}, z_{\text{QSO}}, \mathcal{M}_{\text{-DLA}})$ with the form:

$$\omega'(\lambda, \lambda_{\text{obs}}) = \omega(\lambda) s(z(\lambda_{\text{obs}}))^2; \quad (2.14)$$

$$s(z) = 1 - \exp(-\tau_0(1+z)^\beta) + c_0, \quad (2.15)$$

where z is the absorber redshift. The $\omega(\lambda)$ term represents the global absorption noise, and the $s(z)$ corresponds to the absorption effect contributed by the Lyman- α absorbers along the line of sight as a function of the absorber redshift z .

Thus in our earlier model the Lyman- α forest introduces additional fluctuations in the observed spectrum \mathbf{y} . This assumption worked well for low-redshift spectra, because mean absorption due to the Lyman- α forest at low redshifts is relatively small. At high-redshifts however, the suppression of the mean flux induced by many Lyman- α absorbers is substantial, see Figure 2.1. In our earlier model, essentially all high-redshift QSO spectra were substantially more absorbed than the mean emission model μ due to absorption from the Lyman- α forest. To explain this absorption, our model would fit multiple DLAs with large column densities.

We have improved the modelling of the Lyman- α forest by allowing the mean GP model μ to be redshift dependent, having a mean optical depth following the measurement of [66]:

$$\begin{aligned} \tau_{\text{eff}}(z) &= \tau_0(1+z)^\gamma \\ &= 0.0023 \times \exp(1+z)^{3.65}, \end{aligned} \quad (2.16)$$

There are other measurements of τ_{eff} at higher precision than [66], [e.g., Ref. [67]]. However, they are derived from SDSS data while [66] was derived from high resolution spectra. We

therefore choose to use [66] to preserve the likelihood principle that priors should not depend on the dataset in question.

We include the effect of the whole Lyman series with a similar model, but however accounting for the different atomic coefficients of the higher order Lyman lines:

$$\tau_{\text{eff,HI}}(z(\lambda_{\text{obs}}); \gamma, \tau_0) = \sum_{i=2}^N \tau_0 \frac{\lambda_{1i} f_{1i}}{\lambda_{12} f_{12}} (1 + z_{1i}(\lambda_{\text{obs}}))^\gamma \times I_{(z_{1i}(\min(\lambda_{\text{obs}})), z_{\text{QSO}})}(z) \quad (2.17)$$

Here f_{1i} represents the oscillator strength and λ_{1i} corresponds to the transition wavelength from the $n = 1$ to $n = i$ atomic energy level. We model the Lyman series up to $N = 32$, with $i = 2$ being Ly α and $i = 3$ Ly β . The absorption redshift z_{1i} for the $n = 1$ to $n = i$ transition is defined by:

$$1 + z_{1i} = \frac{\lambda_{\text{obs}}}{\lambda_{1i}}. \quad (2.18)$$

The optical depth at the line center is estimated by:

$$\tau_0 = \sqrt{\pi} \frac{e^2}{m_e c} \frac{N_\ell f_{\ell u} \lambda_{\ell u}}{b}, \quad (2.19)$$

where ℓ indicates the lower energy level and u is the upper energy level. For Lyman- α , we have $\lambda_{\ell u} = 1215.7 \text{ \AA}$ and $f_{\ell u} = 0.4164$; for Lyman- β , we have $\lambda_{\ell u} = 1025.7$ and $f_{\ell u} = 0.07912$. Given Eq. 2.19, we have the effective optical depth for the Lyman- β forest:

$$\tau_\beta = \frac{f_{31} \lambda_{31}}{f_{21} \lambda_{21}} \tau_0 = \frac{0.07912 \times 1025.7}{0.4164 \times 1215.7} \times 0.0023 = 0.0004. \quad (2.20)$$

The mean prior of the GP model for each spectrum is re-written as:

$$\boldsymbol{\mu}(z) = \boldsymbol{\mu} \circ \exp(-\tau_{\text{eff,HI}}(\mathbf{z}; \gamma = 3.65, \tau_0 = 0.0023)). \quad (2.21)$$

We will simply write $\tau_{\text{eff,HI}}(\mathbf{z}) = \tau_{\text{eff,HI}}(\mathbf{z}; \gamma = 3.65, \tau_0 = 0.0023)$ in the following text for simplicity. The new $\boldsymbol{\mu}$ is estimated via:

$$\boldsymbol{\mu} = \frac{1}{N_{\text{-NaN}}} \sum_{y_{ij} \neq \text{NaN}} y_{ij} \cdot \exp(+\tau_{\text{eff,HI}}(z_{ij})). \quad (2.22)$$

Eq. 2.22 rescales the mean observed fluxes back to the expected continuum before the suppression due Lyman series absorption, hopefully recovering approximately the true QSO emission function \mathbf{f} . Figure 2.1 shows the re-trained mean quasar emission model for an example quasar. The mean model, $\boldsymbol{\mu}$, is much closer to the peak emission flux above the absorbed forest.

For model consistency, we account for the mean suppression from weak absorbers in our redshift-dependent noise model ω with:

$$\omega'(\lambda, \lambda_{\text{obs}}) = \omega(\lambda) s_F(z(\lambda_{\text{obs}}))^2; \quad (2.23)$$

$$\text{where } s_F(z(\lambda_{\text{obs}})) = 1 - \exp(-\tau_{\text{eff,HI}}(z(\lambda_{\text{obs}}); \beta, \tau_0)) + c_0. \quad (2.24)$$

τ_0 , β , and c_0 are parameters that are learned from the data. Figure 2.2 shows the mean model and absorption noise variance we use, compared to the model from [3].

Note that the mean flux model introduces degeneracies between the parameters of Eq. 2.24. For example, c_0 may be compensated by the overall amplitude of pixel-wise noise vector $\boldsymbol{\omega}$. For this reason, we should not ascribe strict physical interpretations to the optimal values of Eq. 2.24. The optimised $\boldsymbol{\omega}'$ is simply an empirical relation modeling the pixel-wise and redshift-dependent noise in the null model given SDSS data.

After introducing the effective optical depth into our GP mean model, we decrease the number of large DLAs we detect at high redshifts and thus measure lower Ω_{DLA} at high

redshifts (see Section 2.11.3 for more details). This is because, for high redshift quasars, the mean optical depth may be close to unity. To explain this unexpected absorption, the previous code will fit multiple high-column density absorbers to the raw emission model, artificially increasing the number of DLAs detected. With the mean model suppressed, there is substantially less raw absorption to explain, and so this tendency is avoided.

2.6.4 Learning the flux covariance

\mathbf{K} and Ω (Eq. 2.11) are optimised to maximise the likelihood of generating the data, \mathcal{D} . The mean flux model is not optimized, but follows the effective optical depth reported in [66]. Thus we remove the effect of forest absorption before we train the covariance function and train on $\mathcal{D}' = \{\boldsymbol{\lambda}, \mathbf{y} \circ \exp(+\tau_{\text{eff,HI}}(\mathbf{z})) - \mu(\mathbf{z})\}$ to find the optimal parameters for \mathbf{K} and Ω .

We assume the same likelihood as [3] for generating the whole training data set (\mathbf{Y}):

$$\begin{aligned}
 p(\mathbf{Y} \mid \boldsymbol{\lambda}, \mathbf{V}, \mathbf{M}, \boldsymbol{\omega}, \mathbf{z}_{\text{QSO}}, \mathcal{M}_{\text{-DLA}}) \\
 = \prod_{i=1}^{N_{\text{spec}}} \mathcal{N}(\mathbf{y}_i; \boldsymbol{\mu}, \mathbf{K} + \boldsymbol{\Omega} + \mathbf{V}_i),
 \end{aligned}
 \tag{2.25}$$

where \mathbf{Y} means the matrix containing all the observed flux in the training data, and the product on the right hand side says we are combining all likelihoods from each single spectrum. The noise matrix $\boldsymbol{\Omega} = \text{diag } \boldsymbol{\omega}'$ is the diagonal matrix which represents the Lyman- α forest absorption from Eq. 2.24.

\mathbf{M} is a low-rank decomposition of the covariance matrix \mathbf{K} we want to learn:

$$\mathbf{K} = \mathbf{M}\mathbf{M}^\top,
 \tag{2.26}$$

where \mathbf{M} is an $(N_{\text{pixels}} \times k)$ matrix. Without this low-rank decomposition, we would need to learn $N_{\text{pixel}}^2 = 1\,217 \times 1\,217$ free parameters. With Eq. 2.26, we can limit the number of free parameters to be $N_{\text{pixels}} \times k$, where $k \ll N_{\text{pixels}}$; also, it guarantees the covariance matrix \mathbf{K} to be positive semi-definite. Each column of the \mathbf{M} can be treated as an eigenspectrum of the training data, where we set the number of eigenspectra to be $k = 20$. We will optimise the \mathbf{M} matrix and the absorption noise in Eq. 2.24 simultaneously.

A modification performed in this work is to, instead of directly training on the observed flux, optimise the covariance matrix and noise model on the flux with Lyman- α forest absorption removed (de-forest flux):

$$\begin{aligned} \mathbf{y} &:= \mathbf{y} \circ \exp(+\tau_{\text{eff,HI}}(\mathbf{z})); \\ Y_{ij} &:= Y_{ij} \exp(+\tau_{\text{eff,HI}}(\mathbf{z}))_{ij}. \end{aligned} \tag{2.27}$$

We may write this change into the likelihood:

$$\begin{aligned} p(\mathbf{Y} \circ \exp(+\tau_{\text{eff,HI}}(\mathbf{z})) \mid \boldsymbol{\lambda}, \mathbf{V}, \mathbf{M}, \boldsymbol{\omega}, z_{\text{QSO}}, \mathcal{M}_{\text{-DLA}}) \\ = \prod_{i=1}^{N_{\text{spec}}} \mathcal{N}(\mathbf{y}_i \circ \exp(+\tau_{\text{eff,HI}}(\mathbf{z}_i)); \boldsymbol{\mu}, \mathbf{K} + \boldsymbol{\Omega} + \mathbf{V}_i), \end{aligned} \tag{2.28}$$

where $\boldsymbol{\mu}$ is the mean model from Eq. 2.22. The rest of our optimisation procedure follows the unconstrained optimisation of [3].

We use de-forest fluxes for training as we want our covariance matrix to learn the covariance in the true emission function. The emission function (like our kernel \mathbf{K}) is independent of quasar emission redshift, whereas the absorption noise is not. We only implement the mean forest absorption of [66], so we need an extra term to compensate for the variance of the forest around this mean. We thus still train the redshift- and wavelength-

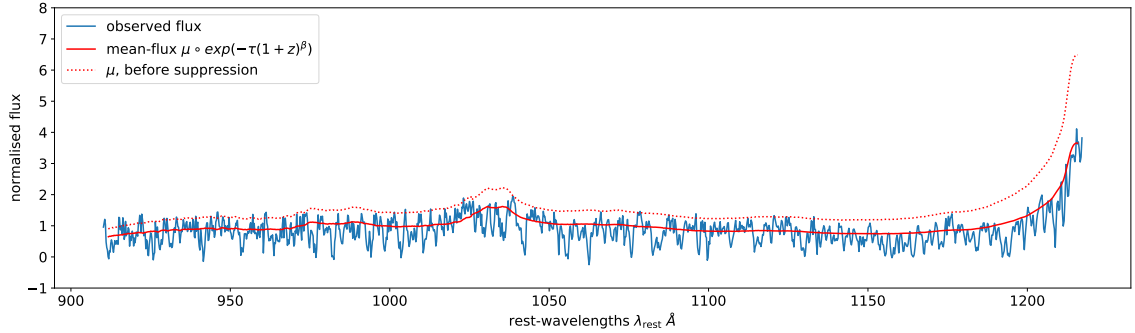


Figure 2.1: The effect of the shift to the GP mean vector from the Lyman- α forest effective optical depth model ($\mu \circ \exp(-\tau_0(1+z)^\beta)$). The dotted red curve shows the mean emission model before application of the forest suppression. The solid red curve is the mean model including the forest suppression.

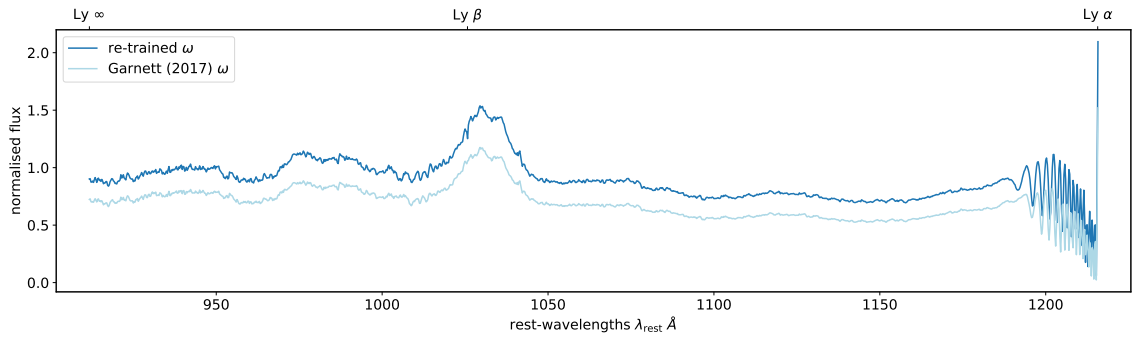


Figure 2.2: The difference between the original pixel-wise noise variance ω [3] and the re-trained ω from Eq. 2.28. The re-trained ω decreases because the fit no longer needs to account for the mean forest absorption.

dependent absorption noise from data. The optimal values we learned for Eq. 2.24 are:

$$c_0 = 0.3050; \tau_0 = 1.6400 \times 10^{-4}; \beta = 5.2714. \quad (2.29)$$

As we might expect, the optimal τ_0 value is smaller than the $\tau_0 = 0.01178$ learned in [3], which implies the effect of the forest is almost removed by applying the Lyman-series forest to the mean model.

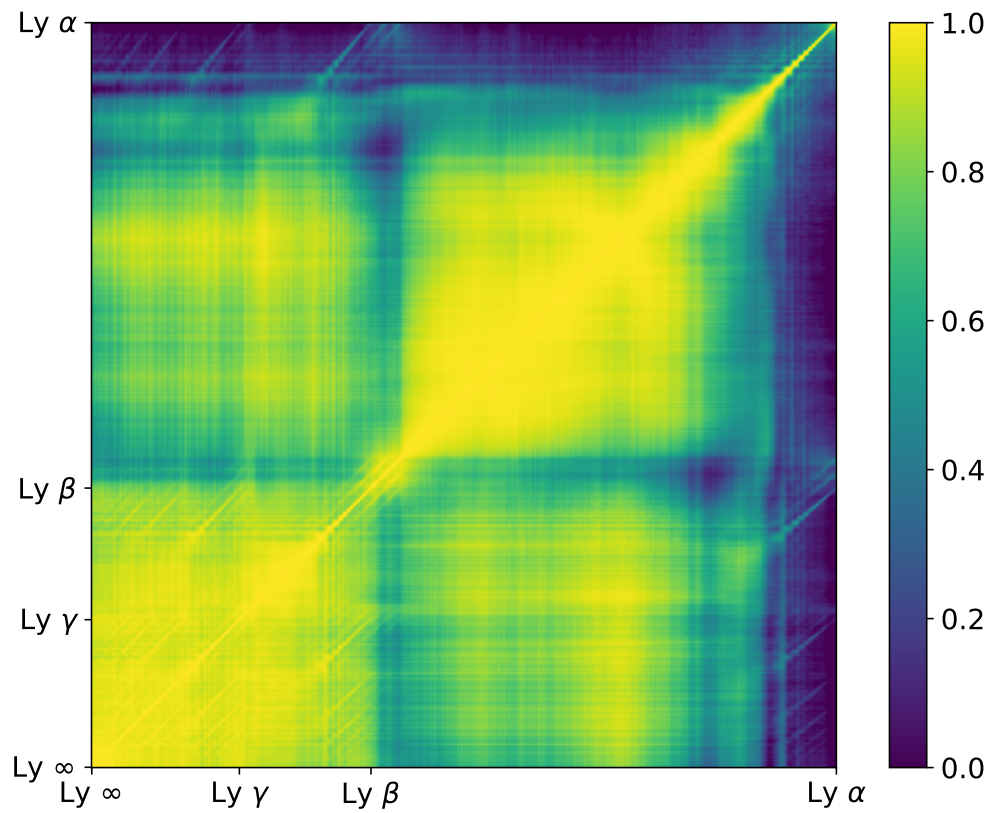


Figure 2.3: The trained covariance matrix \mathbf{M} , which is almost the same as the covariance from [3]. Note that we normalize the diagonal elements to be unity, so this is more like a correlation matrix than a covariance matrix. The values in the matrix are ranging from 0 to 1, representing the correlation between λ and λ' in the QSO emission.

2.6.5 Model evidence

Consider a given QSO observation $\mathcal{D} = (\boldsymbol{\lambda}, \mathbf{y})$ with known observational noise $\nu(\boldsymbol{\lambda})$ and known QSO redshift z_{QSO} . The model evidence for $\mathcal{M}_{\text{-DLA}}$ can be estimated using

$$p(\mathcal{D} \mid \mathcal{M}_{\text{-DLA}}, \nu, z_{\text{QSO}}) \propto p(\mathbf{y} \mid \boldsymbol{\lambda}, \nu, z_{\text{QSO}}, \mathcal{M}_{\text{-DLA}}), \quad (2.30)$$

which is equivalent to evaluating a multivariate Gaussian

$$\begin{aligned} p(\mathbf{y} \mid \boldsymbol{\lambda}, \nu, z_{\text{QSO}}, \mathcal{M}_{\text{-DLA}}) \\ = \mathcal{N}(\mathbf{y}; \boldsymbol{\mu} \circ \exp(-\boldsymbol{\tau}_{\text{eff,HI}}), \mathbf{A}_{\text{F}}(\mathbf{K} + \boldsymbol{\Omega})\mathbf{A}_{\text{F}} + \mathbf{V}). \end{aligned} \quad (2.31)$$

Here $\exp(-\boldsymbol{\tau}_{\text{eff,HI}}) = \text{diag } \mathbf{A}_{\text{F}}$ describes the absorption due to the forest and modifies the mean vector $\boldsymbol{\mu}$, the covariance matrix \mathbf{K} and the noise matrix $\boldsymbol{\Omega}$ to account for the Lyman- α forest effective optical depth.

2.7 A GP Model for QSO Sightlines with Multiple DLAs

In Section 2.6, we learned a GP prior for QSO spectroscopic measurements without any DLAs for our null model $\mathcal{M}_{\text{-DLA}}$. Here we extend the null model $\mathcal{M}_{\text{-DLA}}$ to a model with k intervening DLAs, $\mathcal{M}_{\text{DLA}(k)}$.

Our complete DLA model, \mathcal{M}_{DLA} , will be the union of the models with i DLAs: $\mathcal{M}_{\text{DLA}} = \{\mathcal{M}_{\text{DLA}(i)}\}_{i=1}^k$. We consider only until $k = 4$, as DLAs are rare events and our sample only contains one spectrum with 4 DLAs.

2.7.1 Absorption function

Before we model a quasar spectrum with intervening DLAs, we need to have an absorption profile model for a DLA. Damped Lyman alpha absorbers, or DLAs, are neutral

hydrogen (HI) absorption systems with saturated lines and damping wings in the spectroscopic measurements. Having saturated lines means the column density of the absorbers on the line of sight is high enough to absorb essentially all photons. The damping wings are due to natural broadening in the line.

The optical depth from each Lyman series transition is

$$\tau(\lambda; z_{\text{DLA}}, N_{\text{HI}}) = N_{\text{HI}} \frac{\pi e^2 f_{1u} \lambda_{1u}}{m_e c} \phi(v, b, \gamma), \quad (2.32)$$

where e is the elementary charge, λ_{1u} is the transition wavelength from the $n = 1$ to $n = u$ energy level ($\lambda_{12} = 1215.6701 \text{ \AA}$ for Lyman- α) and f_{1u} is the oscillator strength of the transition. The line profile ϕ is a Voigt profile:

$$\phi(v, b, \gamma) = \int \frac{dv}{\sqrt{2\pi}\sigma_v} \exp(-v^2/2\sigma_v^2) \frac{4\gamma_{\ell u}}{16\pi^2[\nu - (1 - v/c)\nu_{\ell u}]^2 + \gamma_{\ell u}^2}, \quad (2.33)$$

which is a convolution between a Lorentzian line profile and a Gaussian line profile. The σ_v is the one-dimensional velocity dispersion, $\gamma_{\ell u}$ is a parameter for Lorentzian profile, ν is the frequency, and u represents the upper energy level and ℓ represents the lower energy level.

Both profiles are parameterised by the relative velocity v , which means both profiles are distributions in the 1-dimensional velocity space:

$$v = c \left(\frac{\lambda}{\lambda_{1u}} \frac{1}{(1 + z_{\text{DLA}})} - 1 \right). \quad (2.34)$$

The standard deviation of the Gaussian line profile is related to the broadening parameter $b = \sqrt{2}\sigma_v$, and if we assume the broadening is entirely due to thermal motion:

$$b = \sqrt{\frac{2kT}{m_p}}. \quad (2.35)$$

Introducing the damping constant $\Gamma = 6.265 \times 10^8 \text{s}^{-1}$ for Lyman- α , we have the parameter $\gamma_{\ell u}$ to describe the width of the Lorentzian profile

$$\gamma_{\ell u} = \frac{\Gamma \lambda_{\ell u}}{4\pi}. \quad (2.36)$$

Our default DLA profile includes Ly α , Ly β , and Ly γ absorptions. We fix the broadening parameter b by setting $T = 10^4 \text{K}$, which increases the width of the DLA profile by 13 km s^{-1} , small compared to the effect of the Lorentzian wings. Thus, for a given QSO and a true emission function $f(\lambda)$, the function for the observed flux $y(\lambda)$ is

$$y(\lambda) = f(\lambda) \exp(-\tau(\lambda; z_{\text{DLA}}, N_{\text{HI}})) \exp(-\tau_{\text{eff,HI}}(\lambda_{\text{obs}})) + \epsilon, \quad (2.37)$$

where ϵ is additive Gaussian noise including measurement noise and absorption noise.

Suppose we have a DLA at redshift z_{DLA} with column density N_{HI} . We can model the spectrum with an intervening DLA by calculating the DLA absorption function:

$$\mathbf{a} = \exp(-\tau(\boldsymbol{\lambda}; z_{\text{DLA}}, N_{\text{HI}})). \quad (2.38)$$

We apply the absorption function to the GP prior of \mathbf{y} with

$$\begin{aligned} p(\mathbf{y} \mid \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\text{QSO}}, z_{\text{DLA}}, N_{\text{HI}}, \mathcal{M}_{\text{DLA}}) \\ = \mathcal{N}(\mathbf{y}; \mathbf{a} \circ (\mathbf{a}_{\text{F}} \circ \boldsymbol{\mu}), \mathbf{A}(\mathbf{A}_{\text{F}}(\mathbf{K} + \boldsymbol{\Omega})\mathbf{A}_{\text{F}})\mathbf{A} + \mathbf{V}), \end{aligned} \quad (2.39)$$

where $\mathbf{A} = \text{diag } \mathbf{a}$.

For a model with k DLAs with $k \in \mathbb{N}$, we simply take the element-wise product of k absorption functions:

$$\mathbf{a}_{(k)} = \prod_{i=1}^k a(\boldsymbol{\lambda}; z_{\text{DLA}_i}, N_{\text{HI}_i}); \quad (2.40)$$

$$\text{diag } \mathbf{A}_{(k)} = \mathbf{a}_{(k)}.$$

The prior for $\mathcal{M}_{\text{DLA}(k)}$ would therefore be:

$$p(\mathbf{y} \mid \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\text{QSO}}, \{z_{\text{DLA}i}\}_{i=1}^k, \{N_{\text{HI}i}\}_{i=1}^k, \mathcal{M}_{\text{DLA}(k)}) = \mathcal{N}(\mathbf{y}; \mathbf{a}_{(k)} \circ (\mathbf{a}_{\text{F}} \circ \boldsymbol{\mu}), \mathbf{A}_{(k)}(\mathbf{A}_{\text{F}}(\mathbf{K} + \boldsymbol{\Omega})\mathbf{A}_{\text{F}})\mathbf{A}_{(k)} + \mathbf{V}). \quad (2.41)$$

Here we briefly review our notations in Eq. 2.41: $\mathbf{a}_{(k)}$, which is parameterised by $(\{z_{\text{DLA}i}\}_{i=1}^k, \{N_{\text{HI}i}\}_{i=1}^k)$, represents the absorption function with k DLAs in one spectrum. Note that each DLA is parameterised by a pair of $(z_{\text{DLA}}, N_{\text{HI}})$. \mathbf{a}_{F} corresponds to the absorption function from the Lyman series absorptions, which is derived from [66] in the form of Eq. 2.21. The covariance matrix \mathbf{K} and the absorption model $\boldsymbol{\Omega}$ are both learned from data, as described in Section 2.6.4. \mathbf{V} is the noise variance matrix given by the SDSS pipeline, so each sightline would have different \mathbf{V} .

2.7.2 Model Evidence: DLA(1)

The model evidence of our DLA model is given by the integral:

$$p(\mathcal{D} \mid \mathcal{M}_{\text{DLA}(1)}, z_{\text{QSO}}) \propto \int p(\mathbf{y} \mid \boldsymbol{\lambda}, \boldsymbol{\nu}, \boldsymbol{\theta}, z_{\text{QSO}}, \mathcal{M}_{\text{DLA}(1)}) p(\boldsymbol{\theta} \mid z_{\text{QSO}}, \mathcal{M}_{\text{DLA}(1)}) d\boldsymbol{\theta}, \quad (2.42)$$

where we integrated out the parameters, $\boldsymbol{\theta} = (z_{\text{DLA}}, \log_{10} N_{\text{HI}})$, with a given parameter prior $p(\boldsymbol{\theta} \mid z_{\text{QSO}}, \mathcal{M}_{\text{DLA}(1)})$.

However, Eq. 2.42 is intractable, so we approximate it with a quasi-Monte Carlo method (QMC). QMC selects $N = 10\,000$ samples with an approximately uniform spatial distribution from a Halton sequence to calculate the model likelihood, approximating the model evidence by the sample mean:

$$\begin{aligned}
p(D \mid \mathcal{M}_{\text{DLA}(1)}, z_{\text{QSO}}) &\simeq \\
\frac{1}{N} \sum_{i=1}^N p(\mathcal{D} \mid \theta_i, z_{\text{QSO}}, \mathcal{M}_{\text{DLA}(1)}). &
\end{aligned} \tag{2.43}$$

2.7.3 Model evidence: Occam's Razor Effect for DLA(k)

For higher order DLA models, we have to integrate out not only the nuisance parameters of the first DLA model $\mathcal{M}_{\text{DLA}(1)}$, (θ_1) but also the parameters from $\mathcal{M}_{\text{DLA}(2)}$ to $\mathcal{M}_{\text{DLA}(k)}$,

$$\begin{aligned}
p(\mathcal{D} \mid \mathcal{M}_{\text{DLA}(k)}, z_{\text{QSO}}) &\propto \\
\int p(\mathcal{D} \mid \mathcal{M}_{\text{DLA}(k)}, \{\theta_i\}_{i=1}^k) &\times \\
p(\{\theta_i\}_{i=1}^k \mid \mathcal{M}_{\text{DLA}(k)}, \mathcal{D}, z_{\text{QSO}}) & d\{\theta_i\}_{i=1}^k,
\end{aligned} \tag{2.44}$$

which means we are marginalising $\{\theta_i\}_{i=1}^k$ in a parameter space with $2 \times k$ dimensions. The parameter prior of multi-DLAs is a multiplication between a non-informative prior $p(\theta_i \mid \mathcal{M}_{\text{DLA}(1)}, z_{\text{QSO}})$ and the posterior of the $(k - 1)$ multi-DLA model,

$$\begin{aligned}
p(\{\theta_i\}_{i=1}^k \mid \mathcal{M}_{\text{DLA}(k)}, \mathcal{D}, z_{\text{QSO}}) &= \\
p(\{\theta_i\}_{i=1}^{k-1} \mid \mathcal{M}_{\text{DLA}(k-1)}, \mathcal{D}, z_{\text{QSO}}) &p(\theta_k \mid z_{\text{QSO}}, \mathcal{M}_{\text{DLA}(1)}).
\end{aligned} \tag{2.45}$$

We can approximate this integral using the same QMC method. For example, if we want to sample the model evidence for $\mathcal{M}_{\text{DLA}(2)}$, we would need $N = 10\,000$ samples for each parameter dimension $\{\theta_i\}_{i=1}^2$, which results in sampling from two independent Halton

sequences with 10^8 samples in total. If we want to sample up to $\mathcal{M}_{\text{DLA}(k)}$ with N samples for each $\{\theta_i\}$ from $i = 1, \dots, k$, we would need to have:

$$\begin{aligned}
& p(\mathcal{D} \mid \mathcal{M}_{\text{DLA}(k)}, z_{\text{QSO}}) \\
& \simeq \frac{1}{N} \sum_{j^{(1)}=1}^N \frac{1}{N} \sum_{j^{(2)}=1}^N \frac{1}{N} \sum_{j^{(3)}=1}^N \cdots \frac{1}{N} \sum_{j^{(k)}=1}^N
\end{aligned} \tag{2.46}$$

$$p(\mathcal{D} \mid \mathcal{M}_{\text{DLA}(k)}, \{\theta_{1j^{(1)}}, \theta_{2j^{(2)}}, \theta_{3j^{(3)}}, \dots, \theta_{kj^{(k)}}\}, z_{\text{QSO}}),$$

where $\{j^{(1)}, j^{(2)}, j^{(3)}, \dots, j^{(k)}\}$ indicate the indices of QMC samples. The above Eq. 2.46 is thus in principle evaluated with N^k samples.

In practice, we only sample $N = 10\,000$ points from $p(\{\theta_i\}_{i=1}^k \mid \mathcal{M}_{\text{DLA}(k)}, \mathcal{D}, z_{\text{QSO}})$ instead of sampling N^k points, as a uniform sampling of the first DLA model may be reweighted to cover parameter space for the higher order models. A N^{k-1} factor of normalisation is thus left behind in the summation,

$$\begin{aligned}
& p(\mathcal{D} \mid \mathcal{M}_{\text{DLA}(k)}, z_{\text{QSO}}) \\
& \simeq \frac{1}{N^k} \sum_{j=1}^N p(\mathcal{D} \mid \mathcal{M}_{\text{DLA}(k)}, \{\theta_{ij}\}_{i=1}^k, z_{\text{QSO}}) \\
& \simeq \frac{1}{N^{k-1}} \left(\frac{1}{N} \sum_{j=1}^N p(\mathcal{D} \mid \mathcal{M}_{\text{DLA}(k)}, \{\theta_{ij}\}_{i=1}^k, z_{\text{QSO}}) \right) \\
& \simeq \frac{1}{N^{k-1}} \text{mean}_j \left(p(\mathcal{D} \mid \mathcal{M}_{\text{DLA}(k)}, \{\theta_{ij}\}_{i=1}^k, z_{\text{QSO}}) \right).
\end{aligned} \tag{2.47}$$

The additional $\frac{1}{N^{k-1}}$ factor penalises models with more parameters than needed, and can be viewed as an implementation of Occam's razor. This Occam's razor effect is caused by the fact that all probability distributions have to be normalised to unity. A model with more parameters, which means having a wider distribution in the likelihood space, results in a bigger normalisation factor.

The motivation for us to draw N samples from the multi-DLA likelihood function $p(\{\theta_i\}_{i=1}^k \mid \mathcal{M}_{\text{DLA}(k)}, \mathcal{D}, z_{\text{QSO}})$ is that we believe the prior density we took from the posterior density of $\mathcal{M}_{\text{DLA}(k-1)}$ is representative enough even without N^k samples. For example, if we have two peaks in our likelihood density $p(\mathcal{D} \mid \mathcal{M}_{\text{DLA}(1)}, \theta_1, z_{\text{QSO}})$, we expect the sampling for θ_2 in $p(\mathcal{D} \mid \mathcal{M}_{\text{DLA}(2)}, \{\theta_1, \theta_2\}, z_{\text{QSO}})$ would concentrate on sampling the density of the first highest peak in $p(\mathcal{D} \mid \mathcal{M}_{\text{DLA}(1)}, \theta_1, z_{\text{QSO}})$ density. Similarly, while we are sampling for $\mathcal{M}_{\text{DLA}(3)}$, we expect θ_3 and θ_2 would cover the first and the second-highest peaks.

To avoid multi-DLAs overlapping with each other, we inject a dependence between any pair of z_{DLA} parameters. Specifically, if any pair of z_{DLAs} have a relative velocity smaller than 3000 km s^{-1} , then we set the likelihood of this sample to NaN.

2.7.4 Additional penalty for DLAs and sub-DLAs

In Section 2.7.3, we apply a penalty, Occam’s razor, to regularise DLA models using more parameters than needed. This effect is due to the normalization (to unity) of the evidence.

In a similar fashion, and for a similar reason, we apply an additional regularisation factor between the non-DLA and DLA models (including sub-DLAs). This additional factor ensures that when both models are a poor fit to a particular observational spectrum, the code prefers the non-DLA model, rather than preferring the model with more parameters and thus greater fitting freedom. We directly inject this Occam’s razor factor in the model

selection:

$$\Pr(\mathcal{M}_{\text{DLA}} | \mathcal{D}) = \frac{\Pr(\mathcal{M}_{\text{DLA}})p(\mathcal{D} | \mathcal{M}_{\text{DLA}})^{\frac{1}{N}}}{\left(\Pr(\mathcal{M}_{\text{DLA}})p(\mathcal{D} | \mathcal{M}_{\text{DLA}}) + \Pr(\mathcal{M}_{\text{sub}})p(\mathcal{D} | \mathcal{M}_{\text{sub}}) \right)^{\frac{1}{N}} + \Pr(\mathcal{M}_{\text{-DLA}} | \mathcal{D})}, \quad (2.48)$$

where $N = 10^4$ is the number of samples we used to approximate the parameterised likelihood functions. We evaluated the impact of this regularization factor on the area under the curve (AUC) in the receiver-operating characteristics (ROC) plot.⁵ For $N = 10^4$, the AUC changed from 0.949 to 0.960. We considered other penalty values and found that the AUC increased up to $N = 10^4$ and then plateaued.

In addition, we found by examining specific examples that this penalty regularized a relatively common incorrect DLA detection: finding objects in short, very noisy low redshift ($z \sim 2.2$) spectra. In these spectra our earlier model would prefer the DLA model purely because of its large parameter freedom. In particular a high column density DLA, large enough that the damping wings exceed the width of the spectrum, would be preferred. Such a fit exploits a degeneracy in the model between the mean observed flux and the DLA column density when the spectrum is shorter than the putative DLA. The Occam's razor penalty avoids these spurious fits by penalising the extra parametric freedom in the DLA model.

⁵See Section 2.11.1 for how we compute our ROC plot.

2.7.5 Parameter prior

Here we briefly recap the priors on model parameters chosen in [3]. Suppose we want to make an inference for the column density and redshift of an absorber $\theta = (N_{\text{HI}}, z_{\text{DLA}})$ from a given spectroscopic observation, the joint density for the parameter prior would be

$$p(\theta \mid z_{\text{QSO}}, \mathcal{M}_{\text{DLA}(1)}) = p(N_{\text{HI}}, z_{\text{DLA}} \mid z_{\text{QSO}}, \mathcal{M}_{\text{DLA}(1)}). \quad (2.49)$$

Suppose the absorber redshift and the column density are conditionally independent and the column density is independent of the quasar redshift z_{QSO} :

$$\begin{aligned} p(\theta \mid z_{\text{QSO}}, \mathcal{M}_{\text{DLA}(1)}) &= \\ & p(z_{\text{DLA}} \mid z_{\text{QSO}}, \mathcal{M}_{\text{DLA}(1)}) p(N_{\text{HI}} \mid \mathcal{M}_{\text{DLA}(1)}) \end{aligned} \quad (2.50)$$

We set a bounded uniform prior density for the absorber redshift z_{DLA} :

$$p(z_{\text{DLA}} \mid z_{\text{QSO}}, \mathcal{M}_{\text{DLA}(1)}) = \mathcal{U}[z_{\text{min}}, z_{\text{max}}], \quad (2.51)$$

where we define the finite prior range to be

$$z_{\text{min}} = \max \left\{ \begin{array}{l} \frac{\lambda_{\text{Ly}\infty}}{\lambda_{\text{Ly}\alpha}} (1 + z_{\text{QSO}}) - 1 + 3000 \text{ km s}^{-1}/c \\ \frac{\min \lambda_{\text{obs}}}{\lambda_{\text{Ly}\alpha}} - 1 \end{array} \right. \quad (2.52)$$

$$z_{\text{max}} = z_{\text{QSO}} - 3000 \text{ km s}^{-1}/c; \quad (2.53)$$

which means we have a prior belief that the center of the absorber is within the observed wavelengths. The range of observed wavelengths is either from Ly ∞ to Ly α of the quasar rest-frame ($\lambda_{\text{rest}} \in [911.75 \text{ \AA}, 1216.75 \text{ \AA}]$) or from the minimum observed wavelength to Ly α . We also apply a conservative cutoff of 3000 km s $^{-1}$ near to Ly ∞ and Ly α . The

-3000 km s^{-1} cutoff for z_{max} helps to avoid proximity ionisation effects due to the quasar radiation field. Furthermore, the $+3000 \text{ km s}^{-1}$ cutoff for z_{min} avoids a potentially incorrect measurement for z_{QSO} . An underestimated z_{QSO} can produce a Lyman-limit trough within the region of the quasar expected to contain only Lyman-series absorption, and the code can incorrectly interpret this as a DLA.

For the column density prior, we follow [3]. We first estimate the density of DLAs column density $p(N_{\text{HI}} | \mathcal{M}_{\text{DLA}})$ using the BOSS DR9 Lyman- α forest sample. We choose to put our prior on the base-10 logarithm of the column density $\log_{10} N_{\text{HI}}$ due to the large dynamic range of DLA column densities in SDSS DR9 samples.

We thus estimate the density of logarithm column densities $p(\log_{10} N_{\text{HI}} | \mathcal{M}_{\text{DLA}(1)})$ using univariate Gaussian kernels on the reported $\log_{10} N_{\text{HI}}$ values in DR9 samples. Column densities from DLAs in DR9 with $N_{\text{DLA}} = 5854$ are used to non-parametrically estimate the logarithm N_{HI} prior density, with:

$$\begin{aligned} p_{\text{KDE}}(\log_{10} N_{\text{HI}} | \mathcal{M}_{\text{DLA}(1)}) \\ = \frac{1}{N_{\text{DLA}}} \sum_{i=1}^{N_{\text{DLA}}} \mathcal{N}(\log_{10} N_{\text{HI}}; l_i, \sigma^2), \end{aligned} \tag{2.54}$$

where l_i is the logarithm column density $\log_{10} N_{\text{HI}}$ of the i^{th} sample. The bandwidth σ^2 is selected to be the optimal value for a normal distribution, which is the default setting for MATLAB.

We further simplify the non-parametric estimate into a parametric form with:

$$\begin{aligned} p_{\text{KDE}}(\log_{10} N_{\text{HI}} = N | \mathcal{M}_{\text{DLA}(1)}) \simeq \\ q(\log_{10} N_{\text{HI}} = N) \propto \exp(aN^2 + bN + c); \end{aligned} \tag{2.55}$$

where the parameters (a, b, c) for the quadratic function are fitted via standard least-squared fitting to the non-parametric estimate of density $p_{\text{KDE}}(\log_{10} N_{\text{HI}} | \mathcal{M}_{\text{DLA}(1)})$ with the range $\log_{10} N_{\text{HI}} \in [20, 22]$. The optimal values for the quadratic terms were:

$$a = -1.2695; b = 50.863; c = -509.33; \quad (2.56)$$

Note that we have the same values as in [3].

Finally, we choose to be conservative about the data-driven column density prior. We thus take a mixture of a non-informative log-normal prior with the data-driven prior to make a non-restrictive prior on a large dynamical range:

$$\begin{aligned} p(\log_{10} N_{\text{HI}} | \mathcal{M}_{\text{DLA}(1)}) \\ = \alpha q(\log_{10} N_{\text{HI}} = N) + (1 - \alpha) \mathcal{U}[20, 23]. \end{aligned} \quad (2.57)$$

Here we choose the mixture coefficient $\alpha = 0.97$, which favours the data-driven prior. We still include a small component of a non-informative prior so that we are able to detect DLAs with a larger column density than in the training set, if any are present in the larger DR12 sample. Note that $\alpha = 0.97$ is 7% higher than the coefficient chosen in [3], which was $\alpha = 0.90$. Our previous prior slightly over-estimated the number of very large DLAs.

2.7.6 Sub-DLA parameter prior

As reported in [62], the column density distribution function (CDDF) exhibited an edge feature: an over-detection of DLAs at low column densities ($\sim 10^{20} \text{ cm}^{-2}$). This did not affect the statistical properties of DLAs as we restrict column density to $N_{\text{HI}} \geq 10^{20.3} \text{ cm}^{-2}$ for both line densities (dN/dX) and total column densities (Ω_{DLA}). However, to make our

method more robust, here we describe a complementary method to avoid over-estimating the number of low column density absorbers.

The excess of DLAs at $\sim 10^{20} \text{ cm}^{-2}$ is due to our model excluding lower column density absorbers such as sub-DLAs. Since we limited our column density prior of DLAs to be larger than 10^{20} cm^{-2} , the code cannot correctly classify a sub-DLA. Instead it correctly notes that a sub-DLA spectrum is more likely to be a DLA with a minimal column density than an unabsorbed spectrum.

To resolve our ignorance, we introduce an alternative model \mathcal{M}_{sub} to account the model posterior of those low column density absorbers in our Bayesian model selection. The likelihood function we used for sub-DLAs is identical to the one we built for DLA model $\mathcal{M}_{\text{DLA}(1)}$ in Eq. 2.39 but has a different parameter prior on the column densities $p(\log_{10} N_{\text{HI}} | \mathcal{M}_{\text{sub}})$. We restricted our prior belief of sub-DLAs to be within the range $\log_{10} N_{\text{HI}} \in [19.5, 20]$, and, as we do not have a catalogue of sub-DLAs for learning the prior density, we put a uniform prior on $\log_{10} N_{\text{HI}}$:

$$p(\log_{10} N_{\text{HI}} | \mathcal{M}_{\text{sub}}) = \mathcal{U}[19.5, 20]. \quad (2.58)$$

We place a lower cutoff at $\log_{10} N_{\text{HI}} = 19.5$ because the relatively noisy SDSS data offers limited evidence for absorbers with column densities lower than this limit.

2.8 Model Priors

Bayesian model selection allows us to combine prior information with evidence from the data-driven model to obtain a posterior belief about the detection of DLAs $p(\mathcal{M}_{\text{DLA}} | \mathcal{D})$ using Bayes' rule. For a given spectroscopic observation \mathcal{D} , we already have the ability to

compute the model evidence for a DLA ($p(\mathcal{D} | \mathcal{M}_{\text{DLA}})$) and no DLA ($p(\mathcal{D} | \mathcal{M}_{\text{-DLA}})$). However, to compute the model posteriors, we need to specify our prior beliefs in these models. Here we approximate our prior belief $\text{Pr}(\mathcal{M}_{\text{DLA}})$ using the SDSS DR9 DLA catalogue.

Consider a QSO observation $\mathcal{D} = (\boldsymbol{\lambda}, \mathbf{y})$ at z_{QSO} . We want to find our prior belief that \mathcal{D} contains a DLA. We count the fraction of QSO sightlines in the training set containing DLAs with redshift less than $z_{\text{QSO}} + z'$, where $z' = 30\,000 \text{ km s}^{-1}/c$ is a small constant. If N is the number of QSO sightlines with redshift less than $z_{\text{QSO}} + z'$, and M is the number of sightlines in this set containing DLAs in the quasar rest-frame wavelengths range we search, then our empirical prior for \mathcal{M}_{DLA} is:

$$\text{Pr}(\mathcal{M}_{\text{DLA}} | z_{\text{QSO}}) = \frac{M}{N}. \quad (2.59)$$

We can break down our DLA prior $\text{Pr}(\mathcal{M}_{\text{DLA}} | z_{\text{QSO}})$ for multiple DLAs in a QSO sightline $\text{Pr}(\mathcal{M}_{\text{DLA}(k)} | z_{\text{QSO}})$ via:

$$\text{Pr}(\mathcal{M}_{\text{DLA}(k)} | z_{\text{QSO}}) \simeq \left(\frac{M}{N}\right)^k - \left(\frac{M}{N}\right)^{k+1}. \quad (2.60)$$

For example, $\frac{M}{N}$ represents our prior belief of having at least one DLA in the sightline, and $(\frac{M}{N})^2$ represents having at least two DLAs. $\frac{M}{N} - (\frac{M}{N})^2$ is thus our prior belief of having exactly one DLA at the sightline.

2.8.1 Sub-DLA model prior

The column density distribution function (CDDF) of [62] exhibited an edge effect at $\log_{10} N_{\text{HI}} \sim 20$ due to a lack of sampling at lower column densities. We thus construct an alternative model for lower column density absorbers (sub-DLAs, DLAs' lower column

density cousins) to regularise DLA detections. We use the same GP likelihood function as the DLA model \mathcal{M}_{DLA} to compute our sub-DLA model evidence $p(\mathcal{D} | \mathcal{M}_{\text{sub}})$ but with a different column density prior $p(\log_{10} N_{\text{HI}} | \mathcal{M}_{\text{sub}})$.

There is no sub-DLA catalogue available for us to estimate the empirical prior directly. We, therefore, approximate our sub-DLA model prior by rescaling our DLA model prior:

$$\Pr(\mathcal{M}_{\text{sub}} | z_{\text{QSO}}) \propto \Pr(\mathcal{M}_{\text{DLA}} | z_{\text{QSO}}), \quad (2.61)$$

and we require our prior beliefs to sum to unity:

$$\begin{aligned} \Pr(\mathcal{M}_{\text{-DLA}} | z_{\text{QSO}}) + \Pr(\mathcal{M}_{\text{sub}} | z_{\text{QSO}}) \\ + \Pr(\mathcal{M}_{\text{DLA}} | z_{\text{QSO}}) = 1. \end{aligned} \quad (2.62)$$

The scaling factor between the DLA prior and sub-DLA prior should depend on our prior probability density of the column density of the absorbers. Here we assume the density of sub-DLA $\log_{10} N_{\text{HI}}$ is an uniform density with a finite range of $\log_{10} N_{\text{HI}} \in [19.5, 20]$. We believe there are more sub-DLAs than DLAs as high column density systems are generally rarer. We thus assume the probability of finding sub-DLAs at a given $\log_{10} N_{\text{HI}}$ is the same as the probability of finding DLAs at the most probable $\log_{10} N_{\text{HI}}$, which is:

$$\begin{aligned} p(\log_{10} N_{\text{HI}} = N | \{\mathcal{M}_{\text{DLA}}, \mathcal{M}_{\text{sub}}\}) = \\ \alpha q(N | \mathcal{M}_{\text{DLA}}) \mathbb{I}_{(20,23)}(N) \\ + \alpha \max(q(N | \mathcal{M}_{\text{DLA}})) \mathbb{I}_{(19.5,20)}(N) \\ + (1 - \alpha) \mathcal{U}[19.5, 23]. \end{aligned} \quad (2.63)$$

Since $q(N | \mathcal{M}_{\text{DLA}})$ has a simple quadratic functional form, we can solve the maximum value analytically, which is $\max(q(N | \mathcal{M}_{\text{DLA}})) \simeq q(N = 20.03 | \mathcal{M}_{\text{DLA}})$.

We thus can use our prior knowledge about the logarithm of column densities for different absorbers to rescale model priors:

$$\Pr(\mathcal{M}_{\text{sub}} | z_{\text{QSO}}) = \frac{Z_{\text{sub}}}{Z_{\text{DLA}}} \Pr(\mathcal{M}_{\text{DLA}} | z_{\text{QSO}}), \quad (2.64)$$

where the scaling factor is:

$$\frac{Z_{\text{sub}}}{Z_{\text{DLA}}} = \frac{\int_{19.5}^{20} p(N | \{\mathcal{M}_{\text{DLA}}, \mathcal{M}_{\text{sub}}\}) dN}{\int_{20}^{23} p(N | \{\mathcal{M}_{\text{DLA}}, \mathcal{M}_{\text{sub}}\}) dN}, \quad (2.65)$$

which is the odds of finding absorbers in the range of $\log_{10} N_{\text{HI}} \in [19.5, 20]$ compared to finding absorbers in $\log_{10} N_{\text{HI}} \in [20, 23]$. Note that we will treat the model posteriors of the sub-DLA model as part of the non-detections of DLAs in the following analysis sections.

2.9 Catalogue

The original parameter prior in [3] is uniformly distributed in z_{DLA} between the Lyman limit ($\lambda_{\text{rest}} = 911.76 \text{ \AA}$) and the Ly α emission of the quasar. In [62], we chose the minimum value of z_{DLA} to be at the Ly β emission line of the quasar rest-frame (instead of the Lyman limit) to avoid the region containing unmodelled Ly β forest. The primary reason for this was that the original absorption noise model did not include Ly β absorption. With the updated model from Eq. 2.24 we are able to model this absorption. Hence, for our new public catalogue, we sample z_{DLA} to be from Ly ∞ to Ly α in the quasar rest-frame and for the convenience of future investigators our public catalogue contains DLAs throughout the whole available spectrum, including Ly β to Ly ∞ . There is still some contamination in the blue end of high redshift spectra from the Ly β forest and occasional Lyman breaks from a misestimated quasar redshift. In practice we shall see that the contamination is not severe

except for $z_{\text{DLA}} > 3.75$. However, in the interest of obtaining as reliable DLA statistics as possible, when computing population statistics we consider only 3 000 Å redward of Ly β to 3 000 Å blueward of Ly α in the quasar rest frame.

In this paper, we computed the posterior probability of $\mathcal{M}_{\text{-DLA}}$ to $\mathcal{M}_{\text{DLA}(k)}$ models. For each spectrum, the catalogue includes:

- The range of redshift DLA searched $[z_{\text{min}}, z_{\text{max}}]$,
- The log model priors from $\log \Pr(\mathcal{M}_{\text{-DLA}} | z_{\text{QSO}})$, $\log \Pr(\mathcal{M}_{\text{sub}} | z_{\text{QSO}})$, to $\log \Pr(\{\mathcal{M}_{\text{DLA}(i)}\}_{i=1}^k | z_{\text{QSO}})$,
- The log model evidence $\log p(\mathbf{y} | \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\text{QSO}}, \mathcal{M})$, for each model we considered,
- The model posterior $\Pr(\mathcal{M} | \mathcal{D}, z_{\text{QSO}})$, for each model we considered,
- The probability of having DLAs $\Pr(\{\mathcal{M}_{\text{DLA}}\} | \mathcal{D}, z_{\text{QSO}})$,
- The probability of having zero DLAs $\Pr(\mathcal{M}_{\text{-DLA}} | \mathcal{D}, z_{\text{QSO}})$,
- The sample log likelihoods $\log p(\mathbf{y} | \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\text{QSO}}, \{z_{\text{DLA}(i)}\}_{i=1}^k, \{\log_{10} N_{\text{HI}(i)}\}_{i=1}^k, \mathcal{M}_{\text{DLA}(k)})$ for all DLA models we considered, and
- The maximum a posteriori (MAP) values of all DLA models we considered.

The full catalogue will be available alongside the paper: http://tiny.cc/multidla_catalog_gp_dr12q. The code to reproduce the entire catalogue will be posted in https://github.com/rmgarnett/gp_dla_detection/tree/master/multi_dlas.

2.9.1 Running Time

We ran our multi-DLA code on UCR’s High-Performance Computing Center (HPCC) and Amazon Elastic Compute Cloud (EC2). The computation of model posteriors of $\mathcal{M}_{\text{-DLA}}, \mathcal{M}_{\text{sub}}, \{\mathcal{M}_{\text{DLA}(i)}\}_{i=1}^4$ takes 7-11 seconds per spectrum on a 32-core node in HPCC and 3-5 seconds on a 48-core machine in EC2. For each spectrum, we have to compute $10\,000 * 5 + 1$ log likelihoods in the form of Eq. 2.11. If we scale the sample size from $N = 10\,000$ to $100\,000$, it costs 38-52 seconds on a 32-core node in HPCC.

2.10 Example spectra

Here we show a few examples of the fitted GP priors, both to compare our method to others and to aid the reader in understanding concretely how our method works.

Figure 2.5 shows an example where our new code detects three DLAs in a single spectrum, while the older model detected only one DLA as shown in Figure 2.4. Because the mean quasar model includes a redshift dependent term corresponding to intervening absorbers, our new mean model can now fit the mean observed quasar spectrum better. Although we show the sample likelihoods in the $\mathcal{M}_{\text{DLA}(1)}$ parameter space, our current code finds these three DLAs in the six dimensional parameter space $(z_{\text{DLA}(i)}, \log_{10} N_{\text{HI}(i)})_{i=1}^3$.

In Figure 2.6 we show a representative sample of a very common case in our $\mathcal{M}_{\text{DLA}(1)}$ model. The red curve represents our GP prior on the given spectrum, and the orange curve is the curve with fitted DLAs provided by the CNN model presented in [4]⁶.

⁶We used the version of [4]’s catalogue listed in the published paper and found on Google Drive at <https://tinyurl.com/cnn-dlas>.

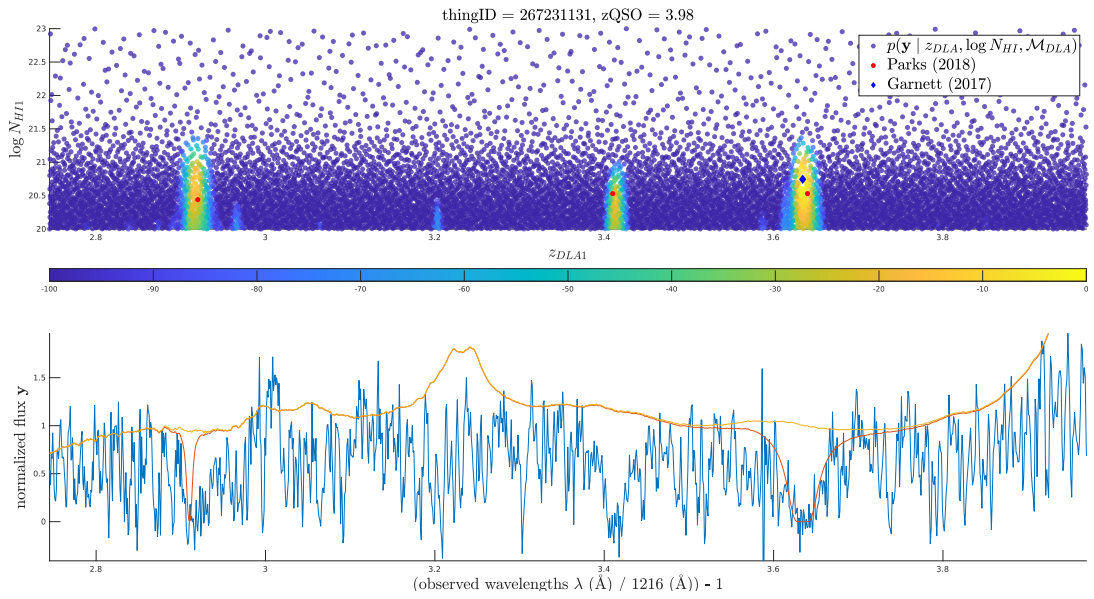


Figure 2.4: An example of finding DLAs using [3]’s model. Here we use the single-DLA per spectrum version of Garnett’s model. **Upper:** sample likelihoods $p(\mathbf{y} \mid \theta, \mathcal{M}_{\text{DLA}})$ in the parameter space $\theta = (z_{\text{DLA}}, \log_{10} N_{\text{HI}})$. Red dots show the DLAs predicted by [4], and the blue squares show the maximum a posteriori (MAP) prediction of the [3]. **Bottom:** the observed spectrum (blue), the null model GP prior (orange), and the DLA model GP prior (Red). So that the upper and bottom panels have the same x-axis, we rescale the observed wavelength to absorber redshift.

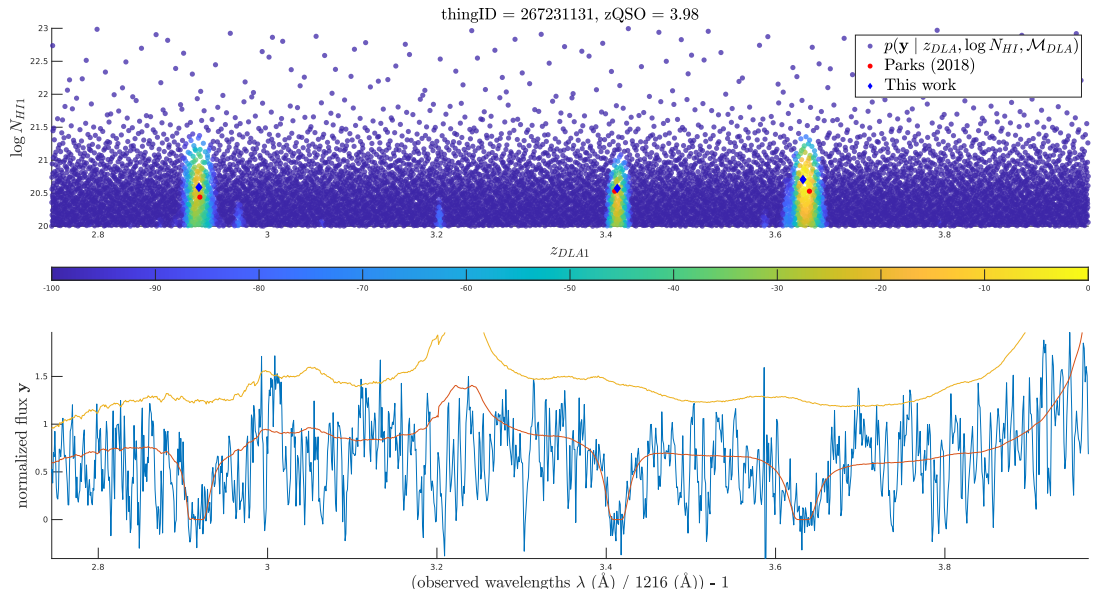


Figure 2.5: The same spectrum as Figure 2.4, but using the multi-DLA model reported in this paper. **Upper:** sample likelihoods $p(\mathbf{y} | \theta, \mathcal{M}_{DLA})$ in the parameter space of the $\mathcal{M}_{DLA(1)}$, with $\theta = (z_{DLA}, \log_{10} N_{HI})$. **Bottom:** the observed spectrum (blue), the null model GP prior before the suppression of effective optical depth (orange), and the multi-DLA GP prior (Red). The orange curve is slightly higher than the one in Figure 2.4 because we try to model the mean spectrum before the forest. However, the DLA quasar model (red curve) matches the level of the observed mean flux better than Figure 2.4 due to the inclusion of a term for the effective optical depth of the Lyman- α forest.

We found [4] underestimated the column densities of the underlying DLAs in the spectra due to not modelling Lyman- β and Lyman- γ absorption in DLAs, while the predictions of N_{HI} in our model are more robust since the predicted N_{HI} is constrained by α , β , and γ absorption. In the spectrum, Lyman- β absorption is clearly visible (although noisy). In Figure 2.6, [4] has actually mistaken the Ly γ absorption line of the DLA for another, weaker, DLA. This demonstrates again the necessity of including other Lyman-series members in the modelling steps. Since [4] broke down each spectrum into pieces during the training and testing phases, it is impossible for the CNN to use knowledge about other Lyman series lines associated with the DLAs. Another example, from a spectrum where we detect 2 DLAs and the CNN detects 4 (although at low significance) is shown in Figure 2.7. Here the CNN has mistaken both the Ly β and Ly γ absorption associated with the large DLA at $z \sim 3$ (near the quasar rest frame) for separate DLAs at $z = 2.4$ and $z = 2.22$ respectively. The large DLA at $z \sim 3$ has been split into two of reduced column density and reduced confidence. The CNN has also missed the second genuine DLA at a rest-frame wavelength of 1025\AA , presumably due to the proximity of an emission line. Our code, able to model the higher order Lyman lines, has used the information contained within them to correctly classify this spectrum as containing two DLAs.

Figure 2.9 shows an example which was problematic in both the models of [3] and [4]. This is an extremely noisy spectrum, where the length of the spectrum is not long enough for us to contain higher order Ly-series absorption or even to see the full length of the putative Lyman- α absorption. By eye, distinguishing a DLA from the noise is challenging.

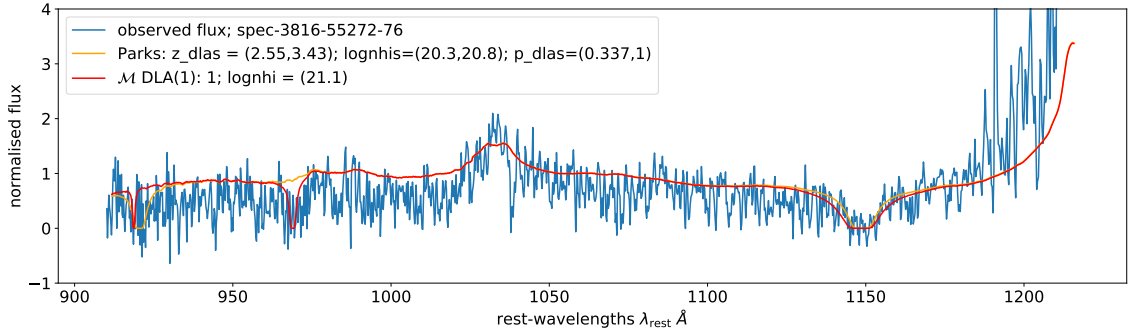


Figure 2.6: **Blue:** the normalised observed flux. The spectral ID represents `spec-plate-mjd-fiber_id`. **Yellow:** Parks' predictions on top of our null model. Our model predicts only one DLA while the CNN model in [4] predicts two DLAs. One of the DLAs predicted by [4] is coincident with the Ly γ absorption from our predicted DLA. `z_dla` corresponds to the DLA redshifts reported in Parks' catalogue, and `lognhi` corresponds to the column density estimations of Parks' catalogue. `p_dla` is the `dla_confidence` reported in Parks. **Red:** Our current model with the highest model posterior and the MAPs of column densities. In this spectrum, we show that it is crucial to include Ly β and Ly γ absorption from the DLA in the DLA profile. It not only helps to localize the DLA, but it also predicts N_{HI} more accurately using information from the Ly β region. The blue line shows the observed flux, the red curve is our multi-DLA GP prior, and the orange curve shows the predicted DLAs from [4] subtracted from our mean model.

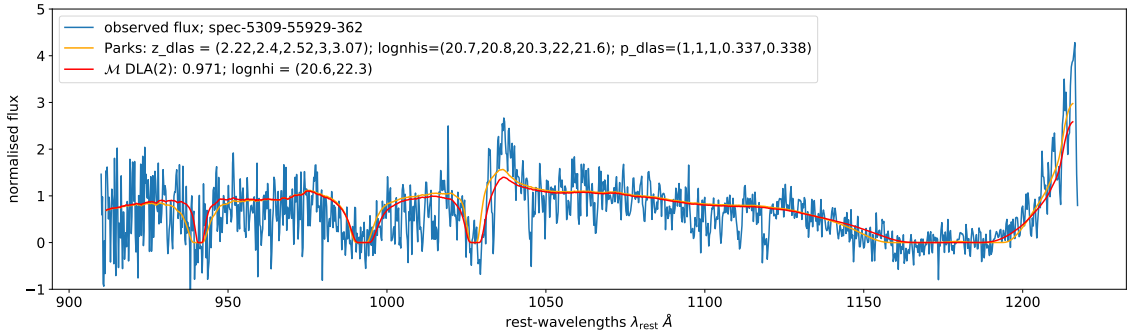


Figure 2.7: A spectrum in which we detect two DLAs. **Blue:** Normalised flux. **Red:** GP mean model with two intervening DLAs. **Yellow:** The predictions from Parks' catalogue. **Pink:** The MAP prediction of [3] on top of the GP mean model without mean flux suppression. The model posterior from [3] is listed in the legend (1) with the MAP value of $\log_{10} N_{\text{HI}}$. The column density estimate for the DLA near $\lambda_{\text{rest}} = 1025\text{\AA}$ has large uncertainty (see Figure 2.8). It is thus possible that this DLA could be a sub-DLA, as preferred by [4].

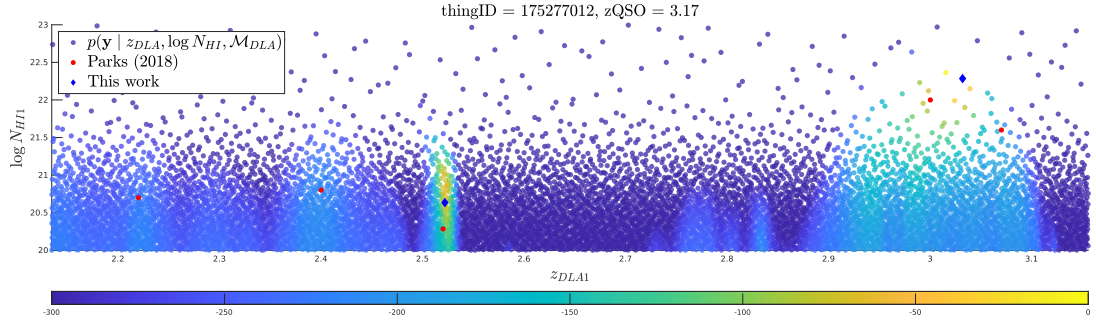


Figure 2.8: The log sample likelihoods for the DLA model of the spectrum shown in Figure 2.7, normalised to range from $-\infty$ to 0. The DLA at $z_{\text{DLA}} \sim 2.52$ could be a sub-DLA (as preferred by [4]), as the $\log_{10} N_{\text{HI}}$ estimate is uncertain. However, we found that the 2-DLA model posterior $\log p(\mathcal{M}_{\text{DLA}(2)} | \mathbf{y}, \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\text{QSO}}) = -638$ is still higher than the model posterior from combining 1-DLA and 1-sub-DLA, which is $\log p(\mathcal{M}_{\text{DLA}(1)} + \mathcal{M}_{\text{sub}} | \mathbf{y}, \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\text{QSO}}) = -691.47$.

If we examine the sample likelihoods from our model (shown in Figure 2.10), we see that the DLA posterior probability is spread over the whole of parameter space; in other words, all models are a poor fit for this noise-dominated spectrum. The model selection is thus really comparing the likelihood function on the basis of how much parametric freedom it has. After implementing the additional Occam’s razor factor between the null model and parameterised models (DLAs and sub-DLAs) described in Section 2.7.4, we found that the large DLA fitted to the noisy short spectrum by [3] was no longer preferred. This indicates that our Occam’s razor penalty is effective. As shown in Figure 2.16, Ω_{DLA} at low redshifts is lower than the measurements in [62], indicating that this class of error is common enough to have a measurable effect on the column density function. We checked that the addition of the Occam’s razor penalty, Ω_{DLA} is insensitive to the noise threshold used when selecting the spectra for our sample.

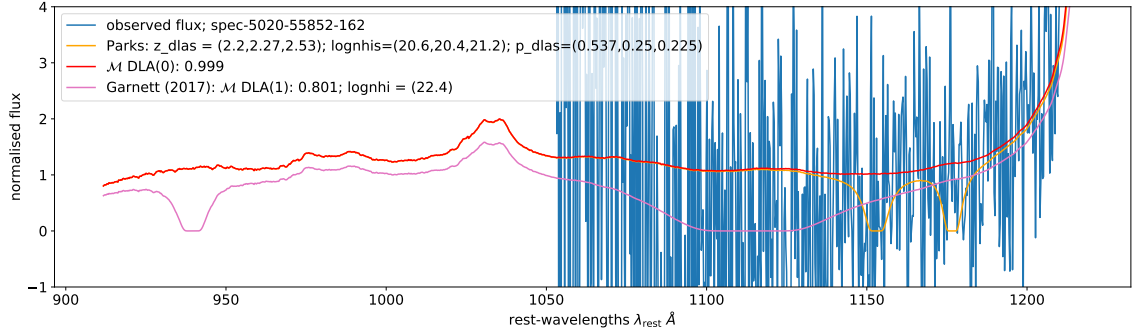


Figure 2.9: A noisy spectrum at $z_{\text{QSO}} = 2.378$ fitted with a large DLA by [3]. **Red:** The model presented in this paper predicts no DLA detection in this spectrum. **Pink:** The MAP prediction of [3] on top the GP mean model without the mean-flux suppression. **Gold:** The prediction of [4] subtracted from our mean model. Note that [4] also indicates a detection of a DLA at $z_{\text{DLA}} = 2.53$, but outside the range of this spectrum.

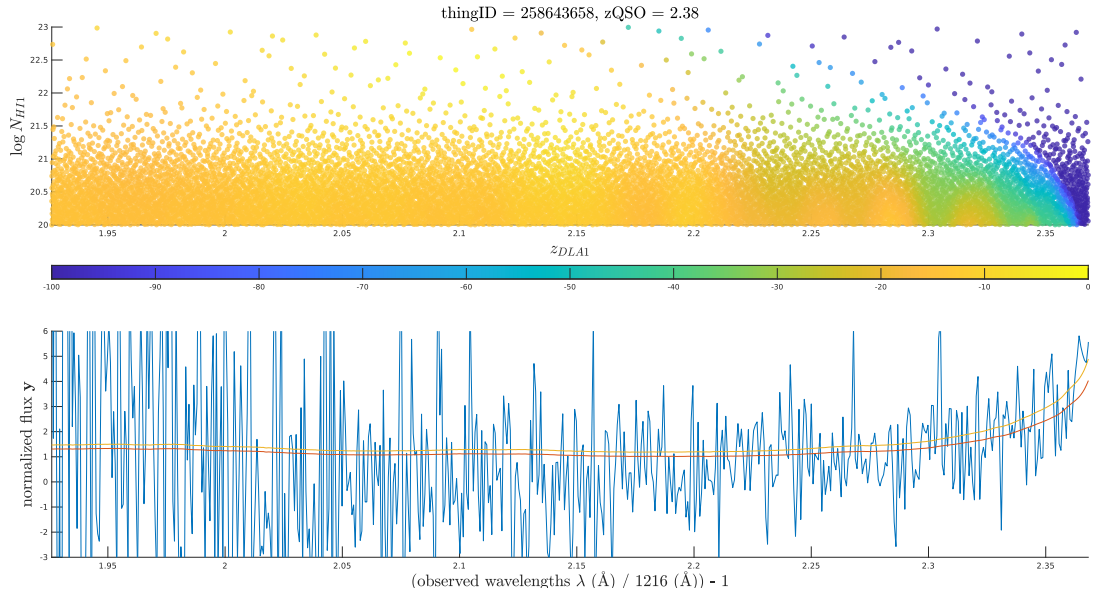


Figure 2.10: **Top:** The sample likelihoods of the spectrum shown in Figure 2.9. The colour bar indicates the normalised log likelihoods ranging from $-\infty$ to 0. **Bottom:** The orange curve indicates the GP mean model before mean-flux suppression, the red curve represents the mean model after suppression, and the blue line is the normalised flux of this spectrum. The x-axis of this spectrum is rescaled to be the same as the z_{DLA} presented in the upper panel.

There are still some very high redshift quasars ($z_{\text{QSO}} \gtrsim 5$) where our code clearly detects too many DLAs in a single spectrum, even at low redshift. We exclude these spectra from our population statistics. At high redshift the Lyman- α forest absorption is so strong as to render the observed flux close to zero. We thus cannot easily distinguish between the null model and the DLA models. It is also possible that at high redshifts the mean flux of the forest is substantially different from the [66] model we assume, and that this biases the fit. Finally, there are few such spectra, and so we cannot rule out the possibility that covariance of their emission spectra differs quantitatively from lower redshift quasars.

2.11 Analysis of the results

In this section, we present results from our classification pipeline, and we also present the statistical properties (CDDF, line densities dN/dX , and total column densities Ω_{DLA}) of the DLAs detected in our catalogue.

2.11.1 ROC analysis

To evaluate how well our multi-DLA classification reproduces earlier results, we rank our DLA detections using the log posterior odds between the DLA model (summing up all possible DLA models $\{\mathcal{M}_{\text{DLA}(i)}\}_{i=1}^k$) and the null model:

$$\begin{aligned} \log(\text{odds}) = & \\ & \log \Pr(\{\mathcal{M}_{\text{DLA}}\} \mid \mathcal{D}, z_{\text{QSO}}) - \log \Pr(\mathcal{M}_{\text{-DLA}} \mid \mathcal{D}, z_{\text{QSO}}), \end{aligned} \tag{2.66}$$

where the ranking is over all sightlines. From the top of the ranked list based on the log posterior odds, we calculate the true positive rate and false positive rate for each rank:

$$\begin{aligned} \text{TPR} &= \frac{\text{TP}}{\text{TP} + \text{FN}}; \\ \text{FPR} &= \frac{\text{FP}}{\text{FP} + \text{TN}}. \end{aligned} \tag{2.67}$$

The true positive rate is the fraction of sightlines where we detect DLAs (ordered by their rank) divided by the number of sightlines with DLAs detected by earlier catalogues. The false positive rate is the number of detections of DLAs divided by the number of sightlines where earlier catalogues did not detect DLAs. In Figure 2.11 we show the TPR and FPR in a receiver-operating characteristics (ROC) plot to show how well our classification performs. We have compared to the concordance DLA catalogue [64] in the hope that it approximates ground truth, there being no completely reliable DLA catalogue.

We also want to know how well our pipeline can identify the number of DLAs in each spectrum. The DR9 concordance catalogue does not count multiple DLA spectra, and so we compare our multi-DLA detections to the catalogue published by [4]. Each DLA detected in [4] comes with a measurement of their confidence of detection (`dla_confidence` or $p_{\text{DLA}}^{\text{Parks}}$) and a MAP redshift and column density estimate. We compare our multiple DLAcatalogue to those spectra with $p_{\text{DLA}}^{\text{Parks}} > 0.98$. The resulting ROC plot is shown in Figure 2.12. We count a maximum of 2 DLAs in each spectrum: 3 or more DLAs in a single sightline are extremely rare and do not provide a large enough sample for an ROC plot. Parks' catalogue is not a priori more reliable than ours, especially in spectra with multiple DLAs, but comparing the first two DLAs is a reasonable way to validate our method's ability to detect multiple DLAs.

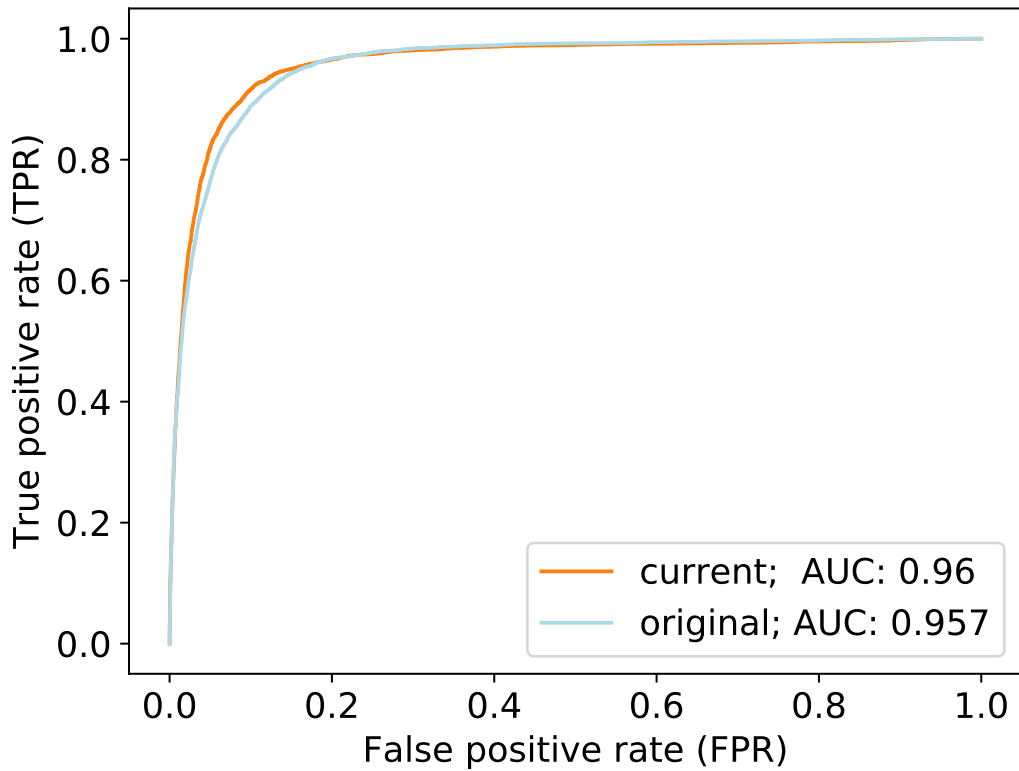


Figure 2.11: The ROC plot made by ranking the sightlines in BOSS DR9 samples using the log posterior odds of containing at least one DLA. Ground truths are from the DR9 concordance catalogue. The orange curve shows the ROC plot of our current multi-DLA model, and the blue curve is derived from [3]. In this plot we consider only the model containing at least one DLA $p(\{\mathcal{M}_{\text{DLA}}\} | \mathcal{D})$, rather than the multiple DLAs models, as the concordance catalogue contains only one DLA per spectrum.

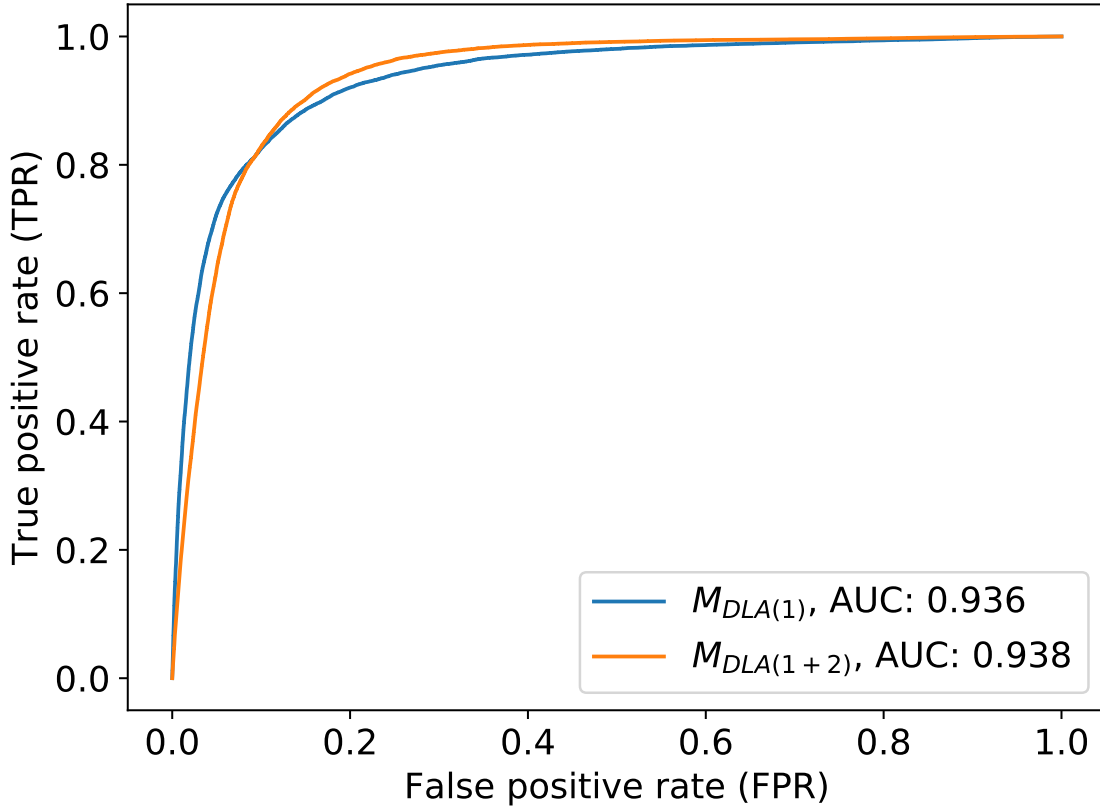


Figure 2.12: The ROC plot for sightlines with one and two DLA detections, by using the catalogue of [4] (with `dla_confidence` > 0.98) as ground truth.

These spectra are counted by breaking down each two-DLA sightline (either in Parks or our catalogue) into two single observations. For example, if there are two DLAs detected in Parks and one DLA detected in our pipeline for an observation \mathcal{D} , we will assign one ground-truth detection to $p(\mathcal{M}_{DLA(1)} | \mathcal{D})$ and assign one ground-truth detection to $p(\mathcal{M}_{\neg DLA} | \mathcal{D})$. On the other hand, if there is only one DLA detected in Parks and two DLAs detected in our pipeline, we will assign one ground-truth detection to $p(\mathcal{M}_{DLA(2)} | \mathcal{D})$ and one ground-truth non-detection to $p(\mathcal{M}_{DLA(2)} | \mathcal{D})$.

In Figure 2.13, we also analyse the *maximum a posteriori* (MAP) estimate of the parameters $(z_{\text{DLA}}, \log_{10} N_{\text{HI}})$ by comparing with the reported values in DR9 concordance DLA catalogue. The median difference between these two is -2.2×10^{-4} (-66.6 km s^{-1}) and the interquartile range is 2.2×10^{-3} (662 km s^{-1}). For the log column density estimate, the median difference is 0.040, and the interquartile range is 0.26. The medians and interquartile ranges of the MAP estimate are very similar to the values reported in [3] with the median of z_{DLA} slightly smaller and the median of $\log_{10} N_{\text{HI}}$ slightly larger. Note that the DR9 concordance catalogue is not the ground truth, so small variations in comparison to [3] can be considered to be negligible. As shown in Figure 2.13, both histograms are roughly diagonal, although the scatter in column density MAP is large. Note that our DLA-detection procedure is designed to evaluate the model evidence across all of parameter space: a single sample MAP cannot convey the full posterior probability distribution. In Section 2.11.2, we thus describe a procedure to propagate the posterior density in the parameter space directly to column density statistics.

2.11.2 CDDF analysis

We follow [62] in calculating the statistical properties of the modified DLA catalogue presented in this paper. We summarise the properties of DLAs using the averaged binned column density distribution function (CDDF), the incident probability of DLAs (dN/dX), and the averaged matter density as a function of redshift ($\Omega_{\text{DLA}}(z)$).

To plot these summary statistics, we need to convert the probabilistic detections in the catalogue to the expected average number of DLAs and their corresponding variances.

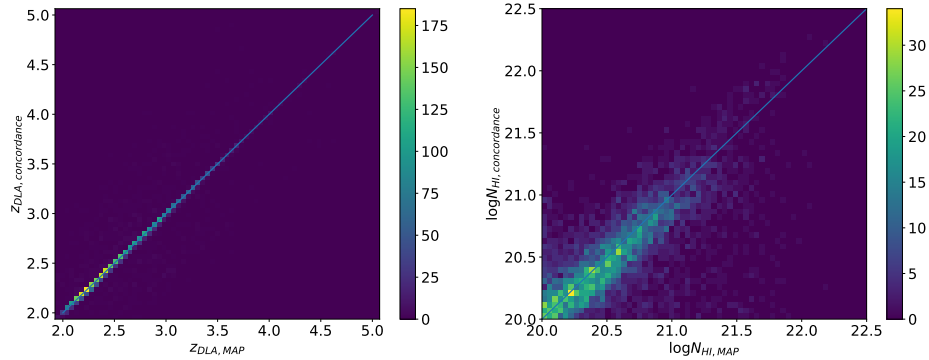


Figure 2.13: The MAP estimates of the DLA parameters $\theta = (z_{\text{DLA}}, \log_{10} N_{\text{HI}})$ for DLAs detected by our model in spectra observed by SDSS DR9, compared to the values reported in the concordance catalogue. The straight line indicates a perfect fit. Note that the concordance $\log_{10} N_{\text{HI}}$ values are not ground truth, so the scatter in column density predictions was expected.

We first describe how we compute the expected number of DLAs in a given column density and redshift bin. Next, we show how we derive the CDDF, dN/dX , and $\Omega_{\text{DLA}}(z)$ from the expected number of DLAs. A sample of n observed spectra contains a sequence of n model posteriors $p_{\text{DLA}}^1, p_{\text{DLA}}^2, \dots, p_{\text{DLA}}^n$ defined by:

$$p_{\text{DLA}}^i = p(\{\mathcal{M}_{\text{DLA}}\} \mid \mathbf{y}_i, \boldsymbol{\lambda}_i, \boldsymbol{\nu}_i, z_{\text{QSO}i}), \quad (2.68)$$

where $i = 1, 2, \dots, n$ is the index of the spectrum, and the DLA model here includes all computed DLA models $\{\mathcal{M}_{\text{DLA}}\} = \{\mathcal{M}_{\text{DLA}(i)}\}_{i=1}^k$, so that $k = 4$ is the maximum possible number of DLAs in each spectrum in our model.

Suppose the region of interest is in a specific bin Θ , an interval in the parameter space of column density or DLA redshift $\Theta \in \{N_{\text{HI}}, z_{\text{DLA}}\}$. To compute the posterior of having DLAs in each spectrum in a given bin Θ , $p_{\text{DLA}}^i(\{\mathcal{M}_{\text{DLA}}\} \mid \Theta)$, we integrate over the sample likelihoods in the bin and multiply the model posterior by the total p_{DLA}^i for

spectrum i :

$$p_{\text{DLA}}^i(\{\mathcal{M}_{\text{DLA}}\} | \Theta) \propto p_{\text{DLA}}^i \times \int_{\underline{\Theta}}^{\overline{\Theta}} p(\mathbf{y}_i | \{\mathcal{M}_{\text{DLA}}\}, \boldsymbol{\lambda}_i, \boldsymbol{\nu}_i, z_{\text{QSO}i}, \theta) d\theta. \quad (2.69)$$

θ is either z_{DLA} or $\log_{10} N_{\text{HI}}$ and $\theta \in \Theta = (\underline{\Theta}, \overline{\Theta})$.

We calculate the posterior probability of having N DLAs by noting that the full likelihood follows the Poisson-Binomial distribution. Consider a sequence of trials with a probability of success equal to $p_{\text{DLA}}^i(\{\mathcal{M}_{\text{DLA}}\} | \Theta) \in [0, 1]$. The probability of having N DLAs out of a total of n trials is the sum of all possible N DLAs subsets in the whole sample:

$$\begin{aligned} \text{Pr}(N) = & \sum_{\text{DLA} \in F_N} \prod_{i \in \text{DLA}} p_{\text{DLA}}^i(\{\mathcal{M}_{\text{DLA}}\} | \Theta) \\ & \prod_{j \in \text{DLA}^c} (1 - p_{\text{DLA}}^j(\{\mathcal{M}_{\text{DLA}}\} | \Theta)) \end{aligned} \quad (2.70)$$

where F_N corresponds to all subsets of N integers that can be selected from the sequence $\{1, 2, \dots, n\}$. The above expression means we select all possible N choices from the entire sample, calculate the probability of those N choices having DLAs and multiply that by the probability of the other $n - N$ choices having no DLAs. If all $p_{\text{DLA}}^i(\{\mathcal{M}_{\text{DLA}}\} | \Theta)$ are equal, the Poisson-Binomial distribution reduces to a Binomial distribution.

The above Poisson-Binomial distribution is not trivial to compute given our large sample size. The technical details of how to evaluate Eq. 2.70 efficiently are described in [62]. In short, we use [68]’s theorem to approximate those spectra with $p_{\text{DLA}}^i(\{\mathcal{M}_{\text{DLA}}\} | \Theta) < p_{\text{switch}} = 0.25$ by an ordinary Poisson distribution, and evaluate the remaining samples with the discrete Fourier transform [69]. Our catalogue contains the posteriors of samples in a

given spectrum. Combined with the above probabilistic description of the total number of DLAs in the entire sample, we are able to obtain not only the point estimation of $\text{Pr}(N)$ but also its probabilistic density interval.

We thus compute the column density distribution function in a given bin $\Theta = N_{\text{HI}} \in [N_{\text{HI}}, N_{\text{HI}} + \Delta N_{\text{HI}}]$ with:

$$f(N) = \frac{F(N)}{\Delta N \Delta X(z)} \quad (2.71)$$

where $F(N) = \mathbb{E}(N \mid N_{\text{HI}} \in [N_{\text{HI}}, N_{\text{HI}} + \Delta N_{\text{HI}}])$ is the expected number of absorbers at a given sightline within a column density interval. Thus, the column density distribution function (CDDF) $f(N)$ is the expected number of absorbers per unit column density per unit absorption distance, within a given column density bin.

The definition of absorption distance $\Delta X(z)$ is:

$$X(z) = \int_0^z (1+z')^2 \frac{H_0}{H(z')} dz', \quad (2.72)$$

which includes the contributions of the Hubble function $H^2(z)/H_0^2 = \Omega_{\text{M}}(1+z)^3 + \Omega_{\Lambda}$, with Ω_{M} is the matter density and Ω_{Λ} is the dark energy density.

The incident rate of DLAs dN/dX is defined as:

$$\frac{dN}{dX} = \int_{10^{20.3}}^{\infty} f(N \mid N_{\text{HI}}, X \in [X, X + dX]) dN_{\text{HI}}, \quad (2.73)$$

which is the expected number of DLAs per unit absorption distance.

The total column density Ω_{DLA} is defined as:

$$\Omega_{\text{DLA}} = \frac{m_{\text{P}} H_0}{c \rho_c} \int_{10^{20.3}}^{\infty} N_{\text{HI}} f(N \mid N_{\text{HI}}, X \in [X, X + dX]) dN_{\text{HI}}, \quad (2.74)$$

where ρ_c is the critical density at $z = 0$ and m_{P} is the proton mass.

2.11.3 Statistical properties of DLAs

Based on the above calculations, we show our CDDF in Figure 2.14, $\frac{dN}{dX}$ in Figure 2.15, and Ω_{DLA} in Figure 2.16.⁷ Note that for determining the statistical properties of DLAs, we limit the samples of z_{DLA} to the range redward of the Lyman- β in the QSO rest-frame, as in [62].

Figure 2.14 shows the CDDF from our DR12 catalogue in comparison to the DR9 catalogue of [5]. Our CDDF analysis combines all spectral paths with QSO redshift smaller than 5, $z_{\text{DLA}} < 5$. The CDDF statistics are dominated by the low-redshift absorbers, as demonstrated in Figure 2.17. The error bars represent the 68% confidence interval, while the grey shaded area encloses the 95% highest density region. The CDDF values in Figure 2.14 are calculated from the posterior distribution directly. We note that there are only two DLAs with MAP $\log_{10} N_{\text{HI}} > 22.5$ in our catalogue with high confidence ($p_{\text{DLA}} > 0.99$). The non-zero values in the CDDF are due to uncertainty in $\log_{10} N_{\text{HI}}$, not positive detections.

[5] contains multi-DLAs, but, as described in Section 2.2 in their paper, they applied a stringent cut on their samples with $\text{CNR} > 3$, where CNR refers to the continuum-to-noise ratio. The CDDF of N12 in the Figure 2.14 is thus a sub-sample of their catalogue. We, on the other hand, use all data even those with low signal-to-noise ratios. Comparing to our previously published CDDF [62], the CDDF in this paper shows DLA detections at low N_{HI} are consistent with [5]. Introducing the sub-DLA as an alternative model successfully regularises detections at $\sim 10^{20} \text{ cm}^{-2}$.⁸

⁷The table files to reproduce Figure 2.14 to Figure 2.16 will be posted in http://tiny.cc/multidla_catalog_gp_dr12q

⁸Note again the artifact at $\sim 10^{20} \text{ cm}^{-2}$ will not affect the analyses of dN/dX or Ω_{DLA} as the definition of a DLAs is absorbers with $N_{\text{HI}} > 10^{20.3} \text{ cm}^{-2}$.

Figure 2.15 shows the line density of DLAs. Our results are again consistent with those of [6] and [5] where they both agree. Our detections are between those two catalogues at low redshift bins and consistent with [6] in the highest redshift bin. Comparing to our previous dN/dX [62], we moderately regularise the detections of DLAs at high redshifts. This change shows that changing the mean model of the GP to include the mean flux absorption prevents the pipeline confusing the suppression due to the Lyman alpha forest with a DLA. While the change of posterior modes in dN/dX is large at high redshift bins, we note that those changes are mostly within 95% confidence interval of our previously published line densities. All analyses shown measure a peak in dN/dX at $z \sim 3.5$. This may be partially due to $z_{\text{DLA}} = 3.5$ the SDSS colour selection algorithm systematic identified by [70], which over-samples Lyman-limit systems (LLS), especially near the quasar, in the redshift range 3.0 – 3.6 [71, 72]. Note however that in our analysis neighbouring redshift bins are highly correlated and so a statistical fluctuation is also a valid explanation. We have checked visually that our sub-DLA model successfully models spectra with a LLS in the proximate zone of the quasar emission peak.

Figure 2.16 shows the total column density Ω_{DLA} in DLAs in units of the cosmic density. Our results are mostly consistent with [5] although we have slightly lower Ω_{DLA} at $z \sim 2$. This is due to our Occam’s razor penalty, which suppresses DLAs in spectra which are not long enough to include the full width of the DLA. Since these are all low redshift quasars, this suppresses DLA detections at $z < 2.3$. As discussed in [5], [73], and [62], the relatively low Ω_{DLA} of [6] is due to the smaller sample size of the SDSS DR5 dataset. We also compare our Ω_{DLA} to that measured by [7] at high redshifts ($z = 4$ and $z = 5$). [7] used a

small but higher signal-to-noise dataset. Our results at $z = 4$ and $z = 5$ are consistent with those from [7]. However, we note that the relatively small sample of [7] may bias it slightly low, as contributions from DLAs with N_{HI} higher than expected to be in the survey will not be included in their Ω_{DLA} estimate. Our Bayesian analysis includes possible contributions of undetected DLAs with column density up to $\log_{10} N_{\text{HI}} = 23$ in the error bars via the prior on the column density.

Compared to our previously published Ω_{DLA} [62], we found a reduction in Ω_{DLA} between $z = 4$ and $z = 5$. This is due to the incorporation of a better mean flux vector model, which reduces the posterior density of high-column density systems for high-redshift absorbers (although within the 95% confidence bars of the earlier work). Our confidence intervals are also substantially smaller for $z_{\text{DLA}} \gtrsim 3.7$ than in [62]. This is due to our inclusion, for the first time, of information from the Lyman- β absorption of the DLAs, which both constrains DLA properties and helps to distinguish DLAs from noise fluctuations.

We have tested the robustness of our method with respect to spectra with different SNRs and found that, as in [62], the statistical properties predicted by our method are uncorrelated with the quasar SNR. Furthermore, the presence of a DLA is uncorrelated with the quasar redshift, fixing a statistical systematic in the earlier work.

As a cross-check of our wider catalogue, we also tested the CDDF, line densities, and total column densities of the DLAs in our catalogue with a full range of z_{DLA} , from Ly ∞ to Ly α . The CDDF was very similar to the CDDF excluding the Ly β region shown in Figure 2.14, but with a moderate increase at high column density. dN/dX was almost

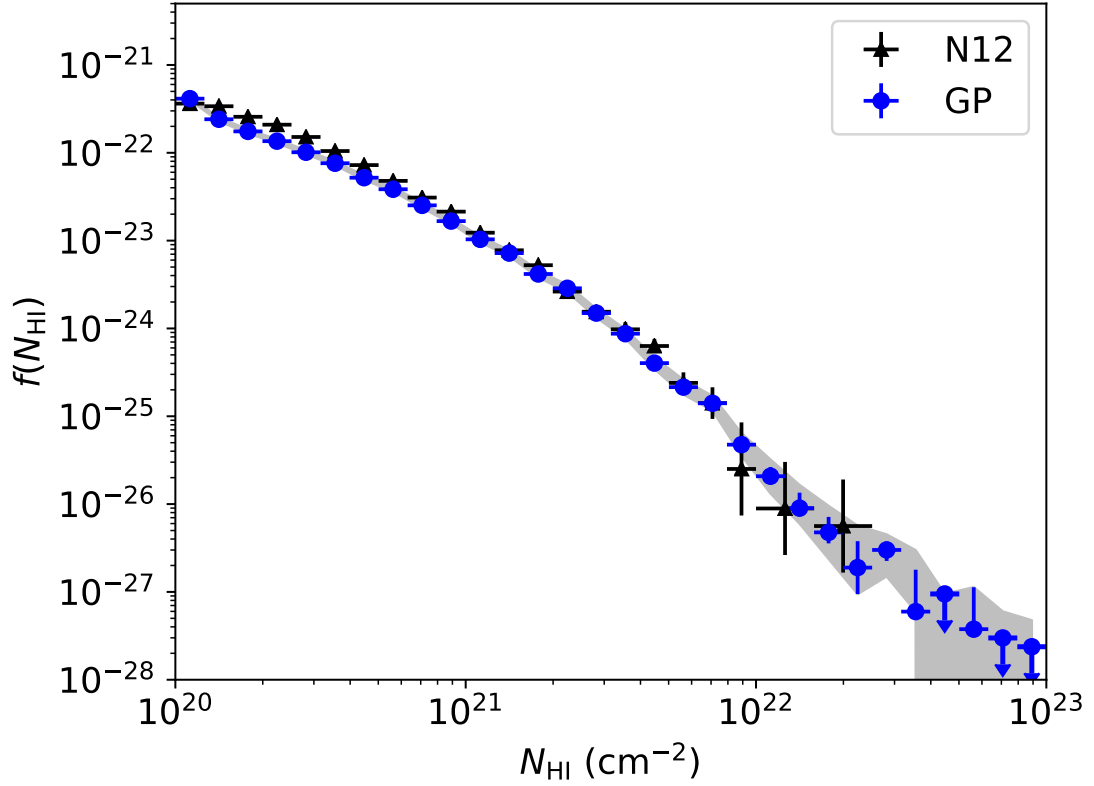


Figure 2.14: The CDDF based on the posterior densities for at least one DLA (blue, ‘GP’). The DLAs are derived from SDSS DR12 spectra using the method presented in this paper. We integrate all spectral lengths with $z < 5$. We also plot the CDDF of [5] (N12; black) as a comparison. The error bars represent the 68% confidence limits, while the grey filled band represents the 95% confidence limits. Note that our CDDF completely overlaps with those of N12 for column densities in the range $10^{21} \text{ cm}^{-2} < N_{\text{HI}} < 10^{22} \text{ cm}^{-2}$.

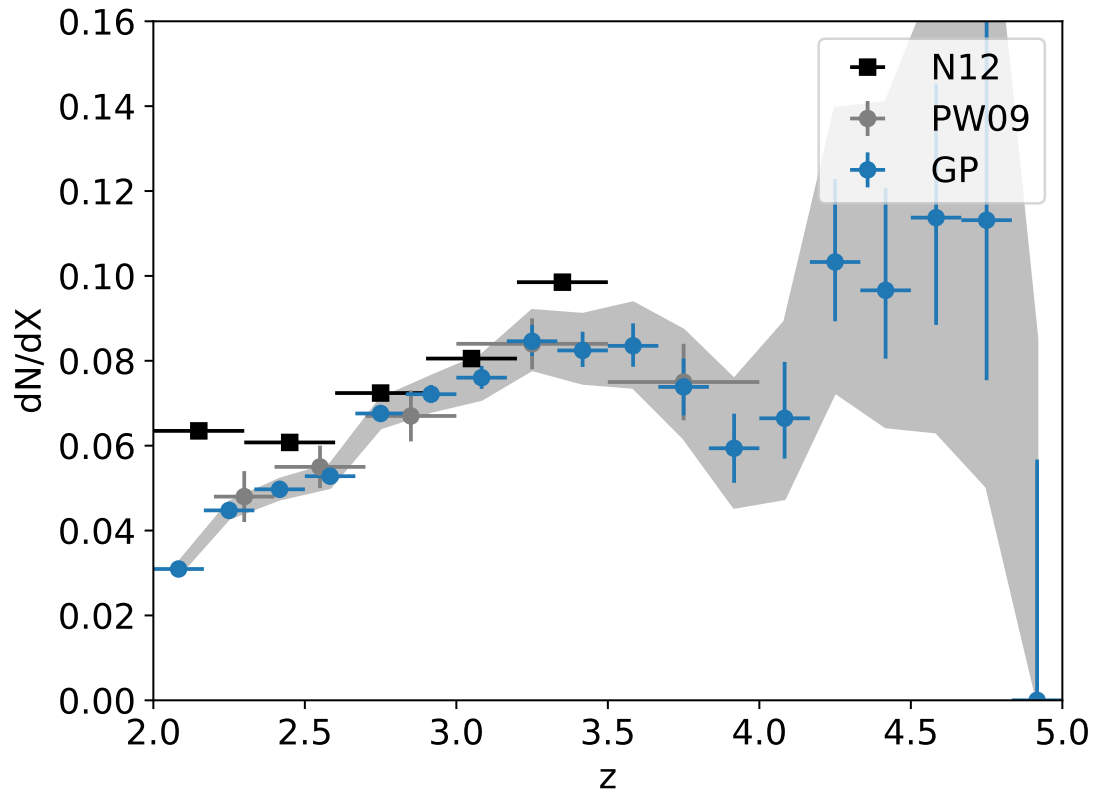


Figure 2.15: The line density of DLAs as a function of redshift from our DR12 multi-DLA catalogue (blue, ‘GP’). We also plot the results of [5] (N12; black) and [6] (PW09; grey). Note that statistical error was not computed in [5].

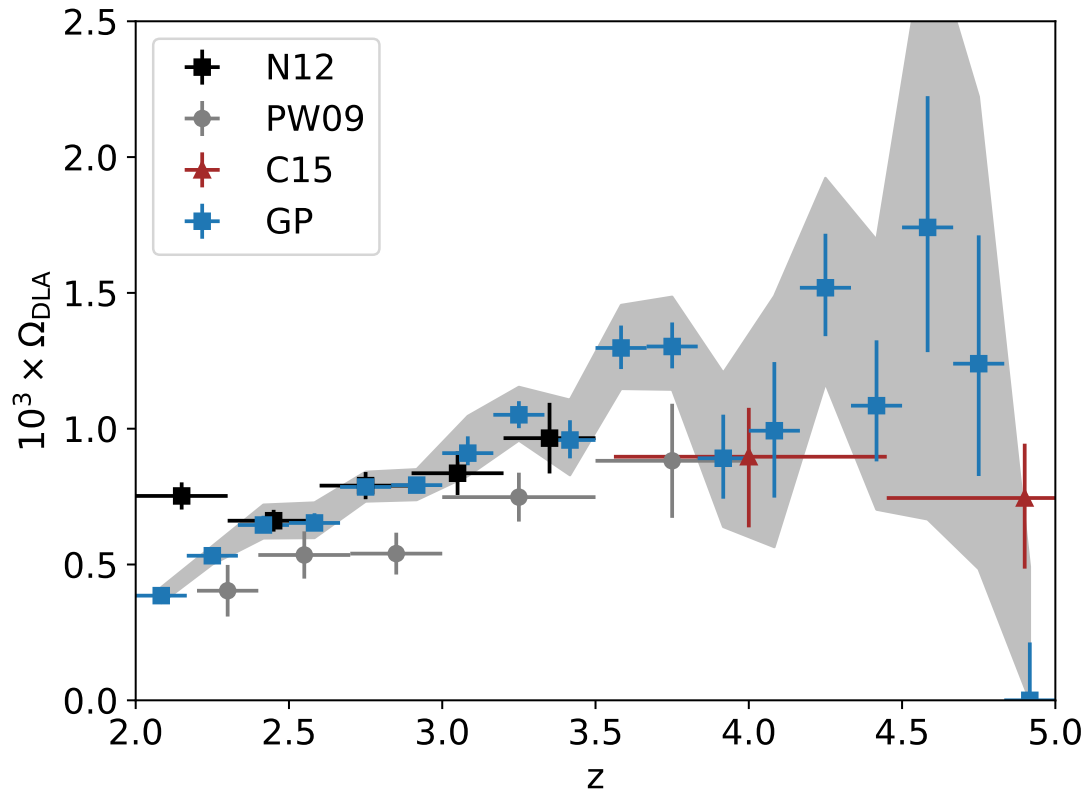


Figure 2.16: The total HI density in DLAs, Ω_{DLA} , from our DR12 multi-DLA catalogue as a function of redshift (blue, ‘GP’), compared to the results of [5] (N12; black), [6] (PW09; grey) and [7] (C15; red).

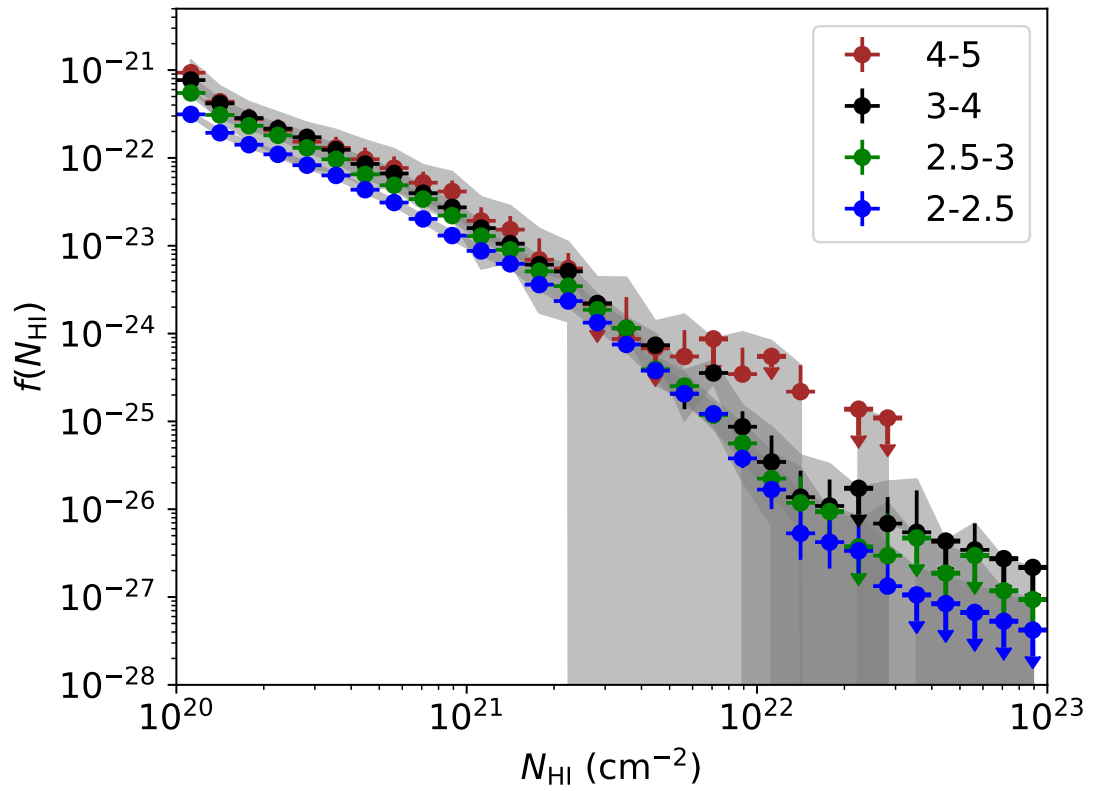


Figure 2.17: The redshift evolution (or non-evolution) of the CDDF. Labels show the absorber redshift ranges used to plot the CDDFs. In column density and redshift ranges with no detection at 68% confidence, a down-pointing arrow is shown indicating the 68% upper limit.

identical to Figure 2.15, indicating that the detection of DLAs is robust even though we extend our sampling range to $\text{Ly}\infty$. However, Ω_{DLA} increases for $3.5 < z_{\text{DLA}} < 4.0$. By visual inspection we found that this is due to the spectra where the quasar redshift from the SDSS pipeline is in error and a Lyman break trough appears at the blue end of the spectrum in a region the code expects to contain only $\text{Ly}\beta$ absorption. As our model does not account for redshift errors, it explains the absorption due to these troughs by DLAs.

2.11.4 Comparison to Garnett’s Catalogue

To understand the effect of the modifications we made to our model in this paper, we visually inspected a subset of spectra with high model posteriors of a DLA in [3] ($p_{\text{DLA}}^{\text{Garnett}}$) but low model posteriors in our current model (p_{DLA}). In particular, we chose spectra with ($p_{\text{DLA}}^{\text{Garnett}} - p_{\text{DLA}} > 0.99$).

A large fraction of these spectra falls within the $\text{Ly}\beta$ emission region. One plausible explanation is that the $\text{Ly}\beta$ emission region has a higher noise variance, which makes it harder to distinguish the DLA and sub-DLA models. We also checked that we are not unfairly preferring the sub-DLA model during model selection. Our model selection uses the sub-DLA model only to regularise the DLA model and does not consider cases where DLAs and sub-DLAs occur in the same spectrum. Thus a spectrum with a clear detection of a sub-DLA could fail to detect a true DLA at a different redshift. In light of this, we also tested if combining multi-DLA models with a sub-DLA affects our results.

We modified the DLA model, assuming that the DLA and sub-DLA models are independent, to include the sub-DLA model prior. We then considered an iterative sampling procedure: First, we sampled the k^{th} DLA likelihood. Next we used the k^{th} DLA parameter

posterior as a prior to sample $\mathcal{M}_{\text{DLA}(k)}$ and combine $\mathcal{M}_{\text{DLA}(k)}$ with the sub-DLA model via sampling a non-informative prior. The full procedure can be written as:

$$\begin{aligned}
 p(\{\theta_i\}_{i=1}^k \mid \mathcal{M}'_{\text{DLA}(k)}, \mathcal{D}, z_{\text{QSO}}) = \\
 (1 + p(\theta_{\text{sub}} \mid \mathcal{M}_{\text{sub}}, z_{\text{QSO}})) \times \\
 p(\{\theta_i\}_{i=1}^k \mid \mathcal{M}_{\text{DLA}(k)}, \mathcal{D}, z_{\text{QSO}}),
 \end{aligned}
 \tag{2.75}$$

For computational simplicity, we only consider the modified model until $\mathcal{M}'_{\text{DLA}(3)}$; the probability of $\mathcal{M}'_{\text{DLA}(4)}$ is expected to be insignificant comparing to the total DLA model posterior, $p(\{\mathcal{M}_{\text{DLA}}\} \mid \mathcal{D}, z_{\text{QSO}})$.

In practice, however, we found that this made a small difference to our results, only marginally modifying the ROC curve and CDDF. Moreover, the ability of the sub-DLA model to regularize low column density DLAs was reduced, so we have preserved our default model.

2.11.5 Comparison to Parks Catalogue

In this section, we compare our results with [4]. We first show the differences between our MAP predictions and Parks' predictions for DLA redshift and column density. We required $p_{\text{DLA}}^{\text{Parks}} > 0.98$. We measured the difference in posterior parameters when both pipelines predicted one DLA. As shown in Figure 2.18, both histograms are roughly symmetric. We measure small median offsets between two pipelines with

$$\begin{aligned}
 \text{median}(z_{\text{DLA}}^{\text{MAP}} - z_{\text{DLA}}^{\text{Parks}}) &= 0.00010; \\
 \text{median}(\log N_{\text{HI}}^{\text{MAP}} - \log N_{\text{HI}}^{\text{Parks}}) &= 0.016.
 \end{aligned}
 \tag{2.76}$$

We also compared our absorber redshift measurements and column density measurements to Parks' catalogue for those spectra which we both agree contain two DLAs.

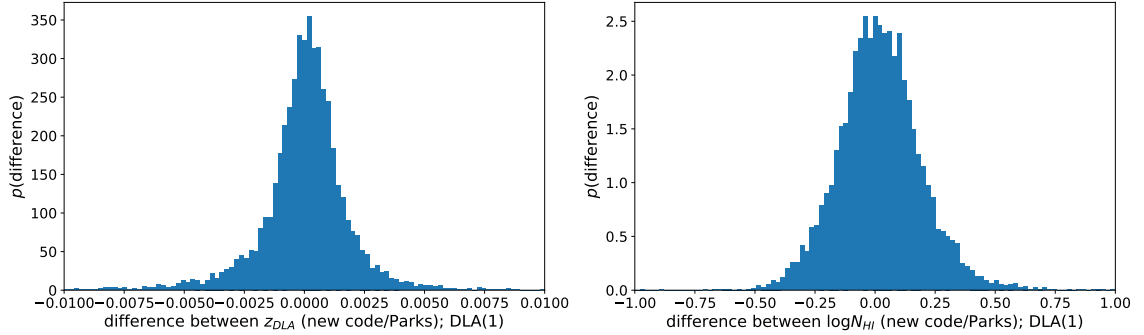


Figure 2.18: The difference between the MAP estimates of the DLA parameters $\theta = (z_{\text{DLA}}, \log_{10} N_{\text{HI}})$, against the predictions of [4]. We consider spectra which both catalogues agree contain one DLA.

The differences between these two have small median offsets of $\Delta z_{\text{DLA}} = 0.000052$ and $\Delta \log_{10} N_{\text{HI}} = 0.006$ (and dominated by low column density systems).

We show the disagreement between multi-DLA predictions for our catalogue and Parks’ catalogue in Table 2.1. Note that though the multi-DLA detections between our method and Parks do not completely agree, the level of disagreement is small: 6.1%. Moreover, if Parks predicts one or two DLAs, our method generally detects one or two DLAs. There are however some spectra where we detected > 2 DLAs, but Parks detected none. To understand the statistical effect of this discrepancy, we compare our DLA properties to those reported by [4]. We plot the CDDF and dN/dX of that catalogue. We assume $p_{\text{DLA}}^{\text{Parks}} > 0.9$ represents a DLA and use z_{DLA} and $\log_{10} N_{\text{HI}}$ reported in their catalogue in JSON format⁹. To compute the sightline path searched over, we assume their CNN model was searching the range $\text{Ly}\infty$ to $\text{Ly}\alpha$ in the quasar rest-frame. Note this differs slightly from [4] Section 3.2 where a sightline search radius ranging from 900\AA to 1346\AA in the quasar rest frame

⁹<https://tinyurl.com/cnn-dlas>

Parks Garnett with Multi-DLAs	0 DLA	1 DLA	2 DLAs	3 DLAs	4 DLAs
0 DLA	138726	6197	142	6	0
1 DLA	3050	8752	335	4	0
2 DLAs	293	570	566	28	0
3 DLAs	30	39	34	21	0
4 DLAs	5	9	6	1	0

Table 2.1: The confusion matrix for multi-DLAs detections between Garnett with multi-DLAs and Parks. Note we require both the model posteriors in Garnett and DLA confidence in Parks to be larger than 0.98. We also require $\log_{10} N_{\text{HI}} > 20.3$.

is given. However, we know the centers of DLAs should be at a redshift between $\text{Ly}\infty$ and $\text{Ly}\alpha$ in the rest frame and modify our search paths accordingly.

Figure 2.20 shows that dN/dX is consistent with [5] for $z_{\text{DLA}} < 3.5$ (although lower than our measurement at higher redshift). The CNN is thus successfully detecting DLAs, especially the most common case of DLAs with a low column density. There are fewer DLAs detected at higher redshift, likely reflecting the increased difficulty for the CNN of distinguishing DLAs from the Lyman- α forest. This is discussed in [4], who note that the CNN finds it difficult to detect a weak DLA in noisy spectra. However, as shown in Figure 2.19, the CDDF measured by the CNN model is significantly discrepant with other surveys for large column densities. Note that the scale is logarithmic: the CNN is failing to detect $> 60\%$ of DLAs with $\log_{10} N_{\text{HI}} > 21$. We noticed that large DLAs were often split into two objects with lower column density, which accounts for many of the discrepancies between our two datasets. We suspect this might be due to the limited size of the convolutional filters used by [4]. If the filter is not large enough to contain the full damping wings of a given DLA, the allowed column density would be artificially limited.

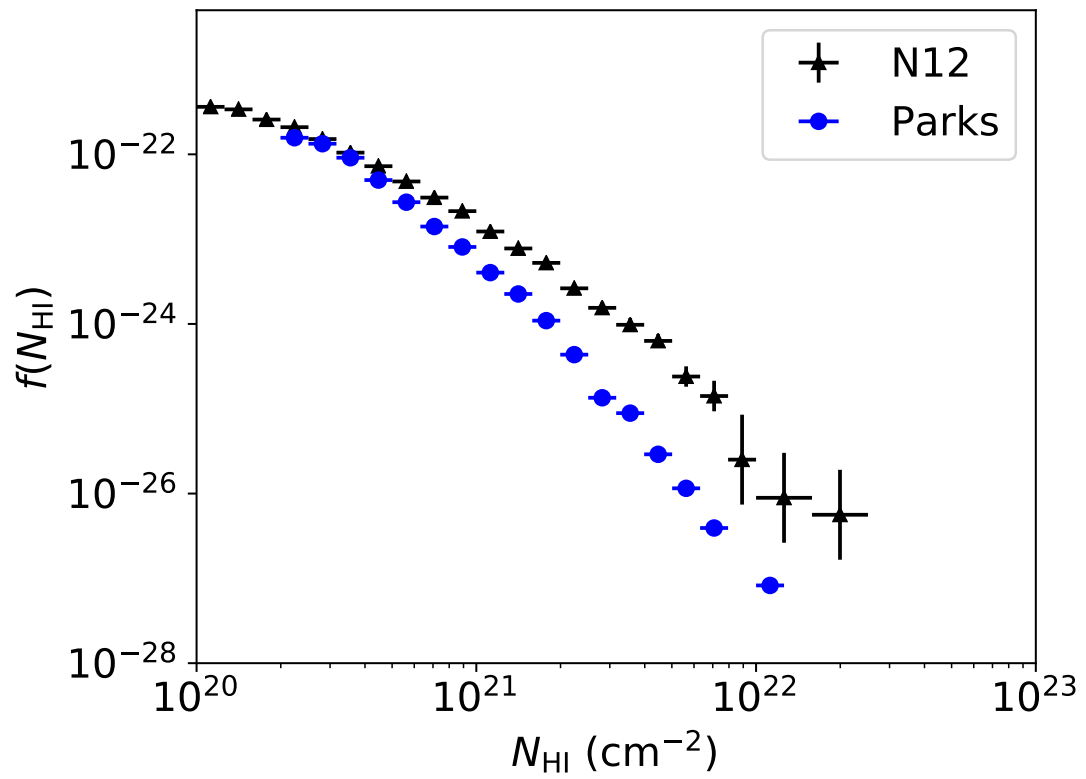


Figure 2.19: The column density distribution function from [4], showing that the CNN algorithm substantially underestimates the number of DLAs in the high- N_{HI} regime.

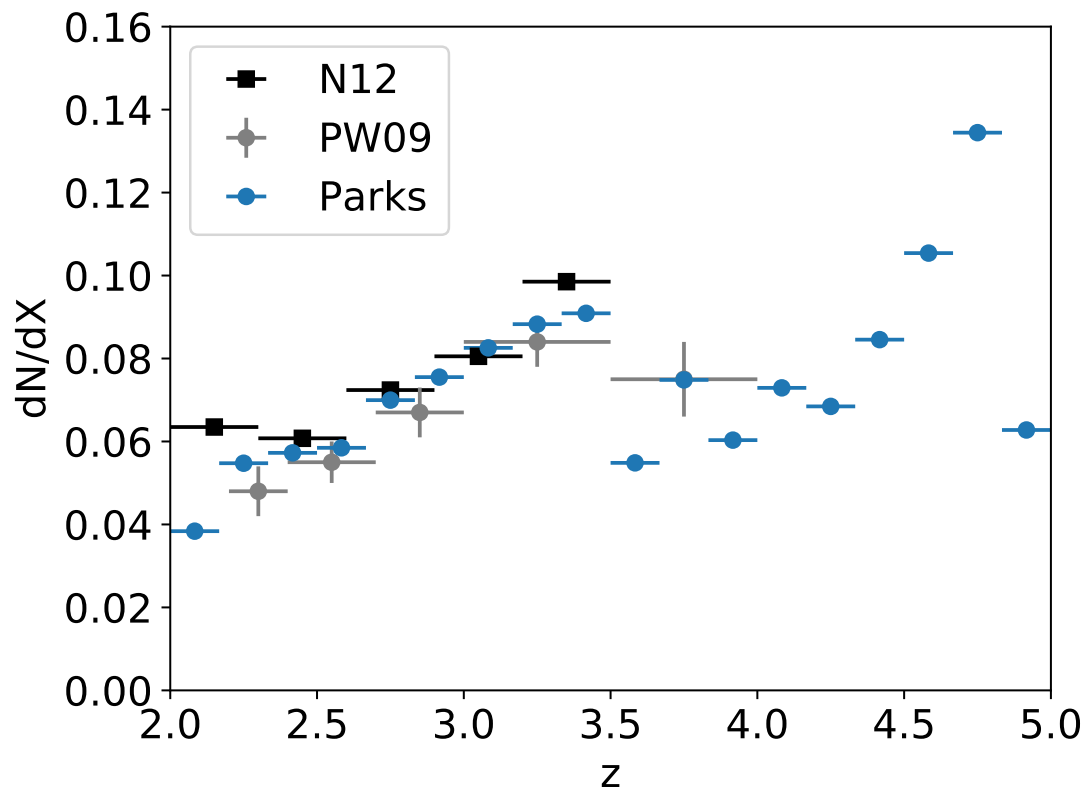


Figure 2.20: dN/dX from [4]. The dN/dX agrees well with other surveys, but there is a moderate deficit of DLAs at high redshifts.

2.12 Conclusion

We have presented a revised pipeline for detecting DLAs in SDSS quasar spectra based on [3]. We have extended the pipeline to reliably detect up to 4 DLAs per spectrum. We have performed modifications to our model for the Lyman- α forest to improve the reliability of DLA detections at high redshift and introduced a model for sub-DLAs to improve our measurement of low column density DLAs. Finally we introduced a penalty on the DLA model based on Occam’s razor which meant that spectra for which both models are a poor fit generally prefer the no-DLA model.

Our results include a public DLA catalogue, with several examples shown above and further examples easily plotted using a python package. We have visually inspected several extreme cases to validate our results and compared extensively to several earlier DLA catalogues: the DR9 concordance catalogue [64] and a DR12 catalogue using a CNN [4]. Our new pipeline had very good performance validated against both catalogues.

Based on the revised pipeline, we also presented a new measurement of the abundance of neutral hydrogen from $z = 2$ to $z = 5$ using similar calculations to [62]. The statistical properties of DLAs were in good agreement with our previous results [62] and consistent with [5], [6], and [7]. The modifications made, including introducing a sub-DLA model, adjusting the mean flux, and penalizing complex models with Occam’s razor, remove over-detections of low column density absorbers and make more robust predictions for the properties of DLAs at $z > 4$. Similarly to previous work, we detect only a small increase in the CDDF for $2 < z < 4$, and a similarly moderate increase in the line density of DLAs and Ω_{DLA} over this redshift range.

Acknowledgements

The authors thank Bryan Scott and Reza Monadi for valuable comments. This work was partially supported by an Amazon Machine Learning Research allocation on EC2 and UCR's HPCC. SB was supported by the National Science Foundation (NSF) under award number AST-1817256. RG was supported by the NSF under award number IIS-1845434.

Data availability

All the code to reproduce the data products is available in our GitHub repo: https://github.com/rmgarnett/gp_dla_detection/tree/master/multi_dlas. The final data products are available in this Google Drive: http://tiny.cc/multidla_catalog_gp_dr12q, including a MAT (HDF5) catalogue and a JSON catalogue. README files are included in each folder to explain the content of the catalogues.

Chapter 3

Damped Lyman-alpha Absorbers from Sloan Digital Sky Survey DR16Q with Gaussian processes

3.1 Abstract

We present a new catalogue of Damped Lyman- α absorbers from SDSS DR16Q, as well as new estimates of their statistical properties. Our estimates are computed with the Gaussian process models presented in [3, 8] with an improved model for marginalising uncertainty in the mean optical depth of each quasar. We compute the column density distribution function (CDDF) at $2 < z < 5$, the line density (dN/dX), and the neutral hydrogen density (Ω_{DLA}). Our Gaussian process model provides a posterior probability distribution of the number of DLAs per spectrum, thus allowing unbiased probabilistic

predictions of the statistics of DLA populations even with the noisiest data. We measure a non-zero column density distribution function for $N_{\text{HI}} < 3 \times 10^{22} \text{ cm}^{-2}$ with 95% confidence limits, and $N_{\text{HI}} \lesssim 10^{22} \text{ cm}^{-2}$ for spectra with signal-to-noise ratios > 4 . Our results for DLA line density and total hydrogen density are consistent with previous measurements. Despite a small bias due to the poorly measured blue edges of the spectra, we demonstrate that our new model can measure the DLA population statistics when the DLA is in the Lyman- β forest region. We verify our results are not sensitive to the signal-to-noise ratios and redshifts of the background quasars although a residual correlation remains for detections from $z_{\text{QSO}} < 2.5$, indicating some residual systematics when applying our models on very short spectra, where the SDSS spectral observing window only covers part of the Lyman- α forest.

3.2 Introduction

Damped Lyman- α absorbers (DLAs) are strong Lyman- α absorption features discovered in quasar spectral sightlines. At the densities required to produce neutral hydrogen column densities above the DLA threshold, $N_{\text{HI}} > 10^{20.3} \text{ cm}^{-2}$ [49], the gas of DLAs is self-shielded from the ionising effect of the ultra-violet background (UVB) [50] but diffuse enough to have a low star formation rate [72]. DLAs contain a large fraction of the neutral hydrogen budget after reionisation [52, 5, 53, 7], which make them a direct probe of the distribution of neutral gas.

Numerical simulations tell us DLAs are associated with a wide range of halo masses, with a peak value in the range of $10^{10} - 10^{11} M_{\odot}$ [54, 55, 18, 74]. Through

cross-correlating the DLAs with the Lyman- α forest, [75] measured a DLA bias factor $b_{\text{DLA}} = 2.17 \pm 0.2$. This implies a median host halo mass of $\sim 10^{12} M_{\odot}$, assuming all DLAs arise from halos of the same mass and. However, a model which assumes a power-law distribution function of DLA cross-section as a function of halo mass is only in marginal tension with the data [57]. Furthermore, a later measurement from SDSS-DR12 [76] found a bias factor $b_{\text{DLA}} = 1.99 \pm 0.11$, and a median host halo mass $\sim 4 \times 10^{11} M_{\odot}$, in good agreement with simulations. Alternative measurements by cross-correlating with CMB lensing data are broadly consistent with both simulated DLAs and Lyman- α clustering [77, 78].

In the cosmology context, the Lyman- α forest is a successful probe of matter clustering between $2 < z < 6$ [79, 80, 81, 82, 83, 84]. However, high column density absorbers such as DLAs will bias cosmological parameter estimates from Lyman- α and thus need to be masked out [85]. Simulations have been performed to study the effect of damped absorbers on the Lyman- α 1-D and 3-D flux power spectrum [86, 87], and a recent Bayesian fitting method has been proposed to better understand how DLA contaminants affect cosmological inference using the BAO peak [88].

In this work, we present new estimates for the column density distribution function (CDDF), the abundance of DLAs, and the average neutral hydrogen density at $z = 2 - 5$ for DLAs in the Sloan Digital Sky Survey IV quasar catalogue from Data Release 16 (SDSS-IV/eBOSS DR16) [89, 90]. We compute DLA population statistics using the Gaussian process (GP) model presented in [8], a modified version of the machine learning framework from [3]. We retrain our model on SDSS DR12 [91, 92, 93, 61] and generate a DLA catalogue

from DR16Q [90]. We compute DLA population statistics from the DLA catalogue, which update the estimates we made in [62, 8].

The pipeline presented in [3] provided for the first time probabilistic detections of DLAs in each spectrum, which comes with a posterior distribution on putative DLAs for the column density and the absorber redshift. With the aid of a full posterior probability distribution for the number of DLAs in each quasar spectrum, “soft” detections in noisy data become available. We propagate uncertainties from each individual spectrum into the global population, without setting any hard threshold on the minimum required probability for the presence of DLAs. We are thus able to include even noisy spectra in our sample of DLAs.

[8] added an alternative model for sub-DLAs, which regularised excessive detections at low column density. We also included absorption from the mean optical depth in the Lyman- α forest in the GP mean function. This helped prevent the pipeline from using DLAs to compensate for Lyman- α forest absorption in the spectrum, essential at high redshift. In this work, we further improve this aspect of our model. We marginalise out uncertainty in the effective optical depth in each spectrum using the measured mean optical depth as a prior when computing the evidence for the null, DLA, and sub-DLA models.

Several other DLA search methods for SDSS spectra have been implemented. These range from visual inspection surveys [59], visually guided Voigt profile fitting [20, 6], and template fitting [94, 5], to machine learning based methods such as a convolutional neural network (CNN) approach [4] and an unpublished Fisher discriminant analysis [63]. The CNN method [4] was also run to identify DLAs as part of the SDSS DR16 quasar

catalogue [90]. We compare the DLAs detected by our GP model and the DLAs in DR16Q in Section 3.7.

Machine learning methods have also been proposed to classify broad absorption lines (BALs), including a line-finder based convolutional neural network (CNN) [95] and a hybrid of a CNN with a principal component analysis [96].

Section 3.3 will briefly outline our modelling decisions and the changes to the model made in this work. Section 3.3.1 describes the cuts we applied to SDSS DR16Q. We recap our modelling details in Section 3.3.2. We present our results in Section 3.4, including the CDDF in Section 3.4.1 and the incidence rate of DLAs and total HI density in Section 3.4.2. In Section 3.5, we discuss the possible remaining systematics in our method. Section 3.6 shows population statistics for DLAs in $\text{Ly}\infty$ to $\text{Ly}\beta$. In Section 3.7, we briefly compare our DLA catalogue to the DLAs presented in the SDSS DR16Q catalogue, which implemented a CNN model [4] to classify DLAs. We conclude in Section 3.8.

3.3 Methods

Here we briefly recap our Gaussian process (GP) based framework for detecting DLAs using *Bayesian model selection*. We summarise the general approach, while more comprehensive mathematical details may be found in [3, 8]. A quasar sightline has spectroscopic observations $\mathcal{D} = (\boldsymbol{\lambda}, \mathbf{y})$, where $\boldsymbol{\lambda}$ is a vector of rest wavelength bins, and \mathbf{y} is a vector of observed flux at these wavelength bins. Suppose we have built likelihood functions for a set of models $\{\mathcal{M}_i\}$. We can evaluate the posterior probability of a model, \mathcal{M} , given

a quasar observation, \mathcal{D} , based on Bayes' rule:

$$\Pr(\mathcal{M} | \mathcal{D}) = \frac{p(\mathcal{D} | \mathcal{M})\Pr(\mathcal{M})}{\sum_i p(\mathcal{D} | \mathcal{M}_i)\Pr(\mathcal{M}_i)}, \quad (3.1)$$

where $p(\mathcal{D} | \mathcal{M})$ is the model evidence of the quasar spectrum \mathcal{D} given model \mathcal{M} , $\Pr(\mathcal{M})$ is the prior probability of model \mathcal{M} , and the denominator on the right-hand-side is the sum of posterior probabilities of all models in consideration.

Concretely, we have the model without DLAs ($\mathcal{M}_{\text{-DLA}}$), the model with k DLAs ($\{\mathcal{M}_{\text{DLA}(i)}\}_{i=1}^k$), and the model with sub-DLAs (\mathcal{M}_{sub}). We set $k = 3$ here, allowing up to 3 DLAs per spectrum. We consider a posterior probability of a sub-DLA, \mathcal{M}_{sub} , not to be a DLA detection, as in [8]. Section 3.3.2 describes the details of how we compute the model evidence for each model.

Table 3.1 lists mathematical notation and definitions of parameters used throughout the paper.

3.3.1 Data

Our GP model requires a training set *without DLAs* for training the null model, $\mathcal{M}_{\text{-DLA}}$. We use the DLAs in SDSS DR12Q detected by [8] as our true DLA labels. Here we list the subset of DR12 quasars omitted from our training sample:

- Quasars with $z_{\text{QSO}} < 2.15$, which have almost no Lyman- α forest, are removed.
- BAL: quasars with a broad absorption line (BAL) probability larger than 0.75 ($\text{BAL_PROB} \geq 0.75$) are removed, as suggested by [90]. BAL_PROB is derived from QuasarNET [95].

Table 3.1: Mathematical notations and definitions

Notation	Description
$\mathcal{M}_{\text{-DLA}}$	Null model, model without DLAs or subDLAs
\mathcal{M}_{DLA}	Model with DLAs ($20 \leq \log_{10} N_{\text{HI}} \leq 23$)
\mathcal{M}_{sub}	Model with subDLAs ($19.5 \leq \log_{10} N_{\text{HI}} < 20$)
$p(\mathcal{D} \mathcal{M})$	Model evidence, marginalised likelihood
$\text{Pr}(\mathcal{M})$	Model prior
$(\beta_{\text{MF}}, \tau_{0,\text{MF}})$	Parameters of power-law relation of effective optical depth model
$\tau_{\text{eff,HI}}(z; \beta_{\text{MF}}, \tau_{0,\text{MF}})$	Power-law model of effective optical depth
$p(\beta_{\text{MF}})$	Prior of β_{MF} , assumed to be a normal distribution
$p(\tau_{0,\text{MF}})$	Prior of $\tau_{0,\text{MF}}$, assumed to be a normal distribution
$p(z_{\text{DLA}} z_{\text{QSO}}, \mathcal{M}_{\text{DLA}})$	Prior of redshift of DLAs, a uniform distribution
$p(N_{\text{HI}} \mathcal{M}_{\text{DLA}})$	Prior of column density of a DLA, a data-driven distribution
\mathbf{y}	Vector of normalised observed flux
$\boldsymbol{\lambda}$	Vector of wavelength pixels in restframe
ν	Vector of instrumental noise variance
$\boldsymbol{\mu}$	Vector of GP mean model
$\boldsymbol{\Sigma}$	Matrix of GP covariance
\mathbf{A}_{F}	Matrix of mean flux suppression from the effective optical depth (diagonal matrix)
\mathbf{K}	Matrix to describe covariance of quasar emission spectrum (2281×2281 matrix, 20×2281 parameters)
$\boldsymbol{\Omega}$	Matrix of Lyman series absorption noise (diagonal matrix)

- `CLASS_PERSON == 30`: quasars classified as BALs by human visual inspection are removed.
- `ZWARNING`: spectra flagged with `ZWARNING` for pipeline redshift estimation are removed, but extremely noisy spectra with `TOO_MANY_OUTLIERS` are kept.

We have in total 89,408 spectra without DLAs for training the null model.

We also use the same above criteria to select the DR16Q spectra for applying our model. In addition to the above criteria, the DR16Q quasar sample to which our model is applied is a subset of the full DR16Q sample chosen following additional conditions:

- `IS_QSO_FINAL == 1`: We require this flag in the quasar sample, specifying that a spectrum is robustly classified as a quasar.
- `CLASS_PERSON == 3` or `0`: This flag specifies that the spectrum was classified by a human as a quasar (3) or was not visually identified (0).
- `SOURCE_Z`: as suggested in Section 3.2 of [90], spectra with $Z > 5$ and `SOURCE_Z == PIPE` have a suspect redshift estimate and should not be used without a careful visual re-inspection. We thus remove these spectra from our analysis.

Integral to our method is a reliable quasar redshift estimate. It is not trivial to reliably estimate quasar redshifts in the large samples provided by DR16Q,¹ and so we are careful to use the redshift estimates suggested by [90]. To ensure our quasar redshifts are as homogeneous as possible, we use `Z_PCA`, the recommended redshift estimate method for statistical analyses of a large ensemble of quasars. We also remove the spectra where

¹Indeed, we have extended our GP framework to provide a quasar redshift estimate [1].

redshift measurements disagree with each other by more than 0.1, which means we remove samples with $|z_i - z_j| > 0.1$ for $z_i, z_j \in \{\text{Z_PIPE}, \text{Z_PCA}, \text{Z}, \text{Z_VI}\}$. If Z_VI is not present, we use only the other three redshift estimates. Our final DR16Q sample size contains 159 807 Lyman- α quasar spectra.

3.3.2 Gaussian process model

Consider a distant quasar with a known redshift, z_{QSO} . Each spectroscopic observation gives us the observed flux, \mathbf{y} , on a set of wavelength pixels in observed-frame wavelengths, $\boldsymbol{\lambda}_{\text{obs}}$. Since the quasar redshift is assumed to be known, we shift into the rest frame, $\boldsymbol{\lambda} = \boldsymbol{\lambda}_{\text{obs}}/(1 + z_{\text{QSO}})$. Standard errors are provided with each observed flux pixel, $\sigma(\lambda_i)$, with λ_i the i th pixel in $\boldsymbol{\lambda}$, and we define the noise variance of each observed flux pixel as $\nu_i = \sigma(\lambda_i)^2$. Given the observed flux of a quasar, we normalise all flux measurements by dividing the median flux observed between $[1425\text{\AA}, 1475\text{\AA}]$ in the rest-frame, a wavelength range redwards of the Ly α emission and avoiding major emission lines.

For each quasar observation, we have data $\mathcal{D} = (\boldsymbol{\lambda}, \mathbf{y}, \boldsymbol{\nu}, z_{\text{QSO}})$. We want to build a likelihood function to describe this data:

$$p(\mathbf{y}|\boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\text{QSO}}),$$

which is the likelihood of the flux \mathbf{y} given all other observed quantities. We model this likelihood as a Gaussian process:

$$p(\mathbf{y}|\boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\text{QSO}}) = \mathcal{N}(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\mu}$ is the mean vector of the GP, and $\boldsymbol{\Sigma}$ is the covariance matrix of the GP. We will use bold lowercase italics for vectors and bold uppercase letters for matrices.

Learning the GP null model

A GP is fully specified by its first two central moments: the mean function, $\mu(\lambda)$, and the covariance kernel, $K(\lambda, \lambda')$, [31]. Our task now is to learn the mean function and the covariance function from the training set. Suppose we have a set of quasar observations without any intervening DLAs, $\{\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_{N_{\text{spec}}}\}$, where N_{spec} is the number of quasars in the training set. We can then learn the mean function by taking a precision weighted average:

$$\mu_j = \frac{\sum_i y_{ij} \neq \text{NaN} (y_{ij} / \nu_{ij})}{\sum_i y_{ij} \neq \text{NaN} (1 / \nu_{ij})}, \quad (3.2)$$

where the summation is over i index. j indicates j th pixel in the observed flux, i represents i th spectrum, and we only average over the non-NaN values. Note this differs from [8], where we used the mean rather than the precision weighted average. The precision weighted average can be viewed as a result of using an uninformative prior on μ_j and an independent Gaussian likelihood for each y_{ij} . If we have a set of normally disturbed flux pixels with each flux pixel follows $y_{ij} \sim \mathcal{N}(\mu_j, \nu_{ij})$ with known variance ν_{ij} and an unknown μ_j with an uninformative prior, the posterior will be a normal distribution with a new mean equals a precision weighted average.

Instead of training on the raw observed flux \mathbf{y} directly, we follow [8] to train the mean function and the kernel on the flux after removing the average effect of the Lyman- α forest, the de-forest flux:

$$\begin{aligned} y_{ij} &\leftarrow y_{ij} \cdot \exp(\tau_{\text{eff,HI}}); \\ \nu_{ij} &\leftarrow \nu_{ij} \cdot \exp(2 \cdot \tau_{\text{eff,HI}}), \end{aligned} \quad (3.3)$$

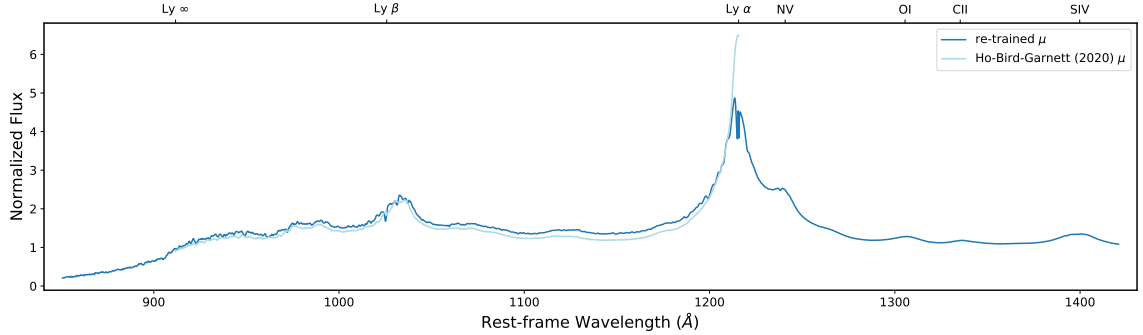


Figure 3.1: Our GP mean function using a precision weighted average of the rest-frame wavelengths. We extended our model compared [8] (light blue), both bluewards past the Lyman break at 912Å and redwards past the SIV emission line.

which means we replace observed flux and its variance with the flux and variance before the suppression of Lyman- α forest. The effective optical depth is parameterised as:

$$\tau_{\text{eff,HI}}(z(\lambda_{\text{obs}}); \beta_{\text{MF}}, \tau_{0,\text{MF}}) = \sum_{i=2}^N \tau_{0,\text{MF}} \frac{\lambda_{1i} f_{1i}}{\lambda_{12} f_{12}} (1 + z_{1i}(\lambda_{\text{obs}}))^{\beta_{\text{MF}}}, \quad (3.4)$$

where λ_{1i} is the transition wavelength from Lyman- α to the i th member in the Lyman series, f_{1i} represents the oscillator strength, z_{1i} is the absorber redshift, and we set $N = 31$.

The absorber redshift is written as:

$$\begin{aligned} 1 + z_{1i}(\lambda, z_{\text{QSO}}) &= \frac{\lambda_{\text{obs}}}{\lambda_{1i}} \\ &= \frac{\lambda(1 + z_{\text{QSO}})}{\lambda_{1i}}. \end{aligned} \quad (3.5)$$

We parameterise the effective optical depth by a power-law relation with $\tau_{0,\text{MF}}$ and β_{MF} parameters. Here we specify a subscript “MF” to annotate the parameters modified by mean flux suppression. Fig 3.1 shows our new GP mean function, compared to [8].

Taking this Lyman- α mean flux into account introduces a dependence on quasar redshift into the mean function of the GP for each quasar:

$$\begin{aligned} \mu(\lambda, z_{\text{QSO}}; \beta_{\text{MF}}, \tau_{0,\text{MF}}) = \\ \mu(\lambda) \cdot \exp(-\tau_{\text{eff,HI}}(z(\lambda, z_{\text{QSO}}); \beta_{\text{MF}}, \tau_{0,\text{MF}})). \end{aligned} \quad (3.6)$$

$\mu(\lambda)$ is the mean function we learned from Eq 3.2. We learn the mean function on a dense grid of wavelengths on a chosen rest-frame wavelength range:

$$\lambda \in [850.75 \text{ \AA}, 1420.75 \text{ \AA}] \quad (3.7)$$

with a linearly equal spacing of $\Delta\lambda = 0.25\text{\AA}$. [8] only modelled the null model in the Lyman- α region, $[911.75\text{\AA}, 1215.75\text{\AA}]$. We extend the red end of our model to include a part of the metal line region until 1420.75 \AA . This empirically improved the column density estimation of DLAs near the Lyman- α emission peak, as otherwise part of the damping wing would go beyond 1215.75 \AA when a large DLA is very close to the quasar.

The mean function is thus written as a mean vector $\boldsymbol{\mu}(z_{\text{QSO}}; \beta_{\text{MF}}, \tau_{0,\text{MF}}) = \mu(\boldsymbol{\lambda}, z_{\text{QSO}}; \tau_{0,\text{MF}}, \beta_{\text{MF}})$ and the kernel is written as a matrix $\Sigma(\lambda, \lambda') = \boldsymbol{\Sigma}$. The covariance matrix's optimisation procedure is described in [3, 8]. We factorise the covariance matrix as in [8]:

$$\boldsymbol{\Sigma}_i = \mathbf{A}_F^\top (\mathbf{K} + \boldsymbol{\Omega}) \mathbf{A}_F + \text{diag } \boldsymbol{\nu}_i. \quad (3.8)$$

The \mathbf{K} matrix is a positive-definite symmetric matrix corresponding to the covariance between each quasar flux pixel. $\boldsymbol{\Omega}$ is a diagonal matrix describing the absorption noise:

$$\text{diag } \boldsymbol{\Omega} = \boldsymbol{\omega} \circ (1 - \exp(-\tau_{\text{eff,HI}}(\boldsymbol{z}; \beta, \tau_0)) + c_0)^2. \quad (3.9)$$

ω is freely optimisable while the Lyman- α flux term, $(1 - \exp(\tau_{\text{eff,HI}}(\mathbf{z}; \beta, \tau_0)) + c_0)^2$, includes the redshift dependent noise variance with which we model the Lyman- α forest. The optimised absorption noise parameters used here are:

$$\tau_0 = 0.000119 \quad \beta = 5.15 \quad c_0 = 0.146. \quad (3.10)$$

The \mathbf{A}_F is a diagonal matrix reflecting the mean vector suppression for each spectrum corresponding to the mean flux in the Lyman- α forest:

$$\text{diag } \mathbf{A}_F = \exp(-\tau_{\text{eff,HI}}(z_{\text{QSO}}; \beta_{\text{MF}}, \tau_{0,\text{MF}})). \quad (3.11)$$

The parameters of this matrix follow the values given in [97], which used a power-law relation to measure the effective optical depth in the Lyman- α forest in SDSS DR12:

$$\tau_{0,\text{MF}} = 0.00554 \quad \beta_{\text{MF}} = 3.182, \quad (3.12)$$

with associated uncertainty for each parameter:

$$\sigma_{\tau_{0,\text{MF}}} = 0.00064 \quad \sigma_{\beta_{\text{MF}}} = 0.074. \quad (3.13)$$

The instrumental noise is encoded in the diagonal matrix $\text{diag } \nu_i$, where i simply denotes the i th quasar observation: The final covariance matrix learned from our data is shown in Fig 3.2. Comparing the kernel matrix we learned in this work to [8], the current kernel is less noisy and contains several distinct features of emission lines. The reduction in the noise is due to a larger training set, SDSS DR12Q catalogue, is used for optimising the kernel.

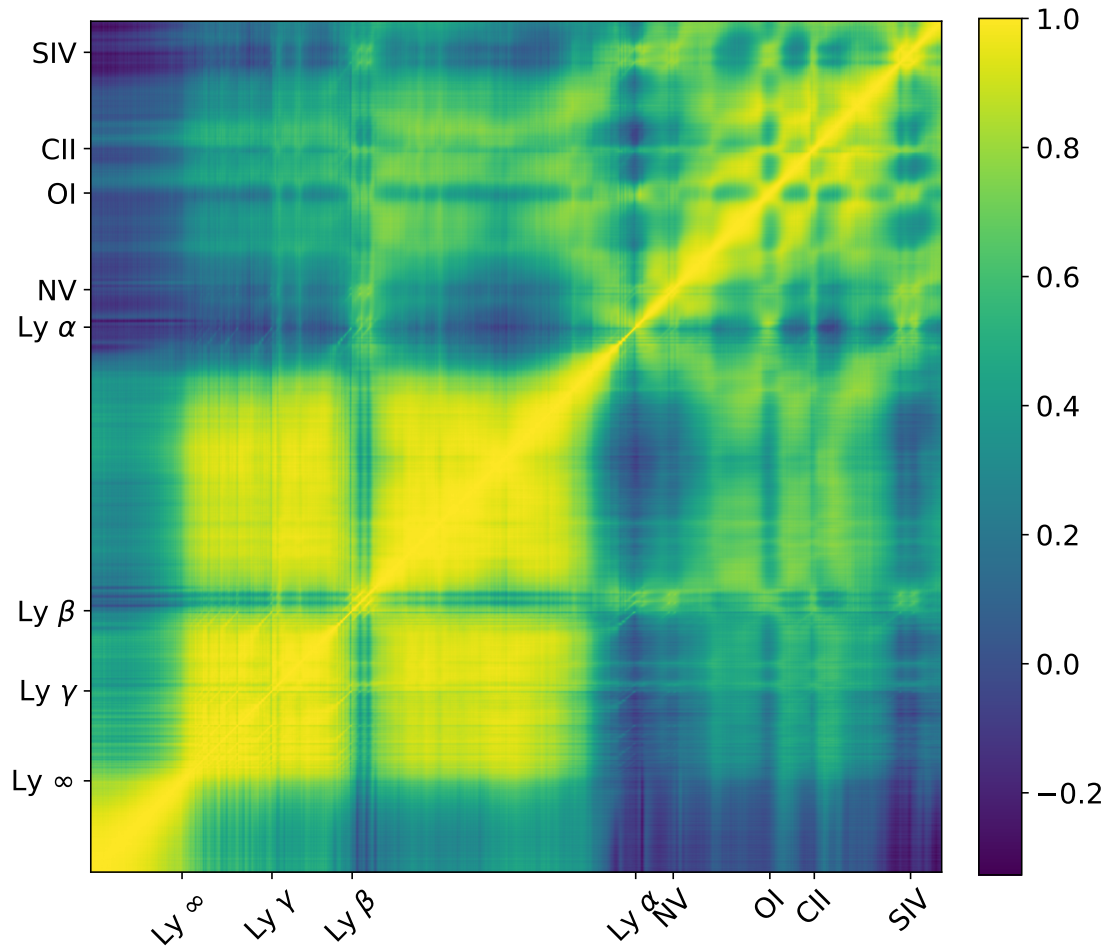


Figure 3.2: The correlation matrix learned from data, which is the covariance matrix \mathbf{K} normalised by the diagonal elements. Note that the correlation in the plot is pixel-by-pixel, and the matrix dimension is 2281×2281 . Different emission lines and the Lyman break are visible in the plot.

After having learned the GP null model, we can write down the null model likelihood function:

$$\begin{aligned}
p(\mathbf{y} \mid \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\text{QSO}}, \beta_{\text{MF}}, \tau_{0,\text{MF}}, \mathcal{M}_{\text{-DLA}}) = \\
\mathcal{N}(\mathbf{y}; \boldsymbol{\mu}(z_{\text{QSO}}; \beta_{\text{MF}}, \tau_{0,\text{MF}}), \mathbf{A}_{\text{F}}^{\top}(\mathbf{K} + \boldsymbol{\Omega})\mathbf{A}_{\text{F}} + \text{diag } \boldsymbol{\nu}_i),
\end{aligned}
\tag{3.14}$$

where the notation $\mathcal{M}_{\text{-DLA}}$ specifies that our null GP model is conditioned on a training set *without DLAs*.

Model evidence for the null model

Once we have trained our GP null model, $\mathcal{M}_{\text{-DLA}}$, according to Section 3.3.2, we need to integrate out the nuisance parameters associated with Lyman- α forest absorption to get the model evidence.

In [8], we only took the mean values of the meanflux parameters $(\beta_{\text{MF}}, \tau_{0,\text{MF}})$ without their uncertainties, so the model evidence straightforwardly equals to Eq 3.14 without integration. In this work, we take the uncertainties of meanflux suppression into account and integrate them out, according to [97] prior. The model evidence thus will be:

$$p(\mathcal{D} \mid \mathcal{M}_{\text{-DLA}}, \boldsymbol{\nu}, z_{\text{QSO}}) \propto p(\mathbf{y} \mid \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\text{QSO}}, \mathcal{M}_{\text{-DLA}}),
\tag{3.15}$$

where we integrate out $(\beta_{\text{MF}}, \tau_{0,\text{MF}})$

$$\begin{aligned}
p(\mathbf{y} \mid \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\text{QSO}}, \mathcal{M}_{\text{-DLA}}) = \\
\int p(\mathbf{y} \mid \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\text{QSO}}, \beta_{\text{MF}}, \tau_{0,\text{MF}}, \mathcal{M}_{\text{-DLA}}) \\
p(\beta_{\text{MF}})p(\tau_{0,\text{MF}})d\beta_{\text{MF}}d\tau_{0,\text{MF}}
\end{aligned}
\tag{3.16}$$

with

$$p(\beta_{\text{MF}}) = \mathcal{N}(\beta_{\text{MF}} = 3.182, \sigma_{\beta_{\text{MF}}} = 0.074) \quad (3.17)$$

$$p(\tau_{0,\text{MF}}) = \mathcal{N}(\tau_{0,\text{MF}} = 0.00554, \sigma_{\tau_{0,\text{MF}}} = 0.00064).$$

We then use Quasi-Monte Carlo (QMC) to integrate out the meanflux parameters with 30 000 samples of $(\beta_{\text{MF}}, \tau_{0,\text{MF}})$. QMC takes samples from a so-called low-discrepancy sequence, leading to faster convergence. Here we draw 30 000 samples generated from a scrambled Halton sequence, which gives samples approximately uniformly distributed on a unit square $[0, 1]^2$. We then use inverse transform sampling to transform the Halton sequence to the distribution described in Eq 3.17.

Model evidence for the DLA model

Suppose we have a trained GP null model in Eq 3.14, the DLA likelihood function will be the null model likelihood function multiplied by Voigt profiles for each line in the Lyman series of the absorber:

$$\begin{aligned} p(\mathbf{y} \mid \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\text{QSO}}, \beta_{\text{MF}}, \tau_{0,\text{MF}}, \\ \{z_{\text{DLA } i}\}_{i=1}^k, \{N_{\text{HI } i}\}_{i=1}^k, \mathcal{M}_{\text{DLA}(k)}) \\ = \mathcal{N}(\mathbf{y}; \mathbf{a}_{(k)} \circ \boldsymbol{\mu}(z_{\text{QSO}}; \beta_{\text{MF}}, \tau_{0,\text{MF}}), \\ \mathbf{A}_{(k)}(\mathbf{A}_{\text{F}}(\mathbf{K} + \boldsymbol{\Omega})\mathbf{A}_{\text{F}})\mathbf{A}_{(k)} + \text{diag } \boldsymbol{\nu}_i). \end{aligned} \quad (3.18)$$

Here $\mathbf{A}_{(k)} = \text{diag } \mathbf{a}_{(k)}$ and $\mathbf{a}_{(k)}$ is the function with k voigt profiles, which represents k DLAs:

$$\mathbf{a}_{(k)} = \prod_{i=1}^k a(\boldsymbol{\lambda}; z_{\text{DLA } i}, N_{\text{HI } i}). \quad (3.19)$$

$a(\boldsymbol{\lambda}; z_{\text{DLA}}, N_{\text{HI}})$ is a Voigt profile parameterised by the DLA’s redshift, z_{DLA} , and the column density of the DLA, N_{HI} . The Voigt profile parameterisation used in this work is the same as [3]. We set the maximum number of DLAs per spectrum at $k = 3$ in this work, as there are rarely more than three DLAs per spectrum. As described in [3], the default Voigt profile we use in this work includes Ly α , Ly β , and Ly γ absorption, which allows us to constrain the DLA column density better when the Lyman- β forest is in the observation window.

To get the model evidence, according to Eq 3.18, we need to integrate out the prior over the DLA parameters and the meanflux parameters $(\beta_{\text{MF}}, \tau_{0,\text{MF}})$. For convenience, we denote the parameters which need to be integrated out by $\theta = \{\{z_{\text{DLA}_i}\}_{i=1}^k, \{N_{\text{HI}_i}\}_{i=1}^k, \beta_{\text{MF}}, \tau_{0,\text{MF}}\}$.

For the model with a single DLA, we have four parameters $\theta = \{z_{\text{DLA}}, N_{\text{HI}}, \beta_{\text{MF}}, \tau_{0,\text{MF}}\}$.

The model evidence is:

$$p(\mathbf{y} \mid \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\text{QSO}}, \mathcal{M}_{\text{DLA}}) = \int p(\mathbf{y} \mid \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\text{QSO}}, \theta, \mathcal{M}_{\text{DLA}}) p(\theta \mid z_{\text{QSO}}, \mathcal{M}_{\text{DLA}}) d\theta. \quad (3.20)$$

By assuming each parameter is independent of each other, we factorise the parameter prior as:

$$p(\theta \mid z_{\text{QSO}}, \mathcal{M}_{\text{DLA}}) = p(z_{\text{DLA}} \mid z_{\text{QSO}}, \mathcal{M}_{\text{DLA}}) p(N_{\text{HI}} \mid \mathcal{M}_{\text{DLA}}) p(\beta_{\text{MF}}) p(\tau_{0,\text{MF}}), \quad (3.21)$$

where we assign the [97] prior for the meanflux parameters as in Eq 3.17. We use the same prior for column density, $p(N_{\text{HI}} \mid \mathcal{M}_{\text{DLA}})$, as [8]. This was trained using kernel density estimation on the $\log_{10} N_{\text{HI}}$ distribution from [64] DR9 DLAs with an addition of a 3% uniform prior.

The z_{DLA} prior is uniform within the search range for DLAs. We set this search range to be from Lyman- β to Lyman- α . Removing DLAs detected in the Lyman- β forest

ensures the purity of DLA samples in deriving the statistical properties of the DLA population. However, to generate a complete catalogue, we also consider a search range from the Lyman limit to Lyman- β .

We used the same model and priors for the sub-DLA model as in [8]. The sampling range of the redshifts of sub-DLAs is the same as for the DLA model. Model priors are the same as [8], based on the DLA catalogue in SDSS DR9 [63].

3.3.3 Example spectra

In this section, we show some example spectra to demonstrate our proposed model. Figure 3.3 shows an example with prominent DLA features. As shown in the parameter space (middle plot), the posterior distribution is peaked at the *maximum a posteriori* (MAP) values of those two DLAs. Our GP model estimates the parameters of the DLAs with small uncertainties. As shown in the top plot, our MAP values agree with the column densities measured by the CNN model reported in the DR16Q catalogue.

Figure 3.4 shows an extremely noisy spectrum, for which our GP model is very uncertain about the effective Lyman- α absorption in the spectrum. The DLA models are degenerate with the absorption from the Lyman- α forest. Without modelling the uncertainty in the mean flux, the GP model does not know that the drop in the spectrum can be explained by Lyman- α forest absorption. It instead fits a big DLA with $N_{\text{HI}} = 10^{22.9} \text{ cm}^{-2}$ as its preferred explanation for the drop in flux.

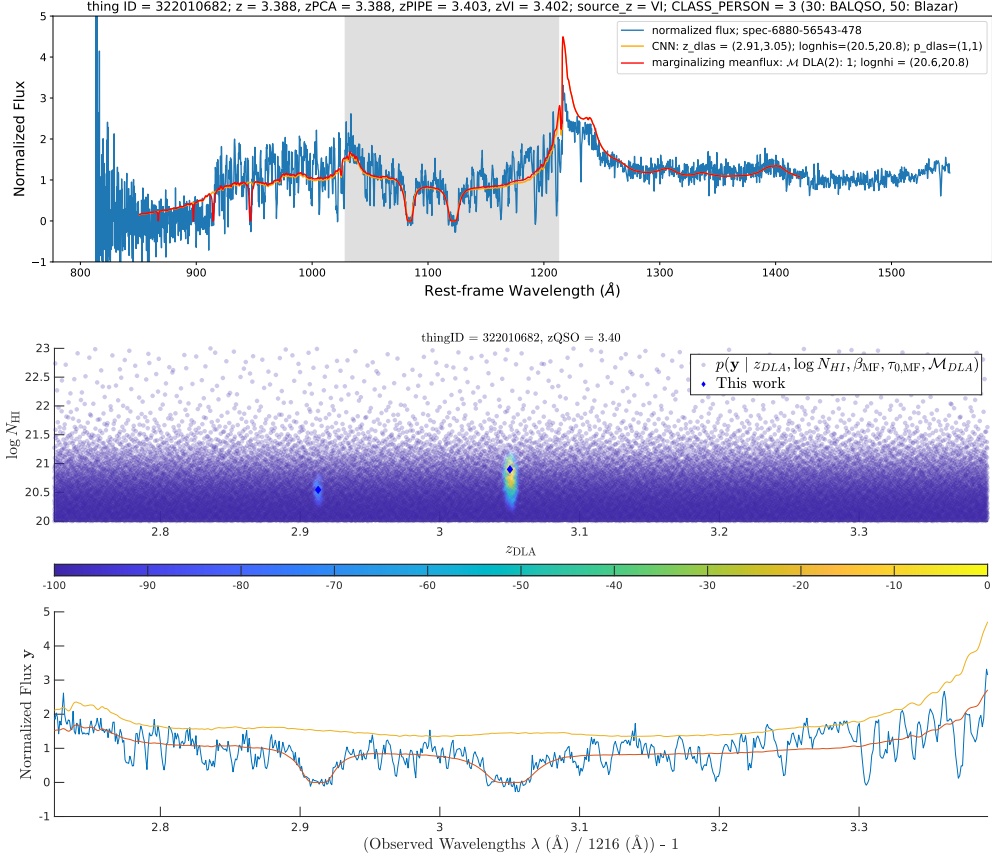


Figure 3.3: An example of a spectrum with distinct DLA features. **(Top):** The normalised observed spectrum in rest-frame wavelengths (blue) with the GP model (red) and the detection from the CNN model reported in DR16Q (orange). The title shows a series of column values in SDSS DR16Q catalogue, including SDSS identifier, best available redshift, PCA redshift, SDSS pipeline redshift, redshift from visual inspection, source for the best available redshift, and object classification from visual inspection (0: not inspected; 1: star; 3: quasar; 4: galaxy; 30: BAL quasar; 50: Blazar(?)). Shaded area (grey) shows the sampling range of z_{DLA} , which is from $\text{Ly}\beta + 3000 \text{ km s}^{-1}$ to $z_{\text{QSO}} - 3000 \text{ km s}^{-1}$. The legend shows the spectrum is from spec-6880-56543-478 (spec-plate-mjd-fiber_id). The CNN model (orange) detected two DLAs, with redshifts of $z_{\text{DLA}} = 2.91, 3.05$ and column densities of $\log_{10} N_{\text{HI}} = 20.5, 20.8$, at DLA confidence = 1 for each DLA. Our GP model (red) also detected two DLAs with the model posterior $p(\mathcal{M}_{\text{DLA}(2)} | \mathcal{D}) = 1$ and column densities $\log_{10} N_{\text{HI}} = 20.6, 20.8$. **(Middle):** The sample likelihoods of detecting DLAs in the parameter space, $\theta \in (z_{\text{DLA}}, \log_{10} N_{\text{HI}})$. Colour bar shows the normalised log likelihoods, $\log p(\mathbf{y} | z_{\text{DLA}}, \log_{10} N_{\text{HI}}, \tau_{0,\text{MF}}, \beta_{\text{MF}}, z_{\text{QSO}}, \mathcal{M}_{\text{DLA}})$, with the maximum log likelihood to be zero. We also show the maximum a posteriori estimates of DLAs in the blue squares. The posterior distribution sharply peaks at the parameter space, indicating the detection of these DLAs in high confidence. **(Bottom):** The observed flux (blue) as a function of absorber redshifts with the GP model (red) and the GP model before the meanflux suppression (yellow). The position on the x-axis directly corresponds to the x-axis in the middle plot.

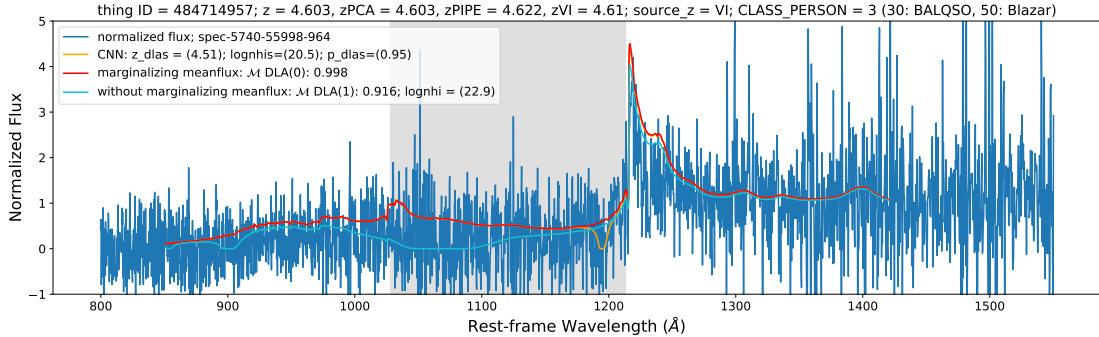


Figure 3.4: An example of a noisy spectrum with an uncertain meanflux. The normalised observed spectrum in rest-frame wavelengths (blue) with the GP model (red) and the detection from the CNN model reported in DR16Q (orange). We also plot the result without marginalising the uncertainty of meanflux prior (cyan). Shaded area (grey) shows the sampling range of z_{DLA} , which is from $\text{Ly}\beta + 3000 \text{ km s}^{-1}$ to $z_{\text{QSO}} - 3000 \text{ km s}^{-1}$. Our proposed model (red) indicates no DLA in the spectrum, with the null model posterior $p(\mathcal{M}_{\text{-DLA}} | \mathcal{D}) = 0.998$. On the other hand, if our model ignores the uncertainty of $(\beta_{\text{MF}}, \tau_{0,\text{MF}})$, it would falsely detect a DLA with $p(\mathcal{M}_{\text{DLA}} | \mathcal{D}) = 0.916$ with $\log_{10} N_{\text{HI}} = 22.9$ (cyan). When marginalising over the uncertainty in effective optical depth, our proposed model (red) avoids detecting a false-positive large DLA.

3.3.4 Selection on the strength of Occam’s razor

As we use more parameters to compute the DLA or sub-DLA model, the model selection will prefer to fit a Voigt profile to the GP if all candidate models are poorly fit. Thus, the DLA or sub-DLA model’s evidence is sometimes too strong compared to the null model.

The most common poor fit situations are quasar spectra with $z_{\text{QSO}} < 2.5$ and with low signal-to-noise ratios (SNR). As SDSS optical spectra have a fixed observing window, quasar spectra with $z_{\text{QSO}} < 2.5$ have an incomplete Lyman- α forest. The constraining power of the quasar becomes weaker as only part of the data fits into our modelling window, $[850.75 \text{ \AA}, 1420.75 \text{ \AA}]$. Thus the DLA model and the null model are closer in likelihood space.

To avoid this situation, we introduced an additional Occam’s razor in [8], which is injected in the model selection as:

$$\Pr(\mathcal{M}_{\text{DLA}} | \mathcal{D}) = \frac{\Pr(\mathcal{M}_{\text{DLA}})p(\mathcal{D} | \mathcal{M}_{\text{DLA}})^{\frac{1}{N}}}{\left(\Pr(\mathcal{M}_{\text{DLA}})p(\mathcal{D} | \mathcal{M}_{\text{DLA}}) + \Pr(\mathcal{M}_{\text{sub}})p(\mathcal{D} | \mathcal{M}_{\text{sub}}) \right)^{\frac{1}{N}} + \Pr(\mathcal{M}_{\text{-DLA}} | \mathcal{D})}, \quad (3.22)$$

Here N is the Occam’s razor penalty, and we used $N = 10\,000$ in [8]. We previously validated the Occam’s razor strength by matching it to the DR9 concordance catalogue [63].

In this work, however, we modify our null model to consider uncertainty from the mean flux measurement, which means it has more parameters. Thus, the null model gains more constraining power, so a weaker Occam’s razor may be preferable. To make our model posteriors more consistent with human identifications, we decided to conduct a visual inspection on a small subset of the spectra.

We first train a model without Occam’s razor and select at random from this model 239 putative large DLAs with $N_{\text{HI}} > 10^{22} \text{ cm}^{-2}$ and 243 putative small DLAs with $10^{20} \leq N_{\text{HI}} < 10^{21} \text{ cm}^{-2}$. We visually inspect each spectrum and compute the model posteriors with a range of strengths for Occam’s razor, $N = \{1, 10, 100, 1\,000, 30\,000\}$. We then treat each spectrum as a multiple-choice problem: if we think the model posterior of a given Occam’s razor describes the given spectrum well, then we record one vote for this value of Occam’s razor. Multiple selections are allowed for each spectrum as the model posteriors are often very close. After collecting votes, the winning value of Occam’s razor was $N = 1\,000$, a ten times reduction from our earlier value.

For quasar spectra with $z_{\text{QSO}} > 2.5$ there are enough data points in the Lyman- α range that the strength of Occam’s razor has a small effect. We will discuss the effect of Occam’s razor in Section 3.5. We suggest incorporating variations due to Occam’s razor into the uncertainty in population statistics for conservative usage.

3.3.5 Summary of the modifications

Here we summarise the modifications we made in this work, comparing to the model of [8]:

1. Our training set is SDSS DR12 quasar spectra with DLAs detected by [8] removed. We considered a DLA to be detected if the posterior probability of a spectrum containing a DLA is larger than 0.9, $P(\mathcal{M}_{\text{DLA}} | \mathcal{D}) > 0.9$.
2. The wavelength range modelled goes from $\lambda_{\text{rest}} = 850.75 \text{ \AA}$ to $\lambda_{\text{rest}} = 1420.75 \text{ \AA}$.
3. The effective optical depth prior, (τ_0, β) , is updated from [66] to [97].
4. The uncertainty in the mean flux suppression parameters, τ_0 and β , is marginalised while computing the model evidence.

The first modification gives us a training set size containing 89,408 spectra without DLAs. The larger training set better learns the covariance structure of quasar emission lines, which allows the second modification: expanding the model to cover the Lyman break and SiIV line. The expansion enables the model to use the metal lines to constrain the correlations of the emission lines in the Lyman- α forest. When using the previous modelling range, $[911.75\text{\AA}, 1215.75\text{\AA}]$, we found that we often detected DLAs with high N_{HI} in the red end

of the spectrum, where the code inserts a DLA at the quasar redshift to compensate for an oddly shaped Lyman- α emission line. This was possible because when we cut the spectrum at a rest frame wavelength 1215.75\AA , half of the damping wings were removed, allowing for more model freedom and dubious N_{HI} estimation.

Third, to make the mean flux suppression prior for (τ_0, β) consistent with the DR12Q training set, we switched to the mean flux measurement based on BOSS DR12Q [97]. Our last modification is marginalising the uncertainty of [97]’s parameters while marginalising the DLA parameters.

To compute the statistical properties of DLAs, we need to convert the posterior distribution of a DLA in each spectrum into the expected number of DLAs per redshift or column density bin, for which we use the method described in [62]. We briefly summarise the modelling decisions we made to produce the DLA samples in the result section:

- Search range: from Lyman- β to Lyman- α .²
- Maximum number of DLAs: three.
- Maximum z_{QSO} : quasar redshifts < 7 .
- DLA redshift $2 < z_{\text{DLA}} < 5$.

²For the CDDF in [8], we used a sampling range from $\text{Ly}\beta + 3\,000\text{ km s}^{-1}$ to $\text{Ly}\alpha - 30\,000\text{ km s}^{-1}$ to avoid finding DLAs in the proximity zone. Here, we instead use $\text{Ly}\beta + 3\,000\text{ km s}^{-1}$ to $\text{Ly}\alpha - 3\,000\text{ km s}^{-1}$. This has a very moderate effect on our results, however, we provide a check of systematics due to removing DLAs near to the quasar redshift in Section 3.5.2.

3.4 Results

3.4.1 Column density distribution function

Figure 3.5 shows the CDDF we estimate from DR16Q spectra. In the following sections, the CDDF is computed for $N_{\text{HI}} \in [10^{20}, 10^{23}]$, while the DLA incidence rate dN/dX and the total HI density in DLAs Ω_{DLA} are computed for $N_{\text{HI}} \in [10^{20.3}, 10^{23}]$. Ho20 refers to [8], a DR12 DLA catalogue that used a modified GP model from [3].

The CDDF is a histogram of column densities normalised by the effective spectral path that could contain DLAs. We count all spectral path with an absorber with $z_{\text{DLA}} < 5$. Error bars denote the 68% confidence limits, and the grey band represents the 95% confidence limits. Note that the uncertainties here are the statistical uncertainties associated with the GP model. They do not include uncertainty due to potential systematics. Section 3.5 will describe how possible systematics would affect the CDDF.

As shown in Figure 3.5, we observe non-zero column density until $3 \times 10^{22} \text{ cm}^{-2}$. Our DR16 measurement is mostly consistent with our previous DR12 measurement until $N_{\text{HI}} \leq 9 \times 10^{21}$. For $N_{\text{HI}} \geq 3 \times 10^{22}$, both our DR12 and DR16 measurement are consistent with zero at 95% confidence level, though there is one bin from DR16 not consistent with zero (see Table .6).

We also measure no turn over for the CDDF at the high column end, $N_{\text{HI}} \sim 10^{21.5} \text{ cm}^{-2}$. It was suggested in [98] that molecular hydrogen sets a maximum N_{HI} so that steepen the CDDF at the high end. The latest simulated CDDF from SIMBA [99], which

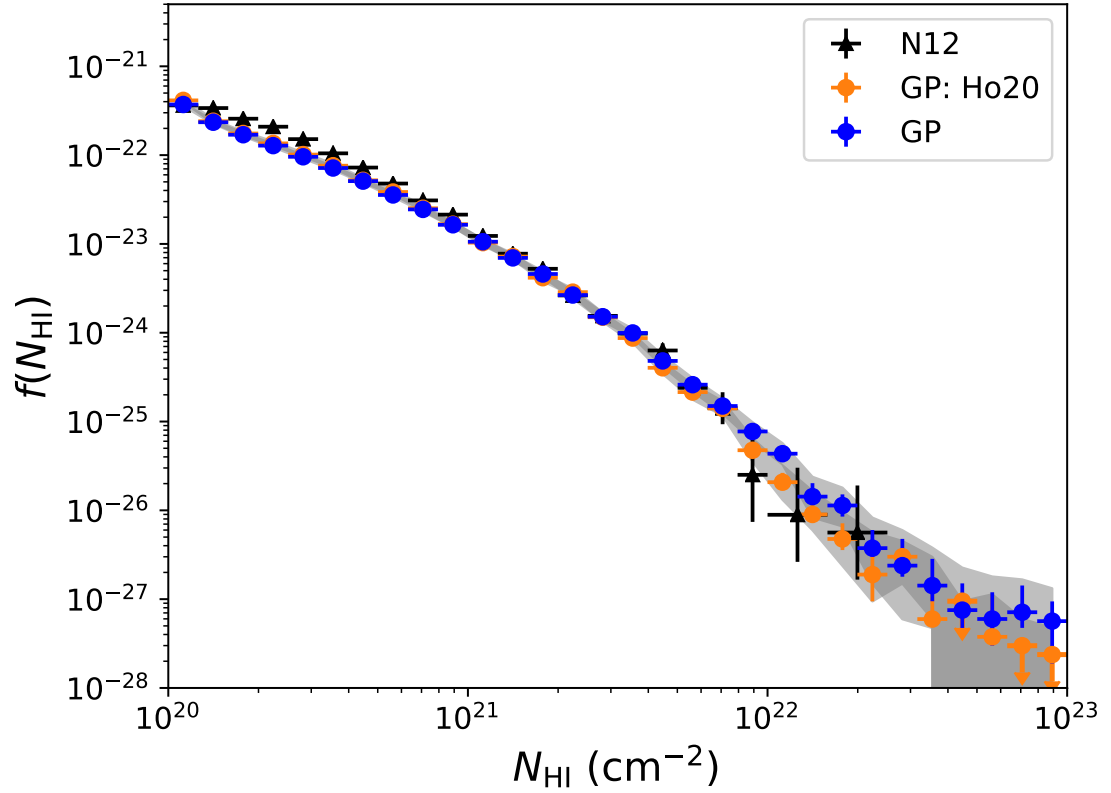


Figure 3.5: The CDDF, integrated over all $z < 5$ spectral path, derived from SDSS DR16Q spectra with our proposed Gaussian process models (GP; blue). The CDDF measurements from [8] (GP: Ho20; orange) are plotted as a comparison. Error bars show the 68% confidence limits, while grey areas show the 95% confidence limits. Black dots are from [5] (N12).

included molecular hydrogen formation in their star formation recipe, predicts no turn over at the high end, consistent with our measurements.

In Figure 3.6, we plot the CDDF with different Occam’s razor strengths. When the Occam’s razor strength is weak ($N = 30$), model selection will find DLAs even though the SNR is low, so we get more absorbers at both high and low column density ends. On the other hand, if the razor strength is strong ($N = 30\,000$), model selection will prefer to avoid finding DLAs at low SNR spectra, which results in a decrease.

However, in general, in Figure 3.6, we observe the razor strength only marginally affects the CDDF. Thus the small tension at the low end, $N_{\text{HI}} \in [10^{20}, 10^{20.3}]$, between our CDDF and N12 is more likely due to other reasons than Occam’s razor.

We show the redshift evolution of the CDDF in Figure 3.7. The downward pointing symbols indicate the 68% upper confidence limit when the data is consistent with zero at 68% confidence limits. As we can anticipate, for high-redshift quasars with $z_{\text{QSO}} > 4$, since the flux is highly absorbed, we detect DLAs with larger uncertainties, and the number of large DLAs is consistent with zero.

In both [62] and [8], we found that the CDDF is getting shallower at $z > 4$. However, given our detection for $N_{\text{HI}} > 4 \times 10^{21}$ at $z > 4$ is highly uncertain and consistent with zero detection, this trend is not significant in our current dataset. Instead, the detection of DLAs with $N_{\text{HI}} < 4 \times 10^{21}$ at $z > 4$ is consistent with the measurements at $z \in [2.5, 4]$. We find no strong evidence for an evolution of the slope of the CDDF at $z > 4$.

One possible reason why we found the CDDF was shallower at $z > 4$ in [62] and [8] is absorption due to the Lyman- α forest. When the spectrum is highly absorbed, there

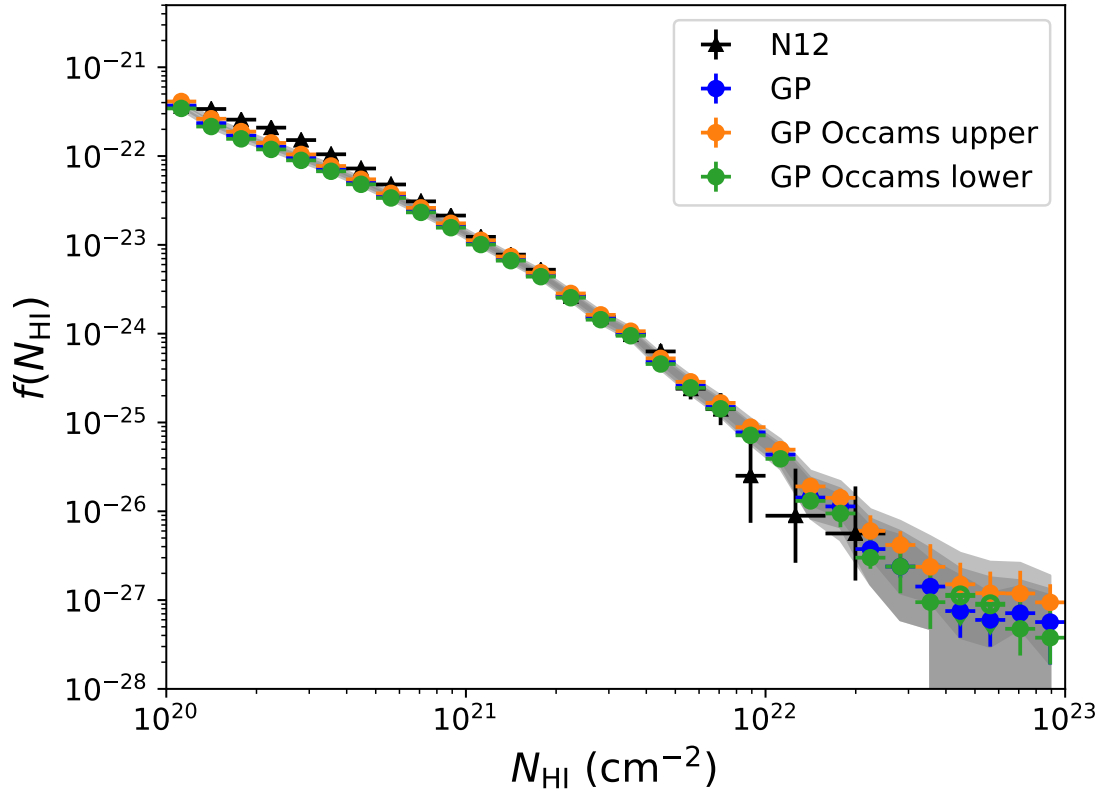


Figure 3.6: The CDDFs with different Occam’s razor strengths, which discussed in Section 3.3.4. Occam’s upper (orange) represents $N = 30\,000$ while Occam’s lower (green) represents $N = 30$. We present our main result (GP; blue) with an optimal strength $N = 1\,000$, which we selected from visually inspecting a subset of the dataset. Note that the difference between different Occam’s strengths is well within 95% confidence limits. Black dots are from [5] (N12).

is a degeneracy between a large DLA and the forest’s absorption. In [62], we did not model the GP mean as a function of effective optical depth, so it is possible the model was trying to use DLAs to compensate the excess absorption due to the forest, which results in a shallower CDDF at $z > 4$. In [8], the slope of the CDDF is less shallow at $z > 4$, as we modelled the effective optical depth into our GP mean. In this work, we integrated out the measurement uncertainty of the mean flux, and the slope is almost indistinguishable from the CDDF at $z \in [2.5, 4]$. This may indicate that, to understand the DLAs at $z > 4$ better, we need a better measurement for the effective optical depth at $z > 4$.

One interesting feature in Figure 3.7 is the drop in the amplitude of the CDDF at $z \in [2, 2.5]$. As we will discuss in Section 3.4.2, the drop of CDDF at the low redshifts also shows in the DLA incident rate, dN/dX . We will discuss this in more detail in the next section.

3.4.2 Redshift evolution of DLAs

Figure 3.8 shows the incident rate of DLAs, dN/dX , as a function of absorber redshift. Our results are consistent with [6] and [8] and are slightly lower than [5]. dN/dX is sensitive to the weaker DLAs, so the difference between [5] and [6] is likely due to the false positive rate.

[6] performed a visually-guided Voigt profile fitting on SDSS-DR5. Though their sample size is smaller, with the help of the human eye, their method is likely less prone to false positives than the automated template fitting used in [5]. This difference may also explain the drop in amplitude of the CDDF at $z \in [2, 2.5]$. [6] and [5] have a larger

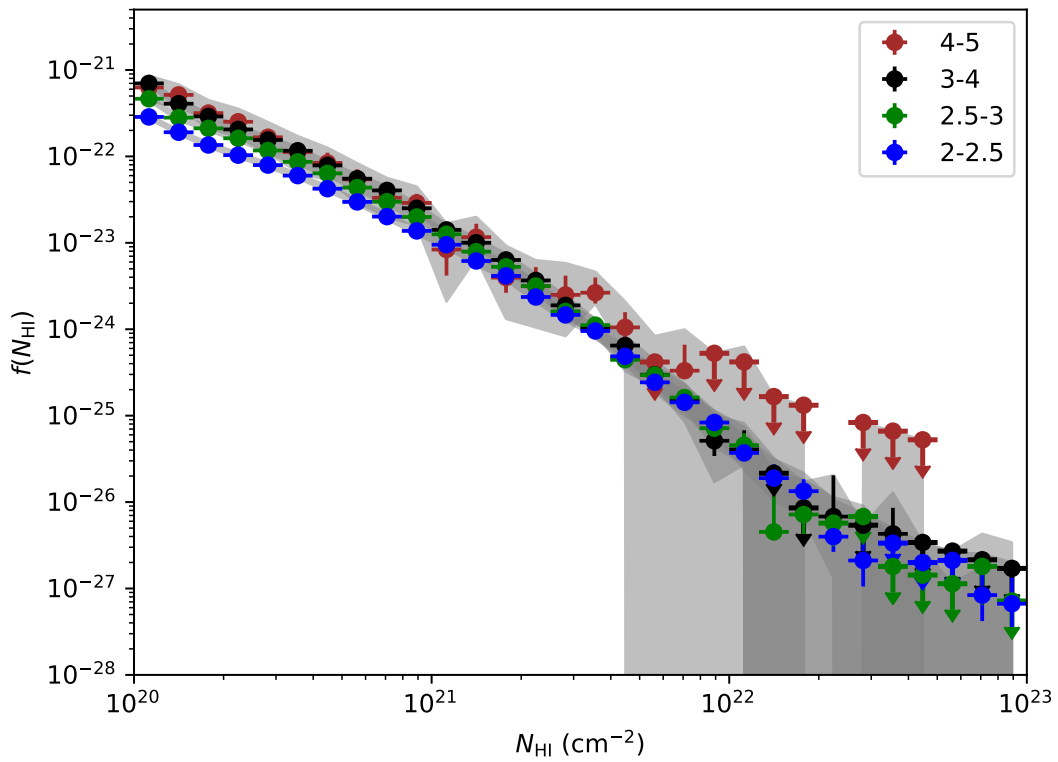


Figure 3.7: The CDDF derived from DLAs in a variety of redshift bins. Labels show the redshift bins in used. We show 68% confidence limits in error bars and 95% confidence limits in grey areas. If the bin is consistent with no detection at 68% limits, we show a down-pointing arrow indicating the 68% confidence upper limit.

discrepancy at $z \in [2, 2.5]$, and [5] may overestimate the weak absorbers at this redshift range where the spectra are short. Our measurement is consistent with [6], which implies that we detect fewer weak absorbers in low redshift bins and explains the small tension in the CDDF at low- N_{HI} .

One noticeable feature in dN/dX , which we have not discussed before, is the decrease of the line density from $z = 3.5$ to $z = 4.0$ and another increase at $z > 4.0$. This feature is also shown in our Ho20 measurement. The drop of dN/dX at $z \in [3.5, 4]$ is consistent with [6] at 95% confidence limits. One interesting question is whether the increase from $z \in [4.0, 4.5]$ is real. The measurements at $z > 4$ still have large error bars, so it is hard to say whether dN/dX at $z > 4$ is an increase or a flat line. More data, especially with high SNR, are needed to determine the trend of line density at $z > 4$.

In Figure 3.9, we show the total HI density in DLAs in terms of cosmic density. Our results are mostly consistent with [5] at $z \in [2.5, 3.5]$. At higher redshift bins, $z > 3.5$, our measurements are consistent with [6] and [7]. [7] used high signal-to-noise spectra from a smaller survey, so they have larger error bars. Comparing to Ho20, our current Ω_{DLA} has more mass at low-redshifts ($z_{\text{DLA}} \sim 2$) and less mass at $z_{\text{DLA}} \in [3.5, 4]$. The trend of Ω_{DLA} in DR16 is shallower than Ho20.

We also plot the Ω_{DLA} measured by [9] in Figure 3.9. We see our DR16 measurement is consistent with [9] even at $z > 3.5$. There is a slight tension at $z < 2.5$, which may be because some low-redshift spectra are too short and noisy to measure column density confidently using our model. SDSS spectra with $z_{\text{QSO}} < 2.5$ only covers a region from the Ly α to Ly β or shorter. When the signal-to-noise is low, it is difficult to identify DLAs even

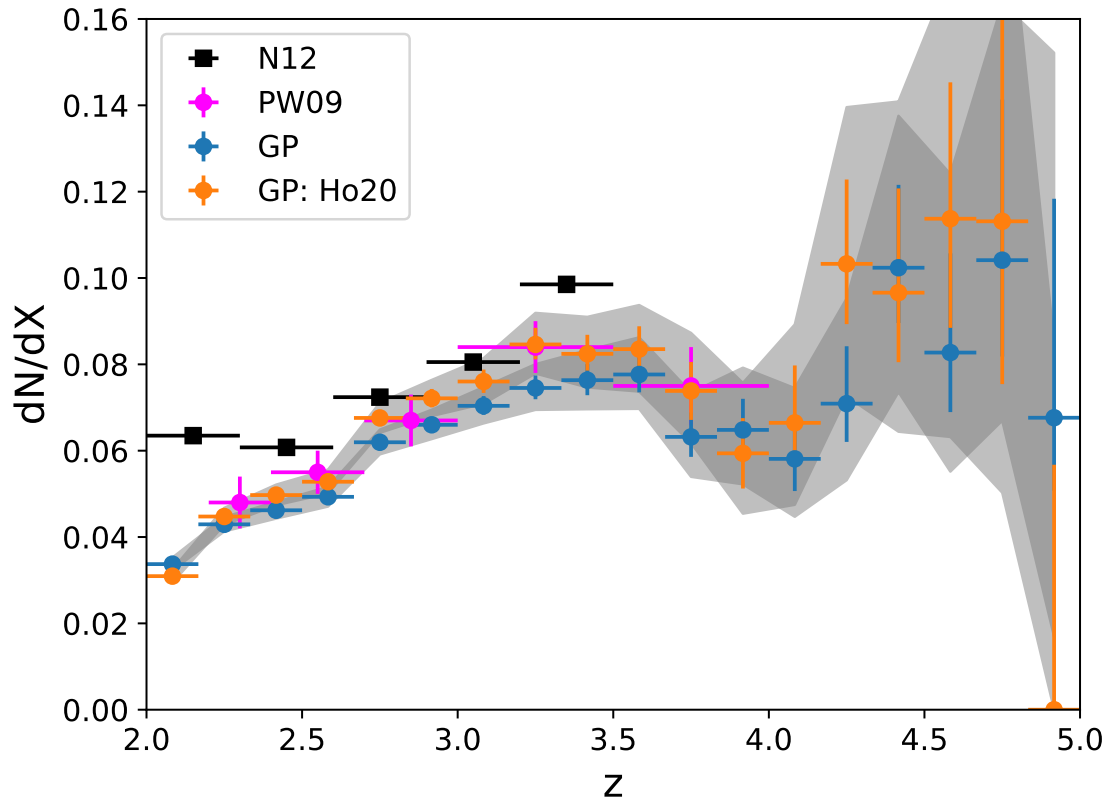


Figure 3.8: The incident rate of DLAs as a function of redshift, integrated over $\log_{10} N_{\text{HI}} > 20.3$ spectra from our catalogue (GP; blue). We also plot the line densities from [5] (N12; black) and [6] (PW09; pink), and [8] (GP: Ho20; orange) as comparisons.

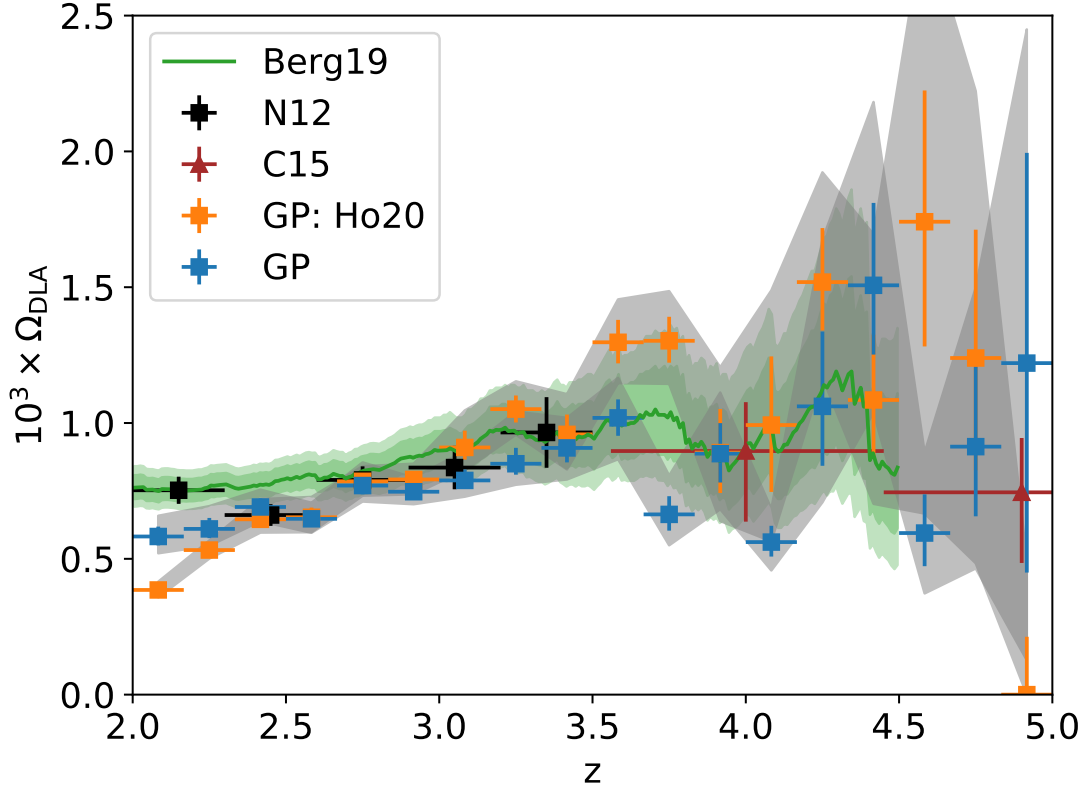


Figure 3.9: The total HI density in DLAs, integrated over DLAs with $\log_{10} N_{\text{HI}} > 20.3$ in our catalogue (GP; blue). For comparison, we plot the measurements from [9] (Berg19; green line and shaded area), [5] (N12; black), [7] (C15; red), and [8] (GP: Ho20; orange).

using human eyes. As shown in Fig 3.10, a different selection of Occam’s razor could moderately affect the two bins with $z < 2.33$. The strength of Occam’s penalty corresponds to a prior belief in detecting a DLA in a short and noisy spectrum, as discussed in Section 3.3.4.

Note that, from Figure 3.7 we see there are no solid detections for $N_{\text{HI}} > 3 \times 10^{21}$ DLAs at $z > 4$. In [62], Ω_{DLA} was skewed towards high values at $z > 4$ even without real detections of large DLAs. Our result in Figure 3.9 does not have this issue. This may

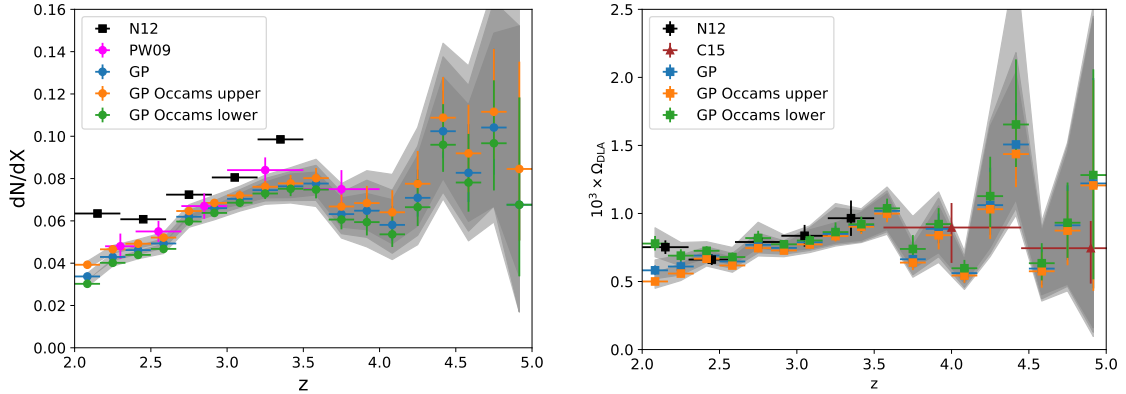


Figure 3.10: The line density (**left**) and Ω_{DLA} (**right**) in DLAs as a function of redshifts with different Occam’s razor strengths. Occam’s upper (orange) represents $N = 30\,000$ while Occam’s lower (green) represents $N = 30$. The main result (GP; blue) is computed with $N = 1\,000$.

indicate our proposed method of integrating out the uncertainty on meanflux measurement helps us avoid the forest biasing the posterior density of N_{HI} towards the high end.

In general, we observe an increase of Ω_{DLA} from $z = 2$ to $z = 3.5$, and a slight decrease from $z = 3.5$ to $z = 4$. For $z_{\text{QSO}} > 4$, the measurement error is larger and less correlated between redshift bins, as in Ho20. This is reasonable given the lower quasar number density at high redshift.

Figure 3.10 shows the line density and Ω_{DLA} with various Occam’s razor strengths. As expected, the razor strengths only moderately affect the statistics of low redshift spectra. For dN/dX , our results are consistent with [6], even with the weakest razor. N12 still detects somewhat more weak DLAs than we do, even though we only apply a small penalty for these short and noisy spectra.

3.5 Checks for systematics

3.5.1 Effect of signal to noise ratios

Figure 3.11 and Figure 3.12 show the abundance of DLAs from subsets of our catalogue with various signal-to-noise cuts (SNR), $\text{SNR} > 2$ and > 4 . We define our SNR as the median of the ratio between the flux and the instrumental noise within the quasar spectrum redwards of the Lyman- α emission peak. This specific choice is to avoid introducing correlations between the detected DLAs and the SNR. With this definition, 80% of the quasar spectra have $\text{SNR} > 2$, and 46% of the spectra have $\text{SNR} > 4$.

We verify that, in Figure 3.11, the CDDF is not sensitive to the SNR when $N_{\text{HI}} < 10^{22} \text{ cm}^{-2}$. However, we note that the highest non-zero column density at 95% confident limits changed from $N_{\text{HI}} < 3 \times 10^{22} \text{ cm}^{-2}$ to $N_{\text{HI}} \lesssim 10^{22} \text{ cm}^{-2}$ for samples with $\text{SNR} > 4$. This is likely because there are too few high column density absorbers to constrain the CDDF sufficiently at the high end in the smaller high SNR sample.

We find that our Ω_{DLA} measurement exhibits no systematic correlation with the SNR cuts. We notice a dependence of SNR on dN/dX at $z \in [2.0, 2.5]$, which is due to the difficulty of finding DLAs in short and noisy spectra. As discussed in Section 3.3.4, it is hard to find features in these spectra, and the observing window cannot fully cover a high- N_{HI} DLA profile with damping wings. To secure our samples' purity, we use an Occam's razor penalty which may also introduce this SNR dependence at $z \in [2, 2.5]$.

As mentioned in [100], the colour and magnitude criteria used in SDSS for quasar target selection is biased against dusty DLAs, which harbour a certain amount of cold

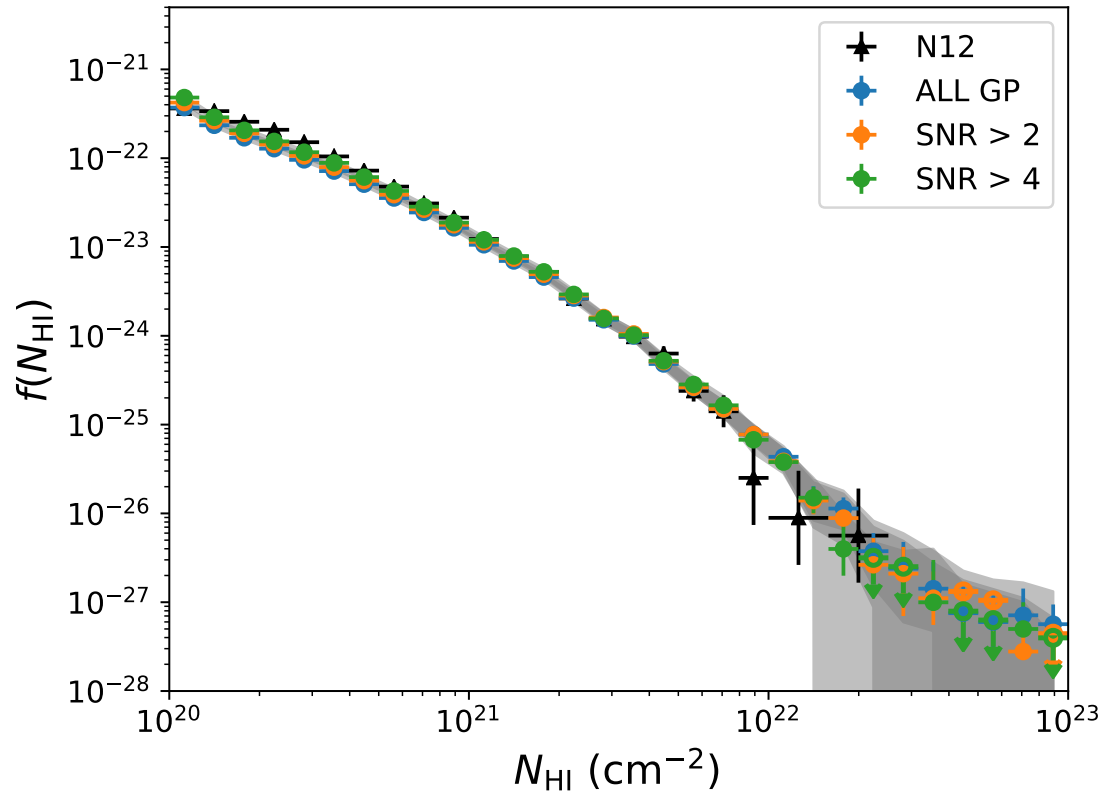


Figure 3.11: The CDDF of DLAs for a subset of samples with different minimal SNRs. SNR > 2 (orange) excludes 20% of the noisiest spectra, and SNR > 4 (green) excludes 54% of the spectra. 68% confidence limits are drawn as error bars, while 95% confidence limits are shown as a grey filled band.

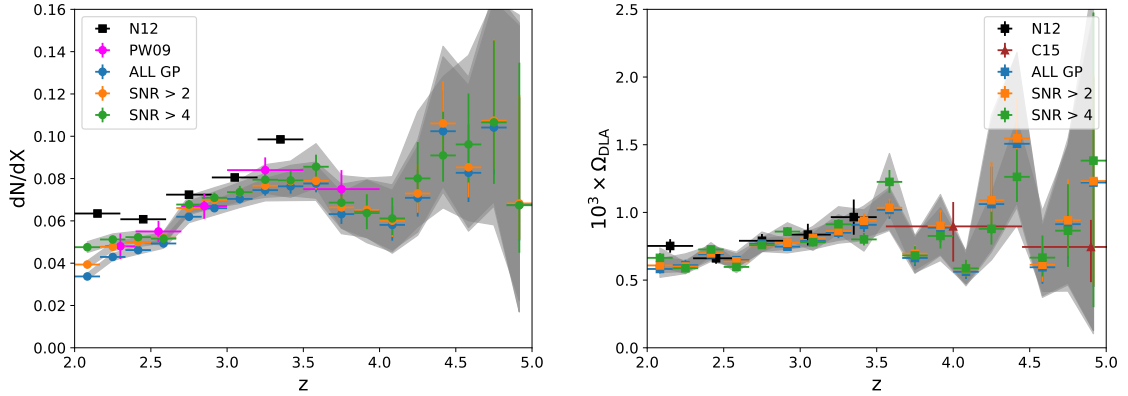


Figure 3.12: The line density (**left**) and total N_{HI} mass (**right**) in DLAs as a function of absorber redshift from subsets of samples with different minimal SNRs. SNR > 2 (orange) excludes 20% of the noisiest spectra, and SNR > 4 (green) excludes 54% of the spectra.

neutral gas. [100] showed that in SDSS DR7 this caused Ω_{DLA} to be underestimated by 10 – 50% at $z \sim 3$. Also, redder quasars containing metal rich dusty DLAs will have lower SNR in the blue part of the spectrum and thus may be excluded from the sample of [94], who enforced $\text{CNR} > 3$. This effect is likely to be substantially reduced in our sample, if present at all, as we use all quasars irrespective of SNR. We also define SNR using the region redwards of the $\text{Ly}\alpha$ emission peak specifically to avoid this kind of selection effect, and we are using DR16, which has a different and more complex selection function. More quantitatively, the XQ-100 targets in [9] use only radio-selected quasars, or quasars previously found by other techniques, and so avoids any SDSS colour selection bias. Figure 3.9 shows that our Ω_{DLA} mostly agrees with [9], implying that colour effects in our sample are smaller than those in DR7.

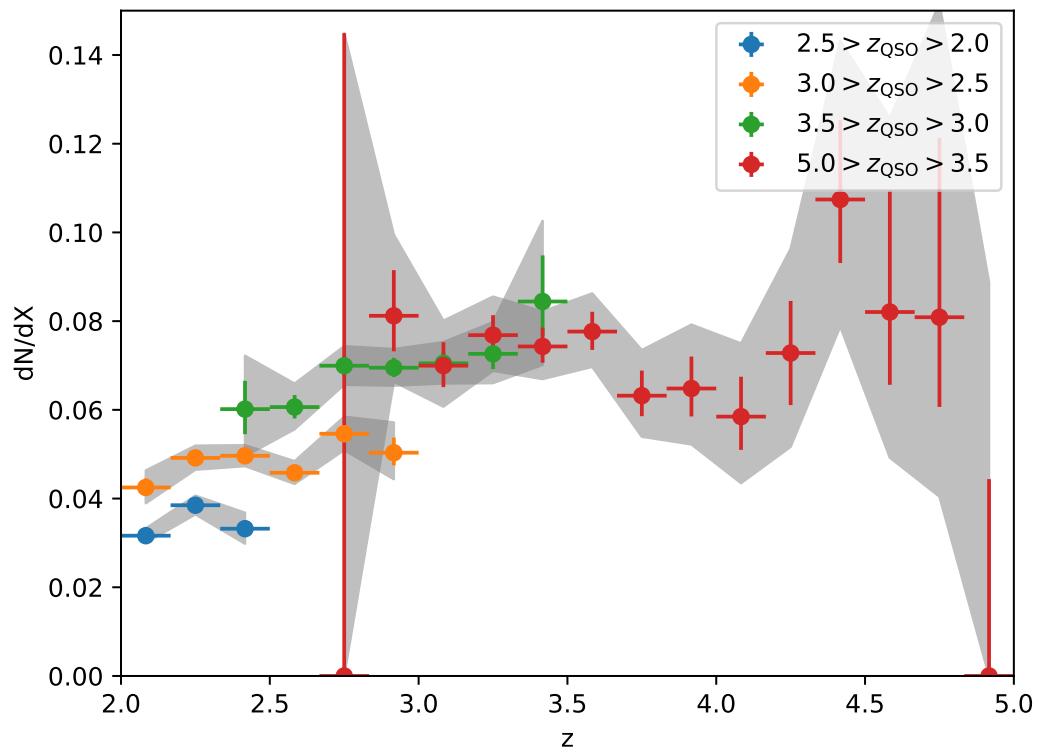


Figure 3.13: The redshift evolution of the incident rate of DLAs, cutting with different quasar redshift intervals. Any correlation between the absorber properties and the background quasars redshifts might indicate systematics.

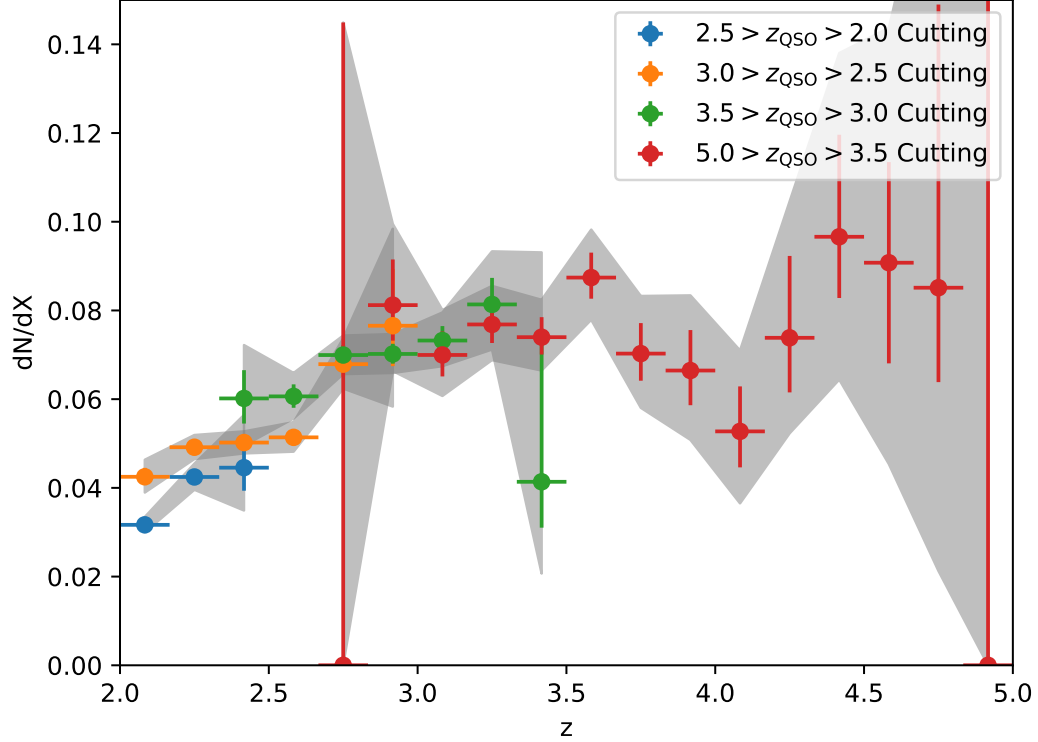


Figure 3.14: The redshift evolution of the incident rate of DLAs , cutting with different quasar redshift intervals. Unlike Figure 3.13, we remove the putative absorbers near the Lyman- α emission line with $|z_{\text{QSO}} - z_{\text{DLA}}| < 30\,000 \text{ km s}^{-1}$.

3.5.2 Effect of quasar redshifts

In Figure 3.13, we test our measured dN/dX with different quasar redshift bins. In a perfect scenario without systematics, we expect that the absorber properties be uncorrelated with the background quasars, as they are widely separated in physical space. However, Figure 3.13, shows some residual correlation between absorber properties and the redshifts of the background quasars for DLAs in spectra with $z_{\text{QSO}} < 3$.

In Figure 3.14, we have investigated removing the sampling range near the quasar redshift, $|z_{\text{QSO}} - z| < 30\,000 \text{ km s}^{-1}$. We found removing the putative absorbers near the Lyman- α emission is sufficient to remove the correlation between quasar redshifts and DLA properties at $z < 3$. A small tension still exists for the $z = 2$ bin within $2.5 > z_{\text{QSO}} > 2.0$ for dN/dX , which may be due to the effect discussed above for SNR, as these very short spectra are often also noisy.

3.5.3 Additional noise test

To understand the implication of applying Occam’s razor to the model posteriors, we conduct a test based on adding noise to a DLA spectrum. We choose a quasar spectrum that we are very confident contains a DLA and add additional Gaussian noise with zero mean and standard deviation σ to the flux and noise variance.

We then examine changes in the DLA model posterior $p(\{\mathcal{M}_{\text{DLA}}\} | \mathcal{D})$. This test will mimic the effect of SNR on the model’s ability to detect the underlying DLAs. For Occam’s razor $N = 30\,000$, the model posterior is $p(\{\mathcal{M}_{\text{DLA}}\} | \mathcal{D}) \simeq 0.9$ for $\sigma \leq 1.5$, which corresponds to $\text{SNR} \simeq 0.9$. On the other hand, for a model without Occam’s razor, the model posterior is $p(\{\mathcal{M}_{\text{DLA}}\} | \mathcal{D}) \simeq 0.9$ for $\sigma \leq 3$, which means $\text{SNR} \simeq 0.5$. A strong Occam’s razor thus introduces false negatives in very noisy spectra. However, by visually inspecting the flux with $\sigma = 3$ we determined that it is almost impossible for humans to identify the underlying DLA. Therefore, we choose to follow the value ($N = 1000$) we determined in Section 3.3.4.

We were unable to quantify the number of false positives, as our simple assumption of Gaussian noise rarely produces correlated structures that resemble DLAs. In practice,

false positives are likely caused by oscillatory structure embedded in the noise, present when the SNR is extremely low.

3.6 Results with DLAs in the Lyman β region

We have shown the CDDF, dN/dX , and Ω_{DLA} of our GP model in Section 3.4. In this section, instead of using a sampling range from Ly β to Ly α , we only compute the population statistics of DLAs detected *within the Ly β forest region*. We set the sampling range to be Lyman limit +30 000 km s $^{-1}$ to Lyman- β . We cut off a wider velocity width at the blue end to avoid counting DLAs detected right on the edge of the Lyman break.

Figure 3.15 shows the CDDF for DLAs in the Lyman β region. As we can see from the figure, it is mostly consistent with the CDDF from Ly β -Ly α for $N_{\text{HI}} < 10^{21}$, and it starts to diverge for $N_{\text{HI}} > 20^{22}$. We visually inspected those spectra and found that they are mostly due to fitting large DLAs on the spectra's noisy left edges. This may indicate that additional regularisation is still needed to avoid spurious detections at the blue end of high redshift spectra. In particular, if the redshift measurement is slightly inaccurate, parts of the Lyman break move into our modelling window.

We also show the dN/dX and Ω_{DLA} for DLAs in the Lyman- β forest region in Figure 3.16. dN/dX in the Lyman- β region is broadly consistent with other measurements, with the detection consistent with zero at $z_{\text{DLA}} > 4$.

For Ω_{DLA} , in the right panel of Figure 3.16, we observe our measurement is biased high and highly uncertain for $z_{\text{DLA}} > 3.5$. This may be because our current model can only poorly estimate the column density from the Lyman- β region from high-redshift quasar

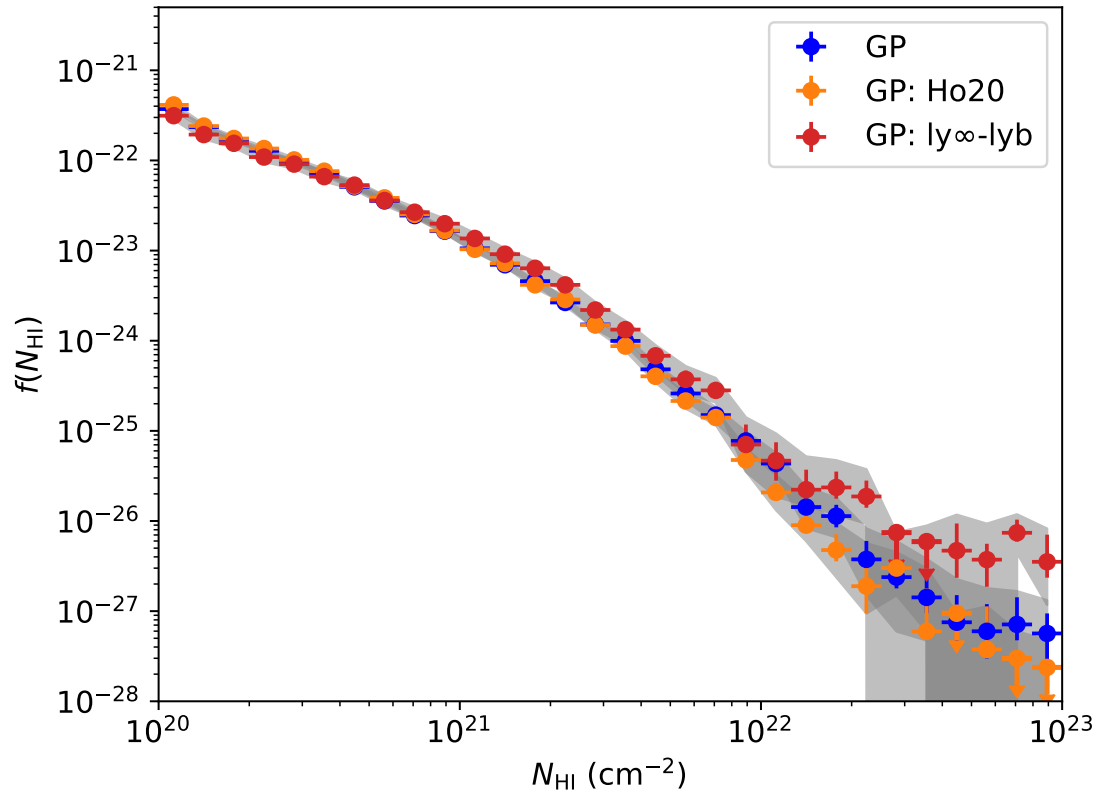


Figure 3.15: Comparing the CDDFs between the sampling range from $\text{Ly}\beta\text{-Ly}\alpha$ (Blue; GP) and $\text{Ly}_{\infty}\text{-Ly}\beta$ (Red; GP: $\text{ly}_{\infty}\text{-lyb}$). The error bars are 68% confidence limits, and the shaded areas are 95% confidence limits. [8] (GP: Ho20; Orange) also used a sampling range from $\text{Ly}\beta\text{-Ly}\alpha$.

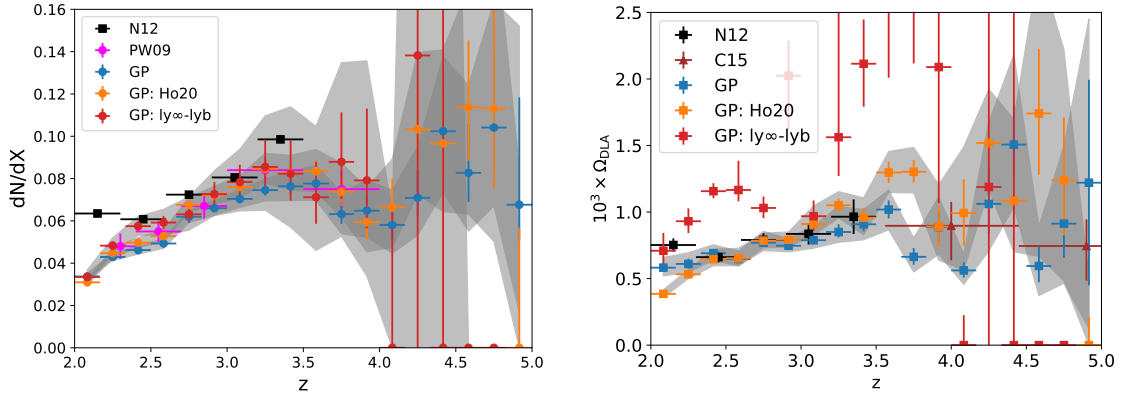


Figure 3.16: **(Left)** The comparison of dN/dX with different sampling ranges, $\text{Ly}\beta\text{-Ly}\alpha$ (blue) and $\text{Ly}\infty\text{-Ly}\beta$ (red). Other plot settings are the same as Figure 3.8. **(Right)** The comparison of Ω_{DLA} with difference sampling ranges, $\text{Ly}\beta\text{-Ly}\alpha$ (blue) and $\text{Ly}\infty\text{-Ly}\beta$ (red). Other plot settings are the same as Figure 3.9.

spectra, perhaps due to the high level of absorption from the Lyman- β and Lyman- α forests at these redshifts. Alternatively, it could again reflect that the mean flux measure is not certain at these redshifts, so the degeneracy between large DLAs and the effective Lyman- α /Lyman- β absorption is not fully broken by sampling $(\tau_{0,\text{MF}}, \beta_{\text{MF}})$.

3.7 Comparison to the CNN model

SDSS DR16Q includes DLA measurements using the convolutional neural network (CNN) model of [4]. The DLAs from the CNN model are recorded as `CONF_DLA`, `Z_DLA`, and `NHI_DLA` columns in the DR16Q catalogue.³

To compare our model and the CNN model, we restrict the z_{DLA} sampling range of the CNN DLAs to be the same as our GP DLAs. Table 3.2 shows the confusion matrix. On the existence of DLAs, which means the binary classification of having at least one

³This column is the log column density of the given DLA.

Table 3.2: The confusion matrix for multi-DLAs detections between the GP and the CNN model [4]. Note we require both the model posteriors of our GP model and DLA confidence in Parks to be larger than 0.98. We also require $\log_{10} N_{\text{HI}} > 20.3$. The maximum number of DLAs is fixed to three, and everything larger than three is considered three.

GP with Multi-DLAs	CNN	0 DLA	1 DLA	2 DLAs	3 DLAs
0 DLA	142759	5686	93	2	
1 DLA	2397	8007	208	1	
2 DLAs	117	234	333	5	
3 DLAs	8	6	11	4	

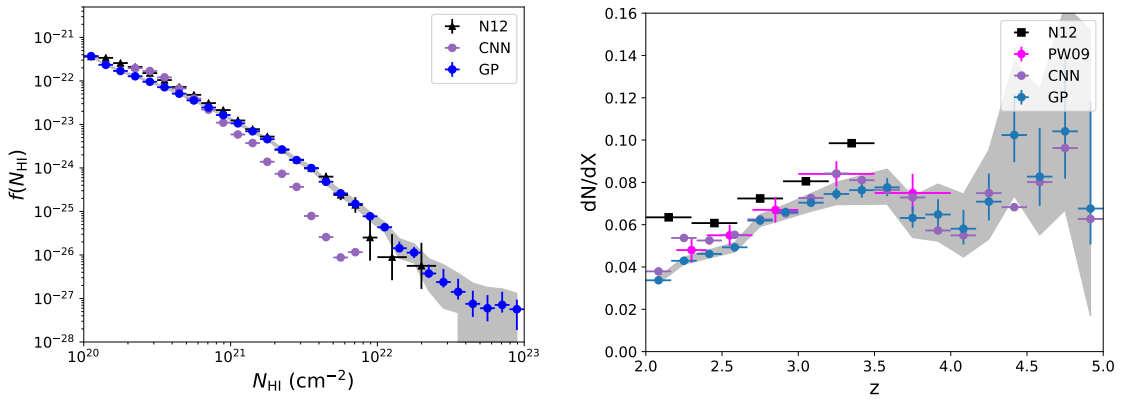


Figure 3.17: **(Left)** The CDDF of the DLAs detected by the CNN model presented in [4]. The z_{DLA} and $\log_{10} N_{\text{HI}}$ values are taken from the SDSS DR16Q catalogue in column `Z_DLA` and `NHI_DLA`. We require the confidence of DLAs to be larger than 0.98 and set the search range of the CNN DLAs to be the same as our search range, which is Lyman- β +3000 km s $^{-1}$ to $z_{\text{QSO}} - 3000$ km s $^{-1}$. **(Right)** The line density of the DLAs detected by the CNN model. All three measurements, GP, PW09, and CNN are consistent on the line density.

DLA or no DLA, the GP model is $\sim 94.8\%$ in agreement with the CNN model. If we only consider only spectra with $\text{SNR} > 6$, the rate of agreement climbs to $\sim 96.5\%$.

We have also checked the CDDF of the CNN DLAs, as shown in Figure 3.17. The sampling range is restricted to be the same as ours, and we only count the DLAs with `CONF_DLA` larger than 0.98. The CDDF of the CNN model under-detects DLAs with $N_{\text{HI}} > 7 \times 10^{20}$, compared to N12. We have discussed this issue in Figure 19 of [8]. The

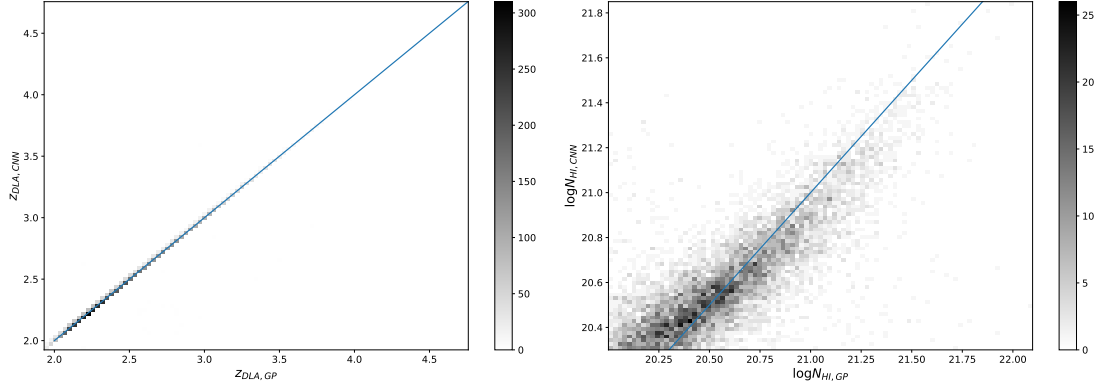


Figure 3.18: The 2D histograms for z_{DLA} (**left**) and $\log_{10} N_{\text{HI}}$ (**right**) estimated by the GP code and the CNN. We use the maximum a posteriori (MAP) estimate for parameter estimation for the GP code. The colourbars indicate the number of DLAs within the bin. The blue line is a straight line that shows the diagonal line of the 2D histogram.

CDDF of the CNN model in the DR16Q catalogue shows improvements in detecting more high column density systems comparing to [4], but it is still an order of magnitude lower than N12 for $N_{\text{HI}} > 2 \times 10^{21}$. Thus the lack of high column density systems in the CNN DLAs, as identified in [8], is still present in the latest catalogue.

The dN/dX of the CNN model, in contrast, mostly agree with our GP measurements. Bins with $z > 4.5$ are even consistent at the $1\text{-}\sigma$ level. Since dN/dX is sensitive to low column density systems, it shows these two codes find consistent small DLAs, but differ in their column density estimates.

We compare DLAs detected by the CNN and GP codes on a spectrum-by-spectrum basis in Figure 3.18. As anticipated, the CNN and the GP code have a perfect agreement in z_{DLA} , but the CNN predicts slightly lower $\log_{10} N_{\text{HI}}$ than the GP code, consistent with the CDDF plot in Figure 3.17.

We visually inspected 319 quasar spectra, where the CNN code strongly disagrees with the GP code’s detections. As expected, most cases are spectra with low SNRs, where even human experts will have difficulty identifying DLAs. Besides those low-SNR cases, in general, the CNN code has false negatives on DLAs overlapping with sub-DLAs or DLAs very close to each other. There are 24 out of 319 cases which show a clear pattern where the CNN missed the DLAs when multiple absorption systems are overlapping or nearby.⁴ Some of these are ambiguous detections, but 9 out of 24 have apparent damping wings on the absorber.

We show two examples in Figure 3.19. The first one shows a sub-DLA intervening on the right of the DLA damping wings. Though the damping wings are disturbed by the sub-DLA, the pattern of a DLA is still visible. The second example shows two DLAs close to each other, but not close enough to overlap. We suspect these non-detections for the CNN code are due to the lack of training data for multiple absorption systems (sub-DLAs or DLAs) close to each other. Since these overlapping cases are rare in the real dataset, we think one might need to implement simulated DLAs/sub-DLAs to augment the CNN training set.

3.8 Conclusion

We have presented a new estimate of the abundance of DLAs from $z = 2$ to $z = 5$ and a DLA catalogue built from SDSS DR16Q spectra [90] using our Gaussian process model [3, 8]. We verify our results are in good agreement with previous measurements

⁴We put the figures for these 24 spectra in here http://tiny.cc/overlapping_dlas for future investigators.

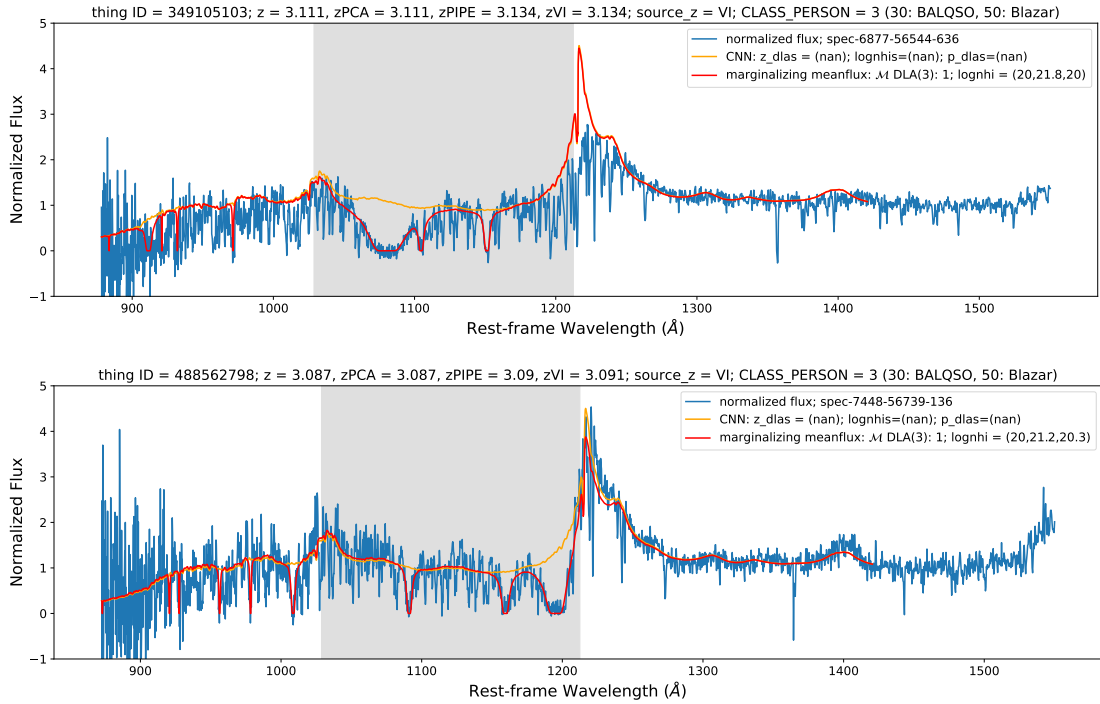


Figure 3.19: Examples showing **(top)** the case of a sub-DLA overlapping a DLA and **(bottom)** the case of a DLA near to another DLA. The red line indicates the GP code predictions, and we describe the $\log_{10} N_{\text{HI}}$ in the legend. We intervene the DLAs from the CNN model in the DR16Q catalogue onto our null model in the orange line. Both spectra have high enough SNR: the upper one has $\text{SNR} = 3.45$ while the bottom one has $\text{SNR} = 7.52$. The damping wings and the Lyman- β absorption lines of the DLAs are visible in the plots.

from [5], [6], and [7]. We newly integrate out the uncertainty in the measured mean flux, which improves our modelling of DLA detection uncertainties for $z_{\text{QSO}} > 4$ without biasing towards high N_{HI} detections.

We note, nevertheless, that there is a residual dependence on low-redshift spectra with $z_{\text{QSO}} < 2.5$. This could be due to unmodelled systematics or simply because the low-redshift optical spectra are incomplete in the Lyman series range, so we can not securely detect DLAs in low z_{QSO} . Incorporating spectra with shorter observed wavelengths could potentially verify these detections at $z_{\text{QSO}} < 2.5$.

Our measurement shows the abundance of DLAs and neutral hydrogen increases moderately over $2 < z < 4$, while the trend beyond $z = 4$ is unclear due to statistical uncertainties. Larger datasets and better mean flux measurements are needed to give more robust constraints for DLA detections at $z > 4$.

Data Availability

Our DLA catalogue is publicly available at http://tiny.cc/gp_dla_dr16q, including both MATLAB catalogue and JSON catalogue. A sub-DLA candidate catalogue is available in JSON format. README files are included to describe the data formats of both catalogues. The data files for DLA population statistics are also included, including CDDE, dN/dX , and Ω_{DLA} with or without SNR cuts. A tutorial for manipulating the MATLAB catalogue is publicly available at https://github.com/jibanCat/gp_dla_detection_dr16q_public/tree/master/notebooks as a notebook file. Our GP code

is also publicly available at https://github.com/jibanCat/gp_dla_detection_dr16q_public/.

Chapter 4

Multi-fidelity Emulation for the Matter Power Spectrum using Gaussian Processes

4.1 Abstract

We present methods for emulating the matter power spectrum by combining information from cosmological N -body simulations at different resolutions. An emulator allows estimation of simulation output by interpolating across the parameter space of a limited number of simulations. We present the first implementation in cosmology of multi-fidelity emulation, where many low-resolution simulations are combined with a few high-resolution simulations to achieve an increased emulation accuracy. The power spectrum's dependence on cosmology is learned from the low-resolution simulations, which are in turn calibrated

using high-resolution simulations. We show that our multi-fidelity emulator predicts high-fidelity counterparts to percent-level relative accuracy when using only 3 high-fidelity simulations and outperforms a single-fidelity emulator that uses 11 simulations, although we do not attempt to produce a converged emulator with high absolute accuracy. With a fixed number of high-fidelity training simulations, we show that our multi-fidelity emulator is $\simeq 100$ times better than a single-fidelity emulator at $k \leq 2 h\text{Mpc}^{-1}$, and $\simeq 20$ times better at $3 \leq k < 6.4 h\text{Mpc}^{-1}$. Multi-fidelity emulation is fast to train, using only a simple modification to standard Gaussian processes. Our proposed emulator shows a new way to predict non-linear scales by fusing simulations from different fidelities.

4.2 Introduction

Current and next generation large scale structure surveys, such as DES¹ [101], LSST (Rubin Observatory)² [102], EUCLID³ [103], DESI⁴ [104], and the Roman Space Telescope (WFIRST) [105] will probe gravitational clustering and galaxy formation at small scales with high accuracy. Thus, the future of cosmology relies on exploiting the information in non-linear structure formation at small scales, where numerical N -body simulations must be used to give accurate theoretical predictions.

Cosmological linear perturbation theory provides accurate analytic predictions on the clustering of mass up to $k \sim 0.1 h\text{Mpc}^{-1}$. Despite the success of the standard model of cosmology, several fundamental physics puzzles are still unanswered: the accelerated

¹<https://www.darkenergysurvey.org>

²<https://www.lsst.org>

³<https://sci.esa.int/web/euclid>

⁴<https://www.desi.lbl.gov>

expansion of the Universe [106], the nature of dark matter [107], and the sum of the neutrino masses [108]. To answer these questions and constrain cosmological parameters using future surveys, theoretical predictions from numerical simulations must be accurate on smaller scales than are accessible to linear theory. As a primary summary statistic, the matter power spectrum needs to be at percent-level precision for $k \lesssim 10 h\text{Mpc}^{-1}$ [26].

Modelling non-linear gravitational clustering is done using N -body simulations, where a dark matter fluid is sampled by macro-particles and evolved using a smoothed gravitational force. Each macro-particle is representative of an ensemble of microscopic dark matter particles. Generations of computational physicists have improved the accuracy of the gravitational evolution, and created quicker and more scalable algorithms to drive the mass resolution of the simulations ever higher [109, 110, 111, 112, 113].

The mass resolution necessary to robustly predict the power spectrum at $k \sim 10 \text{Mpc}/h$ pushes the computational limits of contemporary supercomputers. To adequately sample a high-dimensional input parameter space with Markov chain Monte Carlo (MCMC), millions of samples are needed, while a limited number (at best a few hundred to a few thousand) of high-fidelity simulations are computationally possible.

An efficient way to perform accurate cosmological inference with a limited number of simulations is to use *emulators*. Emulators are flexible statistical models, usually built with Gaussian processes, which learn the mapping from input cosmological parameters to summary statistics. This reduces the number of costly forward simulations by effectively interpolating the function outputs.

Emulators have been applied extensively in the field of cosmological inference. [27, 114] proposed a cosmic calibration project to make percent-level predictions on the matter power spectrum using a Bayesian emulator. [28, 115, 116] implemented this cosmic emulator in their Coyote Universe suite using 37 high-resolution simulations. [117, 118] designed the Mira-Titan Universe suite to train emulators to make precise theoretical predictions using 36 simulations. The latest Euclid preparation [10] runs 250 simulations (3000^3 particles) to prepare their emulator for the matter power spectrum. Besides Gaussian processes, [119] used a neural network to build a cosmic emulator from 6 380 N -body simulations spanning 580 cosmologies.

Beyond the matter power spectrum, emulators have been trained to predict the halo mass function [120], the concentration-mass relation for dark-matter haloes [121], the galaxy power spectrum [122], the galaxy correlation function [123], the halo bias [124], weak lensing peak counts [125], the cosmic shear covariance [126], weak lensing voids [127], the 21 cm signal [45], and the Lyman- α 1D flux power spectrum [128]. They also have been used for inferring beyond- Λ CDM cosmologies [129, 130] and $f(R)$ gravity cosmologies [131].

While all these emulators successfully predict summary statistics using high-fidelity simulations, one question which remains is how to minimize the number of necessary training simulations to achieve a given accuracy. Here we demonstrate that building cosmological emulators from simulations can be improved with multi-fidelity models. Multi-fidelity models [41] minimize the computational cost by combining the predictive power of simulations at different resolutions. They fuse the expensive but accurate *high fidelity* data with cheaply-obtained *low fidelity* approximations. One standard model used by the multi-

fidelity emulation is a *multi-output Gaussian process* [132]. A multi-output Gaussian process (multi-output GP) generalizes a single-output GP to multiple outputs, while building a cross-covariance function to model the shared information between outputs. In this paper, low and high fidelity correspond to simulations at different resolutions. High-fidelity simulations have a finer mass resolution while low-fidelity simulations have a coarser mass resolution.

To train the multi-fidelity emulator using as few high-resolution simulations as possible, we also propose a method for selecting high-fidelity training samples, based on minimizing the loss computed among the low-fidelity simulations. By optimizing the low-fidelity emulator’s loss, we show that one can efficiently train a multi-fidelity emulator by avoiding worst-case combinations of the high-fidelity training samples.

Computational astrophysicists have used methods similar to multi-fidelity modelling to minimize the cost of performing high-resolution simulations [133, 134]. A notable example is Richardson extrapolation [135], a numerical method to improve a simulation’s accuracy by combining a sequence of simulations with varied spatial resolutions and fixed cosmologies. More recently, generative adversarial networks (GAN) have been used to produce high-resolution density fields [136] and particle displacements [137] from low-resolution (but larger volume) input data. In principle, such ‘super-resolution’ simulations could be implemented as a multi-fidelity emulator’s high-fidelity training set, allowing an emulator to be built to a scale not directly accessible to simulations.

[138, 139] proposed using Bayesian optimization to improve emulator accuracy by a sequential choice of new simulation points designed to globally optimize the emulator function. Similar approaches to iterative selection of training data in a cosmological

parameter space have been presented by [140, 141]. Computer scientists and engineers, including [142, 143, 144, 145, 146], have extensively studied combining multi-fidelity methods with Bayesian optimization.⁵ Multi-fidelity Bayesian optimization arises when a cheaper approximation to the object function exists.

We present a multi-fidelity emulator for the matter power spectrum, as output by the cosmological simulation code MP-GADGET [148, 149]. In this paper, we target percent level *relative accuracy*: how well our emulators can reproduce the matter power spectra at our highest fidelity. We defer producing an emulator which allows percent level accurate reconstruction of observations or a hypothetical ideal simulation to future work. The main goal of this paper is to demonstrate that our multi-fidelity techniques can be used to reduce the computational budget required for an emulator.

We use two fidelities in a 256 Mpc/h box: a fast but low resolution version with 128^3 dark-matter particles and a slow but high resolution version with 512^3 particles. Even with only 3 high-fidelity simulations and 50 low-fidelity simulations, we show that we can predict the high-resolution matter power spectrum at percent-level accuracy on average at $k \leq 6.4 h\text{Mpc}^{-1}$ at $z = 0$, with a total computational cost $\lesssim 4$ high-fidelity simulations. Although we only show our application to the matter power spectrum, the methods presented in this paper could apply to other summary statistics, e.g., the halo mass function or the Lyman- α 1D flux power spectrum.

[150] showed that the lack of AGN feedback affects a dark matter-only simulation significantly (compared to the error requirements of upcoming surveys) at $k > 0.1 h\text{Mpc}^{-1}$. Furthermore, baryon cooling can alter the power spectrum at $k \sim 10 h\text{Mpc}^{-1}$ [151]. How-

⁵[147] has a subsection that provides a short review on multi-fidelity Bayesian optimization.

ever, as techniques exist to model this effect in post-processing [152], we defer extending our technique to hydrodynamical simulations including AGN feedback to future work. Here we validate that a multi-fidelity emulator is useful in the simplest case: dark matter-only N -body simulations.

We build two types of multi-fidelity emulators. One uses the linear autoregressive model of [41] (first-order autoregressive model, AR1), which we will call the “linear multi-fidelity model.” The second multi-fidelity emulator uses the non-linear fusion model of [153] (nonlinear auto-regressive Gaussian process, NARGP), and which we call the “non-linear multi-fidelity emulator.”⁶ [41] model the scaling factor between fidelities as a scalar, while [153] allow the scaling factor to depend on input parameters. Our implementation of AR1 and NARGP is based on `emukit` [155],⁷ an open-source package for emulation and decision making under uncertainty, with the modifications mentioned above.⁸

In Section 4.3, we briefly describe the simulation code, MP-GADGET, for training the emulator. We recap the general formalism of a single-fidelity Gaussian process emulator in Section 4.4. Section 4.5 describes the formalism of a multi-fidelity emulator (`MFEulator`). We explain our sampling strategy in Section 4.6. Section 4.7 shows the results, with comparisons between multi-fidelity emulation and single-fidelity emulation. We summarize the runtime for the MP-GADGET simulations in Section 4.8. We conclude with a summary of key contributions and potential applications of our work in Section 4.9.

⁶AR1 and NARGP are acronyms used in [153, 154]. In this paper, AR1 and linear multi-fidelity emulator are interchangeable, and NARGP and non-linear multi-fidelity emulator are interchangeable.

⁷<https://github.com/EmuKit/emukit>

⁸For a detailed comparison between AR1 and NARGP, see [154]. An example code for the comparison between AR1 and NARGP can be found in Emukit’s examples.

Our code for multi-fidelity emulation in the matter power spectrum is publicly available at https://github.com/jibanCat/matter_multi_fidelity_emu.

4.3 Simulations

We prepare our training set by running dark matter-only simulations using the massively parallel N -body code MP-GADGET [156].⁹ MP-GADGET is a publicly available N -body+Hydro cosmological simulation code derived from GADGET3 [148]. It is parallelized using a hybrid OpenMP/MPI strategy and has successfully performed a hydrodynamical simulation using all 8032 *Frontera* nodes, a total of 449792 cores, demonstrating its good scalability properties. The gravitational forces are computed using a Fourier transform based particle-mesh algorithm on large scales and a Barnes-Hut tree on small scales.

We initialise our simulations from the linear power spectrum produced by CLASS [157] at $z = 99$ using the Zel’dovich approximation [158]. The dark matter particles then evolve through gravitational dynamics. The matter power spectra are computed from the output snapshots of MP-GADGET, and used as our emulation targets. In this paper, we fix the IC noise in the nodes and change only the cosmology for the emulator training. We do not use the “paired and fixed” technique [159], but it would be easy to do so using only low resolution simulations as these pairings are designed to remove variance on large scales.

The matter power spectrum, $P(k)$, is a compressed summary statistic of the overdensity field, $\delta(x)$, evaluated as an angle average of the Fourier-transformed overdensity

⁹<https://github.com/MP-Gadget/MP-Gadget/>

Table 4.1: Notations and definitions

Notation	Description
HR	High-resolution simulation, 512^3 particles
LR	Low-resolution simulation, 128^3 particles
$x_{i,t}$	Input cosmological parameters at i th simulation at fidelity t
$y_{i,t}$	Matter power spectrum at i th simulation at fidelity t , at log scale
n_t	Number of simulations at fidelity t
$N_{\text{ptl,side}}$	Number of particles per box side

field:

$$P(|\mathbf{k}|) = \langle \hat{\delta}^*(k) \hat{\delta}(k) \rangle, \quad (4.1)$$

$$\hat{\delta}(\mathbf{k}) = \int d^3\mathbf{r} \delta(\mathbf{r}) e^{-2\pi i \mathbf{k} \cdot \mathbf{r}}. \quad (4.2)$$

We measure the power spectrum with a cloud-in-cell mass assignment, which is deconvolved. The Fourier transform is taken on a mesh the same as the PM grid of the simulation, which has a resolution of 2 times the mean inter-particle spacing.

For a multi-fidelity problem, our data are from simulations at different resolutions. Since low resolution simulations are cheaper to obtain (but are only approximations to the high resolution results), we typically have a limited number of high-fidelity data and many low-fidelity approximations.

To make the text of this section consistent with the following sections, we provide some notation to bridge the terminology, summarized in Table 4.1. We have data from s different fidelities (simulation resolutions). For each fidelity, we have pairs of inputs and outputs $\mathcal{D}_t = \{x_{i,t}, y_{i,t}\} = \{\mathbf{x}_t, \mathbf{y}_t\}$, where $t = 1, \dots, s$ denotes the fidelity level from low to high, and $i = 1, \dots, n_t$ where n_t is the number of data pairs at fidelity t and i indexes

each individual simulation. The data pairs $\mathcal{D}_t = \{\mathbf{x}_t, \mathbf{y}_t\}$ for our emulation setup are the cosmological parameters of the simulations and the power spectrum outputs. Here we have $s = 2$ for two mass resolutions: 128^3 and 512^3 dark matter-only simulations. We will denote 128^3 as low-resolution (LR, $t = 1$) and 512^3 as high-resolution (HR, $t = 2$).

Each fidelity will have a different number of simulations, n_t . Practically, the number of LR simulations will be much larger than the number of HR simulations, $n_1 > n_2$. The compute time for LR ($N_{\text{ptl,side}} = 128$) is ~ 20 core hours and ~ 2000 core hours for HR ($N_{\text{ptl,side}} = 512$). We will empirically show we only need 3 HR and 50 LR to train a multi-fidelity emulator with an average emulator error per k smaller than 1%.

We do not emulate the matter power spectrum across redshifts, conditioning on a given redshift bin z_0 . We generally focus on $z_0 = 0$, but will discuss multi-fidelity emulators at $z_0 = 1$ and $z_0 = 2$ in Section 4.7.4.

4.3.1 Latin hypercube sampling

As [28] mentioned, a space-filling Latin hypercube design is well suited for GP emulators of the matter power spectrum. For a training set with d -dimensional inputs and N simulations, an N^d grid is created first, and simulations are placed on this grid so that only one simulation is present in any row or column. The Latin hypercube design improves on random uniform sampling by ensuring that the chosen points do not crowd together in any subspace.

We apply a Latin hypercube design on the input parameter space, $\{h, \Omega_0, \Omega_b, A_s, n_s\}$. We vary the Λ CDM cosmological parameters $\{h, \Omega_0, \Omega_b, A_s, n_s\}$, which are the Hubble parameter $h = H_0/(100 \text{ km s}^{-1} \text{ Mpc}^{-1})$, the total matter density Ω_0 , the baryon density Ω_b ,

primordial amplitude of scalar fluctuations A_s , and the scalar spectral index n_s . We use the same set of Λ CDM cosmological parameters as [10], allowing us to compute the relative errors of our simulations with respect to EuclidEmulator2.

We use bounded uniform priors for the input parameters:

$$\begin{aligned}
 h &\sim \mathcal{U}[0.65, 0.75]; \\
 \Omega_0 &\sim \mathcal{U}[0.268, 0.308]; \\
 A_s &\sim \mathcal{U}[1.5 \times 10^{-9}, 2.8 \times 10^{-9}]; \\
 n_s &\sim \mathcal{U}[0.9, 0.99]; \\
 \Omega_b &\sim \mathcal{U}[0.0452, 0.0492].
 \end{aligned}
 \tag{4.3}$$

The dark energy density is $\Omega_\Lambda = 1 - \Omega_0$. The prior volume surrounds the WMAP 9-year cosmology [160]. The code to handle the simulation input files and Latin hypercube design is publicly available at <https://github.com/jibanCat/SimulationRunnerDM>.

4.3.2 Preprocessing of the simulated power spectrum

A numerical simulation is constrained by its box size and number of particles. The mass resolution limits the smallest scale (the highest k) of the power spectrum. Thus, high-fidelity simulations can model smaller scales, not fully resolved in low-fidelity simulations, as shown in Figure 4.1.

For k larger than the mean particle spacing, $P(k)$ differs substantially from the resolved value, due to artifacts of the macro-particle sampling. The scale of the mean particle spacing is

$$k_{\text{spacing}} = 2\pi \frac{N_{\text{ptl,side}}}{L_{\text{box}}},
 \tag{4.4}$$

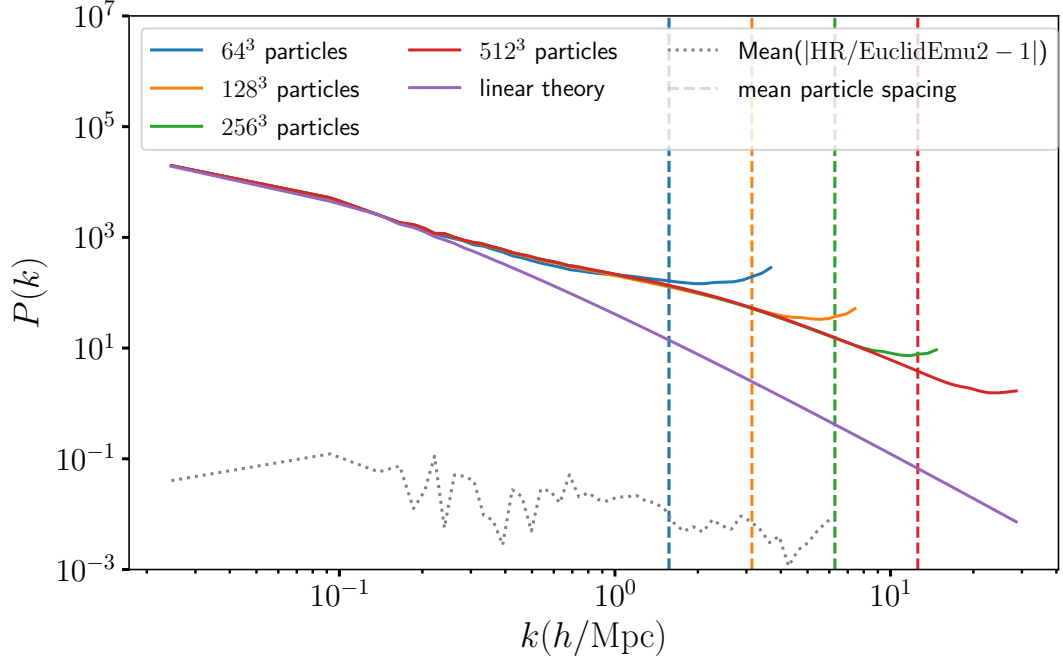


Figure 4.1: The matter power spectrum output by MP-GADGET at different mass resolutions. The vertical dash lines indicate the mean particle spacing k_{spacing} for a given mass resolution. **(Blue)**: The matter power spectrum from a dark-matter only MP-GADGET simulation with 64^3 particles. **(Orange)**: The matter power spectrum from MP-GADGET with 128^3 particles. **(Green)**: The matter power spectrum from MP-GADGET with 256^3 particles. **(Red)**: The matter power spectrum from MP-GADGET with 512^3 particles. **(Purple)**: Linear theory power spectrum. The cosmology parameters are $h = 0.675$, $\Omega_0 = 0.278$, $\Omega_b = 0.0474$, $A_s = 1.695 \times 10^{-1}$, $n_s = 9.405 \times 10^{-1}$. The dotted line shows the relative error of HR (512^3 simulations) compared with EuclidEmulator2 [10], averaged over four different cosmologies.

where $N_{\text{ptl,side}}$ is the number of particles per side of the box. For instance, if we have 512^3 particles in the box, then $N_{\text{ptl,side}} = 512$. L_{box} is the size of the simulation box in units of Mpc/h.

We use the same set of matter power spectrum k bins for all fidelities. The available information at small scales is sparse for the low-fidelity spectrum. To resolve the issue, we fix the k bins to high fidelity and linearly interpolate the low-fidelity power spectrum in a \log_{10}

scale, $\log_{10} P(k)$, onto the high-fidelity k bins. The maximum k is set to be $\simeq 6.4 h\text{Mpc}^{-1}$ when using $N_{\text{ptl,side}} = 128$ as our low-fidelity training set. However, in practice we found that 128^3 and 512^3 simulations shared similar k bins with small offsets at small scales.

We do not model the high-fidelity spectrum with k larger than the maximum k of the low-fidelity spectrum:

$$\max k_{t=2} = \max k_{t=1}, \quad (4.5)$$

where t indicates the fidelity level and $t = 2$ is the highest fidelity. If we do not have any data at a given k from low-fidelity, we cannot extract the correlations between fidelities without other more significant assumptions. In other words, the maximum k we can model is limited by the data available from the low-fidelity simulations, which always have a lower maximum k than high-fidelity simulations. We note that it is possible to get a higher maximum k by particle folding or by increasing the size of the PM grid size used for estimating the power spectrum, although we do not do that here.

We do model the low-fidelity $P(k)$ even on scales smaller than the mean particle spacing, $k > k_{\text{spacing}}$. We made this particular decision because we have a prior belief that even though $P(k > k_{\text{spacing}})$ is highly biased, it still captures some information about how $P(k)$ depends on cosmological parameters. Thus, we should be able to exploit the correlations between fidelities and improve the emulator accuracy at those scales.

To summarize, we:

1. Use the same set of k bins across different fidelities.
2. Preserve all available $P(k)$ from low-fidelity, even scales smaller than the simulation's mean particle spacing.

4.4 Single-fidelity emulators

Here we briefly recap how we train a single-fidelity emulator. Readers familiar with this material may wish to skip to Section 4.5. The notation we use in this section follows those of [153, 154]. Consider a supervised learning problem, in which we wish to learn the mapping relation, f , between a set of input and output pairs $\mathcal{D} = \{x_i, y_i\} = \{\mathbf{x}, \mathbf{y}\}$, where $i = 1, \dots, n$:

$$\mathbf{y} = f(\mathbf{x}), \quad \text{with } \mathbf{x} \in \mathbb{R}^d, \quad (4.6)$$

where d is the dimension of the input space. A Gaussian process (GP) [31] is a probabilistic framework modelling the observations, \mathbf{y} , as drawn from a noisy realization of a single random function f with a likelihood $p(\mathbf{y} | f)$. It models the distribution over f

$$p(f) = \mathcal{GP}(f; \mu, K), \quad (4.7)$$

with μ the GP mean prior function, which is usually assumed to be a zero mean prior, and K the covariance kernel function specified by a vector of hyperparameters, $\boldsymbol{\theta}$. For a given set of inputs, x_1, x_2, \dots, x_n , the kernel function evaluated on these points produces a symmetric, positive-definite covariance matrix $K_{ij} = K(x_i, x_j; \boldsymbol{\theta})$ with $K \in \mathbb{R}^{n \times n}$.

The choice of the covariance kernel depends on our prior knowledge about the data. The hyperparameters of a chosen kernel are optimized by maximizing the marginal log-likelihood:

$$\log p(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta}) = -\frac{1}{2} \log |\mathbf{K}| - \frac{1}{2} \mathbf{y}^\top \mathbf{K}^{-1} \mathbf{y} - \frac{n}{2} \log 2\pi. \quad (4.8)$$

For an emulator, the main purpose is to predict an output $f_* = f(x_*)$ from a new input point x_* , given the provided data \mathcal{D} .

$$\begin{aligned}
 p(f_* | \mathcal{D}, x_*) &= \mathcal{N}(f_* | \mu_*(x_*), \sigma_*^2(x_*)), \\
 \mu_*(x_*) &= \mathbf{k}_{*n} \mathbf{K}^{-1} \mathbf{y}, \\
 \sigma_*^2(x_*) &= K(x_*, x_*) - \mathbf{k}_{*n} \mathbf{K}^{-1} \mathbf{k}_{*n}^\top,
 \end{aligned} \tag{4.9}$$

where μ_* is the posterior mean and σ_* is the standard deviation of the uncertainty in the estimate of the predictions. The vector \mathbf{k}_{*n} is the covariance between the new point and trained data, $\mathbf{k}_{*n} = [K(x_*, x_1), \dots, K(x_*, x_n)]$.

4.4.1 Cosmological emulators

Consider we have a set of dark matter-only simulations with fixed box size and mass resolution. At each redshift bin z_0 , we can compute the matter power spectrum, $P(k, z = z_0)$, given a set of input parameters. We will use the log power spectrum, $\log_{10} P(k, z = z_0)$, as our training data.

The training data, $\mathcal{D} = \{x_i, y_i\}$, are defined as

$$\begin{aligned}
 x_i &= [h_i, \Omega_{0i}, \Omega_{bi}, A_{si}, n_{si}]; \\
 y_i &= \log_{10} P(k, z = z_0),
 \end{aligned}$$

where $i = 1, \dots, n$ indicates the i th simulation we run with this specific set of input parameters.

The rest of the modelling is choosing an appropriate covariance function $K(x, x')$. We use a squared exponential kernel and use automatic relevance determination (ARD)

weights for each input dimension. ARD assigns each input dimension, x_i , a separate hyperparameter, w_i :

$$K(x, x'; \boldsymbol{\theta}) = \sigma^2 \exp \left(-\frac{1}{2} \sum_{i=1}^d w_i (x_i - x'_i)^2 \right) \quad (4.10)$$

where $i = 1, \dots, d$ indicates the dimension of the input space $x \in \mathbb{R}^d$. σ^2 is the variance parameter for the squared exponential kernel, $\{w_i\}_{i=1}^d$ are the ARD weights. $\{w_i\}_{i=1}^d$ are inverse length scales, which define the degree of smoothness at a given input dimension. We note that we assign independent hyperparameters, $\boldsymbol{\theta} = \{\sigma^2, w_1, \dots, w_d\}$, for each k mode.¹⁰ A larger w_i corresponds to a smaller length scale, reflecting that the learned function varies more in the i th dimension. On the other hand, a smaller w_i implies a larger length scale, indicating that the learned function is smoother along the i th dimension. ARD allows each dimension of the learned function to have a different degree of smoothness.

We do not decompose the power spectrum into principle components for training the emulators, as described by [27, 114] because we want to compare single-fidelity emulators to the multi-fidelity emulators, and an `MFEEmulator` only has a limited number of high-resolution simulations available. In our default case, we only have 3 high-resolution simulations for an `MFEEmulator`, and it is not sensible to perform dimension reduction on three power spectra.

To ensure that our single-fidelity emulator is not unfairly disadvantaged in the comparison with our multi-fidelity emulator by poorly constrained hyperparameters, we built a single-fidelity emulator which shared kernel parameters across all k modes and empirically verified that it had similar performance.

¹⁰[140] refers to this approach as the many single-output approach (MS).

4.5 Multi-fidelity emulator

In this section, we describe how we train a multi-fidelity emulator. We outline the modelling assumptions in Section 4.5.1. Section 4.5.2 describes the formalism of the linear multi-fidelity emulator proposed by [41], a multi-output GP with a linear correlation between fidelities. Section 4.5.3 outlines the non-linear multi-fidelity emulator of [153], which models the correlation between fidelities as a function of cosmological parameters. We follow the notation and formalism of [41, 153, 154].

4.5.1 General assumptions

Here we outline our modelling assumptions, following the assumptions made in [41]:

1. **Correlations between the code fidelities:** For an N -body simulation, the simulation cost depends on the mass resolution. We assume a simulation with a low mass resolution can approximate a simulation with a high mass resolution. The matter power spectrum from different fidelities is strongly correlated at large scales since all fidelities are resolved and the mass resolution has negligible effects. At small scales, however, we expect different fidelities are only weakly correlated.
2. **Smoothness:** For an emulation problem, we assume that neighbouring inputs give similar outputs. For example, suppose two sets of input parameters to MP-GADGET are close to each other. In that case, we assume that an N -body simulation will provide a similar outcome.

3. **The prior belief on each fidelity is a Gaussian process:** We assume a prior belief that the mapping from code input to output is a Gaussian process for each fidelity.

The first assumption is the core assumption of a multi-fidelity emulator. Different levels of the same code are simulating the same physical reality. It is thus reasonable to assume different code fidelities should correlate at some level. However, a naive simulation, for example, $N_{\text{ptl,side}} = 16$ could only barely approximate a HR with $N_{\text{ptl,side}} = 512$. Therefore, we should also assume the correlation between fidelities depends on the distance between two fidelities in the dimension of mass resolution.

There is thus a trade-off between the strength of correlation and the computational expense: for example, a simulation with $N_{\text{ptl,side}} = 256$ provides more information about a HR ($N_{\text{ptl,side}} = 512$), but running a 256^3 simulation is 8 times most expensive than running a LR ($N_{\text{ptl,side}} = 128$).

One can select an optimal choice of simulation cost by balancing the computational time and the emulation accuracy. Here we choose $N_{\text{ptl,side}} = 128$ for our low-fidelity simulations because:

1. The maximum k is $\simeq 6.4 h\text{Mpc}^{-1}$, which includes enough non-linear scales to test the emulation accuracy;
2. A 128^3 simulation is 64 times cheaper than a HR, and thus the resolution difference between $N_{\text{ptl,side}} = 128$ and $N_{\text{ptl,side}} = 512$ is large enough to demonstrate whether simulations with lower costs can accelerate the training of an emulator.

In Section 4.7.4, we will show our method is applicable to simulations with different resolutions, $N_{\text{ptl,side}} = 64$ and 256. Empirically, we found that using $N_{\text{ptl,side}} = 256$ as low-fidelity is similar to $N_{\text{ptl,side}} = 128$, while $N_{\text{ptl,side}} = 64$ gives a worse emulation accuracy.

The second assumption, the smoothness assumption, is the general assumption of a GP emulator. A GP emulator will have poor accuracy if the code does not behave similarly with similar input. The smoothness assumption is also the assumption behind the Latin hypercube sampling scheme [for a detailed discussion, see Ref. [28]].

A multi-fidelity emulation could in principle be implemented using other models (see [161] for different data-fit models for surrogates). We chose to use GPs simply because their Bayesian approach supports uncertainty quantification and there is a well-developed community around GP emulation.

4.5.2 Linear multi-fidelity emulator (AR1)

We have multi-fidelity data \mathcal{D}_t as described in Section 4.3. A multi-fidelity emulator is essentially inferencing the highest fidelity model conditioned on data from all model fidelities. The final goal of a multi-fidelity emulator is to find a mapping relation f such that, from an arbitrary input vector x_* , we can always find the highest fidelity code output:

$$y_{s,*} = f(x_*). \quad (4.11)$$

As described by [41], a linear autoregressive model can be applied in a multi-fidelity setting by assuming a hierarchical order between fidelities:

$$f_t(x) = \rho_t f_{t-1}(x) + \delta_t(x), \quad (4.12)$$

where f_t is the function emulated by a GP at t fidelity and f_{t-1} is the function emulated at the previous fidelity level ($t - 1$). The linear component of Eq 4.12 is ρ_t , which models the correlation between fidelities as a linear relation. δ_t is a GP modelling the bias term:

$$\delta_t \sim \mathcal{GP}(\mu_{\delta_t}, K_t). \quad (4.13)$$

We modify Eq 4.12 so inference is performed on each k bin independently. For $k = k_j$, we have independent kernel and scaling parameters for each $k = k_j$ mode. For simplicity, we will drop the $k = k_j$ notation in the rest of the paper:

$$f_t(x) = \rho_t(f_{t-1}(x) - \mu_{t-1}) + \delta_t(x). \quad (4.14)$$

The mean of the bias term, μ_{δ_t} , is assumed to be the zero function. For the low-fidelity part, we subtract the sample mean of the logarithm training power spectra, $\log_{10} P(k)$, and model the low fidelity part of the power spectra as a zero mean GP:

$$(f_1(x) - \mu_1) \sim \mathcal{GP}(0, K_1(x_1, x'_1; \theta_1)). \quad (4.15)$$

As shown in Figure 4.1, the low-fidelity power spectrum is biased high. We pass variations of the low-fidelity power spectrum around its mean to the next fidelity to avoid passing biased outputs. In practice, we found this slightly improves emulation accuracy for multi-fidelity models.

For the highest fidelity bias function, $\delta_s(x)$, we model the power spectrum using a zero mean GP without subtracting the sample mean. We do not have enough points at the highest fidelity for the sample mean to be a good estimate of the true mean. Except for $t = 1$, $f_t(x)$ is completely determined by $f_{t-1}(x)$, $\delta_t(x)$, and ρ_t .

As mentioned by [41], there is a Markov property implied in the covariance structure of Eq 4.12:

$$\text{cov}\{f_t(x), f_{t-1}(x') \mid f_{t-1}(x)\} = 0, \quad (4.16)$$

which is true for all $x \neq x'$. Eq 4.16 indicates that if we have $f_{t-1}(x)$, then other input parameters $f_{t-1}(x')$ do not contribute to training $f_t(x)$.

The Markovian property also suggests that an efficient training set $\{\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_s\}$ for a multi-fidelity GP is a nested structure:

$$\mathbf{x}_1 \subseteq \mathbf{x}_2 \subseteq \dots \subseteq \mathbf{x}_s. \quad (4.17)$$

The above notation says that, given an input point x at fidelity t , there must be an input x in its lower fidelity u , where $u < t$ and $t, u \in \{1, 2, \dots, s\}$. The reason for using a nested experimental design is that since we have $\mathbf{x}_{t-1} \subseteq \mathbf{x}_t$, we can immediately get an accurate posterior $f_{t-1}(x)$ at the x location without interpolating at the $t - 1$ level. However, in practice we found our multi-fidelity emulators performed well even without a nested design in the input space.¹¹

At a given fidelity t , the posterior at a test input x_* could be written as

$$p(f_{*t} \mid \mathcal{D}, x_*) = \mathcal{N}(f_{*t}; \mu_{*t}(x_*), \sigma_{*t}^2(x_*)), \quad (4.18)$$

¹¹Without a nested design in input space, we found, for a multi-fidelity emulator using 50 LR and 3 HR, the non-nested one is only 5% worse than the nested one on the relative errors.

where we denote predictions from new inputs as subscript $*$. The predictive mean and variance are

$$\begin{aligned}\mu_{*t} &= \rho_t \cdot \mu_{*t-1}(x_*) + \mu_{\delta_t} \\ &+ \mathbf{k}_{*n_t} \mathbf{K}_t^{-1} [\mathbf{y}_t - \rho_t \cdot \mu_{*t-1}(\mathbf{x}_t) - \mu_{\delta_t}]; \\ \sigma_{*t}^2 &= \rho_t^2 \cdot \sigma_{*t-1}^2(x_*) + K(x_*, x_*) - \mathbf{k}_{*n_t} \mathbf{K}_t^{-1} \mathbf{k}_{*n_t}^\top,\end{aligned}\tag{4.19}$$

where $\mathbf{k}_{*n_t} = [K_t(x_*, x_1), \dots, K_t(x_*, x_{n_t})]$ is a vector of covariance between the new location and the training locations at fidelity t . $K_t = K_t(\mathbf{x}_t, \mathbf{x}'_t)$ is the covariance matrix of training locations at fidelity t .

Covariance kernel

For a linear multi-fidelity emulator, we place an independent squared exponential kernel on each k_j . The mathematical form of the kernel is the same as Eq 4.10.

Having ARD weights means we assign different length scales to each dimension so that the kernel can be trained anisotropically. We found that using ARD in the highest fidelity did not improve the model's accuracy. Thus, we decided to assign an isotropic kernel for δ_s . For a two-fidelity emulator ($s = 2$), we have 6 hyperparameters in low-fidelity for each k bin; 5 of them are the length scale parameters and 1 is the variance parameter. We have 3 hyperparameters for each k bin in high fidelity, with one scale factor ρ_t between fidelities, one variance parameter, and one length scale parameter. We have 49 bins in k , so the total number of trainable hyperparameters is 441.

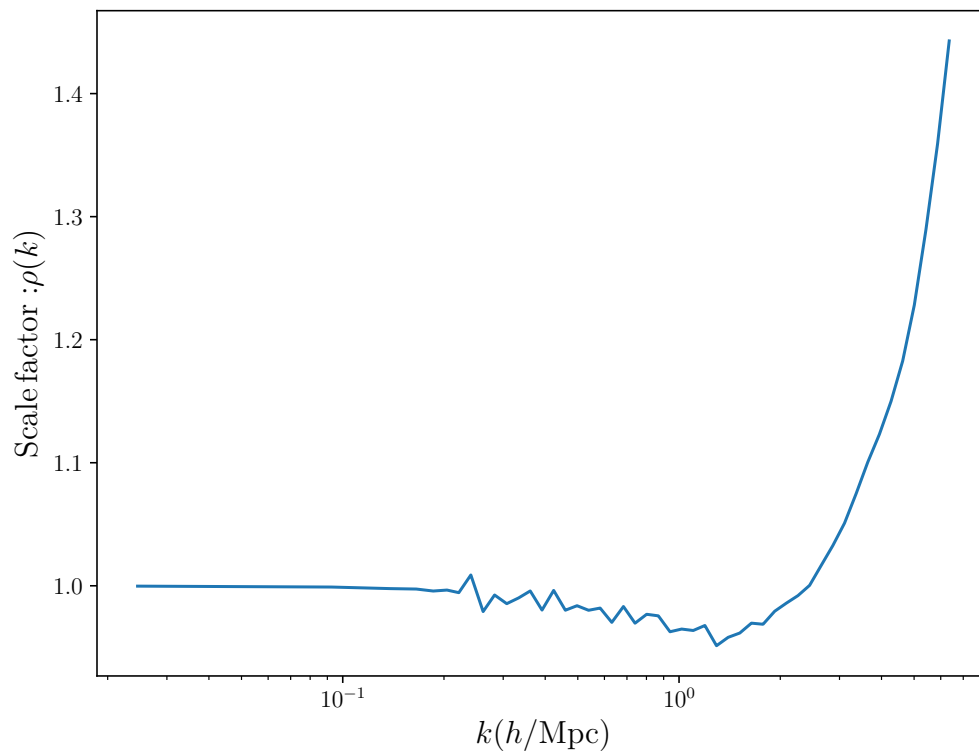


Figure 4.2: The learned scale factor between fidelities in the linear multi-fidelity model, ρ , as a function of k . This scale factor is learned from 50 low-fidelity simulations and 3 high-fidelity simulations.

Figure 4.2 shows the learned scale factor, ρ .¹² ρ is roughly unity at large scales $k \leq 2 h\text{Mpc}^{-1}$, but its value increases dramatically after $k > 2 h\text{Mpc}^{-1}$. Non-linear physics becomes important and the low-fidelity simulations become less reliable at small scales, making the relationship between fidelities non-trivial. We want to emphasize that the scale factor, ρ , is learned from the multi-fidelity emulator. We did not enforce ρ to be a specific shape during the training. Because we learn the mapping from LR to HR using the training data, it is expected that LR runs deviate from HR power spectra. The purpose of multi-fidelity emulation is to correct these deviations.

4.5.3 Non-linear multi-fidelity emulator (NARGP)

The linear multi-fidelity model in Eq 4.12 assumes the scale factor ρ_t is independent of input parameters, x , and so does not model the cosmological dependence of the scale factor ρ_t . The non-linear multi-fidelity model proposed by [153] drops this assumption, allowing the scale factor, $\rho_t(\cdot)$, to be a function of both input cosmology and output from the previous fidelity. As for the linear multi-fidelity model, we model the non-linear multi-fidelity GP independently for each k :

$$f_t(x) = \rho_t(x, f_{t-1}(x) - \mu_{t-1}) + \delta_t(x), \quad (4.20)$$

where $\rho_t(\cdot)$ is a function of both input parameters x and the previous fidelity's output. $\rho_t(\cdot)$ is modelled as a GP. Eq 4.20 results in a more complicated distribution over f_t , a deep Gaussian process [162]. To avoid added computational and statistical complexity, we follow

¹²The multi-fidelity scale factor shown Figure 4.2 is ρ_2 , which is ρ_t when $t = 2$. For simplicity, we use ρ to refer to ρ_2 for our multi-fidelity emulators.

the same approximation as [153] and replace f_{t-1} in Eq 4.20 with its posterior, f_{*t-1} . The result is a regular Gaussian process,

$$f_t \sim \mathcal{GP}(0, K_t), \quad (4.21)$$

whose kernel can be furthermore decomposed:

$$\begin{aligned} K_t(x, x') = & K_{t_\rho}(x, x'; \boldsymbol{\theta}_{t_\rho}) \cdot K_{t_f}(f'_{*t-1}(x), f'_{*t-1}(x'); \boldsymbol{\theta}_{t_f}) \\ & + K_{t_\delta}(x, x'; \boldsymbol{\theta}_{t_\delta}), \end{aligned} \quad (4.22)$$

where $f'_{*t-1} \equiv f_{*t-1}(x) - \mu_{t-1}$ for simplicity. The first kernel K_{t_ρ} models the cosmological dependence of the scale factor ρ . Next, K_{t_f} models the covariance of the output passing from the previous fidelity to the current level. The final term K_{t_δ} models the model discrepancy between fidelities. For the lowest fidelity, the matter power spectrum is only modelled with K_{t_δ} .

Each kernel in Eq 4.22, $(K_{t_\rho}, K_{t_f}, K_{t_\delta})$, is modelled as a squared exponential kernel. Suppose we assign a different length scale parameter for each x dimension. K_{t_ρ} will have $d+1$ hyperparameters, K_{t_f} will have 2 hyperparameters, and K_{t_δ} will have $d+1$ hyperparameters. As for the linear emulator, we found no improvement in accuracy in practice by using ARD for the high-fidelity model. Thus, we have 2 hyperparameters for each kernel in high fidelity and $d + 1$ hyperparameters for low-fidelity. To be explicit, in the high-fidelity model, K_{t_ρ} has 2 hyperparameters, K_{t_f} has 2 hyperparameters, and K_{t_δ} has 2 hyperparameters. For $d = 5$, we have 6 hyperparameters for low-fidelity and 6 for high-fidelity models at each k bin.

Halo Model Interpretation

The formulation of our multi-fidelity emulator bears a marked resemblance to the equations which form the basis of HALOFIT [46], and are themselves motivated by the halo model [47, 48]. This correspondence allows us to provide a physical interpretation of our results. In the halo model, matter clustering is schematically divided into two components: a two-halo term and a one-halo term. The two-halo term arises from correlations between halos on large scales, while the one-halo term, which has a weaker dependence on cosmology, is sensitive to the density profile inside each halo. We can model this by splitting the non-linear power spectrum

$$P_{NL}(k) = P_Q(k) + P_H(k). \quad (4.23)$$

The quasilinear term $P_Q(k)$ is a two-halo term, while $P_H(k)$ is a one-halo term. The two-halo term can be modelled by the linear theory power spectrum filtered by a window function $W(M, k)$:

$$P_Q(k) = P_L(k) \left(\int W(M, k) dM \right)^2. \quad (4.24)$$

The window function depends on the halo mass function and halo bias, encodes how virialisation displaces the linear matter field, and tends to unity on large scales.

There is a clear connection between this model and the form of our multi-fidelity emulator. Eq 4.12 (AR1) and Eq 4.20 (NARGP) move between fidelities via two terms: a scaling factor ρ and an additive factor δ_t . The correlations between fidelities are strong on large scales, and so $\rho \rightarrow 1$ as $k \rightarrow 0$. ρ is analogous to the quasilinear window function, except that it filters not the linear theory power spectrum P_L , but the low-fidelity N -body

model $f_{t-1}(x)$. In the context of the halo model, it extrapolates the existing quasilinear halo filtering to include lower mass halos not included in the low-fidelity simulation.

The additive factor δ_t , which is important on small scales, is analogous to the one-halo term. It models the difference in halo shot noise and internal halo profiles between resolutions. Notice that δ_t , like the one-halo term, depends only weakly on cosmology, as evidenced by it requiring only one length-scale hyperparameter.

4.6 Sampling Strategy for High-Fidelity Simulations

In this section, we will describe how we select the training simulations for our multi-fidelity emulators. We will first describe the nested structure implemented in multi-fidelity emulators in Section 4.6.1. Section 4.6.2 explains how we find the optimal choice of high-fidelity training simulations.

4.6.1 Nested training sets

The proposed sampling scheme for training and testing is shown in Figure 4.3. The corresponding output power spectra are shown in Figure 4.4. In Figure 4.3, the sampling is done using two different Latin hypercubes:

1. Training simulations: a Latin hypercube with 50 points. HR points are a subset of LR points.
2. Testing simulations: another Latin hypercube with 10 points.

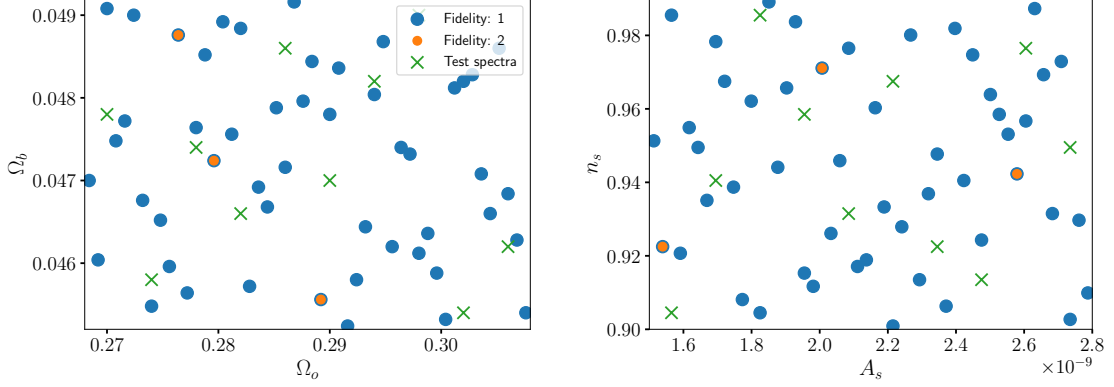


Figure 4.3: Two 2-D cross-sections of the 5-D samples of input parameters. The input parameters are designed with a nested structure, $\mathbf{x}_1 \subseteq \mathbf{x}_2$, between HR and LR. **(Blue):** \mathbf{x}_1 , 50 sampling points in LR. **(Orange):** \mathbf{x}_2 , 3 sampling points in HR. The selection of these 3 points is chosen by the procedure described in Section 4.6.2, which minimizes the LR error in the low-fidelity only emulator. **(Green):** 10 points from the HR testing set, which is a different Latin hypercube than \mathbf{x}_1 .

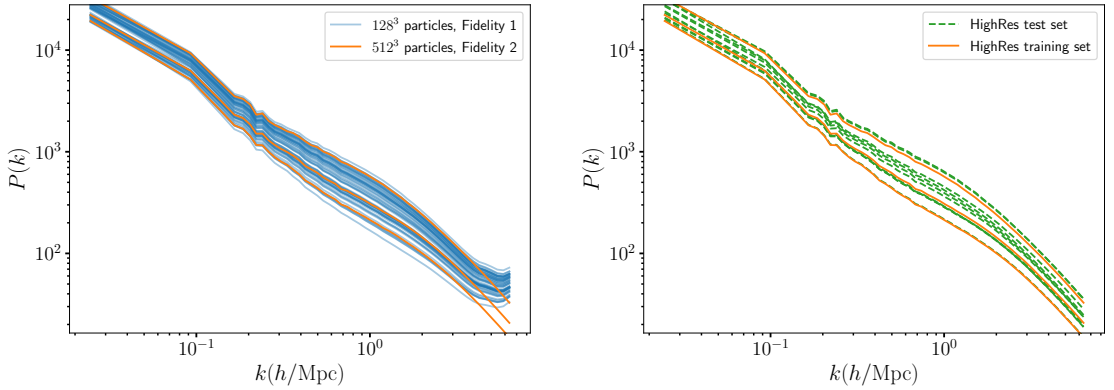


Figure 4.4: Training (left) and testing (right) data for the multi-fidelity emulator. **(Left):** 50 low-fidelity training simulations (blue) and 3 high-fidelity simulations (orange) used in a 50LR-3HR emulator. A HR is a 512^3 simulation and a LR is a 128^3 simulation. Both HR and LR are in a box with 256 Mpc/h per side. The 50 low-fidelity training simulations are drawn from a 5D Latin hypercube, $(h, \Omega_0, \Omega_b, A_s, n_s)$. The 3 high-fidelity simulations are a subset of the low-fidelity simulation hypercube. **(Right):** 10 high-fidelity test simulations (green dashed) and 3 high-fidelity training simulations (orange).

3. We use the notation “ X LR- Y HR emulator” to represent a multi-fidelity emulator trained on X number of low resolution simulations and Y number of high resolution simulations.

The first hypercube with 50 points ensures that we will have a nested experimental design. The second hypercube is to ensure we will not test on the training simulations during the validation phase. In practice, we found that the emulation accuracy roughly converged with ~ 30 LR points.

4.6.2 Optimizing the loss of low-fidelity simulations

For a multi-fidelity problem, we want to minimize the required high-fidelity training simulations to achieve a given accuracy. We search for the optimal subset of LR points to simulate at HR by picking the subset that would minimize the low fidelity training set’s single-fidelity emulator errors. In our experiments with two fidelities, $s = 2$, there are $\binom{n_1}{n_2}$ possible combinations for \mathbf{x}_2 , which are input parameters for the high-fidelity data, $\mathcal{D}_2 = \{\mathbf{x}_2, \mathbf{y}_2\}$.

Retraining low-fidelity only emulators on all possible subsets of the low fidelity grid is computationally intensive. For example, selecting two samples out of 50 points means that we have to train $\binom{50}{2} = 1225$ low-fidelity emulators. To save computational cost, we employed a greedy optimization strategy. Instead of exploring all possible subsets, we grew the subset one point at a time, fixing the previously chosen points. As a further optimisation, we used the same set of kernel hyperparameters for all k bins.

Consider \mathcal{S} , a potential \mathcal{D}_2 with $\mathbf{x}_2 \subset \mathbf{x}_1$. We train a low-fidelity only emulator based on Eq 4.8 using the n_2 low-fidelity points in \mathcal{S} and get a GP:

$$p(f_* | \mathcal{S}, x_*) = \mathcal{N}(f_* | \mu_*^{(i)}(x_*), \sigma_*^{(i)}(x_*)^2), \quad (4.25)$$

which is the posterior as described in Eq 4.9.

With the trained low-fidelity only emulator in Eq 4.25, we can test this single-fidelity emulator's performance by predicting the rest of the data in the low-fidelity Latin hypercube. To evaluate the accuracy, we compute the mean squared error by averaging over the test data:

$$\text{MSE} = \mathbb{E}[(y_* - \mu_*^{(i)}(x_*))^2], \quad (4.26)$$

where $\{(x_*, y_*)\}$ are the low-fidelity data pairs from the rest of the Latin hypercube,

$$\{(x_*, y_*)\} \in \{\mathcal{D}_1 - \mathcal{S}\}. \quad (4.27)$$

This simply means that we test the single-fidelity emulator on the available data not included in the training subset.

Suppose we repeat the training of single-fidelity emulators until we train all possible subsets in the low-fidelity hypercube. We will now have $\binom{n_1}{n_2}$ trained single-fidelity emulators. Each single-fidelity emulator will provide a mean squared error, which is the test error that the emulator generates against the low-fidelity hypercube test data. To select the optimally trained emulator, we compute

$$\mathcal{S}^* = \arg \min_{\mathcal{S}^*} (\mathbb{E}[(y_* - \mu_*^{(i)}(x_*))^2]), \quad (4.28)$$

where we find the subset \mathcal{S}^* which minimizes the mean squared errors on the test set. We use \mathcal{S}^* as our high-fidelity training set \mathcal{D}_2 under the nested experimental design. To be explicit:

$$\mathbf{x}_2 = \mathbf{x}_{\mathcal{S}^*} \subset \mathbf{x}_1, \quad (4.29)$$

where \mathbf{x}_2 are the selected high-fidelity input points, $\mathbf{x}_{\mathcal{S}^*}$ are the input points from the selected subset \mathcal{S}^* (which minimize the low-fidelity emulator mean squared error), and \mathbf{x}_1 are the low-fidelity input points.

This strategy assumes that the effect of a sampling scheme on a low-fidelity emulator is the same as that on a corresponding multi-fidelity emulator. For example, suppose $\Delta\Omega_b$ is crucial for learning how the low-fidelity power spectrum y_1 changes for inputs x_1 . In that case, we expect that information about $\Delta\Omega_b$ can also effectively change the high-fidelity spectrum y_2 .

The above assumption could be violated if the power spectra at small scales, which are not included in the low-fidelity data, behave very differently from those at large scales. This could happen if the smoothness length scale acts very differently between low-fidelity and high-fidelity data for a given input dimension. For example, imagine that a parameter, θ , has a small effect on the outcomes of low-fidelity simulations, but a large effect on the outcomes of high-fidelity simulations.

Figure 4.5 shows the mean squared errors computed from 64^3 single-fidelity emulators and 256^3 single-fidelity emulators. First, note that the selection of the training simulations affects the emulator accuracy. Second, the low-fidelity emulator errors are correlated with their higher fidelity counterparts. This suggests that a low-fidelity emulator can

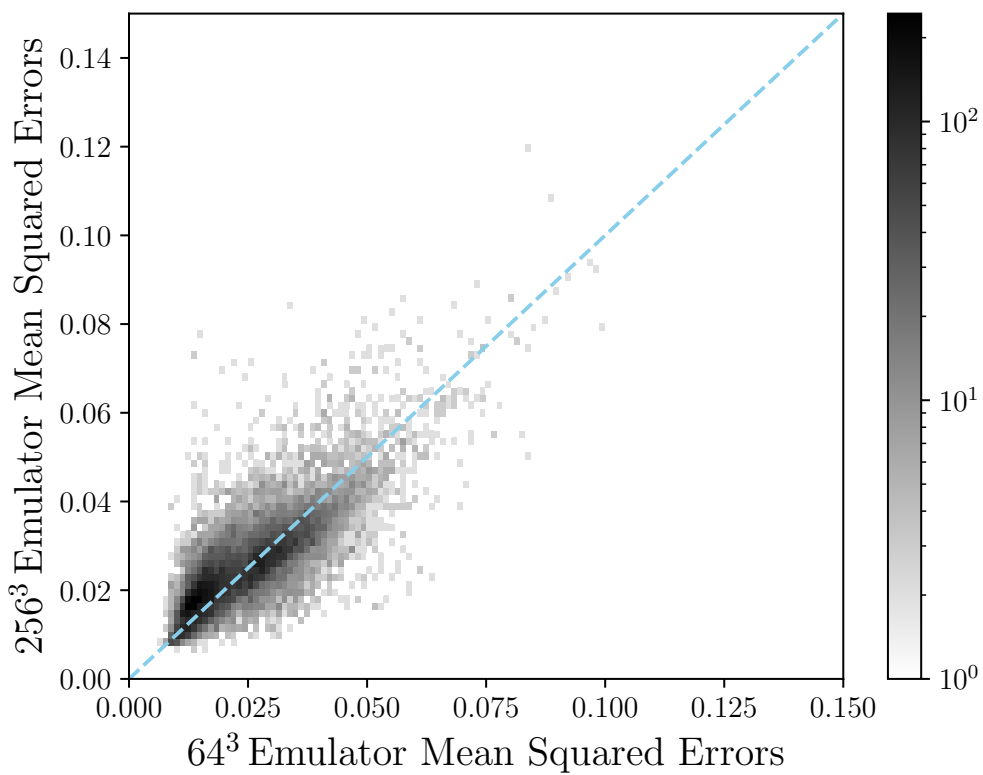


Figure 4.5: Emulator mean squared errors evaluated from 64^3 emulators and 256^3 emulators. We compute all subsets of 3 samples from a 50 samples Latin hypercube, $\binom{50}{3} = 19\,600$ subsets in total. Colorbar is in log scale. The blue dashed line represents a perfect linear relationship.

serve as a guide for placing high-fidelity training simulations. The HR parameter choices used in Section 4.7 were selected with an earlier version of our model using 64^3 particle simulations. We checked that using either 64^3 or 128^3 for selection gave almost the same emulation accuracy for a non-linear 50 LR-3 HR emulator, though one of the selected samples is different.

In practice, we find the procedure above can prevent us from selecting the HR combination that will give us the worst multi-fidelity emulation result. Although we have tested that our procedure works for the matter power spectrum, we would suggest that when emulating a new summary statistic (e.g., the halo mass function), the reader investigates the effectiveness of this method using small test cases. We may in future work investigate using Bayesian optimization [e.g., Ref. [143, 144, 145]] to select the optimal HR samples for multi-fidelity training.

4.7 Results

This section shows the interpolation accuracy of multi-fidelity methods and compares our multi-fidelity emulators to single-fidelity emulators. Section 4.7.1 compares test set emulator errors for the linear multi-fidelity emulator (AR1) and non-linear multi-fidelity emulator (NARGP). Section 4.7.2 compares a multi-fidelity emulator to two kinds of single-fidelity emulators: high-fidelity only and low-fidelity only. We also compare the emulator accuracy as a function of core hours for both multi-fidelity emulators and single-fidelity emulators.

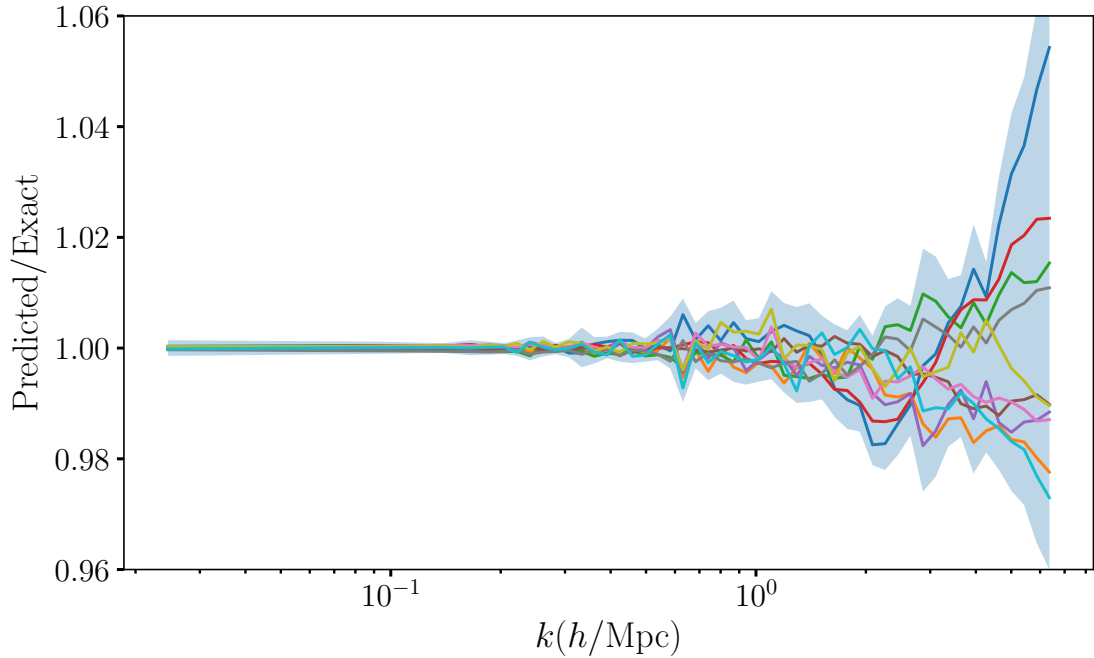


Figure 4.6: Predicted divided by exact power spectrum from a 50 LR-3 HR emulator using a linear multi-fidelity method (AR1). Different colours correspond to 10 test simulations spanning a 5-D Latin hypercube. The shaded area indicates the worst-case $1 - \sigma$ emulator uncertainty. There is one test simulation driving the larger error compared to the non-linear one in Figure 4.7.

To test how much a multi-fidelity emulator can improve with more training simulations, Section 4.7.3 shows the emulator errors with more LR or HR training simulations. Finally, Section 4.7.4 checks the performance of the multi-fidelity method for other emulation settings.

4.7.1 Comparison of Linear and Non-Linear Emulators

Figure 4.6 and Figure 4.7 show the predicted power spectrum divided by the exact power spectrum for simulations in the testing set. Both emulators, linear (AR1) and non-

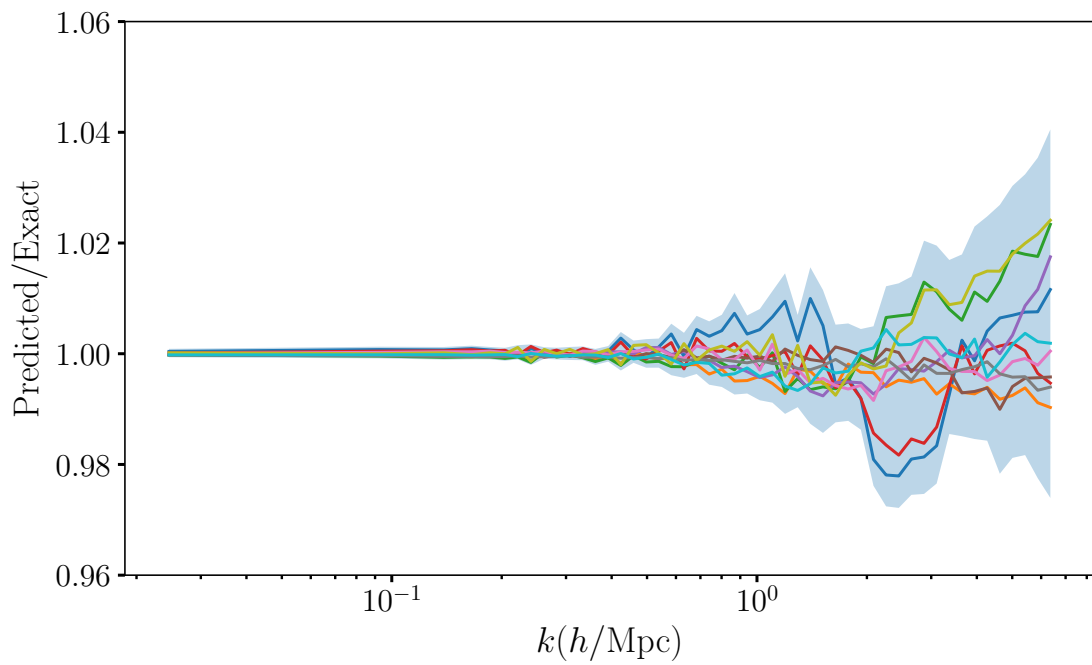


Figure 4.7: Predicted divided by exact power spectrum from a 50 LR-3 HR emulator using a non-linear multi-fidelity method (NARGP). Different colours correspond to 10 test simulations spanning a 5-D Latin hypercube. The shaded area indicates the worst-case $1 - \sigma$ emulator uncertainty. Note that the y-scale in this plot is the same as Figure 4.6.

linear (NARGP), are trained with 50 low-fidelity simulations and 3 high-fidelity simulations. We will call these emulators “50 LR-3 HR emulators” for simplicity. A non-linear (linear) multi-fidelity emulator requires at least 3 (2) HR simulations for training and has $\lesssim 2\%$ ($\lesssim 5\%$) worst-case accuracy per k bin. For a linear multi-fidelity emulator, the minimum required number of HR simulations is 2, reflecting the lower number of hyperparameters in the kernel.

Figure 4.8 shows a comparison between a linear multi-fidelity emulator and a non-linear multi-fidelity emulator in relative emulator error. We include linear and non-linear 50 LR-3 HR emulators. We define the relative emulator error:

$$\text{Emulator Error} = \left| \frac{P_{\text{pred}}}{P_{\text{true}}} - 1 \right|. \quad (4.30)$$

P_{pred} is the predicted power spectrum from the multi-fidelity emulator, and P_{true} is the power spectrum from the high-fidelity test simulation.

Figure 4.8 shows that the linear 50 LR-3 HR emulator predicts an average error $< 1\%$ per k bin for $k \leq 4 h\text{Mpc}^{-1}$ and $< 2\%$ per k bin for $4 < k \leq 6.4 h\text{Mpc}^{-1}$. The non-linear multi-fidelity emulator predicts an average error $\lesssim 1\%$ per k bin, which implies we only need 3 HR to achieve a percent-level accurate emulator using the non-linear multi-fidelity method. At $k \leq 3 h\text{Mpc}^{-1}$, both emulators predict mostly the same accuracy, but the non-linear one performs better at smaller scales $k > 3 h\text{Mpc}^{-1}$.

We found that the non-linear multi-fidelity emulator outperforms the linear one in all aspects. For simplicity, we will only show the non-linear multi-fidelity models in the following sections, but we note that a linear multi-fidelity model is still useful when only

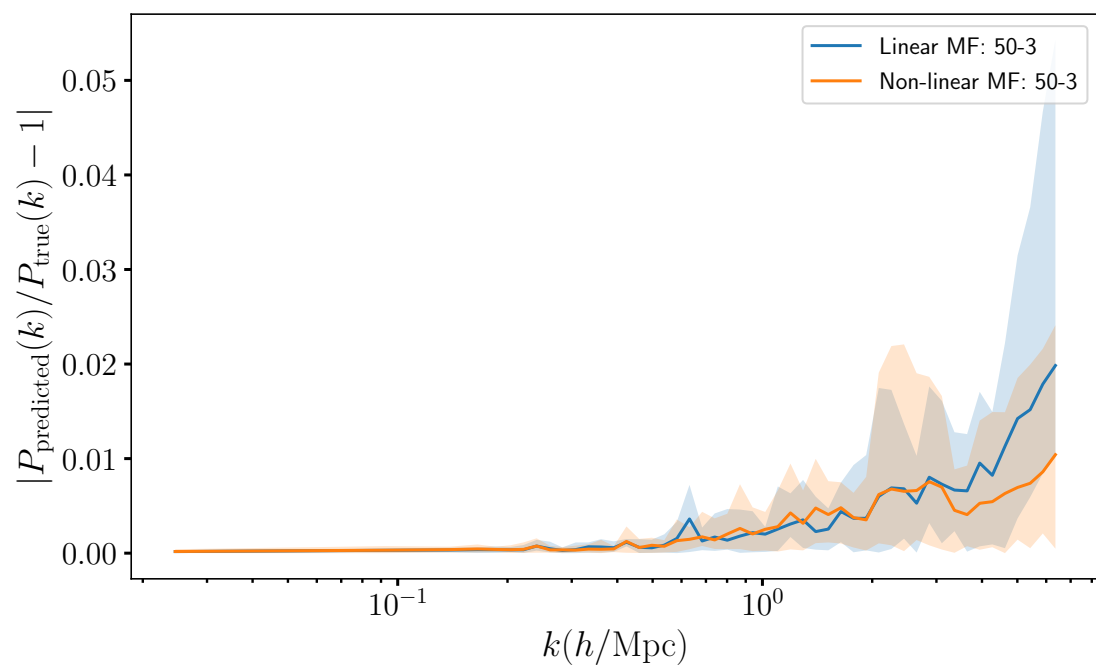


Figure 4.8: Relative emulator errors from a 50 LR-3 HR emulator using linear multi-fidelity (blue) and non-linear multi-fidelity (orange). Solid lines represent the average error from test simulations, $\frac{1}{10} \sum_{i=1}^{10} |\frac{P_{\text{pred},i}}{P_{\text{true}}} - 1|$. Shaded areas show the maximum and minimum test errors.

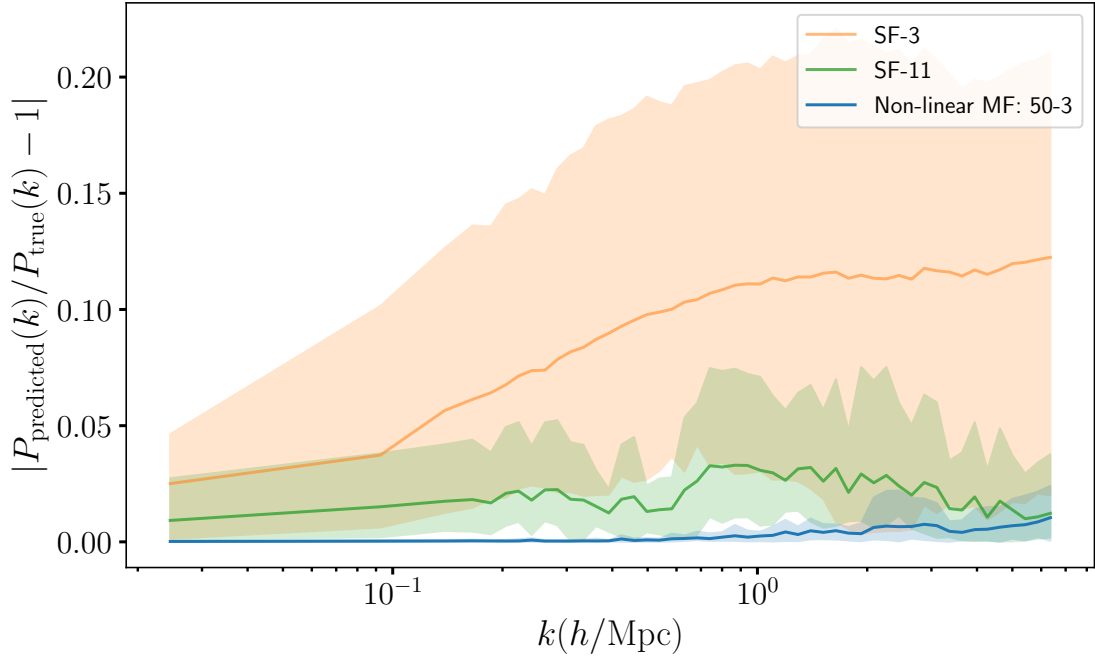


Figure 4.9: **Non-linear multi-fidelity emulator (blue)** with 50 LR and 3 HR simulations, compared to **single-fidelity emulators** with 3 HR (orange) and with 11 HR (green). Shaded area indicates the maximum and minimum emulation errors. The computational cost for a 50 LR-3 HR emulator $\simeq 9\,000$ core hours while the single-fidelity emulator with 11 HR requires $\simeq 25\,000$ core hours. However, a 50 LR-3 HR emulator still outperforms an 11 HR emulator.

two HR simulations are available. We also found that, for the linear model, changing from 50 LR-3 HR emulator to 50 LR-2 HR emulator only slightly degrades the overall accuracy.

4.7.2 Comparison to single-fidelity emulators

Comparison to high-fidelity only emulators

Figure 4.9 shows a comparison between a non-linear 50 LR-3 HR emulator and high-fidelity only emulators. The high-fidelity only emulators are single-fidelity emulators trained solely on HR simulations. The non-linear multi-fidelity emulator outperforms the single-

fidelity emulator with 11 HR at all k modes. It also predicts a worst-case error smaller than the worst-case error from the 11 HR single-fidelity emulator. At $k \leq 2 h\text{Mpc}^{-1}$, the multi-fidelity emulator performs much better than the single-fidelity emulators. Since LR simulations can predict accurate power spectrum at large scales $k \leq 2 h\text{Mpc}^{-1}$, we expect a single-fidelity emulator requires ~ 50 HR to compete with the 50 LR-3 HR emulator on large scales. A HR is $\simeq 64$ times more expensive than a LR, thus the core time for a 50 LR-3 HR emulator is $\simeq 4$ HR. The non-linear multi-fidelity outperforms a single-fidelity 11 HR emulator with $\simeq 3$ times lower computational cost.

The error reduction rate is the relative error of a single-fidelity emulator divided by the error of a multi-fidelity emulator. Both linear and non-linear 50 LR-3 HR emulator show an error reduction rate of $\simeq 100$ for $k \leq 0.5 h\text{Mpc}^{-1}$, $\simeq 100$ times better than the single-fidelity counterpart using 3 HR. At smaller scales $k > 3 h\text{Mpc}^{-1}$, the multi-fidelity emulators are $\simeq 20$ times (non-linear), and $\simeq 10$ times (linear) better than their single-fidelity counterpart.

Comparison to low-fidelity only emulators

Figure 4.10 shows a single-fidelity emulator trained on 50 LR simulations, compared to a non-linear 50 LR-3 HR emulator. Figure 4.10 demonstrates how multi-fidelity modelling improves the emulator accuracy at each k scale. At $k \lesssim 3 h\text{Mpc}^{-1}$, multi-fidelity modelling uses 3 HR to correct the resolution and reduce the average emulator error from $\lesssim 5\%$ to $\leq 1\%$. A low-fidelity emulator predicts a biased power spectrum beyond $k = 3 h\text{Mpc}^{-1}$. However, the multi-fidelity method can moderately correct the bias and reduce the error

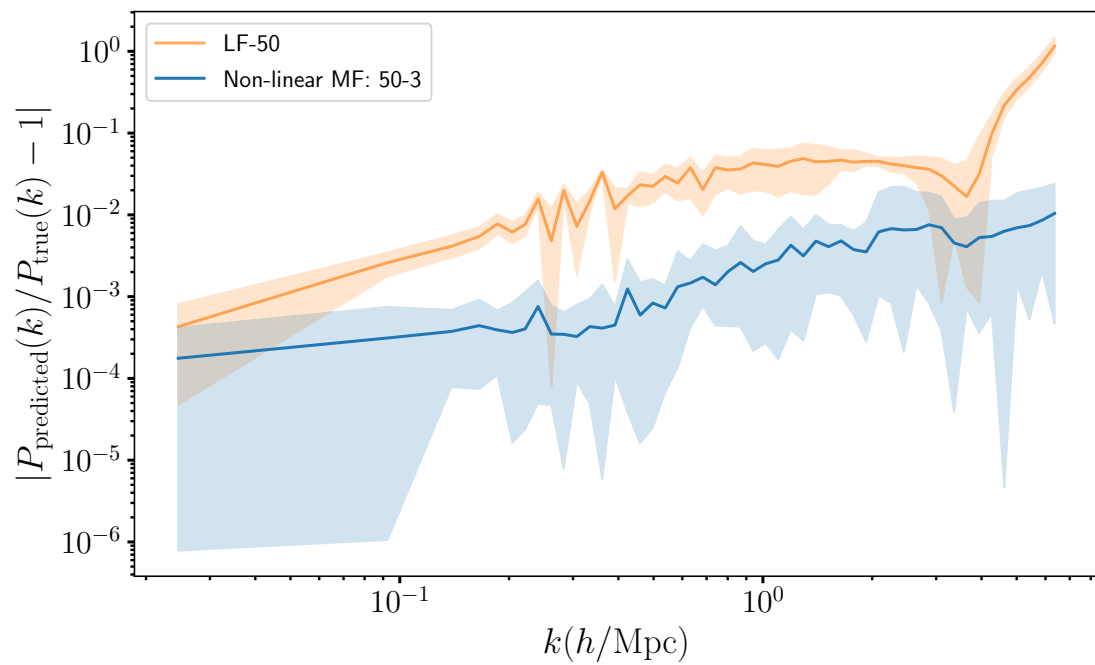


Figure 4.10: Relative emulator errors between a 50 low-fidelity emulator and a non-linear 50LR-3HR emulator. Errors are evaluated on 10 HR simulations. Shaded area indicates the maximum and minimum errors. Note that the y-axis is in \log_{10} scale.

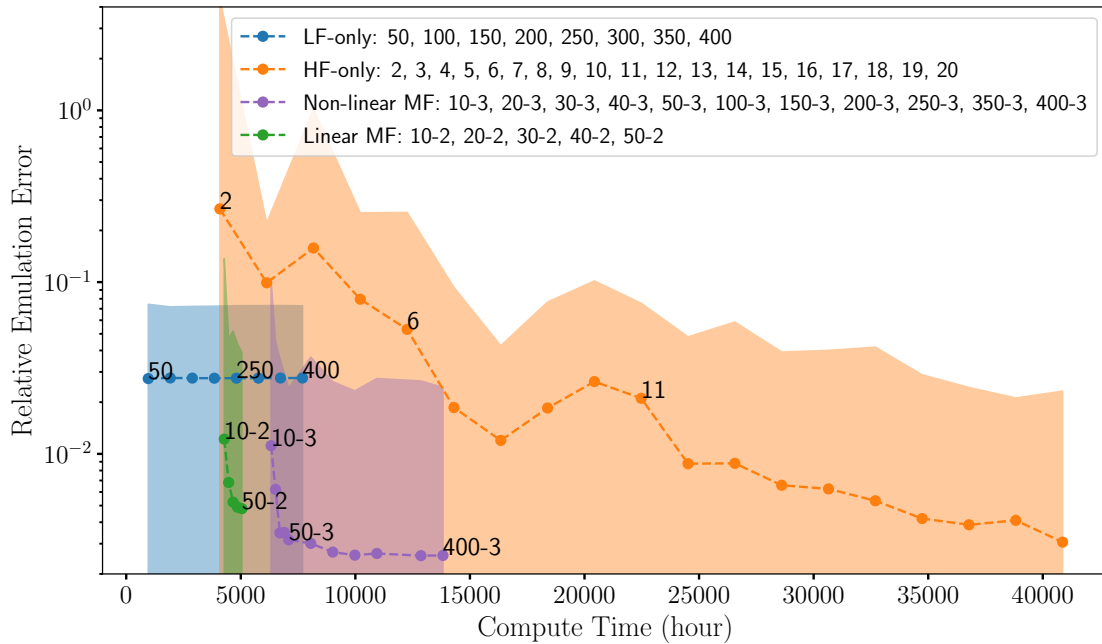


Figure 4.11: Core hours for running the training simulations versus emulation errors for high-fidelity only emulators (orange) and low-fidelity only emulators (blue), linear multi-fidelity emulators (AR1) with 2 HR (green), and non-linear multi-fidelity emulators (NARGP) with 3 HR (purple). The numbers in the labels indicate the number of training simulations used in the emulator. For multi-fidelity emulators, X - Y , X is the number of low-resolution and Y is the number of high-resolution training simulations. The dots show the average errors. The upper shaded areas show the maximum emulator errors among 10 test simulations. The LR samples beyond 100 are drawn from a separate Latin hypercube with 400 samples. For LF-only emulators, we only calculate the relative errors for $k \leq 3$.

to $\lesssim 1\%$. Again, the multi-fidelity technique can use a few HR simulations to calibrate the resolution difference.

Core hours versus emulator errors

Figure 4.11 shows the average relative emulator error as a function of core hours for performing the training simulations. The emulator errors shown in Figure 4.11 are averaged over all k modes, so each emulator corresponds to a single point in the plot.

An ideal emulator will be on the left bottom corner, implying both low cost and high accuracy. The slope of a given emulator in the plot indicates how easily we can improve the emulator with more training data. A steeper (more negative) slope means we can increase the emulator accuracy with a lower cost.

We notice three types of emulators are clustered in separate regions in the plot. The low-fidelity only emulator (LF-only) has the lowest cost and shows no noticeable improvement from increasing training simulations from 50 to 400 LR. The high-fidelity only emulator (HF-only) shows an accuracy improvement with more HR simulations from 3 HR to 11 HR. However, performing one HR requires ~ 2000 core hours, making the HF-only emulator much more expensive than the other two emulators in the plot.

In Figure 4.11, the non-linear multi-fidelity emulator (NARGP) shows a compute time similar to 3 HR simulations but has better accuracy than the HF-only emulator. It also presents a steeper slope than the HF-only emulator, indicating we can efficiently increase the accuracy using low-cost LR simulations. From 10 LR-3 HR emulator to 50 LR-3 HR emulator, it shows that we can decrease the error from ~ 0.02 to ~ 0.003 using an additional ~ 800 core hours. From 50 LR-3 HR emulator to 400 LR-3 HR emulator, we also see a mild decrease of error but not as steep as 10LR-3HR to 50LR-3HR.

We also include the linear model (AR1) to demonstrate the performance of the multi-fidelity method when there are only 2 HR available. The linear model also shows a steep improvement slope from 10LR-2HR to 50LR-2HR. However, we notice that the linear model with 2 HR is slightly worse than the non-linear one with 3 HR.

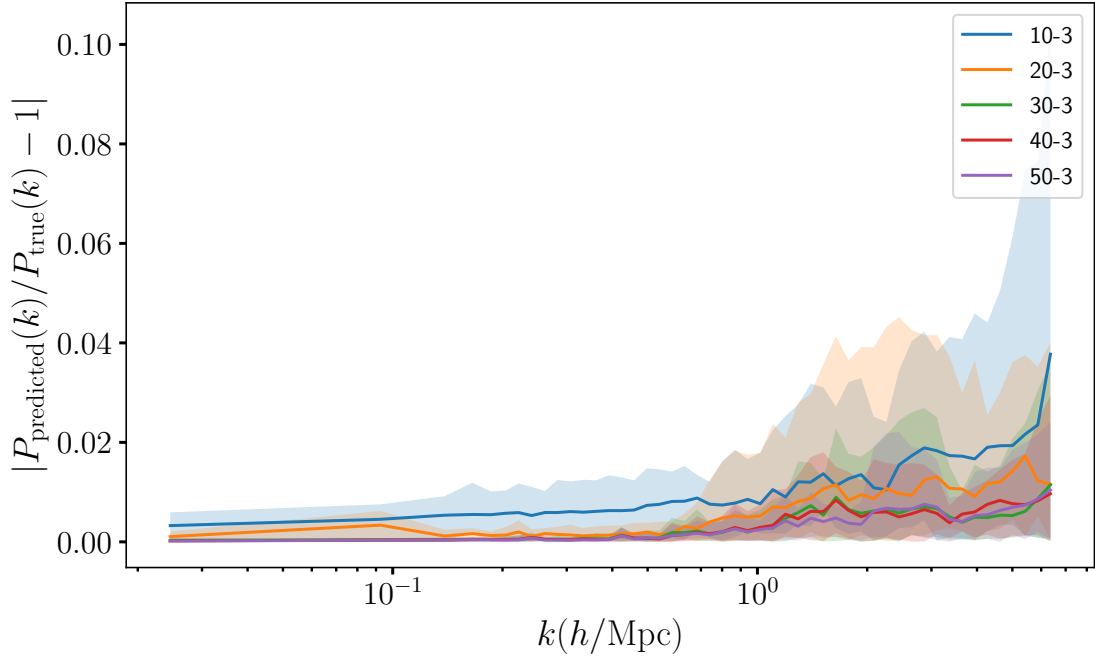


Figure 4.12: Relative emulator error of non-linear N LR-3HR emulator colour coded with different number of LR training simulations, with $N \in \{10, 20, 30, 40, 50\}$. The same as Figure 4.8, solid lines represent the average error from test simulations, $\frac{1}{10} \sum_{i=1}^{10} | \frac{P_{\text{pred},i}}{P_{\text{true}}} - 1 |$, and shaded areas show the maximum and minimum test errors.

Figure 4.11 demonstrates that a multi-fidelity emulator can provide good accuracy with a much lower cost than HF-only emulators. It also points out that we can efficiently improve the accuracy of a multi-fidelity emulator using cheap low-fidelity simulations.

4.7.3 Varying the number of training simulations

Effects of more low-resolution training simulations

The benefit of using a multi-fidelity emulator is that we can improve the emulator accuracy using extra low-fidelity simulations. Figure 4.12 shows the emulator error colour coded by the number of LR training simulations. With more LR training data, the emulator

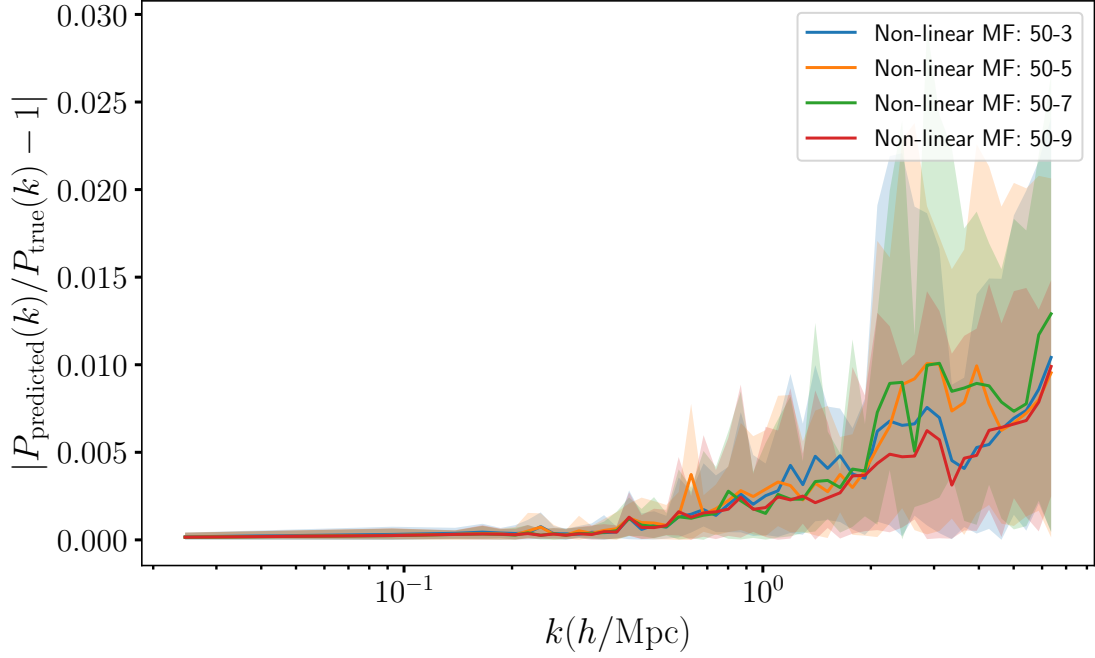


Figure 4.13: Relative emulator errors from non-linear 50 LR- N HR emulator with $N = 3$ (blue), $N = 5$ (orange), $N = 7$ (green), and $N = 9$ (red) HR training simulations. Solid lines are the average test errors. Shaded areas show the maximum and minimum test errors.

performance improves at both large and small scales. We only show the non-linear emulator here for simplicity, but we observe a similar trend in the linear emulator. For $N_{\text{LR-3HR}}$ with $N \in \{10, 20, 30, 40, 50\}$ emulators, the last k bin gives 3.77%, 1.16% , 1.15% , 0.97%, and 1.04% emulator errors, indicating an increase of accuracy with more LR training simulations. Dividing the errors into large and small scales at $k = 1 h\text{Mpc}^{-1}$, the average emulator errors are 0.65%, 0.22%, 0.10%, 0.09%, and 0.09% for $k \leq 1 h\text{Mpc}^{-1}$ and 1.60%, 1.04%, 0.60%, 0.61%, and 0.56% for $k > 1 h\text{Mpc}^{-1}$. The decrease in error is nearly saturated with ~ 40 LR simulations.

Effects of more high-resolution training simulations

In Figure 4.13, we add more HR training simulations to our multi-fidelity emulator. The 50 LR- N HR emulator with $N \in \{3, 5, 7, 9\}$ shows no improvement in average error with more HR, although the worst case error improves noticeably for the 50 LR-9 HR emulator. One reason may be stochasticity in the training set due to simulation modelling error, which is around 1%, and limits the prediction accuracy. In particular, MP-GADGET simulations with 512^3 particles may not be fully converged on small scales, and this limits the emulator’s learning. Another possibility is that the prior from 50 low-fidelity simulations may be too hard to overcome with only 9 HR simulations.

To improve multi-fidelity emulator accuracy further, one could build a more complicated model than the one proposed in this paper. The improvement from the linear to the non-linear model shows that different decisions about the scaling factor ρ could better predict the non-linear structure. However, those complicated models will require more high-fidelity training simulations. We will leave more complex modelling structures to future work.

4.7.4 Effect of other emulation parameters

The resolution of low-fidelity simulations

We have so far tested multi-fidelity emulators using 128^3 simulations (LR) as low-fidelity and 512^3 simulations (HR) as high-fidelity. Figure 4.14 shows non-linear 50 LR-3 HR emulators using different mass resolutions, 64^3 and 256^3 simulations, as low-fidelity.

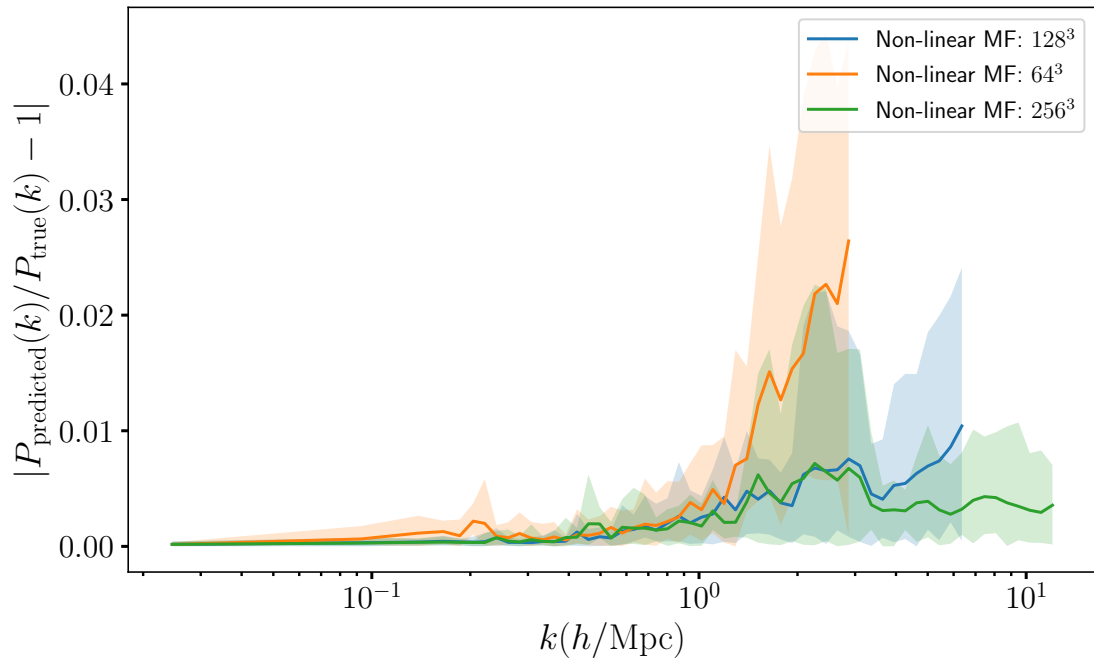


Figure 4.14: Relative emulator errors for 50 LR-3HR emulator emulators using different qualities of LR simulations. **(Blue)**: using 128^3 simulations as low-fidelity training simulations. **(Orange)**: using 64^3 simulations as LR, which are $\simeq 8$ times cheaper than 128^3 simulations. **(Green)**: using 256^3 simulations as LR, which are $\simeq 8$ times most expensive than 128^3 simulations. Shaded area shows the maximum and minimum errors among ten test simulations.

A 64^3 simulation is $\simeq 512$ times cheaper than a HR but has a smaller maximum k with $\max(k) \simeq 3 h\text{Mpc}^{-1}$. It produces percent level accuracy for $k \leq 1 h\text{Mpc}^{-1}$ and has worst-case errors $< 5\%$ at small scales $k \geq 1 h\text{Mpc}^{-1}$. A 256^3 simulation is $\simeq 8$ times cheaper than a HR simulation, so the computational cost for a 50 LR-3 HR emulator is $\simeq 9$ HR simulations. This emulator mildly outperforms the emulator where LR is 128^3 , with an average percent-level emulation until $k \simeq 12 h\text{Mpc}^{-1}$, but at a substantially increased computational cost.

Figure 4.14 demonstrates that one can fuse various qualities of LR with HR simulations to build a multi-fidelity emulator. Figure 4.14 also shows that the multi-fidelity emulator's accuracy depends on the correlation between LR and HR. A 64^3 simulation is only a rough approximation to its 512^3 counterpart, so the emulator that uses 64^3 simulations as low-fidelity is less accurate than the others in Figure 4.14.

Emulation at $z = 1$ and $z = 2$

This section examines the performance of a non-linear emulator at higher redshifts, $z = 1$ and $z = 2$. Figure 4.15 shows the emulator error of a non-linear 50 LR-3 HR emulator at $z = 0, 1, 2$. The mean error at $z = 1$ is smaller than the $z = 0$ error at $k \leq 2 h\text{Mpc}^{-1}$ while it is larger for $k > 2 h\text{Mpc}^{-1}$. This result shows that it is easier to train the correlation between fidelities at large scales $k \leq 2 h\text{Mpc}^{-1}$ while harder to train at small scales $k > 2 h\text{Mpc}^{-1}$. The emulator at $z = 2$ also shows a better performance than $z = 0$ at large scales, $k \leq 2 h\text{Mpc}^{-1}$, but the error diverges to $\sim 10\%$ on smaller scales, $k > 2 h\text{Mpc}^{-1}$. The improved performance on large scales may be because at higher redshifts the matter

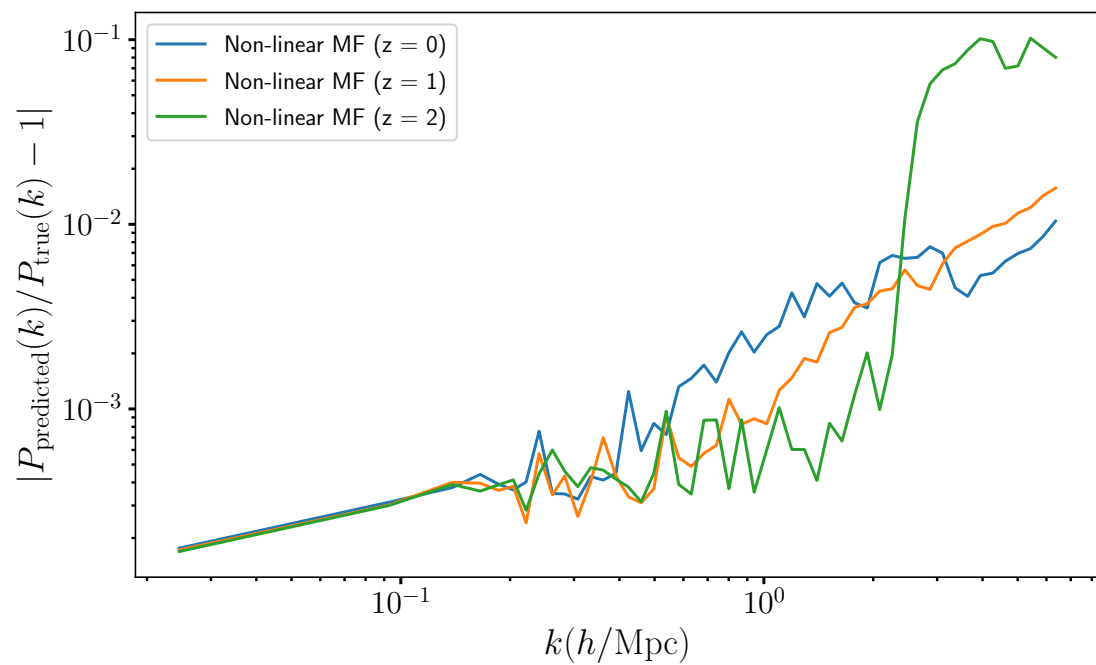


Figure 4.15: Relative emulator errors for a non-linear emulator at different redshifts, $z \in \{0, 1, 2\}$. Note the y-axis is in \log_{10} scale. The larger error in the $z = 2$ emulator at $k > 2 h\text{Mpc}^{-1}$ may be due to a transient near the mean-particle spacing in the LR simulations, see Figure 4.16.

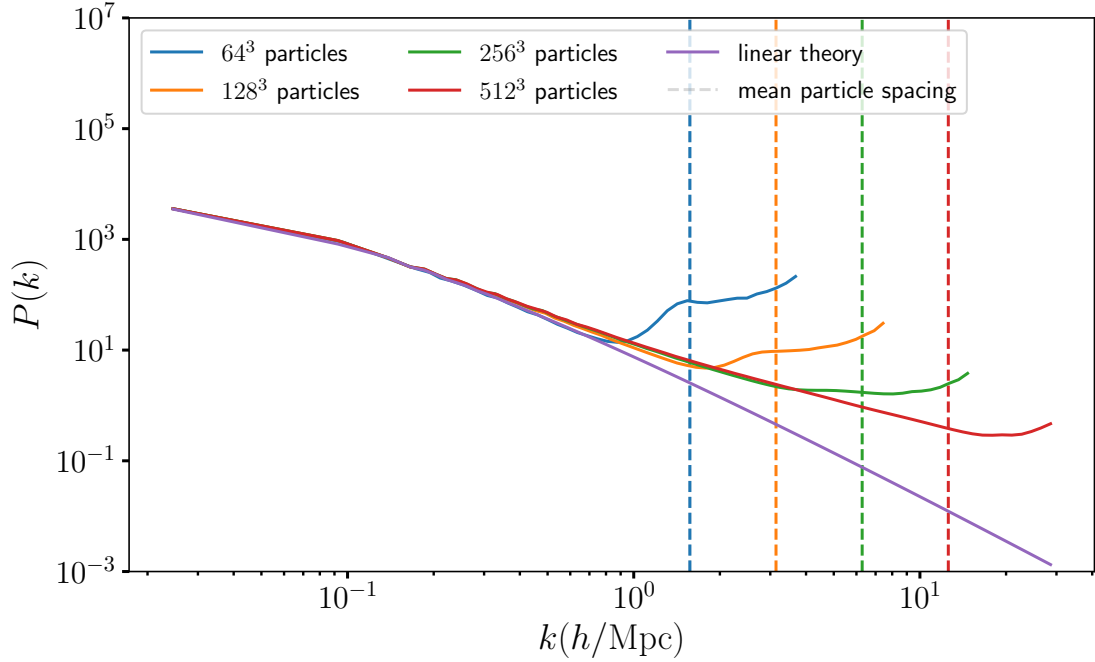


Figure 4.16: The matter power spectrum at $z = 2$, output by MP-GADGET with different mass resolutions. The vertical dash lines indicated the mean particle spacing k_{spacing} for a given mass resolution. **(Blue)**: The matter power spectrum from dark-matter only MP-GADGET simulation with $N_{\text{ptl,side}} = 64$. **(Orange)**: The matter power spectrum from MP-GADGET with $N_{\text{ptl,side}} = 128$. **(Green)**: The matter power spectrum from MP-GADGET with $N_{\text{ptl,side}} = 256$. **(Red)**: The matter power spectrum from MP-GADGET with $N_{\text{ptl,side}} = 512$. **(Purple)**: Linear theory power spectrum.

power spectrum is closer to linear theory and so the correlation between fidelities is easier to learn.

Figure 4.16 shows the matter power spectrum at $z = 2$, with the same cosmological parameters as Figure 4.1 and indicates a potential explanation. At $z = 2$, the low-fidelity simulation contains a systematic at the scale of the mean inter-particle spacing, related to the initial spacing of particles on a regular grid. This systematic is a transient and disappears by $z = 0$. However, at redshifts where it is present it implies that the low

fidelity simulations contain very little cosmological information on scales near their mean interparticle spacing, $k \simeq 3 h\text{Mpc}^{-1}$ and thus cannot significantly improve the emulation accuracy. It may be possible to improve performance at high redshift with the use of other pre-initial conditions such as a Lagrangian glass [163].

4.8 Runtime

We ran our simulations using MP-GADGET on UCR’s High Performance Computing Center (HPCC) and the Texas Advanced Computing Center (TACC). The standard computational setup was 256 MPI tasks per simulation for both HR (512^3 dark matter particles) and LR (128^3 dark matter particles). The runtime was ~ 20 core hours for LR and ~ 2000 core hours for HR, with a fixed boxsize $256 \text{ Mpc}/h$. The computational time for a 64^3 simulation was ~ 1.5 core hours with 64 MPI tasks and ~ 280 core hours for a 256^3 simulation with 256 MPI tasks.

The computational cost for training a non-linear 50 LR-3 HR emulator (NARGP) was $\simeq 0.5$ hours and $\simeq 1.6$ hours for a linear 50 LR-3 HR emulator (AR1) on a single core. For a single-fidelity emulator, it was $\simeq 2$ minutes on one core. The compute time could be further improved by parallelizing the hyperparameter optimization for each k bin. The compute time for optimizing the choice of HR using low-fidelity emulators was ~ 3 hours for selecting 3 HR (on one core). The run time was $\simeq 12$ seconds for evaluating 10 test simulations.

4.9 Conclusions

We have presented multi-fidelity emulators for the matter power spectrum. Multi-fidelity methods fuse together N -body simulations from different mass resolutions to improve interpolation accuracy. Multi-fidelity emulators use many low-fidelity simulations to learn the power spectrum’s dependence on cosmology, correcting for their low resolution by adding a few high-fidelity simulations. The result is equivalent in accuracy to a single-fidelity emulator performed entirely with much more costly high-fidelity simulations. A multi-fidelity emulator’s physical motivation can be understood using the halo model: low-fidelity simulations capture the two-halo term at large scales, while a few high-fidelity simulations are used to learn the (almost cosmology independent) one-halo term at small scales.

We have also proposed a new sampling strategy which uses low-fidelity simulations as a prior to place high-fidelity training simulations. We choose our high-fidelity training samples by optimizing the low-fidelity emulator’s error. In this way, the input parameters at which to run HR simulations can be optimized without knowledge of the HR output. We showed that single-fidelity emulator errors are correlated between different fidelities, indicating that a lower fidelity emulator can serve as a good prior for picking HR simulation points.

Our best multi-fidelity emulator achieved percent level accuracy using only 3 HR simulations and 50 LR simulations, with a total computational cost $\lesssim 4$ HR simulations. We showed it outperforms a single-fidelity emulator with 11 HR simulations. We expect that a

single-fidelity emulator would require ~ 50 HR simulations to compete with the multi-fidelity one at large scales, $k \leq 2 h\text{Mpc}^{-1}$.

In this paper, we used 128^3 simulations as our low-fidelity training sample and 512^3 simulations as high-fidelity, with a fixed $256 \text{ Mpc}/h$ box. However, Figure 4.14 indicates our method still has a good performance when extended to other resolutions. We tested our emulator with a series of 10 HR simulations in a Latin hypercube. Two types of multi-fidelity emulators, linear (AR1) and non-linear (NARGP), are used. We showed that both emulators perform similarly at large scales, while the non-linear one has a better accuracy at small scales.

We focussed on $z = 0$, but also investigated higher redshifts. Higher redshift power spectra behave more linearly than at $z = 0$, so it is easier to learn the large-scale correlation between fidelities. However, the low-fidelity power spectra are less reliable beyond the mean particle spacing at higher redshifts, inducing some difficulty modelling small scales with $k > 2 h\text{Mpc}^{-1}$.

Our multi-fidelity emulators could provide percent-level predictions for future space- and ground-based surveys at a minimum computational cost. All current emulators are single-fidelity, training only on expensive high-fidelity simulations. A single-fidelity emulator requires at least ~ 40 simulations to give percent-level accuracy in a ΛCDM Universe. For example, [28] use 37 simulations to emulate a 5-dimensional ΛCDM model. [10] use ~ 200 high-fidelity simulations (3000^3 dark matter particles) to achieve the upcoming Euclid mission's desired accuracy in an 8 dimensional parameter space.

Our multi-fidelity methods can also be used to improve the existing single-fidelity emulators. For example, suppose we have run 50 high-resolution simulations to build an emulator. We can perform 3 additional super high-resolution simulations and combine them to build a super-resolution multi-fidelity emulator. The choice of these 3 simulations could be selected via the optimization strategy proposed in this paper. Instead of performing super high-resolution simulations, one could use generative adversarial network techniques [see, Ref. [137]] to generate super-resolution simulations and combine them with a multi-fidelity emulator.

Besides increasing the resolution, multi-fidelity methods could also be used to decrease the emulation uncertainty of an existing emulator by extending it with many low resolution simulations. This indicates a low-cost way to enhance current emulators. Multi-fidelity emulators may make possible efficient expansion of the prior parameter volume. Since high-fidelity simulations are only used to calibrate the resolution, they might not need to span the whole parameter space, implying we can expand the sampling range of an existing emulator by extending the low-fidelity sampling range. We will leave this technique to future work.

In this work, we have tested our multi-fidelity emulators with 512^3 resolution and a relatively small box $256 \text{ Mpc}/h$. In future we will apply the framework developed here to create a production quality emulator using higher particle load simulations (e.g., 2048^3 particles) in larger boxes. Other summary statistics, including the halo mass function and the cosmic shear power spectrum, could also be emulated using the same framework.

The multi-fidelity framework may also be extended to hydrodynamical simulations, which are much more costly than their dark matter-only counterparts. No production hydrodynamical emulators including galaxy formation effects such as AGN feedback yet exist.¹³ However, AGN feedback significantly affects the matter power spectrum at $k > 0.1 h\text{Mpc}^{-1}$ [150] and pressure forces can affect the power spectrum at $k \sim 10 h\text{Mpc}^{-1}$ [151]. Thus practical exploitation of the small-scale information from future surveys will require the development of hydrodynamical emulators. By decreasing the computational cost of an emulator by a factor of ≈ 3 and still outperforming a single-fidelity emulator, the work presented here makes emulation development substantially more practical.

Software

We used the `GPY` [165] package for Gaussian processes. For multi-fidelity kernels, we moderately modified the multi-fidelity submodule from `emukit` [155].¹⁴ We used the `MP-GADGET` [149] software for simulations.¹⁵ We generated customized dark matter-only simulations using Latin hypercubes a modified version of `SimulationRunner`.¹⁶

Data Availability

The code to reproduce a 50 LR-3 HR emulator is available at https://github.com/jibanCat/matter_multi_fidelity_emu alongside the power spectrum data.

¹³[36] has a neural net emulator trained with 4233 (magneto-)hydrodynamical simulations in a relatively small box, 25 Mpc/h. [164] has an hydro-emulator using baryonification methods for BACCO simulations.

¹⁴<https://github.com/EmuKit/emukit>

¹⁵<https://github.com/MP-Gadget/MP-Gadget>

¹⁶<https://github.com/sbird/SimulationRunner>

Chapter 5

MF-Box: Multi-fidelity and multi-scale emulation for the matter power spectrum

5.1 Abstract

We introduce **MF-Box**, an extended version of **MFEmulator**, designed as a fast surrogate for power spectra, trained using N-body simulation suites from various box sizes and particle loads. To demonstrate **MF-Box**'s effectiveness, we design simulation suites that include low-fidelity suites (L1 and L2) at 256 Mpc/h and 100 Mpc/h, each with 128^3 particles, and a high-fidelity suite (HF) with 512^3 particles at 256 Mpc/h, representing a higher particle load compared to the low-fidelity suites. **MF-Box** acts as a probabilistic resolution correction function, learning most of the cosmological dependencies from L1 and L2 sim-

ulations and rectifying resolution differences with just 3 HF simulations using a Gaussian process. MF-Box successfully emulates power spectra from our HF testing set with a relative error of $< 3\%$ up to $k \simeq 7 h\text{Mpc}^{-1}$ at $z \in [0, 3]$, while maintaining a cost similar to our previous multi-fidelity approach, which was accurate only up to $z = 1$. The addition of an extra low-fidelity node in a smaller box significantly improves emulation accuracy for MF-Box at $k > 2 h\text{Mpc}^{-1}$, increasing it by a factor of 10. We conduct an error analysis of MF-Box based on computational budget, providing guidance for optimizing budget allocation per fidelity node. Our proposed MF-Box enables future surveys to efficiently combine simulation suites of varying quality, effectively expanding the range of emulation capabilities while ensuring cost efficiency.

5.2 Introduction

Over the past decade, cosmological large-scale structure surveys have evolved increasingly in resolution and size. As observations probe more non-linear structures with high precision, theoretical predictions must be highly accurate to match the observational errors at corresponding small scales. The only way to achieve such accurate predictions is by running N -body simulations. However, including expensive numerical simulations in the cosmological inference will require $\sim 10^6$ likelihood evaluations using simulations, i.e., $\sim 10^6$ numerical simulations in the Markov Chain Monte Carlo (MCMC) sampling, making it impractical to use simulations for Bayesian inference directly.

In the development of statistical surrogate modeling, emulators emerged as a Bayesian approach to analyze simulations and perform fast function predictions [166, 30,

167]. In cosmology, emulators have been widely used as a fast surrogate model to replace the expensive likelihood evaluations in the MCMC sampling. For example, using surrogate models to replace the Boltzmann code in cosmological inference [168, 169, 170, 171, 172, 173]. With a large number of training samples ($\sim \mathcal{O}(10^4 - 10^6)$), these Boltzmann code emulators have successfully improved the speed of the current parameter estimation pipeline. Another approach is using surrogates to replace MCMC to emulate the posterior distribution directly, reducing the overall required number of likelihood evaluations [174].

Unlike the emulators for Boltzmann codes, likelihood evaluations based on numerical simulations, such as cosmological N -body simulations, are more expensive per training sample. Therefore, only a limited number of full-size training simulations ($\sim \mathcal{O}(10^1 - 10^2)$) are computationally available. Emulation based on numerical simulations has been implemented in various cosmological applications: the matter power spectrum [28, 116, 118, 175, 11], baryonification simulations [32, 176], arbitrary cosmology [129], $f(R)$ gravity [177, 178], weak lensing [126, 179, 180], halo mass function [124, 181, 182], 21-cm power spectrum [45] and global signal [183, 184, 185], and Lyman- α forest [128, 138, 130, 186, 187, 188]. All these emulators are self-consistent and can replicate the simulations as surrogate models to accelerate the parameter inference pipeline.

Emulators have also been used in several current surveys. [189] used an emulator on Dark Energy Survey year 3 data (DES Y3) for cosmic shear peak statistics. [190] used an emulator on SDSS quasars and galaxies. Beyond cosmological inference, [191] uses emulation to calibrate the galaxy formation simulations. [192, 193] build emulators to quantify the subgrid feedback effects in the hydrodynamical simulations. Emulators have also been used

in a wide range of disciplines, for example, exoplanet [194], gravitational wave [195], stellar population synthesis [196], heavy-ion physics [197, 198], astrochemistry [199], and biology [200].

The computational costs of cosmological emulators are rapidly increasing, driven by an increase in both survey accuracy and number of model parameters. Over the past few years, cosmological emulators based on N -body simulations have evolved from five-dimensional cosmology (e.g., w CDM in Coyote Universe [28]) to higher dimensions, for example, eight-dimensional w_0w_a CDM+ $\sum m_\nu$ cosmology in [11] and Mira-Titan Universe [118, 201]. The increase in dimensionality means the number of simulations required for training an accurate emulator also needs to increase dramatically. For instance, EuclidEmulator2 requires more than 200 high-resolution simulations with 3000^3 in an eight-dimensional cosmology. Moreover, when the astrophysics effects are not ignorable for cosmological inference [176, 32, 202], more expensive simulations, such as hydrodynamical simulations including baryonic effects, must be used for training realistic emulators. This increase in computational cost poses a challenge for the implementation of emulators in future surveys, making them prohibitively expensive and difficult to adopt unless the efficiency of emulation techniques can be improved.

An efficient approach to reducing the computational cost is building emulators using multi-fidelity emulation (**MFE**mulator), which allows simulations with different particle loads to be combined [16]. [42] showed that it is possible to construct a realistic emulator using hydrodynamical simulations through the **MFE**mulator technique, emulating Lyman- α forest with sub-percent test accuracy using only 6 high-fidelity simulations. In [16, 42], we

assumed the particle load is the only fidelity variable. This is a limitation, as simulation volumes also correlate with the accuracy of a simulation: With a constant particle load, larger box sizes enhance accuracy at larger scales but diminish it at smaller scales due to reduced mass resolution. Smaller volumes with the same particle load can capture finer small-scale details, though a minimum box size requirement exists [29, 26]. Here we show that the cost of training a `MFEulator` can be further reduced by having multiple fidelities which vary both simulation volumes and particle loads.

The multi-fidelity method we use, based on [41, 16], is just one of many multi-fidelity techniques. [161] surveyed the multi-fidelity methods in uncertainty quantification, inference, and optimization. A few popular methods include the control variate technique, which has been applied in cosmology in [203, 204] on reducing the variance of the covariance matrix, and multi-level or multi-stage Markov Chain Monte Carlo [205, 206], which use low-fidelity models to reduce the number of expensive likelihood evaluations in MCMC. Though multi-level MCMC is a promising method, its practical use requires running thousands of N -body simulations in the sampler, which is not yet applicable to cosmological inference. Another similar method is using deep learning methods to learn the mapping from low- to high-resolution simulations to directly generate the snapshots of the ‘super-resolution’ simulations [207, 136, 208]. While this method shows promise, it is currently limited to a single cosmology and is not yet suitable for inference.

The statistical and computer science literature already contains work on multi-fidelity techniques with more than one low-fidelity node. [144, 145] considered a multi-information source framework, which combines more than one information node to achieve

an overall lower variance. In this work, we use a graphical Gaussian process, based on a directed acyclic graph [197], to predict high-fidelity simulations using low-fidelity simulations in two different simulation volumes.

A design using multiple low-fidelity nodes can be helpful in several ways. One example, which we will show in this work, is enhancing the resolution at small scales using an additional low-fidelity node with a smaller box size. A cosmological simulation has strict volume requirements to ensure that the base mode is linear and to beat cosmic variance. However, it also needs high enough particle load (or spatial resolution) to capture the non-linearities at small scales. `MFEulator` provides a way to improve small-scale structures using a simulation suite from a lower particle load. Nevertheless, the non-linear information in a lower particle-load simulation is also limited. An economical way to resolve small scales is to run simulations in small boxes to increase the spatial resolution by sacrificing some large-scale information.

Another approach to minimizing the number of training simulations is Bayesian optimization, where a sequential choice of new training simulations is designed to optimize the likelihood function globally. For example, [138, 139, 140] implemented Bayesian optimization in the cosmological inference. Similar approaches, such as [141, 209, 210, 190], iteratively train emulators on the high likelihood regions of the parameter space, thus minimizing the overall training samples to achieve accurate posterior distribution. Our multi-fidelity emulation is a complimentary technique, which can be combined with Bayesian optimisation for the lowest computational cost.

This paper presents **MF-Box**, extending our previously developed **MFEulator** to allow multiple low-fidelity nodes in a multi-fidelity emulator. **MF-Box** uses the multi-fidelity graphical Gaussian process model (GMGP) [197] to emulate high-fidelity simulations using low-fidelity simulations from two different simulation volumes. A GMGP model is an extension of the traditional KO model [41] and NARGP model [153]. The difference is that a GMGP allows multiple nodes in a fidelity while KO or NARGP models assume one node per fidelity. For example, in our case, the low-fidelity nodes include separate box sizes with the same particle load, resolving different scales of the Universe.

Our references to low- and high-fidelity nodes are based on a relative scale within the context of our multi-fidelity framework. We do not directly compare these definitions to other matter power spectrum emulators. Our primary goal is to demonstrate the effectiveness of **MF-Box** as a probabilistic resolution correction tool. This allows us to correct the resolution of a low-fidelity emulator, approximating higher particle loads using a limited number of high-fidelity simulations.

Consequently, the focus of our discussion on emulation error revolves around predicting unseen high-fidelity simulations in the test set. This choice is intentional, as it allows us to assess how well **MF-Box** can upscale a low-fidelity emulator when predicting high-fidelity simulation outputs. It is worth highlighting that the framework we present here can be adapted for use with various other summary statistics emulators, accommodating different definitions of low- and high-fidelity nodes as needed.

We will also present an analysis of the emulation errors in relation to the computational budget. Previous studies [197, 211] have demonstrated that Gaussian process

emulator errors can be bounded by a power-law function. In this paper, we model the emulation error from MF-Box as a power-law function of the number of training simulations and empirically infer the emulator error function from our MF-Box results. By utilizing this empirical error function, we can estimate the emulation error associated with a given multi-fidelity design, as well as determine the optimal budget allocation for each node. This error analysis serves as a useful guide for future development of MFEmulator techniques.

In Section 5.3, we will describe our simulations and experimental design. Section 5.4 will review the single-fidelity emulator as well as three multi-fidelity emulation methods, namely AR1, NARGP, and MF-Box. Our sampling strategy for selecting input cosmologies for high-fidelity simulations will be outlined in Section 5.5. Empirical inference of the emulation error function will be discussed in Section 5.6. Section 5.7 will present the results of MF-Box, followed by the conclusion in Section 4.9.

5.3 Simulations

We perform dark matter-only simulations using the open source MP-Gadget code [149],¹ an N -body and smoothed particle hydrodynamical (SPH) simulation code derived from Gadget-3 [148] and used to run the ASTRID simulation [39, 212], a large-scale high-resolution cosmological simulation with 250 Mpc/h containing 2×5500^3 particles. The base of MP-Gadget is Gadget-3, but, among other improvements, it has been rewritten to take advantage of shared-memory parallelism and the hierarchical timestepping from Gadget-4 [213]. Detailed descriptions of the simulation code can be found in [39].

¹<https://github.com/MP-Gadget/MP-Gadget/>

Table 5.1: Low- and high-fidelity simulation suites used in our study. The definition of low- and high-fidelity nodes is based on a relative scale specific to our approach and is not intended for direct comparison with other matter power spectrum emulators.

Simulation	Box Volume	N_{part}	Node Hour
L1	$(256 \text{ Mpc/h})^3$	128^3	~ 1.0
L2	$(100 \text{ Mpc/h})^3$	128^3	~ 1.7
HF	$(256 \text{ Mpc/h})^3$	512^3	~ 140
Test	$(256 \text{ Mpc/h})^3$	512^3	~ 140

We start the simulations at $z = 99$ and finish at $z = 0$. The initial linear power spectrum and transfer function are produced by CLASS [157] at $z = 99$ through the Zel’dovich approximation [158]. We assume periodic boundary conditions. We use a Fourier-transform-based particle-mesh method on large scales for the gravitational forces and a Barnes-Hut tree [110] on small scales. Table 5.1 summarizes the simulation volumes and particle loads used in this paper. We use the same set of low-fidelity (L1) and high-fidelity (HF) pairs as in [16], with an additional low-fidelity node (L2) to demonstrate the emulation using simulations from different box sizes. However, the framework presented in this paper is generalizable to more than two low-fidelity nodes. Fig 5.1 shows a visual illustration for the dark-matter only simulations used in this paper.

Our emulation target is the matter power spectrum, $P(k)$, a summary statistic of the over-density field. We measure the matter power spectrum with a cloud-in-cell mass assignment. We use the built-in power spectrum estimator from MP-Gadget; the power spectrum is thus generated on a mesh the same size as the simulation’s PM grid, which is 3 times the mean interparticle spacing. The multi-fidelity emulation framework we introduce

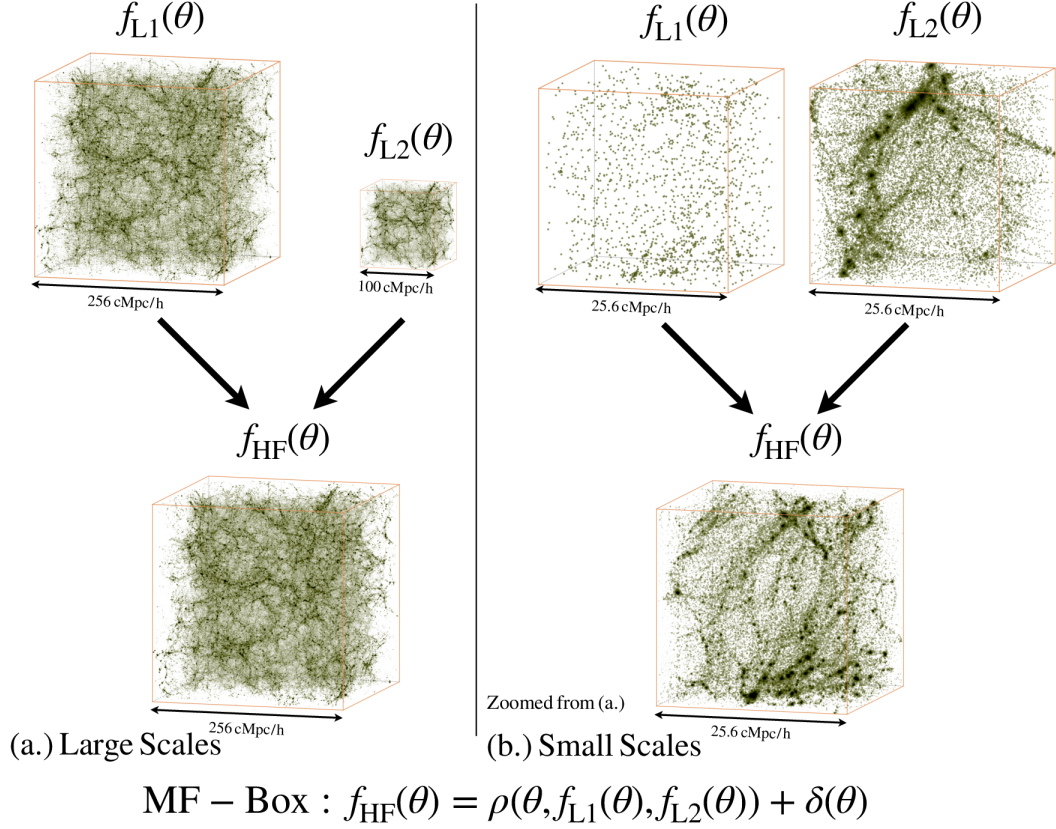


Figure 5.1: Illustration of the MF-Box framework and the dark-matter only simulations performed at $z = 0$. MF-Box provides an emulation framework to connect power spectra (denoted as $f(\theta)$, where θ is the input cosmology) from low-fidelity simulations (L1 and L2) to high-fidelity simulations (HF), providing an efficient emulation framework in predicting HF power spectra using only a few HF simulations augmented with many low-fidelity simulations with various volumes. ρ is a learnable multiplicative resolution correction parameter, and δ is a learnable additive resolution correction parameter. Details of the MF-Box model can be found in Section 5.4.2. The particle loads and box sizes for each simulation are listed in Table 5.1. **(a.)** Large-scale structures of each simulation are shown. Simulations L1 and L2 have the same particle load ($N_{\text{ptl,side}} = 128$), but L1 has a smaller box size (100 Mpc/h). As a result, the large scales of L1 resemble those of the high-fidelity (HF) simulation, while L2 lacks the necessary large-scale information to match HF. **(b.)** Zoomed-in view (25.6 Mpc/h) of the small scales from (a.). L1 lacks structures due to the sparsity of particles at this scale, whereas L2 captures more structures by utilizing a smaller box size. As a result, L1 resembles HF at small scales due to its finer mass resolution.

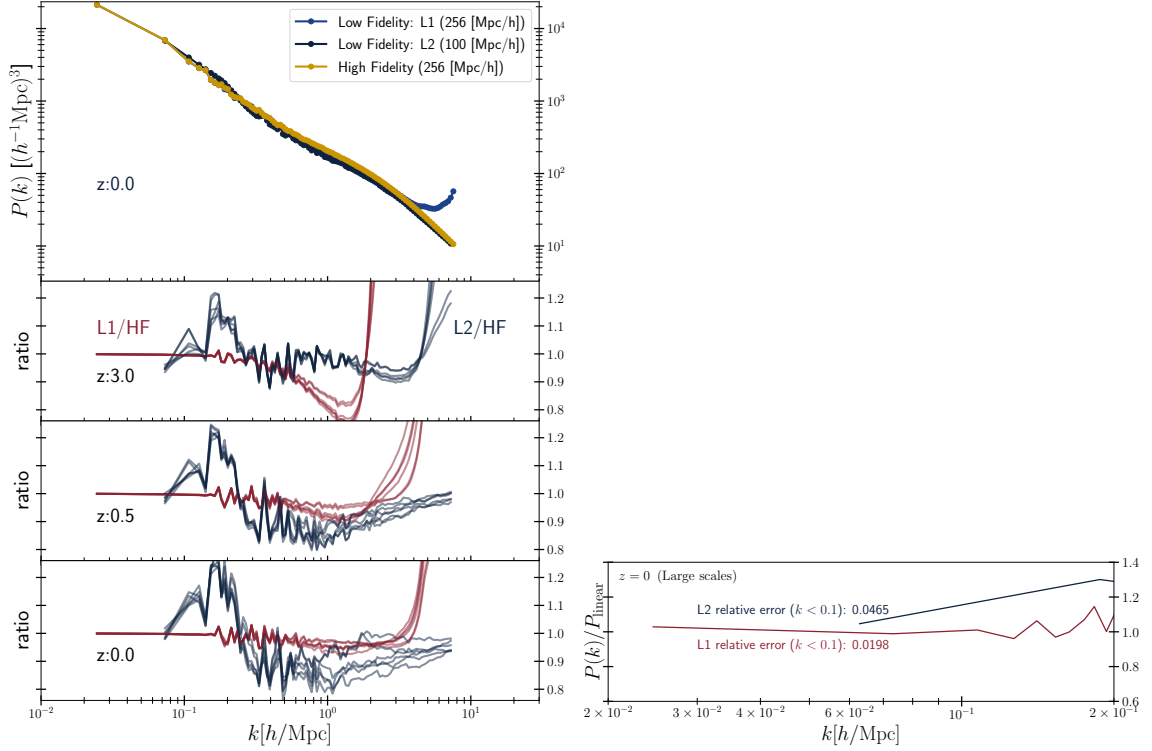


Figure 5.2: Matter power spectra from dark-matter only MP-Gadget simulations with various fidelities, conditioning on the same cosmology. The top panel shows the power spectra from a large-box low-fidelity (L1; blue), a small-box low-fidelity (L2; black), and a large-box high-fidelity simulations (HF; yellow). The numeric values for different fidelities of simulations are tabulated in Table 5.1. The 2nd, 3rd, and bottom panels show the ratios of L1/HF (red) and L2/HF (black) simulations, conditioned on different redshift bins, $z = 3.0, 0.5, 0$. (Bottom panel): We also show the ratio between (L1, L2) and the linear theory power spectrum from CLASS at large scales. The solid lines show the median and shaded areas show the 68% quantiles across 60 different cosmologies.

here is also applicable to other implementations of power spectrum calculations, such as those generated by NBodyKit [214].

Figure 5.2 shows an example of our emulation target: matter power spectra from different resolutions, where the low-fidelity simulations (L1 and L2) have two different box sizes. L1 simulations are in the same box size (256 Mpc/h) as high-fidelity simulations (HF) with the same initial condition seeding; whereas, L2 simulations have a smaller box size (100 Mpc/h) than L1 and HF. In principle, L2 can capture more small-scale structures due to its smaller box size. Indeed, as shown in the 2nd, 3rd, and bottom panels in Figure 5.2, L2 is more accurate than L1 at small scales. For example, at $z = 3$, L2/HF is closer to 1 than L1/HF at small scales ($k > 0.6 h\text{Mpc}^{-1}$).

Note that L2 is not necessarily better than L1 in matching the HF simulations. L1 matches the HF power spectrum extremely well at large scales, while L2 performs better at small scales. Therefore, the accuracy of the different simulations are not in a monotonically increasing sequence. Thus the [41] method we used in [16] cannot be directly applied to this example.



Figure 5.3: Experimental design of low- and high-fidelity simulations in this work. The prior volume is chosen to be the same as EuclidEmulator2 [11]. Crosses (black) are the input parameters for the low-fidelity simulations (both L1 and L2). Circles (red and yellow) are the parameters for high-fidelity simulations, which is a subset of the low-fidelity experimental design. We use max-min Sliced Latin Hypercube (SLHD) [12] for the LF design, containing 20 slices with 3 samples in each slice. Red and Yellow circles show two of the slices, which we select to be the input parameters for HF simulations.

Figure 5.3 shows our experimental design in the input parameter space, corresponding to the prior range of

$$\begin{aligned}
\Omega_0 &\sim \mathcal{U}(0.24, 0.4); \\
\Omega_b &\sim \mathcal{U}(0.04, 0.06); \\
h &\sim \mathcal{U}(0.61, 0.73); \\
A_s/10^{-9} &\sim \mathcal{U}(1.7, 2.5); \\
n_s &\sim \mathcal{U}(0.92, 1),
\end{aligned} \tag{5.1}$$

where Ω_0 is the total matter density parameter in the Universe, Ω_b is the total baryon density parameter, h is the dimensionless Hubble parameter, A_s is the spectral amplitude and n_s is the spectral index.

We generated 60 Latin hypercube samples using max-min Sliced Latin Hypercube [12], including 20 slices with 3 samples in each slice. We will discuss SLHD in Section 5.5.1. SLHD partitions the design into several equal slices (or blocks). Each slice itself is also a Latin hypercube design, as well as the whole design. We thus choose one of the Latin hypercube slices as our high-fidelity input. By using SLHD, we can avoid the design points of the HF node clustered in the corner of the prior volume. We ran L1 and L2 nodes using the same cosmological parameters (although this is not required by the GMGP from [197]).

We summarize the notation used in this paper in Table 5.2.

Table 5.2: Notations and definitions

Notation	Description
HF	High Fidelity
LF	Low Fidelity
θ	Input cosmological parameters
$f(\theta)$	Summary statistics (matter power spectrum in this work) corresponding to input parameters.
$N_{\text{pt,side}}$	Number of particles per box side
AR1	Autoregressive GP [41]
NARGP	Non-linear autoregressive GP [153]
GMGP	Graphical GP [197]
MFEulator	Multi-fidelity cosmological emulator [16]
MF-Box	Multi-fidelity cosmological emulator with different box sizes in low fidelity.

5.4 Emulation

Emulation predicts the output from expensive cosmological simulations. First, a handful of simulations are run at carefully chosen experimental design points as a training set. Next, a surrogate model (an emulator) fits the prepared training set to predict simulation output. The trained emulator will be a proxy for the simulation results, allowing for inexpensive evaluation of a likelihood function.

In Section 5.4.1, we will briefly review emulation using a Gaussian process. Section 5.4.2 will review how we can extend the Gaussian process emulator to model simulations from different qualities using a multi-fidelity emulator, `MFEulator`. Our earlier multi-fidelity technique based on the KO method [41] will be reviewed in Section 5.4.2. Section 5.4.2 will review an extension of the KO method based on a deep Gaussian process,

NARGP [153]. Section 5.4.2 describes a graphical-model Gaussian process model (GMGP) [197], an extension of NARGP to allow more than one node in the same fidelity.

5.4.1 Gaussian process emulator

A Gaussian process (GP) regression model [31] is widely used as a cosmological emulator. A GP provides closed-form expressions for predictions. In addition, a GP naturally comes with uncertainty quantification, which is handy for inference framework and Bayesian optimization. In emulation, a GP can be seen as a Bayesian prior for the simulation response. It is a prior because the emulator model is chosen to ensure smoothness in the simulation response *before* data are collected [30].

Let $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d$ be the input cosmologies for the simulator, and $f(\boldsymbol{\theta})$ be the corresponding output summary statistic. This work assumes that the summary statistic is the non-linear matter power spectrum. A GP regression model is a prior on the response surface of our simulated matter power spectrum:

$$p(f) = \mathcal{GP}(f; \mu, k), \quad (5.2)$$

where $\mu(\boldsymbol{\theta}) = \mathbb{E}[f(\boldsymbol{\theta})]$ is the mean function, and $k(\boldsymbol{\theta}, \boldsymbol{\theta}') = \text{Cov}[f(\boldsymbol{\theta}), f(\boldsymbol{\theta}')] is the covariance kernel function. The mean function is usually assumed to be a constant or zero mean unless there is prior knowledge about the mean function. In this work, we assume a zero mean function. The covariance kernel function is typically chosen as a squared exponential function (radial basis function, RBF) to return a smooth response surface.$

Suppose we run the simulations at n carefully chosen input cosmologies, $\mathcal{D} = \{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n\}$, and we compress each simulation into the corresponding matter power spec-

trum, $\mathbf{y} = \{f(\boldsymbol{\theta}_1), \dots, f(\boldsymbol{\theta}_n)\}$. Conditioning on this training data and optimizing the hyperparameters using maximum likelihood estimation, we can get the predictive distribution of f at a new input cosmology $\boldsymbol{\theta}_*$, $f_* = f(\boldsymbol{\theta}_*)$, through a closed-form expression

$$p(f_* | \mathbf{y}_*, \mathcal{D}, \boldsymbol{\theta}) = \mathcal{N}(f_* | \mu_*(\boldsymbol{\theta}_*), \sigma_*^2(\boldsymbol{\theta}_*)), \quad (5.3)$$

where the mean and variance are

$$\mu_*(\boldsymbol{\theta}_*) = \mathbf{k}(\boldsymbol{\theta}_*, \mathcal{D})^\top \mathbf{K}(\mathcal{D})^{-1} \mathbf{y}; \quad (5.4)$$

$$\sigma_*^2(\boldsymbol{\theta}_*) = k(\boldsymbol{\theta}_*, \boldsymbol{\theta}_*) - \mathbf{k}(\boldsymbol{\theta}_*, \mathcal{D})^\top \mathbf{K}(\mathcal{D})^{-1} \mathbf{k}(\boldsymbol{\theta}_*, \mathcal{D}).$$

The vector $\mathbf{k}(\boldsymbol{\theta}_*, \mathcal{D}) = [k(\boldsymbol{\theta}_*, \boldsymbol{\theta}_1), \dots, k(\boldsymbol{\theta}_*, \boldsymbol{\theta}_n)]$ represents the covariance between the new input cosmology, $\boldsymbol{\theta}$, and the training data. The matrix $\mathbf{K}(\mathcal{D})$ is the covariance of the training data.

Although we do not explicitly state this in the notation, we let $f(\boldsymbol{\theta})$ be a single-value output. If the target summary statistic is a vector, we let the Gaussian process model each bin separately. It will be more apparent why we make this modeling decision in later sections (Section 5.4.2). The primary reason is that the correlation between low-fidelity and high-fidelity summary statistics changes depending on the scales. The multi-fidelity method can only capture scale dependence if we model the scales separately.²

5.4.2 Multi-Fidelity Emulation

We briefly recap the multi-fidelity emulation framework we proposed in [16]. We will first review the Kennedy-O’Hagan model (autoregressive GP; AR1) [41] and NARGP

²An alternative way is to apply a co-kriging kernel to model the covariance for each vector element. We do not do that in this work because we found the single-output GP is enough for our cosmological emulation purpose, so there is no need to introduce another layer of complexity.

(non-linear autoregressive GP) [153] in Section 5.4.2 and Section 5.4.2, respectively. We do not change our AR1 and NARGP modeling presented in [16], except we simplified the notations to only two fidelities. Finally, we will introduce the GMGP model [197], combining simulations from different box sizes.

Kennedy O’Hagan Method

[41] proposed a linear autoregressive GP to model the response surfaces of a sequence of computer codes with increasing fidelity. For simplicity, we assume there are only two fidelities: dark-matter only simulations with fewer particles in low fidelity (LF) and with more particles in high fidelity (HF).

Let $\{\mathbf{y}_{\text{LF}}, \mathbf{y}_{\text{HF}}\}$ be the matter power spectrum in the training set, where $\mathbf{y}_{\text{LF}} = \{f_{\text{LF}}(\boldsymbol{\theta}_i^{\text{LF}})\}_{i=1}^{n_{\text{LF}}}$ and $\mathbf{y}_{\text{HF}} = \{f_{\text{HF}}(\boldsymbol{\theta}_i^{\text{HF}})\}_{i=1}^{n_{\text{HF}}}$. Here n_{LF} and n_{HF} are the number of simulations in the low and high fidelity. The KO method models the multi-fidelity emulator as:

$$f_{\text{HF}}(\boldsymbol{\theta}) = \rho \cdot f_{\text{LF}}(\boldsymbol{\theta}) + \delta(\boldsymbol{\theta}), \quad (5.5)$$

where ρ (the scale parameter) is a trainable parameter describing the amount of common behavior in low- and high-fidelity response surfaces. $\delta(\boldsymbol{\theta})$ is a GP that models the remaining bias, modeling the variability that cannot be captured by correlating LF to HF. In the context of the matter power spectrum, the $\rho \cdot f_{\text{LF}}(\boldsymbol{\theta})$ term dominates at the large scales describing the two-halo term while $\delta(\boldsymbol{\theta})$ dominates at the small scales describing the one-halo term.

We normalize the matter power spectra into a logarithmic scale. The sample mean is subtracted from the LF log power spectra to keep the output close to zero, while the HF

log power spectra are passed directly to the training:

$$\begin{aligned} \mathbf{y}_{\text{LF}} &\leftarrow \log \mathbf{y}_{\text{LF}} - \mathbb{E}[\log \mathbf{y}_{\text{LF}}]; \\ \mathbf{y}_{\text{HF}} &\leftarrow \log \mathbf{y}_{\text{HF}}. \end{aligned} \tag{5.6}$$

Not subtracting the mean spectrum of HF simulations is a compromise decision. Our benchmark multi-fidelity emulator uses only 3 HF samples, and the sample mean of 3 power spectra will often deviate substantially from the true mean spectrum. Instead, we entirely rely on the bias term, $\delta(\boldsymbol{\theta})$, to compensate for the deviation caused by not subtracting the mean.

As mentioned in [16], the ρ parameter has to be scale-dependent (as a function of k) to model the scale-dependent correlation between high- and low-fidelity. Here we use the same method as [16], where we assume Equation 5.5 is a single-output GP model and build a KO model for each k bin of the data. In this way, we can model ρ as a function of k .

We also assign different KO models to different redshifts. We note that it is possible to assume a smooth function to model $\rho(k, z)$, and we may examine this in future work.

Non-linear Autoregressive Gaussian Process (NARGP)

Another multi-fidelity method we used in [16] is the non-linear autoregressive GP, or NARGP, developed by [153]. NARGP is a modification of the KO method to allow non-linearity in the scale parameter, ρ , through a deep GP [162]. In cosmic emulators, it means that we allow ρ to vary as a function of cosmology.

Let $f_{\text{HF}}(\boldsymbol{\theta})$ be the high-fidelity and $f_{\text{LF}}(\boldsymbol{\theta})$ be the low-fidelity power spectra as functions of cosmology, $\boldsymbol{\theta}$. NARGP models the multi-fidelity problem as:

$$f_{\text{HF}}(\boldsymbol{\theta}) = \rho(\boldsymbol{\theta}, f_{\text{LF}}(\boldsymbol{\theta})) + \delta(\boldsymbol{\theta}), \quad (5.7)$$

Here, ρ is modeled as a GP and is a function of the cosmologies, $\boldsymbol{\theta}$, and the output from the previous fidelity, $f_{\text{LF}}(\boldsymbol{\theta})$. We follow the approximation made in [153] to simplify the computation of a deep GP to two separate GPs. The approximation is done by replacing the $f_{\text{LF}}(\boldsymbol{\theta})$ with its posterior, $f_{*,\text{LF}}(\boldsymbol{\theta})$. Eq 5.7 can thus be further reduced to a regular GP with a kernel function K :

$$f_{\text{HF}} \sim \mathcal{GP}(0, K) \quad (5.8)$$

with

$$K(\boldsymbol{\theta}, \boldsymbol{\theta}') = K_{\rho}(\boldsymbol{\theta}, \boldsymbol{\theta}') \cdot K_f(f_{*,\text{LF}}(\boldsymbol{\theta}), f'_{*,\text{LF}}(\boldsymbol{\theta}')) + K_{\delta}(\boldsymbol{\theta}, \boldsymbol{\theta}'). \quad (5.9)$$

We integrate the bias GP and the scale parameter GP here into one single GP with a composite kernel. Each kernel, $(K_{\rho}, K_f, K_{\delta})$, is a squared exponential kernel. K_{δ} models the bias term, and the scale parameter GP is factorized into the K_f , modeling the covariance between LF output posteriors. K_{ρ} models the cosmological dependence of ρ .

Graphical Multi-fidelity Gaussian Process (GMGP)

Here we briefly explain a new multi-fidelity model using a graphical model Gaussian process (GMGP), first introduced in [197]. A similar approach is the multi-information source method [145], which allows multiple low-fidelity nodes (information sources) to resolve a single high-fidelity truth. However, we find the model in [197] is methodologically

closer to what we applied before in [16], and so use this technique for our emulation problem for low-fidelity nodes with different box sizes.

The graphical GP model [197] utilizes a directed acyclic graph to model multi-fidelity data. Instead of assuming the fidelities of a simulation code form a monotonically increasing sequence in accuracy, a GMGP allows the fidelities to have a directed-in tree structure. [197] has a thorough mathematical description for applying GMGP in an arbitrarily directed in-tree structure. Thus each high fidelity node has more than one corresponding low fidelity node, a common situation as there are many ways to approximate high fidelity simulations.

We use the simplest case of the tree structure, illustrated in Fig 5.1, with two low fidelity nodes and one high fidelity node. In the case of N -body simulations, one may vary not only the number of particles, but also the box size of the simulation. Thus we can use a low-fidelity simulation with a smaller box size to improve emulation at the high-fidelity node. We will call this tree “MF-Box” throughout the rest of the paper. In the following text, we will assume L1 is the low-fidelity node that has 128^3 particles. L2 has the same number of particles as L1 but a smaller box size (100 Mpc/h), and HF is the high-fidelity node with 512^3 particles and the same box size as L1 (Table 5.1).

The deep GMGP model (dGMGP) we use from [197] is an extension of NARGP, where [197] implemented a specific kernel structure allowing low-fidelity information from multiple nodes to be passed to the HF node³. For the directed graph in Fig 5.1, the dGMGP

³Since we found NARGP outperformed AR1 in [16] for the matter power spectrum case, we will use dGMGP instead of the GMGP extended from the AR1 model.

model can be written as:

$$f_{\text{HF}}(\boldsymbol{\theta}) = \rho(\{f_t(\boldsymbol{\theta}) : t \in L1, L2\}, \boldsymbol{\theta}) + \delta(\boldsymbol{\theta}). \quad (5.10)$$

Here we pass the cosmologies $\boldsymbol{\theta}$ and the outputs from L1 and L2 to the ρ function. We make the same approximation as in Section 5.4.2, so we can train the deep GP recursively: We first train the low-fidelity emulators on L1 and L2, respectively. Then, we sample the output posteriors from the L1 and L2 emulators and use them as the training input for Eq 5.10.

Similar to NARGP, we use a composite kernel for the high-fidelity GP in the dGMGP:

$$K_{\text{dGMGP}}(\boldsymbol{\theta}, \boldsymbol{\theta}') = K_{\rho}(\boldsymbol{\theta}, \boldsymbol{\theta}') \cdot K_f(f_{*,\text{LF}}(\boldsymbol{\theta}), f_{*,\text{LF}}(\boldsymbol{\theta}')) + K_{\delta}(\boldsymbol{\theta}, \boldsymbol{\theta}'), \quad (5.11)$$

where the above expression is the same as Eq 5.9 except that K_f takes the outputs from both L1 and L2 emulators as inputs,

$$K_f(f_{*,\text{LF}}(\boldsymbol{\theta}), f_{*,\text{LF}}(\boldsymbol{\theta}')) = K_{\text{linear}}(f_{*,\text{LF}}(\boldsymbol{\theta}), f_{*,\text{LF}}(\boldsymbol{\theta}')) + K_{\text{rbf}}(f_{*,\text{L1}}(\boldsymbol{\theta}), f_{*,\text{L1}}(\boldsymbol{\theta}')) \cdot K_{\text{rbf}}(f_{*,\text{L2}}(\boldsymbol{\theta}), f_{*,\text{L2}}(\boldsymbol{\theta}')). \quad (5.12)$$

Here, K_{rbf} is a radial basis kernel, and K_{linear} is a linear kernel, which can be expressed more explicitly as

$$K_{\text{linear}}(f_{*,\text{LF}}, f'_{*,\text{LF}}) = \sigma_1^2 f_{*,\text{L1}} f'_{*,\text{L1}} + \sigma_2^2 f_{*,\text{L2}} f'_{*,\text{L2}},$$

where σ_1^2 and σ_2^2 are the hyperparameters of the linear kernel. A linear kernel in a Gaussian process is equivalent to a Bayesian linear regression.⁴ The multiplication in the kernel

⁴See the kernel cookbook: <https://www.cs.toronto.edu/~duvenaud/cookbook/>.

operation means an “AND” operation, showing high covariance only if both kernels have high values. The addition operator means an “OR” operation, indicating the final covariance is high if either of the kernels gives a high value. The intuition here is that the linear kernel encodes the linear regression part while the multiplication of RBF kernels encodes the non-linear transformation from L1 and L2 nodes to the HF node

5.5 Sampling strategy for high-fidelity simulations

This section describes the method used for selecting the input parameters for our high-fidelity training simulations. Following [197], we employ a Sliced Latin Hypercube Design (SLHD) [215, 12] to assign input parameters for the high-fidelity (HF) nodes. Each slice (or subset) in an SLHD is a Latin hypercube and thus can be served as the design points for the HF node. This approach offers a less computationally intensive and more straightforward implementation compared to the grid search method utilized in our previous work [16]. The details of SLHD will be discussed in Section 5.5.1, and our process for selecting the optimal HF design from the SLHD will be discussed in Section 5.5.2.

5.5.1 Sliced Latin hypercube design (SLHD)

Sliced Latin Hypercube Design (SLHD) is a type of Latin hypercube that can be partitioned by slices or blocks, each of which contains an equal number of design points. Each slice is itself a Latin hypercube. SLHD ensures the space-filling property both in the whole design and in each slice. Therefore, SLHD is an intuitive choice for a multi-fidelity problem.

Suppose we have an SLHD for the LF node. We can use one of the slices to generate simulations for the HF node, which ensures that both the LF and HF nodes are in Latin hypercubes. Another advantage of SLHD is that we can directly obtain a nested experimental design where the LF samples form a superset of the HF samples, i.e., $\theta_{\text{HF}} \subset \theta_{\text{LF}}$. As mentioned in [41], a nested design is an efficient training set for a multi-fidelity model because it allows us to obtain an accurate posterior $f_{\text{LF}}(\theta)$ at location θ without interpolating at the low fidelity.

SLHD, initially proposed by [215], is a technique developed for applying the Latin hypercube design to categorical variables. [12] later developed an efficient method for constructing optimal SLHD designs. The number of categories for categorical variables is usually fixed based on qualitative properties, making it challenging to apply a Latin hypercube design to such variables. However, SLHD addresses this challenge and enables the use of Latin hypercube designs with categorical variables. In SLHD, a Latin hypercube is divided into equal slices along the dimensions associated with categorical variables, while non-categorical dimensions are still sampled with ordinary Latin hypercube sampling. The usage of SLHD in the context of modeling the multi-fidelity problem was demonstrated in [197]. Furthermore, SLHD has also been employed in cosmology, specifically by the Dark Emulator [181].

For implementation, we use the maximin SLHD package, `maximinSLHD`,⁵ in R [12]. We set the number of design points to 3 for each slice and the number of slices to 20. In total, we have 60 design points. We assign the SLHD with 60 points to LF and select one

⁵<https://rdrr.io/cran/SLHD/man/maximinSLHD.html>

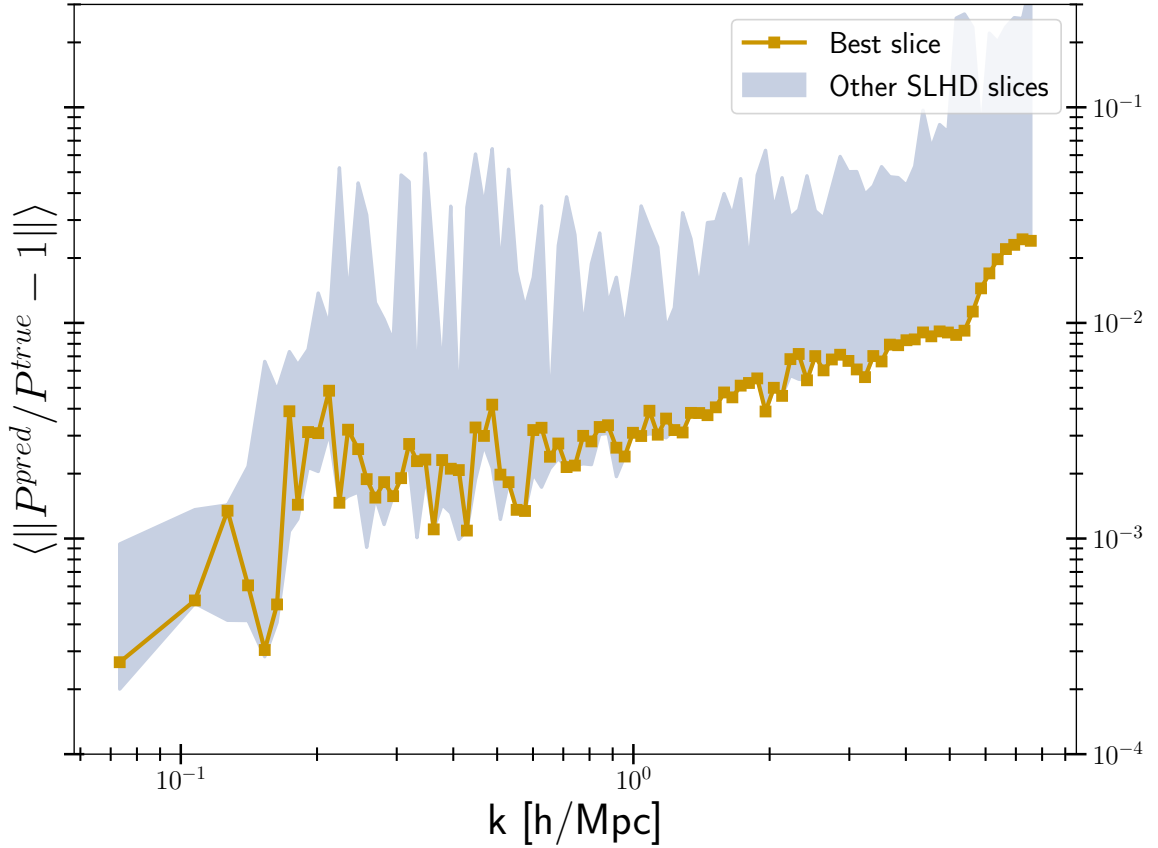


Figure 5.4: MF-Box’s emulation errors, averaged over redshift bins and test simulations, using 60 L1, 60 L2, and 3 HF (see Table 5.1). Here, we show the emulation minimum and maximum errors using different slices from SLHD (blue shaded area), and the best slice found by the grid search method is labeled as yellow.

slice as our HF design. We use 60 LF points in this work because we learned in [16] that ~ 50 simulations are enough for a 5 dimensional emulation problem.

5.5.2 Selecting the optimal slice

Slices in SLHD are Latin hypercubes in smaller sizes. In principle any slice should produce reasonably good emulation, as the points in each slices span parameter space.

However, in practice some slices still perform somewhat better than others, as shown in Figure 5.4. We use a procedure similar to our grid search approach in [16] to avoid choosing the worst slice. The procedure is described below:

1. Prepare SLHD for LF simulation suite.
2. Build low-fidelity only emulators (LFEmu) for each slice, compute the interpolation error for each LFEmu, testing solely on the LF simulation suite.
3. Select the slice which can best minimize the interpolation error.

Note that we do not use any HF simulations in the above procedure. The selection entirely relies on the LF simulation suite. The underlying assumption is that the interpolation error of the low-fidelity node is correlated with the interpolation error of the high-fidelity node. We labeled the selected slice in Figure 5.3. We will use the best slice as our HF training set for the results in Section 5.7.

To summarize, SLHD is a special kind of LHD, with each slice in the SLHD being a Latin hypercube as well as the whole design. We thus can assign HF nodes with a slice (or slices) of SLHD, making both LF and HF nodes Latin hypercubes. In the end, we describe a procedure to avoid choosing the worst slice for training a MFEmulator.

5.6 Computational budget estimation

In this section, we present our approach to quantifying the optimal allocation of simulation budgets across different fidelities. Building upon the error bounds established in [197], we have made modifications to adapt them to our specific context, as described in

Section 5.6.1. We approximate the emulation errors of our MF-Box using the form of [197] and empirically infer the error function of the emulator for various training designs, denoted as (n_{L1}, n_{L2}, n_{HF}) . Our objective is to utilize this empirical error function to determine the most cost-effective strategy for assigning low- and high-fidelity simulations in order to achieve optimal accuracy.

In Section 5.6.1, we present an approximate error function for our MF-Box emulator in predicting high-fidelity simulation outputs. Next, in Section 5.6.2, we show the analysis for assigning optimal computational budgets to low- and high-fidelity simulations, under the assumption that the emulator error follows the approximate error function. In Section 5.6.3, we empirically estimate the approximate error function of the MF-Box by analyzing the average emulator errors obtained from 144 distinct MF-Box training results. Finally, we determine the optimal number of low- and high-fidelity simulations required for achieving accurate power spectra emulation using the MF-Box approach.

5.6.1 Error bounds for Gaussian process emulators

[197] presents an error bound for a multi-fidelity emulator, and for the case of two low-fidelity nodes, the form is given by $\sim \mathcal{O}(\rho_{L1} \cdot n_{L1}^{-\frac{\nu_{L1}}{d}} + \rho_{L2} \cdot n_{L2}^{-\frac{\nu_{L2}}{d}} + n_{HF}^{-\frac{\nu_{HF}}{d}})$, where (ρ_{L1}, ρ_{L2}) are the scale parameters for the L1 and L2 nodes, respectively. $(\nu_{L1}, \nu_{L2}, \nu_{HF})$ are positive spectral indices, and (n_{L1}, n_{L2}, n_{HF}) represent the number of training simulations at the L1, L2, and HF nodes, respectively. While this bound does not directly apply to our case, we utilize the form of the bound as an approximate model for the MF-Box error and empirically determine the parameters by fitting them to the MF-Box emulation results using different multi-fidelity designs, i.e., varying combinations of (n_{L1}, n_{L2}, n_{HF}) .

The equation below represents the error function of the MF-Box emulator we want to infer. Note that our discussion primarily focuses on the emulation error when predicting “high-fidelity” power spectra. This emphasis aligns with the core objective of MF-Box, which is to correct the resolution of low-fidelity simulations for accurate predictions of their high-fidelity counterparts.

$$\begin{aligned}
\Phi(n_{L1}, n_{L2}, n_{HF}) &= \frac{1}{N} \sum_{i=1}^N \left| \frac{f_{HF}(\boldsymbol{\theta}_i) - m_{f_{HF}}(\boldsymbol{\theta}_i)}{f_{HF}(\boldsymbol{\theta}_i)} \right| \\
&\approx \tilde{\Phi}(n_{L1}, n_{L2}, n_{HF}) \\
&= \eta \cdot (\rho_{L1} \cdot n_{L1}^{-\frac{\nu_{L1}}{d}} + \rho_{L2} \cdot n_{L2}^{-\frac{\nu_{L2}}{d}} + n_{HF}^{-\frac{\nu_{HF}}{d}}),
\end{aligned} \tag{5.13}$$

where $N = 10$ test simulations in a Latin hypercube are used to average the emulation relative error. The emulator error function $\Phi(n_{L1}, n_{L2}, n_{HF})$ represents the average relative error of the MF-Box as a function of the number of simulations in L1, L2, and HF nodes. To estimate this error function, we have already averaged the emulation error across k bins, enabling us to obtain an approximation of the error as a function of the design points (n_{L1}, n_{L2}, n_{HF}) . Then, we infer the parameters of this error function from the MF-Box emulation results, as denoted by the \approx sign in Eq 5.13. The normalization factor of the functional form in Eq 5.13 is determined by the free parameter η .

An important term in Eq 5.13 is the one describing how the error scales with an increasing number of simulations, $n_t^{-\frac{1}{d}}$, where $t \in L1, L2, HF$. This scaling term comes from the fact that the fill distance is proportional to $\mathcal{O}(n_t^{-\frac{1}{d}})$, where d is the number of dimensions in a space-filling design [211].

To determine the parameters of $\tilde{\Phi}(n_{L1}, n_{L2}, n_{HF})$, we employ Markov Chain Monte Carlo (MCMC) inference based on 144 distinct MF-Box emulators that were trained with

varying numbers of (n_{L1}, n_{L2}, n_{HF}) . Specifically, we generated MF-Box emulators using [12, 18, 24, \dots , 60] L1/L2 points and [2, 3, \dots , 18] HF points, resulting in a total of 144 emulators. For simplicity, we only considered cases where the number of simulations in L1 and L2 nodes was equal, i.e., $n_{L1} = n_{L2}$, as the costs of L1 and L2 nodes are similar, therefore, choosing between them is not important. To simplify the notation, we employ n_{LF} to represent the number of training points in both the L1 and L2 nodes. Figure 5.5 presents the average relative errors, $\Phi(n_{L1}, n_{L2}, n_{HF})$, for all 144 designs under consideration.

For each pixel in Figure 5.5, we compute the average emulator relative error across 10 test simulations and multiple k bins across a redshift range, $z \in [0, 0.2, 0.5, 1, 2, 3]$. To solve the parameter estimation problem, we employ Markov Chain Monte Carlo (MCMC) inference with a Gaussian likelihood,⁶

$$\begin{aligned} & \tilde{\Phi}(n_{L1}, n_{L2}, n_{HF}) \\ &= \eta \cdot (\rho_{L1} \cdot n_{L1}^{-\frac{\nu_{L1}}{d}} + \rho_{L2} \cdot n_{L2}^{-\frac{\nu_{L2}}{d}} + n_{HF}^{-\frac{\nu_{HF}}{d}}) \\ &\sim \mathcal{N}(\mu = \Phi(n_{L1}, n_{L2}, n_{HF}), \sigma^2 = \Phi_{\text{var}}(n_{L1}, n_{L2}, n_{HF})). \end{aligned} \tag{5.14}$$

Here, $\Phi(n_{L1}, n_{L2}, n_{HF})$ represents the average relative errors, while $\Phi_{\text{var}}(n_{L1}, n_{L2}, n_{HF})$ denotes the variance of the relative errors across 10 test simulations.

The results of our MCMC analysis, including the priors and posteriors, are summarized in Table 5.3. The posteriors show that $\nu_{L1} \simeq \nu_{L2}$ and $\rho_{L1} \simeq \rho_{L2}$, indicating that both L1 and L2 nodes contribute to improving the accuracy of the emulator in a similar manner. In contrast, the power-law index ν_{HF} for the HF node is approximately twice as large as ν_{L1} and ν_{L2} , suggesting that the HF node has a more pronounced impact on

⁶We use the PyMC package version 4 [216] for the MCMC inference.

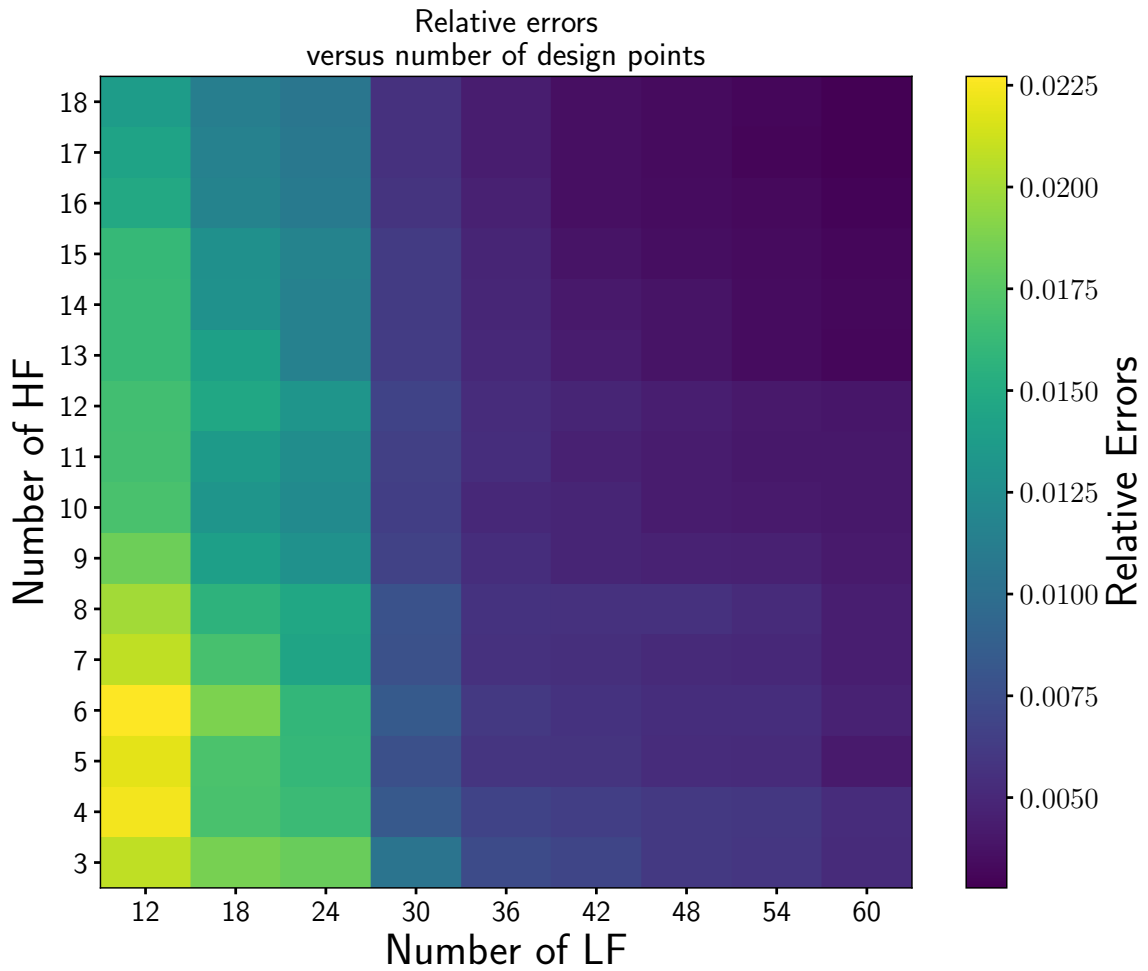


Figure 5.5: Relative errors plotted against the number of LF and HF design points in a MF-Box emulator. Here, LF refers to the combined number of L1 and L2 points, where $LF = n_{L1} = n_{L2}$. The plot reveals a trend of decreasing errors as the number of low-fidelity training simulations increases. However, due to the limited number of high-fidelity points compared to LF points, the decreasing trend is relatively modest.

Table 5.3: MCMC analysis of Eq 5.13: $\frac{1}{N} \sum_{i=1}^N \left| \frac{f_{\text{HF}}(\boldsymbol{\theta}_i) - m_{f_{\text{HF}}}(\boldsymbol{\theta}_i)}{f_{\text{HF}}(\boldsymbol{\theta}_i)} \right| = \Phi(n_{\text{L1}}, n_{\text{L2}}, n_{\text{HF}}) \approx \eta \cdot (\rho_{\text{L1}} \cdot n_{\text{L1}}^{-\frac{\nu_{\text{L1}}}{d}} + \rho_{\text{L2}} \cdot n_{\text{L2}}^{-\frac{\nu_{\text{L2}}}{d}} + n_{\text{HF}}^{-\frac{\nu_{\text{HF}}}{d}})$. The notation $\{\Phi(n_{\text{L1},j}, n_{\text{L2},j}, n_{\text{HF},j})\}_{j=1}^{144}$ means all 144 MF-Box emulator errors used for parameter estimation. The column “Posterior (50%)” reports the medians of the posteriors of the parameters, and “Posterior (25%, 75%)” reports the 25% and 75% quantities of the posterior distributions.

Parameters	Prior	Posterior (50%)	Posterior (25%, 75%)
η	$\text{Normal}(\mu = \text{Mean}(\{\Phi_j\}_{j=1}^{144}), \sigma^2 = \text{Var}(\{\Phi_j\}_{j=1}^{144}))$	0.0308	(0.0290, 0.0327)
ν_{HF}	$\text{LogNormal}(\mu = 0, \sigma = 1)$	9.80	(9.44, 10.2)
ν_{L1}	$\text{LogNormal}(\mu = 0, \sigma = 1)$	5.49	(5.33, 5.67)
ν_{L2}	$\text{LogNormal}(\mu = 0, \sigma = 1)$	5.49	(5.33, 5.67)
ρ_{L1}	$\text{Normal}(\mu = 1, \sigma = 1)$	4.53	(3.97, 5.08)
ρ_{L2}	$\text{Normal}(\mu = 1, \sigma = 1)$	4.54	(3.97, 5.10)

enhancing the emulator’s accuracy compared to the LF nodes. Table 5.3 shows that the parameters in Eq 5.13 are reasonably well-defined. Thus, we will use the median of the posterior as point estimates for the error function for the remainder of this paper.

5.6.2 Optimal number of simulations per node

Eq 5.13 models the emulation error, $\Phi(n_{L1}, n_{L2}, n_{HF})$, which behaves as a combination of power-law functions of the number of simulations in each node, namely LF or HF. The primary goal of an emulator is to better represent the original simulator by minimizing the prediction error, subject to a limited computational budget, denoted by C . By using $\Phi(n_{L1}, n_{L2}, n_{HF})$, we can determine the optimal number of simulations per node, given the computational budget available for running each node.

Consider a two-fidelity emulator consisting of two low-fidelity nodes, L1 and L2, where $\rho_{L1,L2}$ are the scale parameters and (n_{L1}, n_{L2}, n_{HF}) represent the number of simulations in L1, L2, and HF nodes, respectively. Our goal is to minimize the emulation error while subject to a limited budget.

$$n_{L1} \cdot C_{L1} + n_{L2} \cdot C_{L2} + n_{HF} \cdot C_{HF} \leq C, \quad (5.15)$$

where we know the ratios between the costs of HF and LF nodes (L1 and L2) are $\frac{C_{HF}}{C_{L1}} \simeq 140$ and $\frac{C_{HF}}{C_{L2}} \simeq 140/1.7$, from Table 5.1.

The Lagrangian for optimizing the error subjecting to the cost is:

$$\begin{aligned} \mathcal{L}(n_{L1}, n_{L2}, n_{HF}, \lambda) = & \eta(\rho_{L1} \cdot n_{L1}^{-\frac{\nu_{L1}}{d}} + \rho_{L2} \cdot n_{L2}^{-\frac{\nu_{L2}}{d}} + n_{HF}^{-\frac{\nu_{HF}}{d}}) \\ & + \lambda(n_{L1} \cdot C_{L1} + n_{L2} \cdot C_{L2} + n_{HF} \cdot C_{HF} - C), \end{aligned} \quad (5.16)$$

Here, λ is the Lagrange multiplier. To find the optimal number of (n_{L1}, n_{L2}, n_{HF}) minimizing the emulation error, we use the 1st order derivative conditions of the Lagrangian,

$$\begin{aligned}\frac{\partial \mathcal{L}(n_{L1}, n_{L2}, n_{HF}, \lambda)}{\partial n_{L1}} &= 0; \\ \frac{\partial \mathcal{L}(n_{L1}, n_{L2}, n_{HF}, \lambda)}{\partial n_{L2}} &= 0; \\ \frac{\partial \mathcal{L}(n_{L1}, n_{L2}, n_{HF}, \lambda)}{\partial n_{HF}} &= 0,\end{aligned}\tag{5.17}$$

resulting in

$$\begin{aligned}\eta \frac{\nu_{L1}}{d} \rho_{L1} \cdot n_{L1}^{-\frac{\nu_{L1}+d}{d}} &= \lambda C_{L1} \Rightarrow n_{L1} \propto \left(\frac{\nu_{L1} \rho_{L1}}{C_{L1}}\right)^{\frac{d}{\nu_{L1}+d}} \\ \eta \frac{\nu_{L2}}{d} \rho_{L2} \cdot n_{L2}^{-\frac{\nu_{L2}+d}{d}} &= \lambda C_{L2} \Rightarrow n_{L2} \propto \left(\frac{\nu_{L2} \rho_{L2}}{C_{L2}}\right)^{\frac{d}{\nu_{L2}+d}} \\ \eta \frac{\nu_{HF}}{d} n_{HF}^{-\frac{\nu_{HF}+d}{d}} &= \lambda C_{HF} \Rightarrow n_{HF} \propto \left(\frac{\nu_{HF}}{C_{HF}}\right)^{\frac{d}{\nu_{HF}+d}}.\end{aligned}\tag{5.18}$$

Here, the intuition is relatively straightforward: the number of simulations required is inversely proportional to the cost of each simulation at a given fidelity. However, if we observe a strong correlation between fidelities (i.e., if $\rho_{L1, L2}$ is large), then we should use more low-fidelity simulations because they are less expensive.

To ensure that Eq 5.18 identifies local minima instead of maxima, we can verify the positivity of the second-order derivatives of the Lagrangian.

$$\begin{aligned}\frac{\partial^2 \mathcal{L}(n_{L1}, n_{L2}, n_{HF}, \lambda)}{\partial n_{L1}^2} &= \eta \rho_{L1} \frac{\nu_{L1}(\nu_{L1} + d)}{d^2} n_{L1}^{-\frac{\nu_{L1}+2d}{d}} > 0; \\ \frac{\partial^2 \mathcal{L}(n_{L1}, n_{L2}, n_{HF}, \lambda)}{\partial n_{L2}^2} &= \eta \rho_{L2} \frac{\nu_{L2}(\nu_{L2} + d)}{d^2} n_{L2}^{-\frac{\nu_{L2}+2d}{d}} > 0; \\ \frac{\partial^2 \mathcal{L}(n_{L1}, n_{L2}, n_{HF}, \lambda)}{\partial n_{HF}^2} &= \eta \frac{\nu_{HF}(\nu_{HF} + d)}{d^2} n_{HF}^{-\frac{\nu_{HF}+2d}{d}} > 0.\end{aligned}\tag{5.19}$$

The parameters $(\nu_{L1}, \nu_{L2}, \nu_{HF})$, $(\rho_{L1}, \rho_{L2}, \rho_{HF})$, and η are all positive, while the dimension of the input space, d , must be a positive integer. Similarly, the number of simulations (n_{L1}, n_{L2}, n_{HF}) must be positive integers as well. Therefore, all second-order derivatives are positive, indicating that Eq 5.18 minimizes the emulation error.

In the special case where $\nu \equiv \nu_{\text{LF}} = \nu_{\text{HF}}$, Eq 5.18 simplifies to the optimal budget identified in [197]:

$$\frac{n_{\text{LF}}}{n_{\text{HF}}} = \left(\frac{\rho_{\text{LF}} C_{\text{HF}}}{C_{\text{LF}}} \right)^{\frac{d}{\nu+d}}, \quad (5.20)$$

where the ratio of LF/HF training sample sizes is inversely proportional to the cost of each simulation per run and directly proportional to the correlation with the high-fidelity node.

5.6.3 Empirical estimate of the error function

In this section, we present the predicted errors of MF-Box obtained from our MCMC analysis. We explore the impact of different MF-Box designs on error predictions. Finally, we discuss the choices of the optimal number of simulations for MF-Box based on the analysis presented in Section 5.6.2.

We illustrate the predicted emulation errors in Fig 5.6, categorized by MF-Box models with varying LF and HF points. The predictions align with the overall trend of the data, except when n_{LF} is low, where the limited availability of LF training points leads to suboptimal training performance.

Fig 5.7 and Fig 5.8 depict the predicted relative errors as a function of LF and HF points, respectively. Both figures exhibit a power-law trend characterized by a negative spectral index, indicating that the error decreases as the number of training points increases. For example, in Fig 5.7, the $X_{\text{LR-3HR}}$ emulator emulators ($X \in \{12, 18, 24, 30, 36, 42, 48, 54, 60\}$) follow this trend concerning the number of LF points, suggesting that achieving further accuracy improvements becomes challenging once a sufficient number of LF points are used.

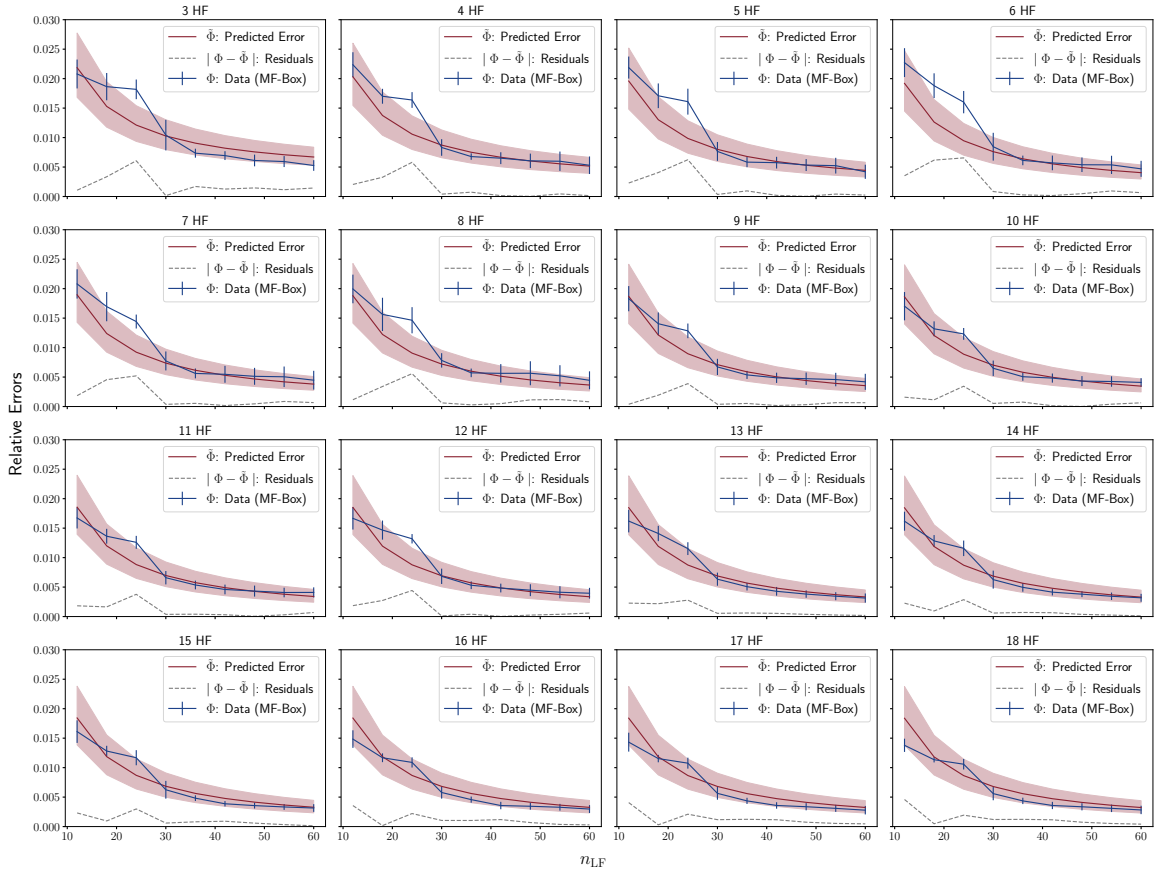


Figure 5.6: Inferred relative errors for all available MF-Box emulators are displayed. Each subplot corresponds to a fixed number of HF points (as indicated in the title) with varying LF points (on the x-axis). The red curves represent the median predictions (50% posterior). Blue lines indicate the average relative errors obtained from the MF-Box emulators, while the error bars represent the standard deviation of relative errors across 10 simulations in the test set. The shaded area depicts the 25% and 75% confidence interval of the predictions based on the inference results. Overall, the relative errors demonstrate a decreasing trend as the number of LF and HF points increases.

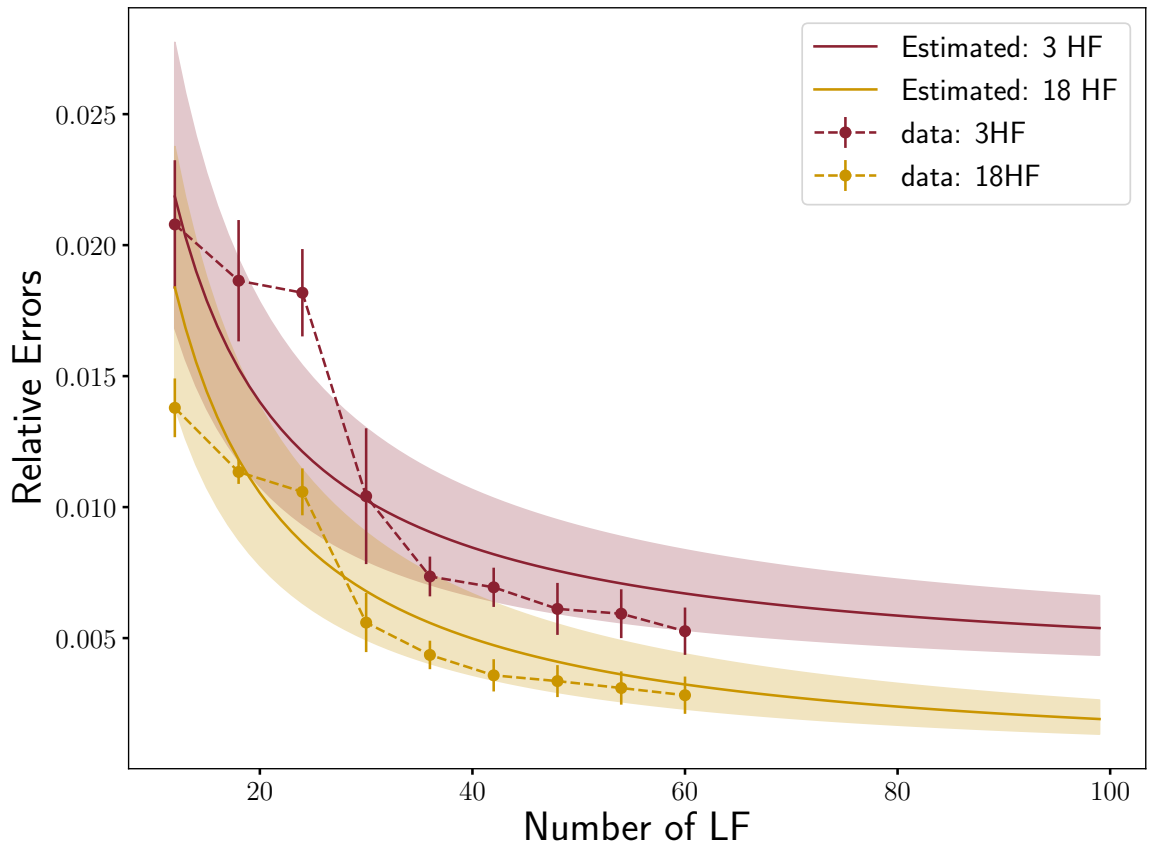


Figure 5.7: Inferred relative errors as a function of LF points. Shaded area shows the 25% and 75% confidence interval of the prediction from the inference result.

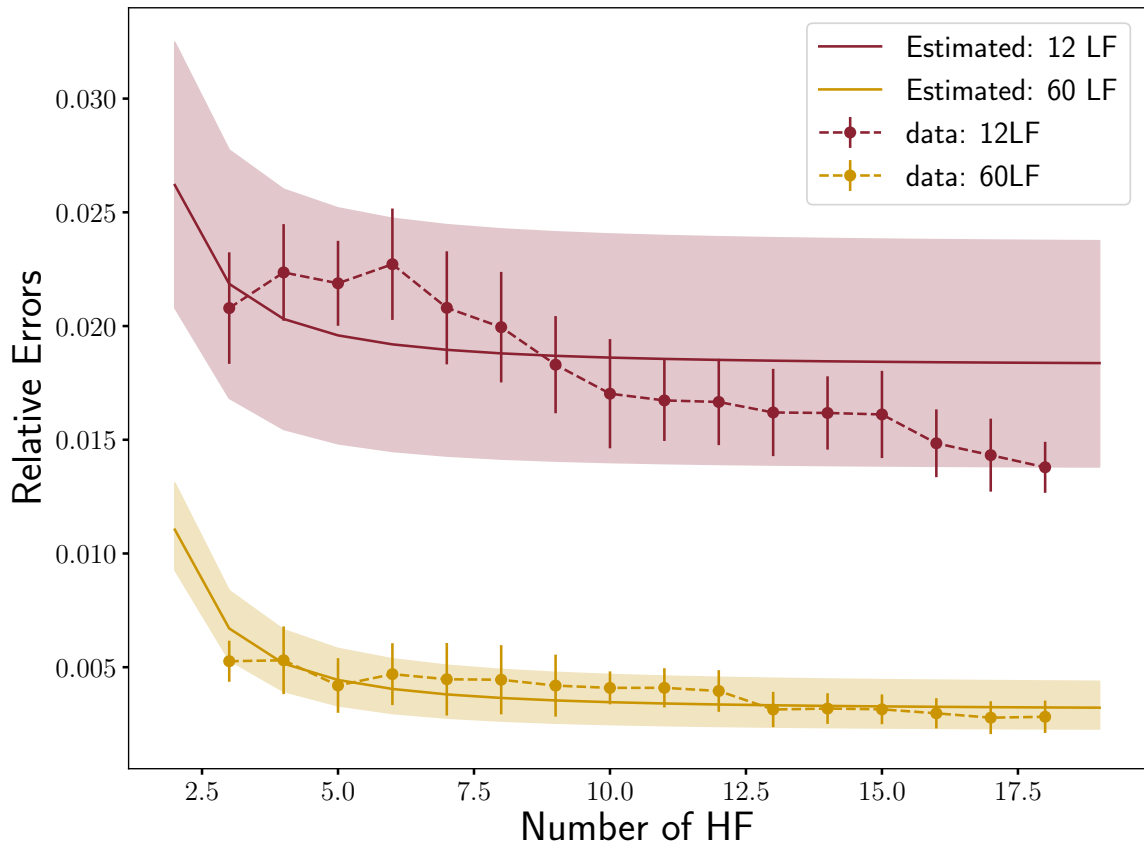


Figure 5.8: Inferred relative errors as a function of HF points. Shaded area shows the 25% and 75% confidence interval of the prediction from the inference result.

How much the error can be reduced by increasing the number of LF points is also influenced by the correlation between LF and HF simulations, which is controlled by the ρ parameter. A higher value of ρ indicates that LF points can more effectively reduce the error.

On the other hand, incorporating additional HF points can also enhance accuracy. In Fig 5.7, increasing the number of points in the HF node from 3 to 18 shifts the power-law function towards lower values, which itself follows the trend in Fig 5.8. Similarly, as more HF points are included in the training, achieving further emulation accuracy becomes more challenging.

Fig 5.9 displays the predicted error functions $\Phi(n_{L1}, n_{L2}, n_{HF})$ for different MF-Box emulator designs. We compile these predictions to create a plot of emulator error versus budget size. The bottom left region of the plot represents the most economical budget setup, where the error is minimized relative to the allocated budget.

Based on the predictions in Figure 5.9, we can determine the optimal number of simulations (n_{L1}, n_{L2}, n_{HF}) for achieving a desired level of average accuracy. For instance, if we aim for at least 1% average error, the optimal choice is $(n_{L1} = 30, n_{L2} = 30, n_{HF} = 3)$, which corresponds to a cost of approximately 500 L1 simulations. Note that a minimum of 3 HF simulations (~ 420 L1 simulations) is required to train a MF-Box in our power spectrum emulation problem. Similarly, if we aim for at least 0.5% average error, the optimal setup becomes $(n_{L1} = 60, n_{L2} = 60, n_{HF} = 4)$. However, a slightly higher cost is required for the setup with $(n_{L1} = 50, n_{L2} = 50, n_{HF} = 5)$, which yields a similar error.

In Figure 5.9, the purple dashed curve represents the predicted error of 60 LR-[2-10] HR emulators, illustrating the trend of increasing the number of HF points while keeping

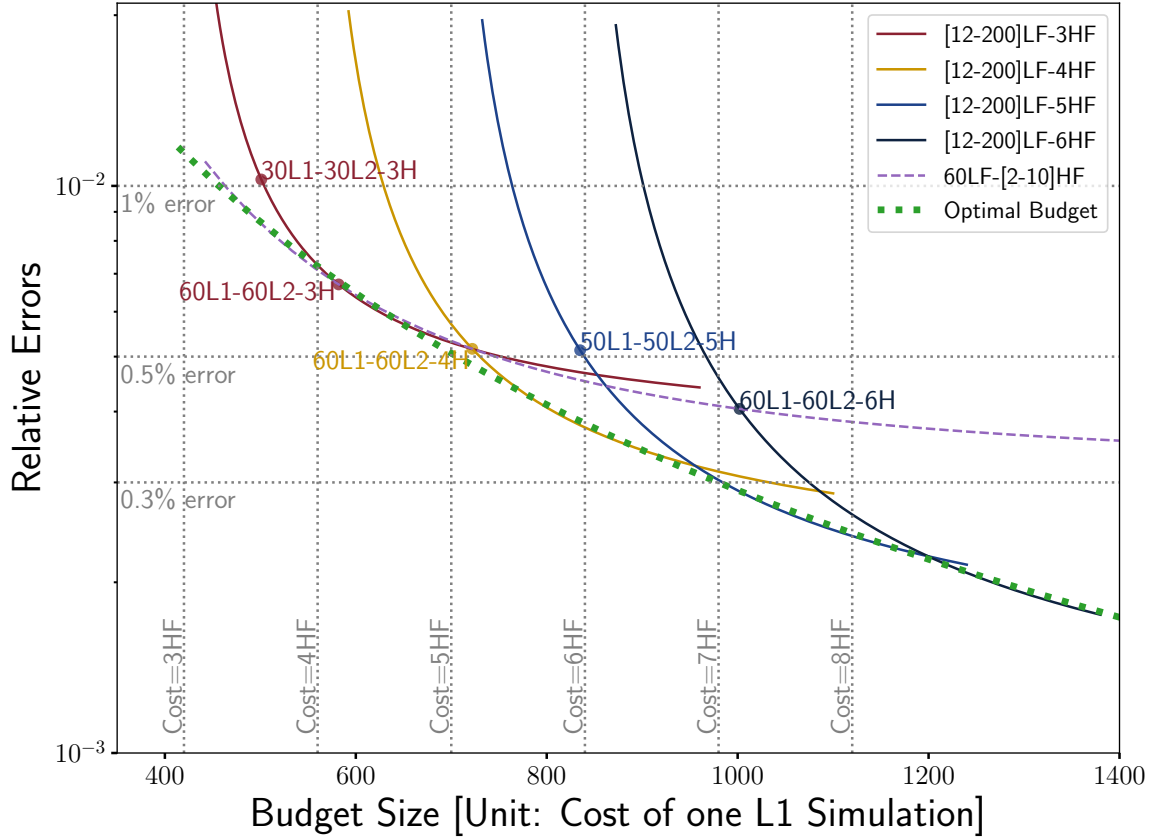


Figure 5.9: The predicted emulator errors as a function of the budget size, in the unit of the number of LF simulations. The predictions are based on the medians of the parameter posteriors presented in Table 5.3. The plot shows the predicted error functions using different combinations of LF and HF nodes. The red, yellow, blue, and black curves represent the predicted error functions with varying LF nodes and a fixed HF node ($n_{\text{HF}} = 3, 4, 5, 6$). In contrast, the purple dashed curve represents the predicted error function with varying HF nodes and a fixed LF node ($n_{\text{LF}} = 60$). The green dotted line illustrates the error function corresponding to the optimal budget (Eq 5.21). The vertical gray dotted lines indicate the budget size in terms of the number of HF simulations. The horizontal gray dotted lines denote the predicted errors at the levels of (1%, 0.5%, 0.3%).

a fixed number of 60 LF nodes. At the point of (60 LF, 3 HF), the error decrease exhibits a similar gradient to [12-200] LR-3 HR emulator, but it shows a steeper gradient after 4 HF points. This result suggests that adding more LF or HF nodes does not necessarily lead to superior performance compared to each other.

Under the assumptions outlined in Section 5.6.2, we can determine an optimal number of simulations (n_{L1}, n_{L2}, n_{HF}) for a MF-Box to achieve the best emulation accuracy within a given computational budget. The optimal ratio between the number of HF and LF simulations can be expressed as:

$$n_{LF}^{-\frac{\nu_{LF}+d}{d}} = n_{HF}^{-\frac{\nu_{HF}+d}{d}} \frac{C_{LF}}{C_{HF}} \frac{\nu_{HF}}{\rho_{LF} \nu_{LF}}. \quad (5.21)$$

Here, LF is either L1 or L2. C_{LF} and C_{HF} represent the computational cost of one simulation in the LF and HF, respectively.

In Figure 5.9, the green dotted line represents the optimal budget according to Eq 5.21. When $n_{HF} = 2.5$, the optimal number of low-fidelity simulations is $(n_{L1}, n_{L2}) = (80, 60)$, which is close to our initial setup of MF-Box with $(n_{L1} = 60, n_{L2} = 60, n_{HF} = 3)$. Moreover, the design of $(n_{L1} = 60, n_{L2} = 60, n_{HF} = 4)$ is also nearly optimal (close to the green dotted line), as demonstrated in Figure 5.9.

In summary, this section introduces an approach to model the average emulation error of MF-Box as a function of LF and HF points using an approximate error model based on power-law functions. Through empirical analysis of 144 MF-Box designs with various configurations, we have inferred this error model. We demonstrate that this empirical model can guide the selection of an optimal design within a given computational budget, facilitating the construction of accurate emulators in a resource-efficient manner.

5.7 Results

This section will demonstrate the emulation accuracy achieved by incorporating simulations with different box sizes through **MF-Box** for correcting the resolution of low-fidelity emulators to predict high-fidelity counterparts. The emulation error in this section is computed using a hold-out test set comprising 10 high-fidelity (HF) simulations, carefully selected from a separate Latin hypercube that was not part of the training set. Here, we will use **MF-Box** to denote the emulators using the GMGP model [197] with the graph structure in Figure 5.1. Section 5.7.1 will show how **MF-Box**'s accuracy improves by adding an L2 node in 100 Mpc/h. Section 5.7.2 will show how **MF-Box**'s accuracy changed as a function of L2 box size, from 100 Mpc/h to 256 Mpc/h. Finally, Section 5.7.3 show the runtime comparison between single-fidelity emulators, **MFEimulator** (including AR1, NARGP) and **MF-Box**.

5.7.1 **MF-Box** accuracy (256 + 100 Mpc/h)

This section shows how the emulation error changed when a suite of small-box simulations is included as a second LF node, L2, through **MF-Box**. More precisely, we use two LF nodes:

- L1: 128^3 simulations with 256 Mpc/h;
- L2: 128^3 simulations with 100 Mpc/h.

The information about the training simulations is summarized in Table 5.1.

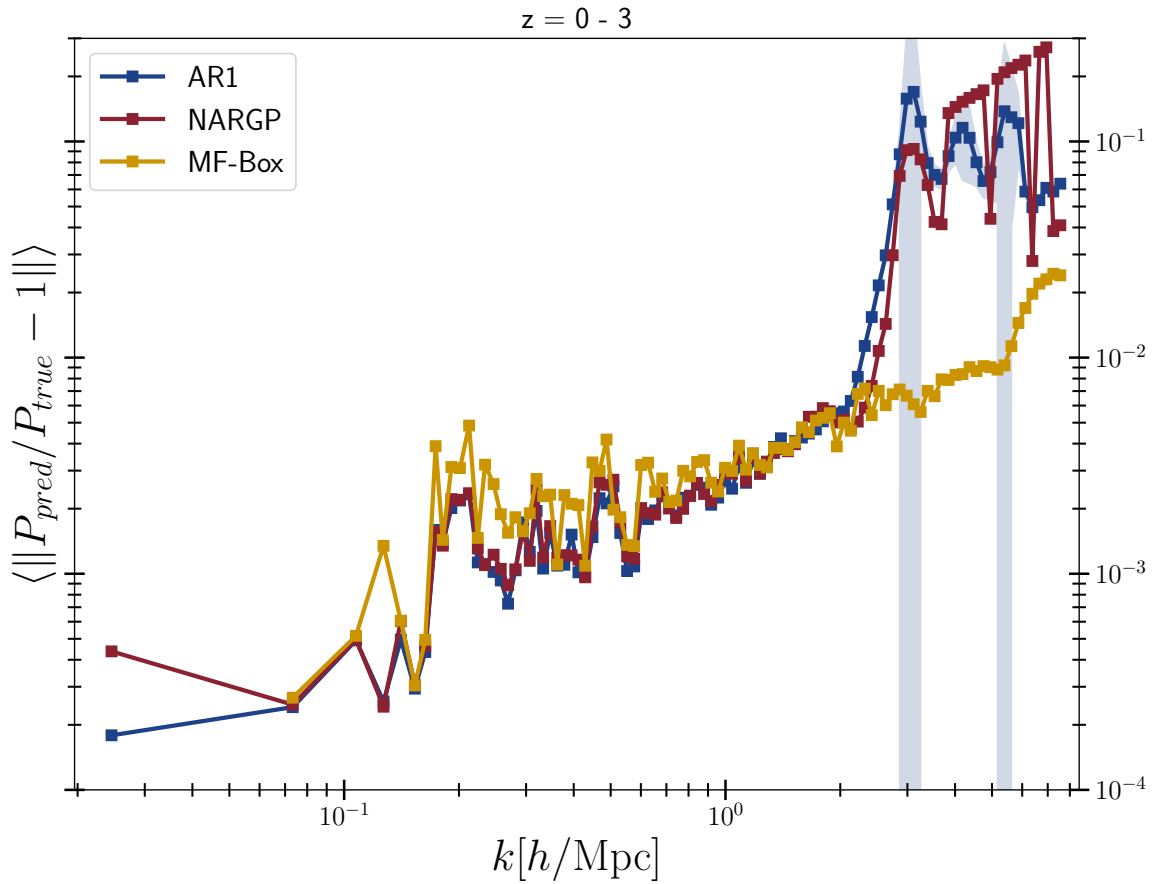


Figure 5.10: Relative errors averaged over $z = [0, 0.2, 0.5, 1, 2, 3]$ for different multi-fidelity models, AR1 (blue), NARGP (red), and MF-Box (yellow). The MF-Box model uses 60 L1 (256 Mpc/h), 60 L2 (100 Mpc/h), and 3 H (256 Mpc/h) simulations for training. Both AR1 and NARGP use 60 L1 and 3 HF for training. The shaded area is the variance among different test simulations.

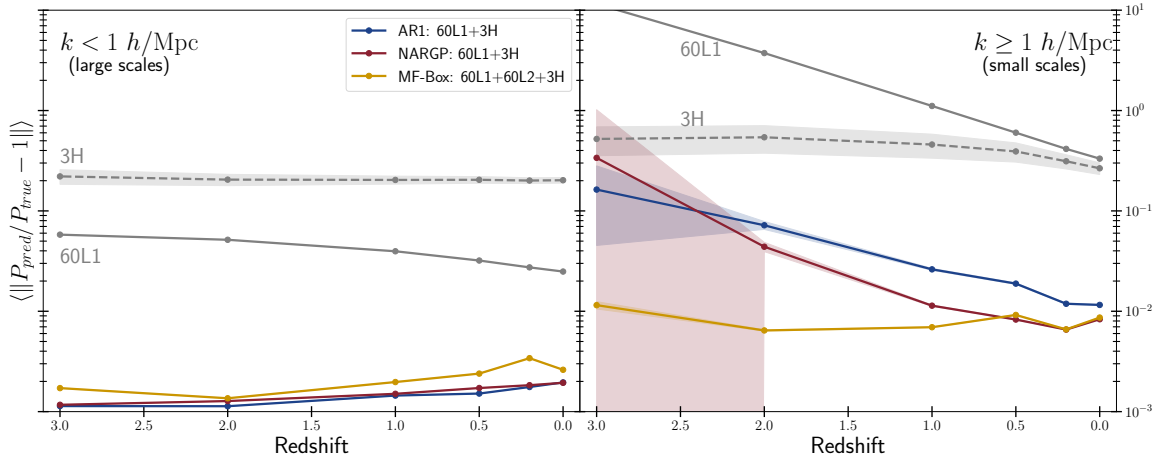


Figure 5.11: Relative errors averaged over all k modes (split into large and small scales) for different multi-fidelity models (AR1 (blue), NARGP (red), and MF-Box (yellow)), broken down into different redshift bins. The grey dashed line is the HF-only emulator using 3 H simulations, and the solid grey line is the LF-only emulator using 60 L1 simulations. The shaded area is the variance among different test simulations. MF-Box improves the emulation at small scales at higher redshifts ($z \geq 1$). We do not include the variance of LFEmu (60L1) because the variance is too large.

Figure 5.10 shows the emulation error averaged over redshift bins, $z \in [0, 3]$, by using different multi-fidelity models, AR1, NARGP, and MF-Box. All three models perform similarly at large scales ($k < 2 h\text{Mpc}^{-1}$). The main difference is MF-Box performs better at $k \geq 2 h\text{Mpc}^{-1}$ while AR1 and NARGP have an error bump at 10% level.

In the right panel of Figure 5.11, we can easily see the 10% error bump exists at $z = 1 - 3$ at small scales ($k \geq 1 h\text{Mpc}^{-1}$). The small-scale improvement in the right panel is not a surprise. The additional low-fidelity node in a smaller box (L2) brings more accurate small-scale statistics than L1, making MF-Box outperform AR1 and NARGP. MF-Box stays $\simeq 1\%$ error within the redshift range $z \in [0, 3]$, in contrast to AR1 and NARGP where the error increases from $\simeq 1\%$ to $\simeq 20\%$ (from $z = 0$ to $z = 3$).

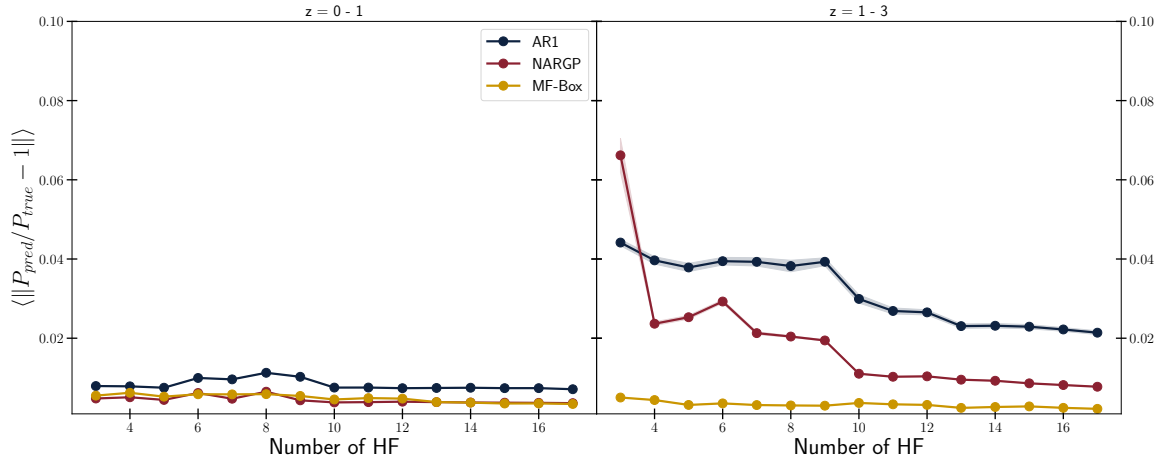


Figure 5.12: Relative error as a function of the number of HF training points for different multi-fidelity methods: AR1 (blue), NARGP (red), and MF-Box (yellow). The range of the number of HF points is relatively small, so the error estimate trend is unclear. However, in general, the emulation error decreases with more HF points. (Left) Averaged relative error for $z \in [0, 0.2, 0.5]$. (Right) Averaged relative error for $z \in [1, 2, 3]$.

The bump in interpolation error in AR1 and NARGP at $z > 1$ is due to the feature at the initial inter-particle spacing at these redshifts, corresponding to the initial particle grid, as mentioned in [16]. The mean particle spacing of the initial condition appears as a delta function in the matter power spectrum at high redshift. This feature eventually disappears, erased by gravitational interactions. The L2 and high fidelity box, however, both have a smaller mean inter-particle spacing and thus show the delta function on smaller scales, beyond those we wish to emulate. Using the information the L2 simulations provide, MF-Box is able to maintain similar accuracy across $z \in [0, 3]$.

The left panel of Figure 5.11 shows the redshift trend at large scales, indicating no significant difference between AR1, NARGP, and MF-Box. The slightly worse accuracy in MF-Box is probably because MF-Box has more hyperparameters to fit, making it slightly more difficult to reach $\sim 0.1\%$ accuracy.

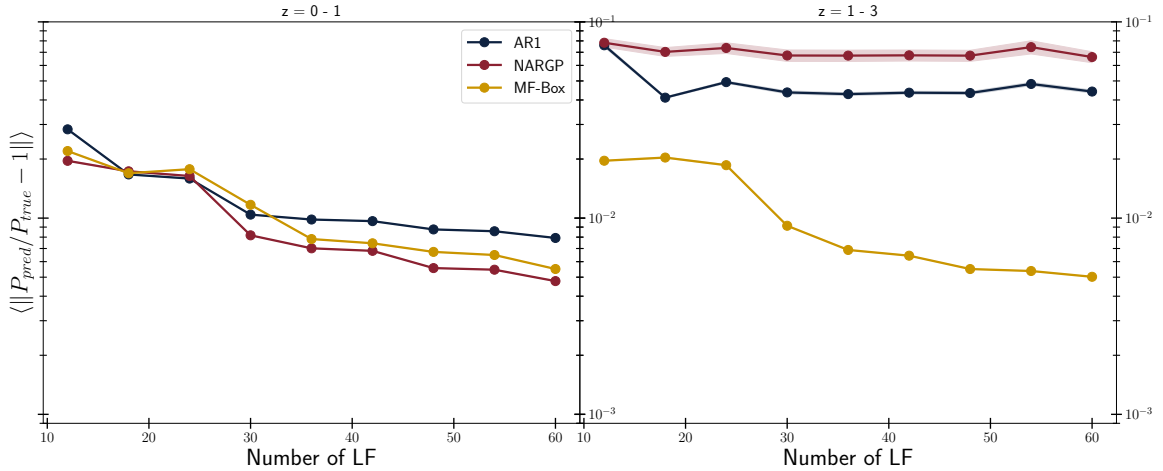


Figure 5.13: Relative errors for AR1 (blue), NARGP (red), and MF-Box (yellow) as a function of LF points, splitting into two redshift bins. (Left) Averaged error for $z \in [0, 0.2, 0.5]$. (Right) Averaged error for $z \in [1, 2, 3]$.

Figure 5.12 shows the AR1, NARGP, and MF-Box accuracies as a function of the number of HF points, splitting into two redshift bins. The left panel shows the accuracy averaged over the low redshift bins, $z \in [0, 0.2, 0.5]$, where NARGP and MF-Box perform similarly and outperform the AR1 model. It is not a surprise that NARGP and MF-Box perform similarly since MF-Box is an extension of NARGP.

The left panel of Figure 5.12 shows that the error is almost flat as a function of HF points. In Section 5.6, we showed that the emulator error is a power-law function of the number of training points. Here, the emulation accuracy is likely limited by the intrinsic accuracy of our 512^3 HF simulations, so it is hard to get improvement at the sub-percent level.⁷ The right panel of Figure 5.12 shows that MF-Box performs better than the other two models by a factor of $\sim 5 - 10$.

⁷As discussed in [16], our HF power spectra are $\sim 0.1 - 10\%$ error compared with EuclidEmulator2.

Figure 5.13 shows the averaged emulation error as a function of LF points. We see a mild improvement at low-redshift bins (left panel) by adding more LF points for all three models. At the higher redshift bins (right panel), AR1 and NARGP cannot be easily improved by adding more LF training simulations. This is likely because the error is dominated by the delta function in L1 at small scales. MF-Box achieves an average error at the 1% level with 30L1+30L2+3HF, as expected from Section 5.6.

In summary, we show that the improvement of MF-Box happens at small scales ($k > 2 h\text{Mpc}^{-1}$) at the higher redshift bins ($z \in [1, 2, 3]$). This is primarily because the L1 node at these redshifts has the delta function feature from the initial particle grid dominating on small scales.

5.7.2 Emulation with various box sizes

In Section 5.7.1, we have learned that we can achieve better emulation performance by incorporating a low-fidelity node in a smaller box. This section examines how MF-Box’s emulation error changed as a function of the L2 box size.

Figure 5.14 shows the emulation error as a function of L2 box size, averaging over all k bins and splitting into two redshift bins. We include AR1, NARGP, and MF-Box. In this section, we use the L2 node as the LF node for both AR1 and NARGP. The left panel shows the error at the low-redshift bin ($z \in [0, 0.2, 0.5]$). AR1 and NARGP have $< 1\%$ error with $L2 = 256 \text{ Mpc}/h$, but the error gets worse when the L2 box size becomes smaller due to the cosmic variance at large scales. On the other hand, MF-Box error stays flat for $L2 \in [100, 224] \text{ Mpc}/h$.

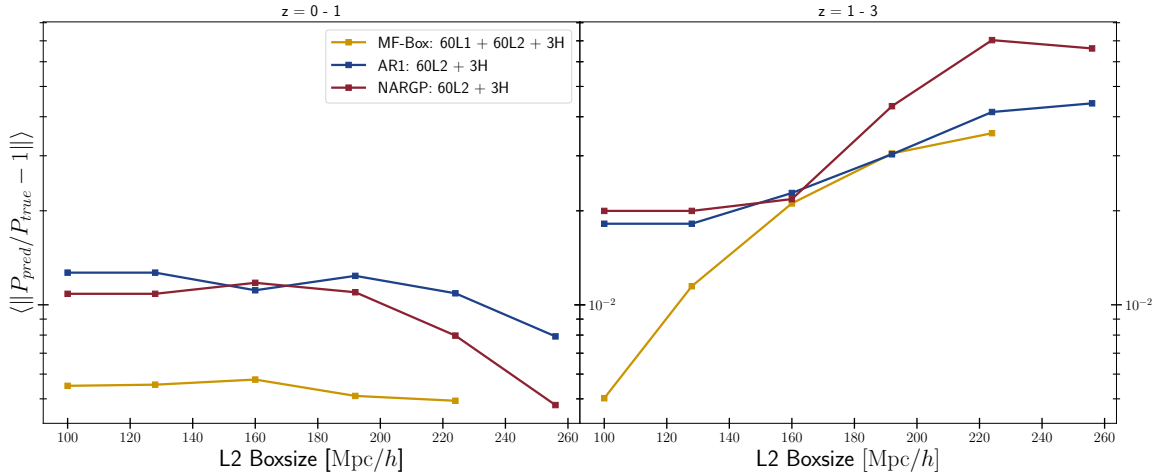


Figure 5.14: Relative errors of multi-fidelity emulation as a function of L2 boxsize, for AR1 (blue), NARGP (red), and MF-Box (yellow). Note that we use L2 instead of L1 for AR1 and NARGP models.

The right panel of Figure 5.14 shows the error versus L2 box size at the high-redshift bin, $z \in [1, 3]$. All models show a decrease in error using a smaller L2 box size in training. This is mainly due to the feature at the initial inter-particle spacing mentioned in Section 5.7.1. If a smaller L2 is used, the feature moves to smaller scales, away from those we are emulating, causing a decline of error from the large L2 box to the small L2 box size.

To help visualize the performance change on different scales, we show in Figure 5.15 the emulation error as a function of k , averaged over all redshift bins. As Figure 5.15 shows, for different L2 sizes, MF-Box accuracy only changes at the small scales with $k > 3 h\text{Mpc}^{-1}$. This is not a surprise because all MF-Box models share the same L1 node (128^2 simulations in $256 \text{Mpc}/h$), and thus the emulation at large scales stays the same. The NARGP shown in Figure 5.15 uses L2 with $100 \text{Mpc}/h$ as a low-fidelity node. Its performance is worse than MF-Box with $L2 = 100 \text{Mpc}/h$ at all k bins.

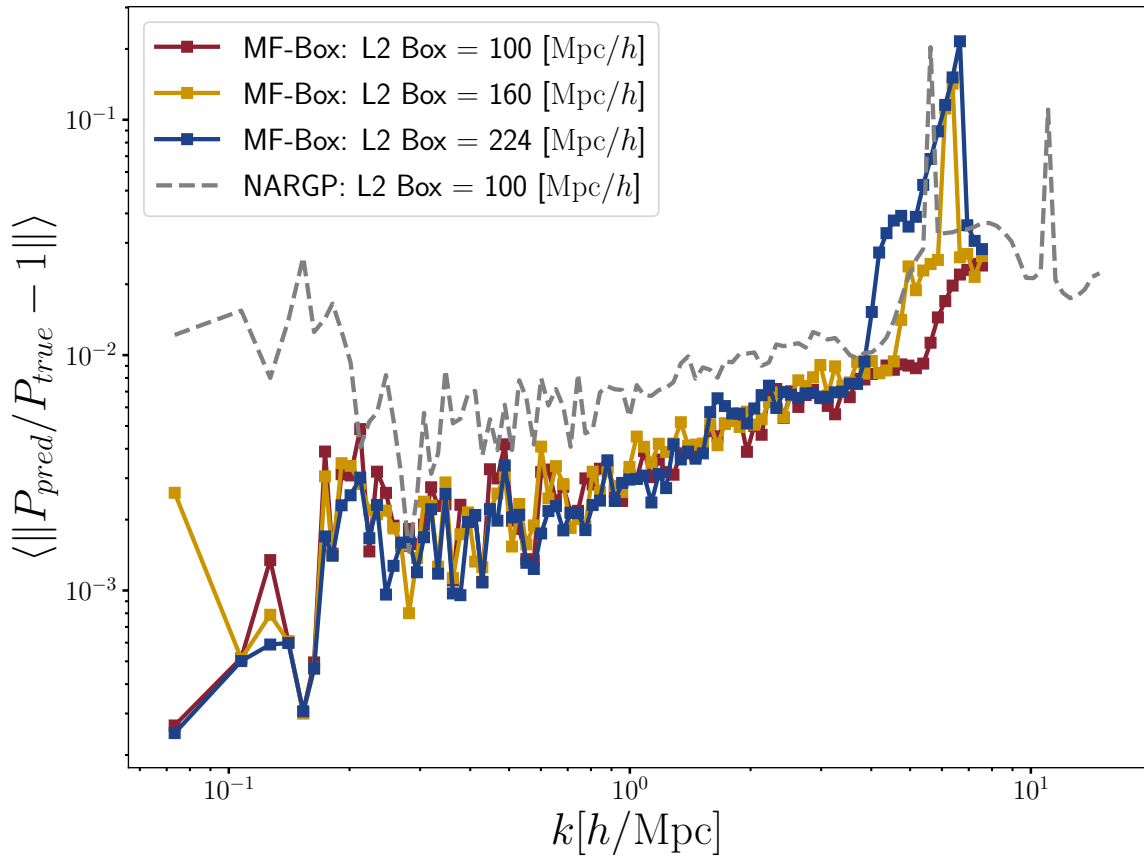


Figure 5.15: Relative errors averaged over redshift bins, as a function of k modes. MF-Box with 224 Mpc/h L2 (blue), MF-Box with 160 Mpc/h L2 (yellow), and MF-Box with 100 Mpc/h L2 (red). The gray dashed line is the NARGP model uses 100 Mpc/h L2.

To sum up, the error of MF-Box changed as a function of L2 box size: using a smaller L2 can result in better MF-Box accuracy. The improvement caused by L2 is mostly at small scales ($k > 2 h\text{Mpc}^{-1}$) at higher redshift bins ($z = 1, 2, 3$).

5.7.3 Runtime comparison

We will compare the costs of each method in this section. Figure 5.16 shows the error of different emulators as a function of node hours for the training simulations. A similar compute time versus accuracy plot can be found in Figure 4 of [16], albeit only for $z = 0$. We performed the MP-Gadget simulations at High-Performance Computing Center (HPCC) at UC Riverside,⁸ each compute node has 32 intel Broadwell cores.

To understand Figure 5.16, we can start with the high-fidelity only emulators ([3-11] HF). This is the emulator we would train before we have multi-fidelity methods. HF-only emulator shows a steady improvement with an increase in run time. However, the error gradient gets flatter with more training points, indicating the difficulty of improving an emulator at a highly accurate regime.

This trend is intuitive because the error of an emulator roughly scales as a power-law function, $(\text{number of training points})^{-\frac{\nu}{d}}$. Each line in Figure 5.16 is a segment of different power-law models. In this view, we can see AR1 and NARGP follow two very similar trends, except one has a lower mean emulation error.

Switching the focus to MF-Box, we can see the mean error of the power law is $\sim 6-8$ times better than AR1 and NARGP. The error for both AR1 and NARGP plateaus,

⁸<https://hpcc.ucr.edu>

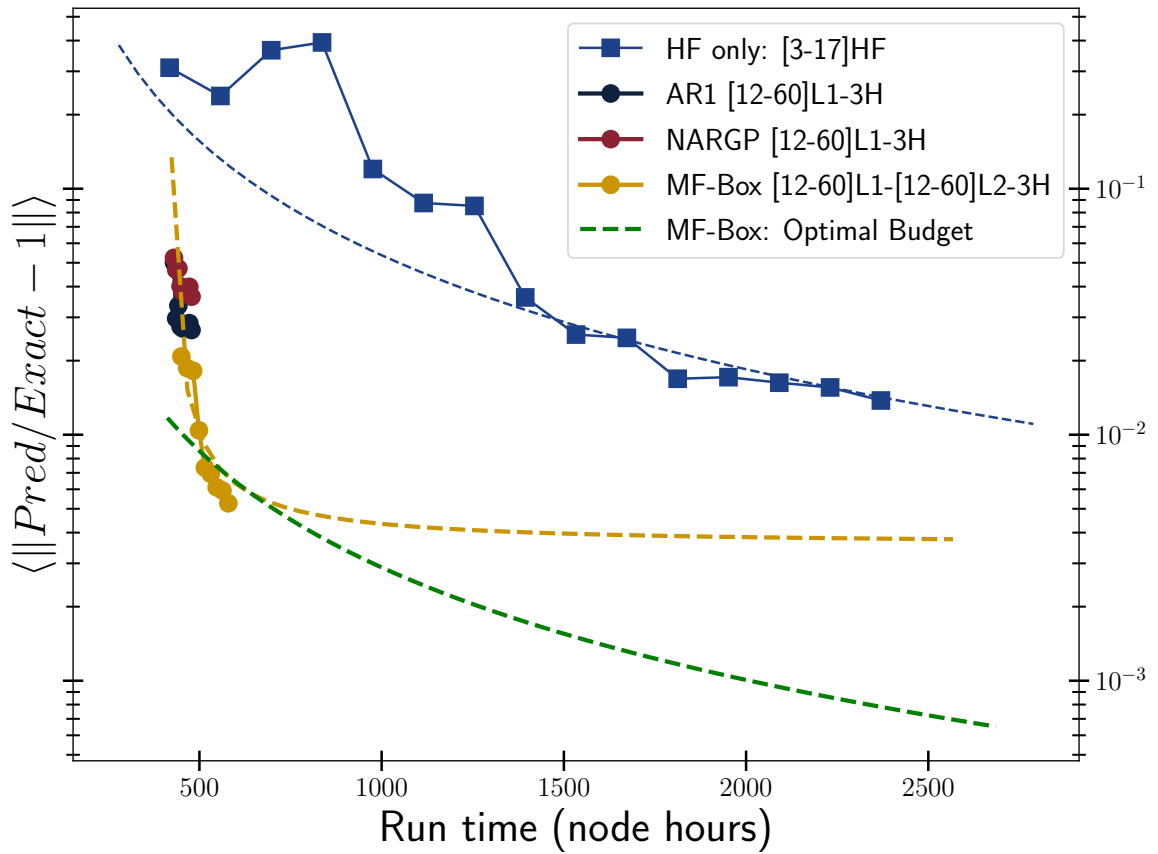


Figure 5.16: Runtime comparison in node hours. We average the error across redshift bins $z = [0, 0.2, 0.5, 1, 2, 3]$ and average across k bins. AR1 and NARGP perform similarly to MF-Box at $z < 1$. Dashed lines are the predicted error based on the error function Eq 5.13, which we inferred in Section 5.6.

implying that adding new simulations will not increase the emulator’s accuracy. The only way to improve the emulation at a similarly good efficiency is using small-box simulations through MF-Box.

Recall the HF/L1 ratios in Figure 5.2. L1 is roughly at $\sim 5\%$ error at large scales. On the other hand, the L2-only emulator is at $\sim 10\%$ error. Using a MF-Box, the information carried by L1 and L2 is corrected to be at $\sim 0.5\%$ level, which is a substantial improvement given that only 3 HF simulations are utilized to establish correlations between fidelities.

5.8 Conclusions

In this work, we show that our multi-fidelity emulation, MF-Box (model structure refers to Figure 5.1, and simulation data refer to Table 5.1), can combine simulations from different box sizes to achieve improved overall emulator accuracy. MF-Box has a higher accuracy improvement per CPU hour than the multi-fidelity method with only one box size. The framework is adaptable to different simulation suites and emulation problems.

We summarize the key contributions of this work below:

1. **Propose a new multi-fidelity emulation, MF-Box, combining information from different simulation box sizes:** Using the in-tree graph of GMGP [197], we can fuse cheap low-fidelity simulations from multiple box sizes in one unified machine-learning model. Simulations in a large box capture large-scale statistics, while the simulations in a small box can improve small-scale statistics. Previously, the cheapest way to improve MFEulator was by increasing the particle load in the low-fidelity

node, which scales as $\sim \mathcal{O}(N_{\text{ptl,side}}^3)$. **MF-Box** opens a new avenue to add additional information to the multi-fidelity emulation framework in a cheaper way.

2. Leverage accurate and systematic-free information from L2 to improve

multi-fidelity emulation accuracy: L2 provides unique information absent in L1, and also acts as a cross-check for L1. Systematic errors or unknown bugs in low-fidelity nodes can limit the effectiveness of multi-fidelity methods, as it relies on existing information. [16] identified such a limitation, noting that systematic errors present in the low-fidelity node can make achieving high accuracy difficult. **MF-Box** helps resolve the systematic in one low-fidelity node by introducing an additional L2 node without the systematic. It is worth noting that systematic errors may exist in both L1 and L2 nodes, but **MF-Box** can help mitigate these errors by cross-checking the information provided by two nodes, as long as the systematic errors are present at different scales.

3. Power-law analysis of emulation errors in multi-fidelity modeling with

MF-Box: In Section 5.6, we present an error analysis of **MF-Box** models. We empirically estimate the emulation error function, which follows a power-law decay with respect to the number of training simulations. This explains why it is difficult to improve single-fidelity emulators which are already percent-level accurate. Multi-fidelity emulation shows advantageous in reducing the overall cost and time required to achieve high accuracy. The estimated error function can also serve as a guide for optimizing resource allocation across fidelity nodes, facilitating the development of accurate emulators in a more efficient use of resources.

MF-Box also opens up opportunities to experiment with different ways to implement multi-fidelity emulation in cosmology. The second low-fidelity node, **L2**, can be anything that brings new information to a multi-fidelity emulator. For example, it could be a node that runs using hydrodynamical simulations, or a node that uses a linear perturbation theory code. One example could be **L1** runs with dark-matter only simulations at high-resolution, **L2** runs with hydrodynamical simulations at low resolution (and in a small box), and an **HF** node as hydrodynamical simulations at high-resolution. This way, the cosmological dependence of the baryonic effects is captured by **L2**, and **L1** gives us highly accurate gravitational clustering. **MF-Box**, using a different box size in an additional low-fidelity node, is just a simple example to demonstrate the flexibility of this method.

The main remaining limitation of our multi-fidelity emulation framework is that the highest fidelity node must be in the training set, and encompass the largest box and highest resolution. In other words, our multi-fidelity framework cannot extrapolate to predict the results of a simulation with a resolution higher than the high-fidelity node.

Future applications of our multi-fidelity emulation include applying the **MF-Box** to the accurate high-resolution simulations, where the resolution can match the future experiments. We may also apply **MF-Box** to different cosmological probes, especially applying to the beyond 2-point statistics, such as weak lensing peak counts and scattering transform coefficients.

Chapter 6

Investigating the mixing between two black hole populations in LIGO-Virgo-KAGRA GWTC-3

6.1 Abstract

We introduce a population model to analyze the mixing between hypothesised power-law and $\sim 35M_{\odot}$ Gaussian bump black hole populations in the latest gravitational wave catalog, GWTC-3, estimating their co-location and separation. We find a relatively low level of mixing, $3.1^{+5.0}_{-3.1}\%$, between the power-law and Gaussian populations, compared to the percentage of mergers containing two Gaussian bump black holes, $5.0^{+3.2}_{-1.7}\%$. Our analysis indicates that black holes within the Gaussian bump are generally separate from the power-law population, with only a minor fraction engaging in mixing and contributing

to the $\mathcal{M} \sim 14M_{\odot}$ peak in the chirp mass. This leads us to identify a distinct population of Binary Gaussian Black Holes (BGBHs) that arise from mergers within the Gaussian bump. We suggest that current theories for the formation of the massive $35M_{\odot}$ Gaussian bump population may need to reevaluate the underlying mechanisms that drive the preference for BGBHs.

6.2 Introduction

Gravitational wave astronomy is shifting focus from in-depth analysis of single events to population inference that addresses key questions in astrophysics [13, 217], fundamental physics [218], and cosmology [219]. Research using the third Gravitational-Wave Transient Catalog (GWTC-3) [220], published by the LIGO Scientific Collaboration [219], Virgo Collaboration [221], and KAGRA Collaboration [222], demonstrated this [223, 224, 225, 226]. GWTC-3 has become a vital tool for understanding binary black hole (BBH) formation physics (e.g., Ref.[227, 228, 225, 229, 13]). Investigating the formation history of BBHs through a single gravitational wave (GW) event is a difficult task. However, the population analysis of numerous merging BBH events can provide insights into their formation channels (e.g., [230, 227]). For example, the lack of black holes with masses $\sim 2 - 5M_{\odot}$ [231, 232, 233, 234] may indicate maximum neutron star masses [235, 236, 237], and also the timescale related to supernova explosions (such as [238, 239, 240]) and mass transfer (e.g., [241]).

LIGO-Virgo-KAGRA's (LVK, hereafter) population analysis of the GWTC-3 catalog indicates distinct substructures within the primary black hole mass spectrum [13].

In the primary mass distribution, two prominent peaks are observed at $m_1 \sim 10M_\odot$ and $m_1 \sim 35M_\odot$ with high significance. Another peak, at $m_1 \sim 20M_\odot$, appears to be less certain. The peak at approximately $10M_\odot$ is postulated to exist above the black hole-neutron star (BH-NS) low-mass gap and can arise from the stellar initial mass function (IMF). The corresponding peak in the binary black hole (BBH) mass function could be attributed to the evolution of binary star systems [13, 241, 242]. Several options have been suggested to explain the peaks at $m_1 \sim 20M_\odot$ and $m_1 \sim 35M_\odot$. A popular explanation for the peak at around $m_1 \sim 35M_\odot$ is that it results from pulsational pair-instability supernovae (PPSNe) originating from stars initially ranging between $100M_\odot$ and $150M_\odot$ [243, 244, 245, 246, 247, 248, 249, 250, 251, 252, 253], though some recent studies suggest the $35M_\odot$ peak is unlikely to be due to the PPSNe [254, 255]. The pair-instability mechanism also predicts a sharp cutoff at masses greater than $40M_\odot$, attributed to the absence of remnants from pair-instability supernovae occurring in stars with initial masses ranging from $150M_\odot$ to $250M_\odot$. Other exotic formation channels that may explain the $35M_\odot$ peak include primordial black holes [256, 257, 258, 259], massive triple stars [260, 261], low-metallicity star progenitors [262], and hierarchical mergers [263].

As the number of BBH detections from gravitational wave observations increases, it becomes feasible to test BBH formation channels by examining the statistical properties of the population of secondary black holes. In situations where black holes merge within dense environments (e.g. star clusters), following a dynamical channel, the underlying mass distribution would likely appear similar as the comparable component masses have a higher binding energy [264, 265, 266], though it is possible for a dynamical channel to produce

unequal component mass binaries through ultra-wide binaries [267]. While the isolation formation channel (“field binaries”) prefers BBHs with comparable masses [268], some isolation channels can produce unequal component masses [269, 270]. This variation could be influenced by a range of uncertain physical processes, such as the binary IMF [271], the evolution of binary star systems [272, 273, 269, 270], and possible mechanisms like mass transfer or inversion [274, 275, 276, 277].

A widely used parameterization for the BBH mass spectrum is the Power-law+Peak model [249], which involves modeling a combination of the primary mass (m_1), the heavier black hole in the BBH, and the mass ratio ($q = m_2/m_1 < 1$), representing the ratio between the secondary and primary masses. The Power-law+Peak model posits a power-law function with a Gaussian peak for the primary mass distribution, while the physical distinction between primary and secondary masses is modeled using a power-law model on the mass ratio. This approach has been utilized in several studies (e.g., [278, 279, 14, 229, 280, 281]). Beyond the Power-law+Peak model, Ref. [282, 266] explore various models for the secondary mass spectrum in BBHs.

In this paper, we construct a population model to estimate the mixing fractions between the populations of black holes from the power-law distribution (corresponding to the peak at $m_1 \sim 10M_\odot$) and those from the Gaussian bump (at $m_1 \sim 35M_\odot$). We want to understand how likely it is for black holes originating from different peaks to mix in the Universe, forming the BBHs we observe. Inferring the mixing fraction between the power-law and Gaussian peak populations can provide new perspective into BBH formation mechanisms. For instance, low mixing between the two populations might indicate that

black holes produced by different formation mechanisms remain separated, with binaries likely forming within their respective populations.

Our population model begins by forward-sampling black hole masses from either a power-law or Gaussian mass spectrum. Each pair in a BBH is then sampled from a mixture of these mass populations. We vary the relative abundance in the mixture model, thus controlling the mixing fraction between the power-law and Gaussian mass populations. Using a mixture of power-law and Gaussian populations ensures our primary mass function aligns well with the Power-law+Peak model from Ref. [13].

Our inference suggests that a significant portion (approximately $5.0_{-1.7}^{+3.1}\%$ of the total population) of the BBHs consist of Binary Gaussian Black Holes (BGBHs), where both black holes originate from the $\sim 35M_{\odot}$ Gaussian bump. We also observe a low mixing fraction between the power-law and the Gaussian bump, $3.1_{-3.1}^{+5.0}\%$, indicating that the Gaussian bump black holes are primarily separate from the power-law population. Another interesting aspect of our model is the alignment of the second chirp mass peak at $\mathcal{M} \sim 14M_{\odot}$ with the mixing between the power-law distribution peak ($\sim 10M_{\odot}$) and the Gaussian distribution peak ($\sim 35M_{\odot}$), calculated as $\mathcal{M} \sim \frac{(10M_{\odot} \times 35M_{\odot})^{3/5}}{(10M_{\odot} + 35M_{\odot})^{1/5}} \simeq 15M_{\odot}$. Among notable features in the black hole mass spectrum, the second peak around $m_1 \sim 20M_{\odot}$ has been identified as marginally significant in primary mass [283, 280, 284, 285]. However, its nature remains debated. Some argue it may be a result of Poisson fluctuations within the power-law function [284], while others suggest that the corresponding second peak in the chirp mass spectrum ($\mathcal{M} \sim 14M_{\odot}$) is more pronounced than the primary mass substructure

[285]. Our inference suggests a way to interpret the $\mathcal{M} \sim 14M_{\odot}$ chirp mass peak as a result of mixing between two populations with different formation mechanisms.

This paper is structured as follows: Section 6.3 introduces our population model for BBHs. Section 6.4 outlines the Bayesian inference approach we employ, taking into account detection efficiency. Section 6.5 presents our inference results and the predicted black hole mass functions. Section 6.7 offers concluding remarks.

6.3 Population Model

In this section, we discuss our BBH population model designed to understand the mixing between different black hole populations. Our population model is detailed in Section 6.3.1, where we discuss the three different subpopulations of BBHs and the forward model for generating samples of BBHs. Next, in Section 6.3.2, we show the exploratory models of the predicted chirp mass and mass ratios according to different parameters of the population model. In Section 6.3.3, we discuss a method for acquiring fiducial parameters for our population mixture model.

6.3.1 Population model: Subpopulations

In this section, we discuss the reasoning behind our population model, which is motivated by the Gaussian bump in the primary mass function. LVK population analysis identified a notable Gaussian peak mixed into the primary mass function [13], with the mixing fraction represented by $\lambda_{\text{peak}} \sim 3.8^{+5.8}_{-2.6}\%$, suggesting roughly 3.8% of the primary mass black holes are from the Gaussian distribution.

To begin, our population model assumes the black holes in BBHs are drawn from either a power-law distribution or a Gaussian distribution. This points to three distinct subpopulations of BBHs in our model: Power-Power (PP), Power-Gaussian (PG), and Gaussian-Gaussian (GG).

- Power-Power (PP) model: A black hole from the Power-law population merging another power-law population black hole.
- Gaussian-Gaussian (GG) model: A black hole from the Gaussian population merging with another Gaussian population black hole. If the fraction of GG events is high relative to PG events, then it is likely Gaussian bump black holes are separate from the rest of the black holes. We name this population of BBHs as BGBHs.
- Power-Gaussian (PG) model: A black hole from the Power-law population merging with a black hole from the Gaussian population. If the fraction of PG events is high relative to GG events, then it is likely that the Gaussian bump black holes are mixed with the rest of the black holes.

A cartoon version representing these three scenarios can be found in Fig 6.1. Measuring a high fraction of GG events alongside a low fraction of PG events would suggest that the Gaussian bump is separate from the power-law population. Conversely, a significantly high fraction of PG events suggests that they are part of a singular, co-located population containing a mixture of black holes. The co-location (and separation) here refers to a broader concept of co-locating (and separation) in the phase space of space or time.

We chose to use a broad definition of co-location (and separation) because the GWTC data only provides measurements of BBH mergers. Therefore, the separation of

the Gaussian peak population in the measured mass spectrum is not necessarily due to spatial separation. There are some degeneracies, such as these black holes being temporally separated (formed at different redshifts). Therefore, in this work, when we say that BH populations are separate, this could imply that they are distributed separately in space or time.

Following Ref. [249], we define the power-law population as

$$\mathcal{B}(m \mid -\alpha, \delta_m, m_{\min}, m_{\max}) = \frac{m^{-\alpha}}{Z_m(m_{\min}, m_{\max})} S(m \mid m_{\min}, \delta_m). \quad (6.1)$$

Here, $-\alpha$ represents the spectral index of the power-law. The $Z_m(m_{\min}, m_{\max})$ is the normalization factor for the power-law

$$Z_m(m_{\min}, m_{\max}) = \frac{m_{\max}^{-\alpha+1} - m_{\min}^{-\alpha+1}}{-\alpha + 1} \quad (6.2)$$

with the smoothing function at the low-mass end

$$S(m \mid m_{\min}, \delta_m) = \left(\exp \left(\frac{\delta_m}{m - m_{\min}} + \frac{\delta_m}{m - m_{\min} - \delta_m} \right) + 1 \right)^{-1} \quad (6.3)$$

where $S(m \mid m_{\min}, \delta_m) = 0$ for $m < m_{\min}$ and $S(m \mid m_{\min}, \delta_m) = 1$ for $m \geq m_{\min} + \delta_m$. That is, a smoothing kernel in the range $m_{\min} \leq m < m_{\min} + \delta_m$. Additionally, we incorporate a parameter for the maximum mass cutoff, m_{\max} . The same smoothing kernel is also applied to the Gaussian distribution

$$\mathcal{G}(m \mid \mu, \sigma, \delta_m, m_{\min}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{m-\mu}{\sigma}\right)^2} S(m \mid m_{\min}, \delta_m). \quad (6.4)$$

Here, μ is the mean and σ is the standard deviation. Our power-law and Gaussian models are the same as the ones in the Power-law+Peak model [249]. We do not explicitly model

the primary and secondary mass spectra, instead we draw black hole masses from one of the population models. Explicitly modelling the primary and secondary requires us to ensure that the primary is the more massive object, which makes our model more complex.

Next, we build a forward model that draws samples of BBHs. We do not directly model the primary and secondary masses, but instead, we model the mass function of the component black holes in binaries. For clarity, when referencing arbitrary masses in a BBH system, we will use m_a and m_b . We will use the chirp mass and mass ratio as observables, which are derived from m_a and m_b samples we draw from the subpopulation model. For PP subpopulation, the two-dimensional (m_a, m_b) probability density is given by:

$$p_{\text{PP}}(m_a, m_b \mid -\alpha, \delta_m, m_{\text{min}}, m_{\text{max}}) \propto \mathcal{B}(m_a \mid -\alpha, \delta_m, m_{\text{min}}, m_{\text{max}})\mathcal{B}(m_b \mid -\alpha, \delta_m, m_{\text{min}}, m_{\text{max}}), \quad (6.5)$$

For GG subpopulation, the probability density is

$$p_{\text{GG}}(m_a, m_b \mid \mu, \sigma, \delta_m, m_{\text{min}}) \propto \mathcal{G}(m_a \mid \mu, \sigma, \delta_m, m_{\text{min}})\mathcal{G}(m_b \mid \mu, \sigma, \delta_m, m_{\text{min}}). \quad (6.6)$$

And for the PG subpopulation, the probability density is

$$p_{\text{PG}}(m_a, m_b \mid -\alpha, \mu, \sigma, \delta_m, m_{\text{min}}, m_{\text{max}}) \propto \mathcal{B}(m_a \mid -\alpha, \delta_m, m_{\text{min}}, m_{\text{max}})\mathcal{G}(m_b \mid \mu, \sigma, \delta_m, m_{\text{min}}). \quad (6.7)$$

Here, we do not repeat the sampling for $m_a \sim \mathcal{G}$ and $m_b \sim \mathcal{B}$ because we assume the shape parameters are the same for m_a and m_b , so the (m_a, m_b) labels are interchangeable in this sampling.

Since we are not modeling the primary and secondary mass directly, we cannot directly use the probability density of (m_a, m_b) as a likelihood function and apply it on

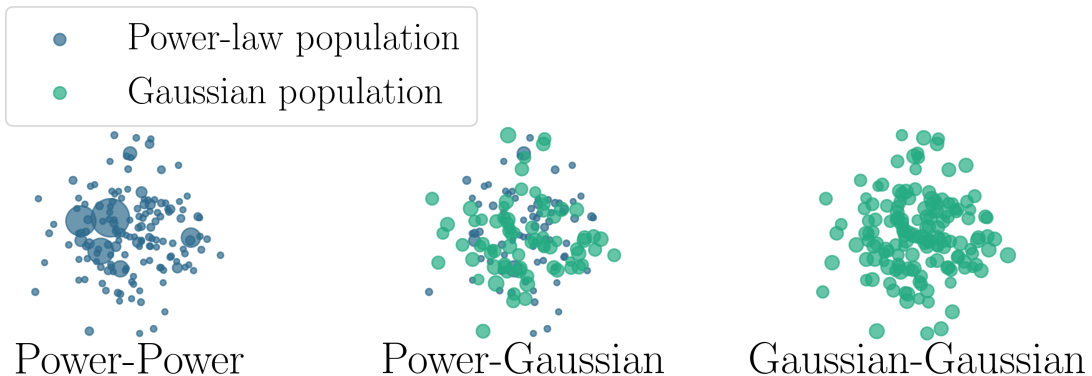


Figure 6.1: A cartoon illustrates the mixing scenarios used in this work. The size of the circles represents the masses of the black holes, while the color indicates the underlying population. If the power-law and Gaussian bump BHs are mixed, as in the PG model, the resulting two-dimensional probability density of chirp mass and mass ratio (\mathcal{M}, q) will exhibit a distinct morphology, as shown in Figure 6.2.

the data. Instead, we convert the (m_a, m_b) parameters into quantities that we observe in gravitational events, i.e., chirp mass and mass ratio, (\mathcal{M}, q) , with $\mathcal{M} = (m_a m_b)^{3/5} / (m_a + m_b)^{1/5}$ and $q = \min(m_a, m_b) / \max(m_a, m_b)$. They have the same statistical information as (m_a, m_b) and are some of the more directly measured parameters in gravitational wave events. In addition, in this paper, we focus on the binary-centric properties, i.e., the mixing fractions of BBH subpopulations; it is thus reasonable to use chirp mass and mass ratio (the properties specific to BBHs) over component masses.

After defining the three different subpopulation models, we can now construct a mixture model to infer the relative abundances of the three subpopulations:

$$\begin{aligned}
p(\mathcal{M}, q \mid \psi_{\text{PP}}, \psi_{\text{PG}}, \psi_{\text{GG}}, -\alpha, \mu, \sigma, \delta_m, m_{\text{min}}) \propto \\
\psi_{\text{PP}} p_{\text{PP}}(\mathcal{M}, q \mid -\alpha, \delta_m, m_{\text{min}}, m_{\text{max}}) + \\
\psi_{\text{PG}} p_{\text{PG}}(\mathcal{M}, q \mid -\alpha, \mu, \sigma, \delta_m, m_{\text{min}}, m_{\text{max}}) + \\
\psi_{\text{GG}} p_{\text{GG}}(\mathcal{M}, q \mid \mu, \sigma, \delta_m, m_{\text{min}}),
\end{aligned} \tag{6.8}$$

where $(\psi_{\text{PP}}, \psi_{\text{PG}}, \psi_{\text{GG}})$ are the relative abundances for PP, PG, and GG subpopulations, with $\psi_{\text{PP}} + \psi_{\text{PG}} + \psi_{\text{GG}} = 1$ and $\psi_{\text{PP}}, \psi_{\text{PG}}, \psi_{\text{GG}} \in [0, 1]$.

The PG subpopulation is the key to measuring the mixing between the power-law and Gaussian bump populations. The Power-law+Peak models the mass ratios from all BBHs as a single power-law, which does not allow us to separate the contributions of PG, PP, and GG to the mass ratio distribution. Even if the mass ratio from Power-law+Peak prefers equal-mass binaries, this preference could be driven by the majority of power-law black holes. Our model allows us to separate the mass ratio contribution of the PG from the rest of the BBHs, providing a more direct measurement of the separation of the Gaussian bump.

6.3.2 Visualizations of the population model

To help gain intuition for the population model, we generate histograms from the Monte Carlo samples of the three subpopulations across the parameter space of (\mathcal{M}, q) , shown in Figure 6.2. Each subpopulation covers a unique region within this (\mathcal{M}, q) space.

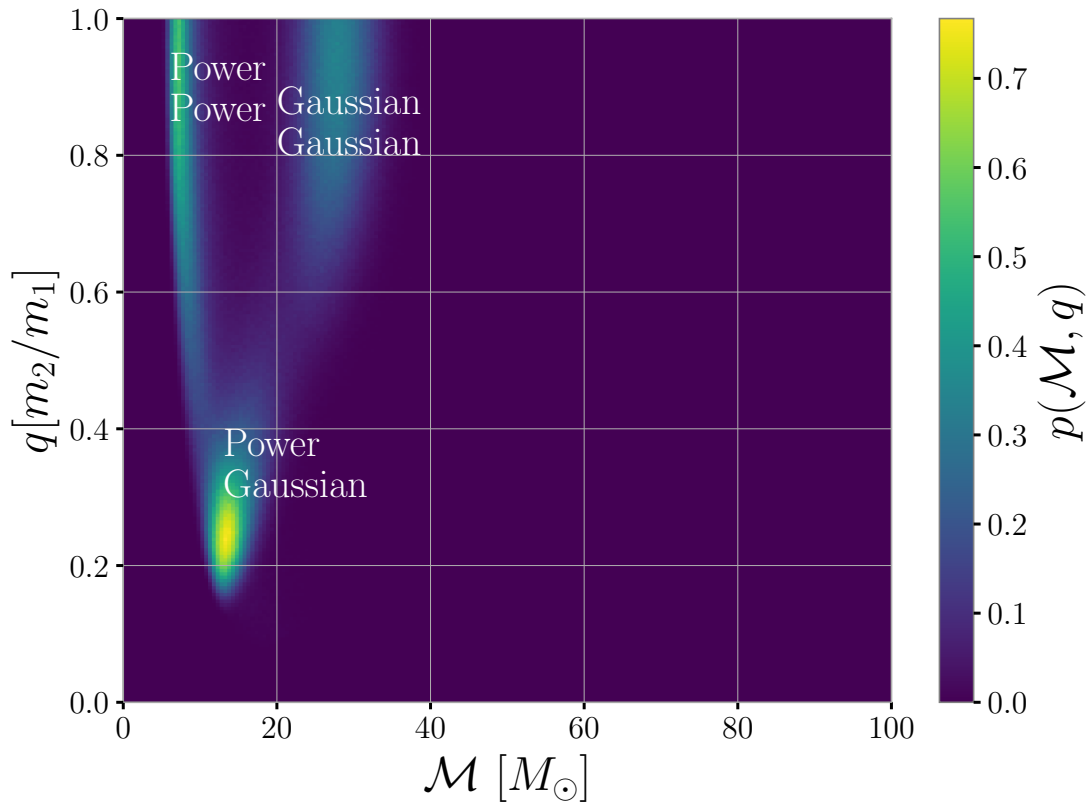


Figure 6.2: Map of likelihood density in chirp mass (\mathcal{M}) versus mass ratio (q) space for three subpopulation models, PP, PG, and GG. The shape parameters, $\boldsymbol{\lambda} = (-\alpha, \mu, \sigma) = (-3.66, 31.59, 5.51)$ used to generate the map come from the average mass spectrum of the GWTC-3's Power-law+Peak model as derived in Section 6.3.3.

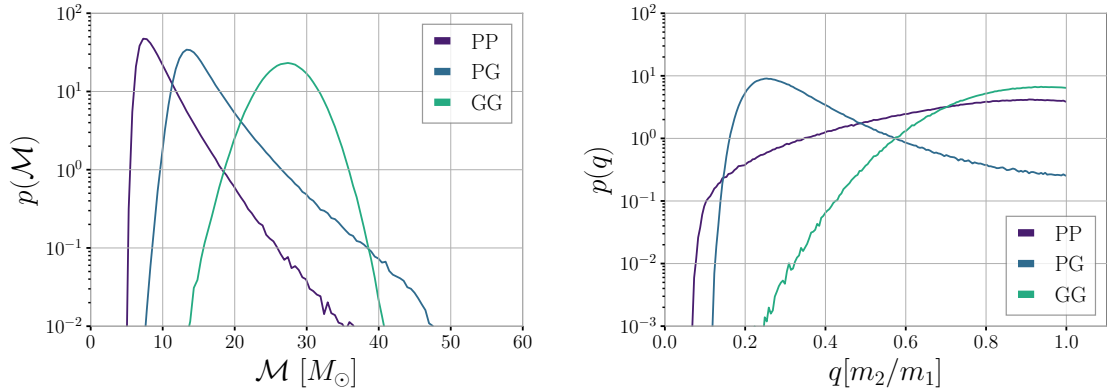


Figure 6.3: The one-dimensional marginal distribution of the two-dimensional density shown in Figure 6.2. The chirp mass spectrum, as shown in the upper panel, features three peaks at $\mathcal{M} \sim 8M_{\odot}$, $14M_{\odot}$, and $28M_{\odot}$. The mass ratio spectrum reveals a bump at $q \sim 0.2$ for the PG population, highly equal-mass binaries in the GG population, and a smooth mass ratio distribution for the PP population.

These distinct areas will aid in determining the mixing fraction of each subpopulation in the gravitational wave data.

Figure 6.3 shows the 1-D marginal distributions derived from each subpopulation model. The chirp mass distributions align with the three peak structures observed in the GWTC-3 chirp mass spectrum, i.e., $(8M_{\odot}, 14M_{\odot}, 28M_{\odot})$. The mass ratio distributions exhibit a bump at $q \sim 0.2$ for the PG population, highly equal-mass binaries in the GG population, and a smooth mass ratio distribution for the PP population.

Figure 6.4 illustrates various potential outcomes of our population model with different values for the spectral index. We have fixed the relative abundances for each subpopulation model at equal weights, $(\psi_{PP}, \psi_{PG}, \psi_{GG}) = (1/3, 1/3, 1/3)$. The spectral index, $-\alpha$, is varied from -7 to -2 . A flatter spectral index results in a more diffuse (less peaked) density of the PP and PG models in the (\mathcal{M}, q) space. Conversely, a steeper

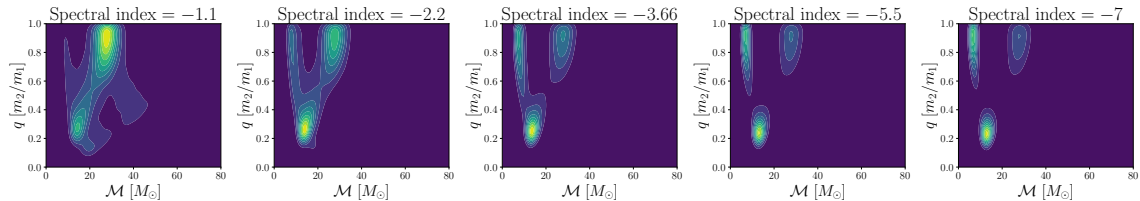


Figure 6.4: The exploratory models with varying spectral indices, $-\alpha$, in the space of chirp mass and mass ratio, (\mathcal{M}, q) . It ranges from a flat spectral index (left panel) to a sharp spectral index (right panel). In these exploratory plots, each subpopulation model has the same relative abundance, $1/3$.

spectral index (e.g., $-\alpha = 7$) results in a more distinct separation of the density of each subpopulation model in the (\mathcal{M}, q) space.

Figure 6.5 presents the 1D marginal distributions with varying spectral indices. The relative abundances are set to $(\psi_{\text{PP}} = 0.92, \psi_{\text{PG}} = 0.03, \psi_{\text{GG}} = 0.05)$, which are close to the inferred relative abundance from the model averaging results in Section 6.5.2. The chirp mass spectrum exhibits three peaks at $\mathcal{M} \sim 8M_{\odot}$, $14M_{\odot}$, and $28M_{\odot}$ for $-\alpha \lesssim -3.7$. For $-\alpha \gtrsim -3.7$, the chirp mass spectrum shows a relatively uniform density across the mass ratio spectrum.

6.3.3 Average mass spectrum

Our model operates on the average mass spectrum for individual black holes, treating black hole masses without distinguishing them as primary or secondary. Thus, the fiducial values for our model’s shape parameters, $\boldsymbol{\lambda} = (-\alpha, \mu, \sigma)$, will be different from those defined by the Power-law+Peak published in GWTC-3. We transform the primary and secondary masses, (m_1, m_2) , from the Power-law+Peak into a single black hole mass spectrum. We then fit a combination of the power-law and Gaussian model to this average

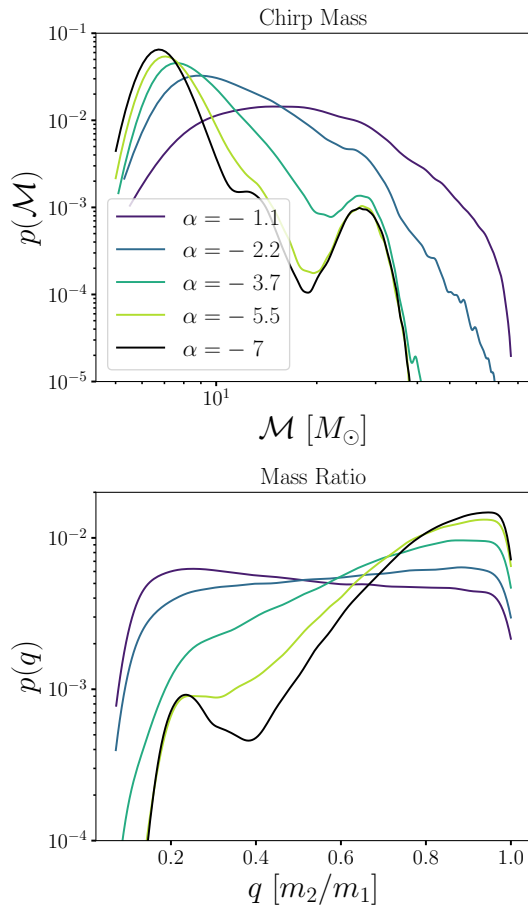


Figure 6.5: The exploratory models with varying spectral indices, $-\alpha$, on the chirp mass ($p(\mathcal{M})$) and mass ratio ($p(q)$) marginal distributions. The relative abundance is fixed to $(\psi_{PP}, \psi_{PG}, \psi_{GG}) = (0.92, 0.03, 0.05)$, matching the maximum a posteriori (MAP) of the model averaging results in Section 6.5.2.

mass spectrum, obtaining the fiducial values for our population model. This single mass spectrum reflects what we aim to represent by (m_a, m_b) . Throughout the paper, this will be referred to it as the “average mass spectrum” for black holes.

The primary mass distribution in the Power-law+Peak is parameterized as a combination of a power-law and a Gaussian distribution:

$$\begin{aligned}
 p(m_1 \mid -\alpha, \delta_m, m_{\max}, m_{\min}, \mu, \sigma, \lambda_p) = & \\
 (1 - \lambda_p)\mathcal{B}(m_1 \mid -\alpha, m_{\max}, m_{\min}, \delta_m) & \quad (6.9) \\
 + \lambda_p\mathcal{G}(m_1 \mid \mu, \sigma, m_{\min}, \delta_m). &
 \end{aligned}$$

Note that we have integrated the smoothing kernel directly into the functions of \mathcal{G} and \mathcal{B} .

Besides the primary mass model, the Power-law+Peak model also includes a mass ratio model conditioned on m_1 , which is parameterized as follows:

$$p(q \mid \beta, m_1, m_{\min}, \delta_m) \propto q^\beta S(qm_1 \mid m_{\min}, \delta_m). \quad (6.10)$$

Here, the S function refers to the smoothing kernel, as previously presented in Eq 6.3.

This function is a conditional probability for the secondary mass, represented as qm_1 . The desired average mass spectrum combines both (m_1, m_2) .

Unfortunately, there is no straightforward analytical method to transform the probability density function of Power-law+Peak model to an average mass spectrum. This difficulty is primarily because of the smoothing kernel at the low-mass end in both primary mass and mass ratio parameterizations. Even if we were to ignore this smoothing kernel, the combination of the probability density functions between m_1 and q via $m_2 = qm_1$ would still lead to the m_2 integral that is not computable analytically. Therefore, we have opted for a numerical strategy to acquire the average mass spectrum from Power-law+Peak.

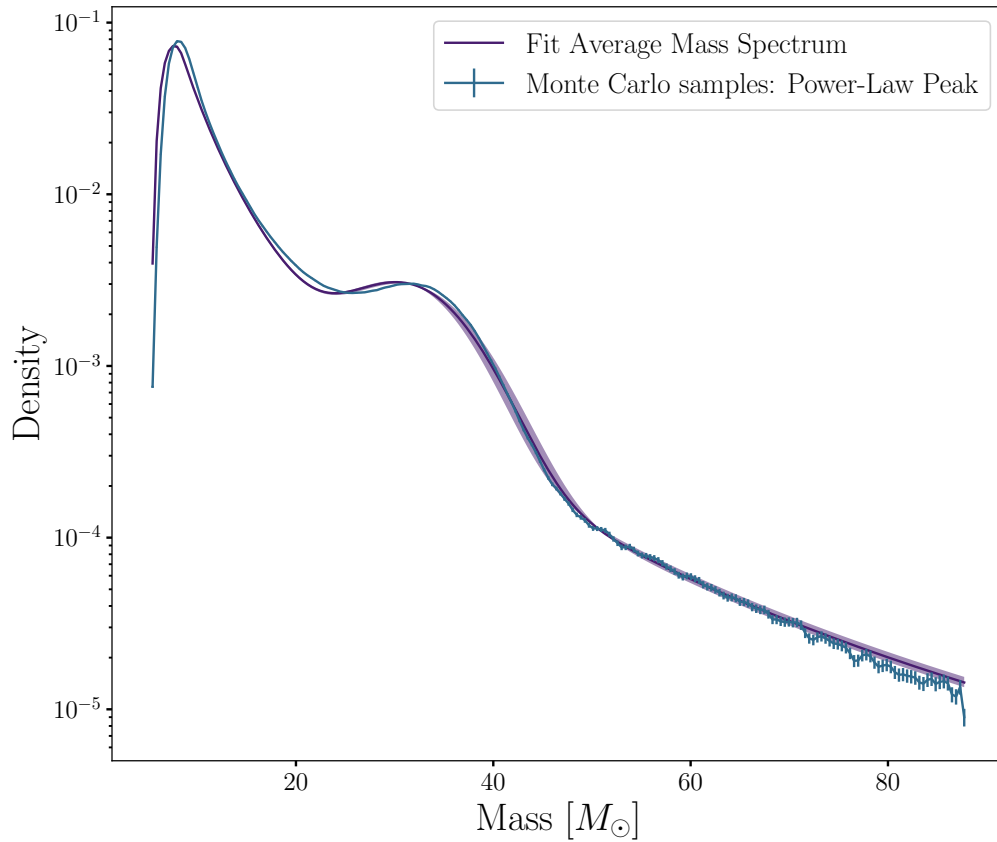


Figure 6.6: The average mass spectrum sampled from GWTC-3 Power-law+Peak model and the best-fit mass spectrum with the fiducial values of our population model. The data points represent the Monte Carlo samples from the Power-law+Peak with best-fit parameters from [13], with the Poisson uncertainty of the Monte Carlo samples. The purple line represents the best-fit power-law function with a Gaussian peak to the sampled average mass spectrum.

We generate Monte Carlo samples for (m_1, m_2) using the fiducial values from Ref. [13] (see Table 6.1), then merge these samples to create a unified average mass spectrum. We then apply a Kernel Density Estimation (KDE) on the combined values of (m_1, m_2) , thereby deriving the average mass spectrum. This spectrum, shown in Figure 6.6, generally preserves the original shape of the Power-law+Peak model. We fit a mix of the power-law and Gaussian model to this numerically-derived average mass spectrum using the same parameterization in Eq 6.9. From this fitting, we obtain the new shape parameters as fiducial values for our population model.

Table 6.1 shows the best-fit parameters from fitting the average mass spectrum. The power-law spectral index, α , rises from around 3.5 to roughly 3.7. This shows a steeper power-law shape in the average mass spectrum than in the primary mass. This is expected, given that the secondary mass is lighter than the primary one, leading to a steeper average power-law. The Gaussian bump shifts from around $33.5 M_\odot$ to approximately $31.5 M_\odot$, with its standard deviation expanding from around $4.6 M_\odot$ to roughly $5.6 M_\odot$. We outline the details of this average mass spectrum fitting procedure in Appendix. 4.

6.4 Population Inference

In this section, we explain the population inference framework implemented in this work. Section 6.4.1 specifies the posterior of the mixing fractions that we aim to infer, taking into account the detection efficiency of the gravitational wave detectors. Section 6.4.2 demonstrates the model averaging approach we have adopted to marginalize over the shape parameters.

Table 6.1: The fiducial shape parameters for our population model, transforming the fiducial values of Power-law+Peak to our average mass spectrum parameterization. The uncertainty in converting the shape parameters from one population model to another can be arbitrarily small, depending on the number of Monte Carlo samples used to construct the KDE for the average mass spectrum. Therefore, we do not include this uncertainty in the table. We do not vary m_{\min} or m_{\max} .

Parameter	Description	Best-fit values	Power-law+Peak values
δm	The δm for the low-mass mass spectrum smoothing	4.62	4.95
m_{\max}	Maximum mass bound for the power-law model	87.73	87.73
m_{\min}	Minimum mass bound for both power-law and Gaussian models	5.06	5.06
$-\alpha$	Spectral index of the power-law	3.66	3.51
μ	Mean of the Gaussian model	31.59	33.56
σ	Standard deviation of the Gaussian model	5.51	4.61
λ_p	Mixing fraction of the Gaussian model	0.034	0.038

6.4.1 Inference Framework

We now discuss how we infer the hyperparameters of the population model. The population inference is determined using a set of N_{obs} gravitational wave events with data \mathbf{d}_i for the i^{th} event. The set of data for the entire catalog will be denoted as $\{\mathbf{d}_i\}$. In this work, we use GWTC Releases 1, 2, and 3 with a selection criteria specified by a False Alarm Rate (FAR) $< 1 \text{ yr}^{-1}$, which includes 73 BBH events. For GWTC-1, we use the re-analysis of the events in GWTC-2.1¹ [286]. Compared to Ref. [13], we do not include GW170817, GW200105_162426, and GW190426_152155, as their chirp masses are below the minimum chirp mass of our population model. For this work, we use only the event posteriors of chirp mass and mass ratio with a combined analysis of C01:IMRPhenomXPHM [287] and C01:SEOBNRv4PHM [288] waveforms.

We define some notation below to align with the notation in the literature. We differentiate between the event parameters, $\boldsymbol{\theta} = \{\mathcal{M}, q\}$, and the population hyperparameters, $\boldsymbol{\Lambda}$. The population hyperparameters encompass the mixing fractions, or relative abundances, which are given by $\boldsymbol{\psi} = \{\psi_{\text{PP}}, \psi_{\text{PG}}, \psi_{\text{GG}}\}$, as well as the shape parameters, symbolized by $\boldsymbol{\lambda} \equiv (-\alpha, \mu, \sigma)$. We use the subscript a denotes which hyperparameter set comes from which subpopulation model, namely $\text{class}_a \in \{\text{class}_{\text{PP}}, \text{class}_{\text{PG}}, \text{class}_{\text{GG}}\}$. For clarity, we use $a = \{\text{PP}, \text{PG}, \text{GG}\}$ to represent each model. The mixing fractions and shape parameters jointly describe the entire population model: $\boldsymbol{\Lambda} \equiv \boldsymbol{\lambda} \cup \boldsymbol{\psi}$.

¹<https://zenodo.org/records/6513631>

Our primary goal is to infer the relative abundance of each subpopulation model, and the mixing fraction posterior is defined as follows [289, 290, 291, 292]

$$p(\boldsymbol{\psi} \mid \{\mathbf{d}_i\}, \{\text{trig}\}, N_{\text{obs}}, \boldsymbol{\lambda}) \propto \frac{p(\boldsymbol{\psi})p(\text{trig} \mid \boldsymbol{\Lambda})^{-N_{\text{obs}}}}{p(\{\mathbf{d}_i\}, \{\text{trig}\}, N_{\text{obs}})} \prod_{i=0}^{N_{\text{obs}}} \mathcal{L}_i^{\text{obs}}. \quad (6.11)$$

Here, we use a Dirichlet prior over the mixing fractions, $\boldsymbol{\psi}$:

$$p(\psi_{\text{PP}}, \psi_{\text{PG}}, \psi_{\text{GG}} \mid \alpha_1, \alpha_2, \alpha_3) = \frac{1}{\mathbb{B}(\alpha_1, \alpha_2, \alpha_3)} \left(\psi_{\text{PP}}^{\alpha_1-1} + \psi_{\text{PG}}^{\alpha_2-1} + \psi_{\text{GG}}^{\alpha_3-1} \right). \quad (6.12)$$

Here, the normalization factor, $\mathbb{B}(\alpha_1, \alpha_2, \alpha_3)$, is a multivariate beta function

$$\mathbb{B}(\alpha_1, \alpha_2, \alpha_3) = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)}{\Gamma(\alpha_1 + \alpha_2 + \alpha_3)}. \quad (6.13)$$

We use a Dirichlet prior to ensure that the mixing fractions sum up to one, $\psi_{\text{PP}} + \psi_{\text{PG}} + \psi_{\text{GG}} = 1$. This reduces the number of parameters we need to infer to two. We opt for a non-informative prior, setting $(\alpha_1, \alpha_2, \alpha_3) = (1, 1, 1)$. This ensures we do not initially favor any specific mixing fraction. The relative abundances, $(\psi_{\text{PP}}, \psi_{\text{PG}}, \psi_{\text{GG}})$, physically represent the fraction of BBHs coming from each mixing scenario.

Our population model adopts the average mass spectrum approach (see Section 6.3.3), so it starts with the shape parameters that are close to the best-fit fiducial values from GWTC-3’s Power-law+Peak. This work primarily focuses on estimating the mixing fraction between the $35M_{\odot}$ Gaussian bump and the power-law population, not accurately estimating the shape parameters. To simplify the computation, the uncertainty in the shape parameters is taken into account through model averaging. We use a set of pre-computed models within a Latin hypercube of shape parameters, which will be detailed in Section 6.4.2.

In Eq. 6.11, we have implicitly marginalized over the total rate of BBH mergers [293]. We have incorporated the concept of “detection” in the formalism by introducing the trigger term, $\{\text{trig}\}$, into our notation. This term represents the criteria that determines if an individual event is selected, typically based on a specific signal-to-noise threshold of the observational instrument. Mathematically, the probability of detection given an actual realization of data is defined as

$$p(\text{trig}|\mathbf{d}_i) = \begin{cases} 0 & \rho(\mathbf{d}_i) < \rho_{\text{threshold}}, \\ 1 & \rho(\mathbf{d}_i) \geq \rho_{\text{threshold}}, \end{cases} \quad (6.14)$$

where $\rho(\mathbf{d}_i)$ defines some deterministic calculation on the data (the signal-to-noise ratio, for example) which classifies data as containing an event or not, based on some threshold value $\rho_{\text{threshold}}$. This notation is inherited from Ref. [294] which gives a detailed explanation of work originally derived in past literature [289, 290, 291, 292]. We use $p(\text{trig} | \mathbf{\Lambda})$ to represent the detection efficiency, which quantifies the proportion of detectable sources based on the population model represented by population parameters, $\mathbf{\Lambda}$. The detection efficiency, $p(\text{trig} | \mathbf{\Lambda})$, can be explicitly expressed as follows:

$$\begin{aligned} p(\text{trig} | \mathbf{\Lambda}) &= \int d\mathbf{d} \int d\boldsymbol{\theta} p(\text{trig} | \mathbf{d})p(\mathbf{d} | \boldsymbol{\theta})p(\boldsymbol{\theta} | \mathbf{\Lambda}) \\ &= \int d\boldsymbol{\theta} p(\text{trig} | \boldsymbol{\theta})p(\boldsymbol{\theta} | \mathbf{\Lambda}). \end{aligned} \quad (6.15)$$

Here, $p(\text{trig} | \boldsymbol{\theta})$ is known as the detection probability and depends on the event parameters, $\boldsymbol{\theta}$, not population parameters, $\mathbf{\Lambda}$. Note that the concept of detection fundamentally relies exclusively on the data itself, $p(\text{trig} | \mathbf{d})$ defined above, and is only connected to the event parameters $\boldsymbol{\theta}$ through the event likelihood $p(\mathbf{d} | \boldsymbol{\theta})$. The detection probability implicitly

marginalizes over this hierarchical relationship. Mathematically, this quantity is defined as

$$p(\text{trig} | \boldsymbol{\theta}) \equiv \int p(\text{trig} | \mathbf{d})p(\mathbf{d} | \boldsymbol{\theta})d\mathbf{d}. \quad (6.16)$$

As an approximation to this integral, we use the calculation for detection probability graphically shown in Fig. (3) of Ref. [295] denoted as $p_{\text{det}}(\boldsymbol{\theta})$ in that work, which is pre-marginalized over extrinsic parameters (sky location and orientation) using standard distributions (uniform on the sky and uniform in orientation) [296, 297, 298]. The details of that calculation can be found in Ref. [295], for example, but essentially amounts to evaluating this sky-location-averaged detection statistic as a function of the SNR for an optimally oriented binary. We neglect the BH spin in this calculation, which should have a small effect on the population-averaged detection rate or the sensitive volume. While spins can drastically increase the detectability of individual events (e.g. [299, 300]), the quantity of interest in hierarchical inference is Eq. 6.15, which includes information about the distribution of spins coming from a population model. As the latest inference on the spin distributions of BBH indicate a distribution clustered around $\chi_{\text{eff}} \sim 0$, this factor should be negligible. See Ref. [299] and their calculation of the impact on the detectable volume for isotropically distributed spins (the distribution most consistent with LVK’s results), which indeed shows negligible impact. This means $p(\text{trig} | \boldsymbol{\theta})$ simply involves an integral over the mass parameters \mathcal{M} and q .

We evaluate Eq. 6.15 with a fixed Power Spectral Density (PSD) function for all events, where we use the analytic `AdVMidHighSensitivityP1200087` [301, 302] PSD throughout this work, but we also discuss the impact of using different PSDs in Appendix .5. The pre-marginalized approach in Ref. [295] factors out the detector dependent quantities

from SNR on $p_{\text{det}}(\boldsymbol{\theta})$ is explained in Ref. [268, 303, 304]. In this work, injections with expected SNR less than 8 ($\rho_{\text{threshold}} = 8$) are not considered detected. To calculate the SNR, we employed the IMRPhenomD [305, 306, 307]. We discuss the priors used in calculating $p_{\text{det}}(\boldsymbol{\theta})$ in Appendix .5. While this semi-analytic method is an approximation of the more accurate method (which involves injecting signals from known distributions into the entire detection pipeline on top of real data), it has been utilized extensively in the literature (e.g., [234, 308, 279, 299]) and shown to accurately capture the salient features of selection bias [223, 298, 309, 310]. The main approximations of this method relevant to this study relate to the non-Gaussian and non-stationarity of gravitational wave detectors, which are reasonable approximations in current detector networks.

In practice, we numerically estimate the detection efficiency by Monte Carlo sampling the event parameters, $\boldsymbol{\theta}$, under a given set of population parameters, $\boldsymbol{\Lambda}$:

$$p(\text{trig} \mid \boldsymbol{\Lambda}) \approx \frac{1}{S} \sum_{i=0}^S p(\text{trig} \mid \boldsymbol{\theta}_i); \quad (6.17)$$

$$\boldsymbol{\theta}_i \sim p(\boldsymbol{\theta} \mid \boldsymbol{\Lambda}),$$

where we generate $S = 500\,000$ samples of event parameters according to $p(\boldsymbol{\theta} \mid \boldsymbol{\Lambda})$. For the event parameters $\boldsymbol{\theta}$, we use primary/secondary masses and luminosity distance, L . For the primary and secondary masses, we sample from the population model. For the luminosity distance, we sample with a prior of $p(L) \propto L^2$, which is uniform in volume.

The mixture model construction allows us to simplify the estimation of detection efficiency to the sum of detection efficiency of each subpopulation model:

$$p(\text{trig} \mid \boldsymbol{\Lambda}) = \sum_{a=\{\text{PP,PG,GG}\}} \psi_a p(\text{trig} \mid \boldsymbol{\lambda}, \text{class}_a); \quad (6.18)$$

$$p(\text{trig} \mid \boldsymbol{\lambda}, \text{class}_a) = \int d\boldsymbol{\theta} p(\text{trig} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{\lambda}, \text{class}_a).$$

This substantially simplifies computing the detection efficiency. In practice, we pre-compute Monte Carlo samples from subpopulation models with a fixed set of shape parameters, $\boldsymbol{\lambda}$. Therefore, we can quickly compute the $p(\text{trig} \mid \boldsymbol{\Lambda})$ via a weighted sum by varying mixing fractions, $\boldsymbol{\psi}$.

For the likelihood of each data set \mathbf{d}_i , $\mathcal{L}_i^{\text{obs}} \equiv p(\mathbf{d}_i \mid \boldsymbol{\Lambda})$, we ultimately compare the distribution of the event parameters to the predictions from our population model. For each event, i , the likelihood is

$$\begin{aligned} \mathcal{L}_i^{\text{obs}} &= \sum_{a=\{\text{PP,PG,GG}\}} p(\text{class}_a \mid \boldsymbol{\Lambda}) p(\mathbf{d}_i \mid \text{class}_a, \boldsymbol{\Lambda}), \\ &= \sum_{a=\{\text{PP,PG,GG}\}} \psi_a p(\mathbf{d}_i \mid \text{class}_a, \boldsymbol{\lambda}), \end{aligned} \quad (6.19)$$

where $p(\text{class}_a \mid \boldsymbol{\Lambda})$ is the prior probability that an event belongs to each subpopulation model. Eq 6.19 describes the likelihood of the data marginalized over the class of the event.

We can explicitly write it as

$$p(\mathbf{d}_i \mid \text{class}_a, \boldsymbol{\lambda}) = \int d\boldsymbol{\theta} p(\mathbf{d}_i \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \text{class}_a, \boldsymbol{\lambda}), \quad (6.20)$$

where the integral can be approximated via importance sampling [311]

$$\begin{aligned} \int d\boldsymbol{\theta} p(\mathbf{d}_i \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \text{class}_a, \boldsymbol{\lambda}) &\approx \\ \frac{1}{S} \sum_{c=0}^S \frac{p(\boldsymbol{\theta}_c \mid \text{class}_a, \boldsymbol{\lambda})}{\pi(\boldsymbol{\theta}_c)} & \end{aligned} \quad (6.21)$$

with

$$\boldsymbol{\theta}_c \sim p(\boldsymbol{\theta} \mid \mathbf{d}_i). \quad (6.22)$$

Here, $p(\boldsymbol{\theta} \mid \mathbf{d}_i)$ is the event posterior provided by LVK's template fitting. We need to divide the posterior by the event parameter prior, $\pi(\boldsymbol{\theta})$, to get the likelihood for each event. The

event parameter priors used by LVK are

$$\begin{aligned}\pi(\mathcal{M}) &\propto \mathcal{M} \\ \pi(q) &\propto \frac{(1+q)^{2/5}}{q^{6/5}}.\end{aligned}\tag{6.23}$$

Finally, $p(\boldsymbol{\theta}_c \mid \text{class}_a, \boldsymbol{\lambda})$ represents evaluating the event posterior samples on the histogram likelihoods shown in Figure 6.2. Thus, Eq 6.21 simply states that we evaluate the Monte Carlo samples of event likelihoods on the numerical probability density derived from the histograms.

6.4.2 Model averaging

We describe our numerical method to integrate the population posterior over $\boldsymbol{\psi}$ with a fixed set of shape parameters $\boldsymbol{\lambda}$ in Eq 6.11. Our primary goal is to infer $\boldsymbol{\psi}$, not accurately estimate the shape parameters, $\boldsymbol{\lambda}$. One way to marginalize over the uncertainty of $\boldsymbol{\lambda}$ in this case is through model averaging. We run a fixed set of 1,000 choices for the shape parameters, $\boldsymbol{\lambda}$, and obtain the MCMC posterior of $\boldsymbol{\psi}$ for each mixture model. For the model averaging approach, we approximate the marginalization by treating each $\boldsymbol{\lambda}$ as a model, and we use the posterior of $\boldsymbol{\lambda}$ to weight the contribution of each model to the population posterior through a Monte Carlo sum with a discrete set of $\boldsymbol{\lambda}$:

$$\begin{aligned}p(\boldsymbol{\psi} \mid \{\mathbf{d}_i\}, \{\text{trig}\}, N_{\text{obs}}) &\simeq \\ \frac{1}{S} \sum_{j=1}^S p(\boldsymbol{\psi} \mid \{\mathbf{d}_i\}, \{\text{trig}\}, N_{\text{obs}}, \boldsymbol{\lambda}^j) &\times\end{aligned}\tag{6.24}$$

$$p(\boldsymbol{\lambda}^j \mid \{\mathbf{d}_i\}, \{\text{trig}\}, N_{\text{obs}})$$

with

$$\boldsymbol{\lambda}^j \sim p(\boldsymbol{\lambda}).\tag{6.25}$$

The shape parameters in the λ space have a prior volume $\alpha \sim \mathcal{U}(1, 6)$, $\mu \sim \mathcal{U}(25, 40)$, and $\sigma \sim \mathcal{U}(3, 8)$. We sample λ using a Latin hypercube, maximizing coverage of the parameter space. This effectively searches the hyperspace of λ and marginalizes out uncertainty in the shape parameters. The posterior $p(\lambda \mid \{\mathbf{d}_i\}, \{\text{trig}\}, N_{\text{obs}})$, is obtained by evaluating the model evidence for all events, where we assume a uniform, model prior for each choice of $(-\alpha, \mu, \sigma)$.

6.5 Results

In this section, we present the inference results of our population model. First, in Section 6.5.1, we discuss the inference results of the population model with fixed shape parameters, as obtained from the average mass spectrum approach. Subsequently, we present the results of model averaging in Section 6.5.2, where we marginalize over the shape parameters.

6.5.1 Fiducial model

We have obtained the fiducial shape parameters, $(-\alpha, \mu, \sigma, \delta_m, m_{\text{max}}, m_{\text{min}})$, from the average mass spectrum as detailed in Section 6.3.3. The posterior distribution for the mixing fraction is shown in Figure 6.7. The 95% posterior confidence intervals for the mixing fractions are $(\psi_{\text{PP}} = 92.9_{-11.1}^{+2.2} \%, \psi_{\text{PG}} < 8.7\%, \psi_{\text{GG}} = 7.1_{-3.0}^{+4.8} \%)$. Interestingly, the mode of the relative abundance of the PG mixing, ψ_{PG} , is consistent with zero. This indicates some evidence for the separation of the two populations, based on the fiducial shape parameters.

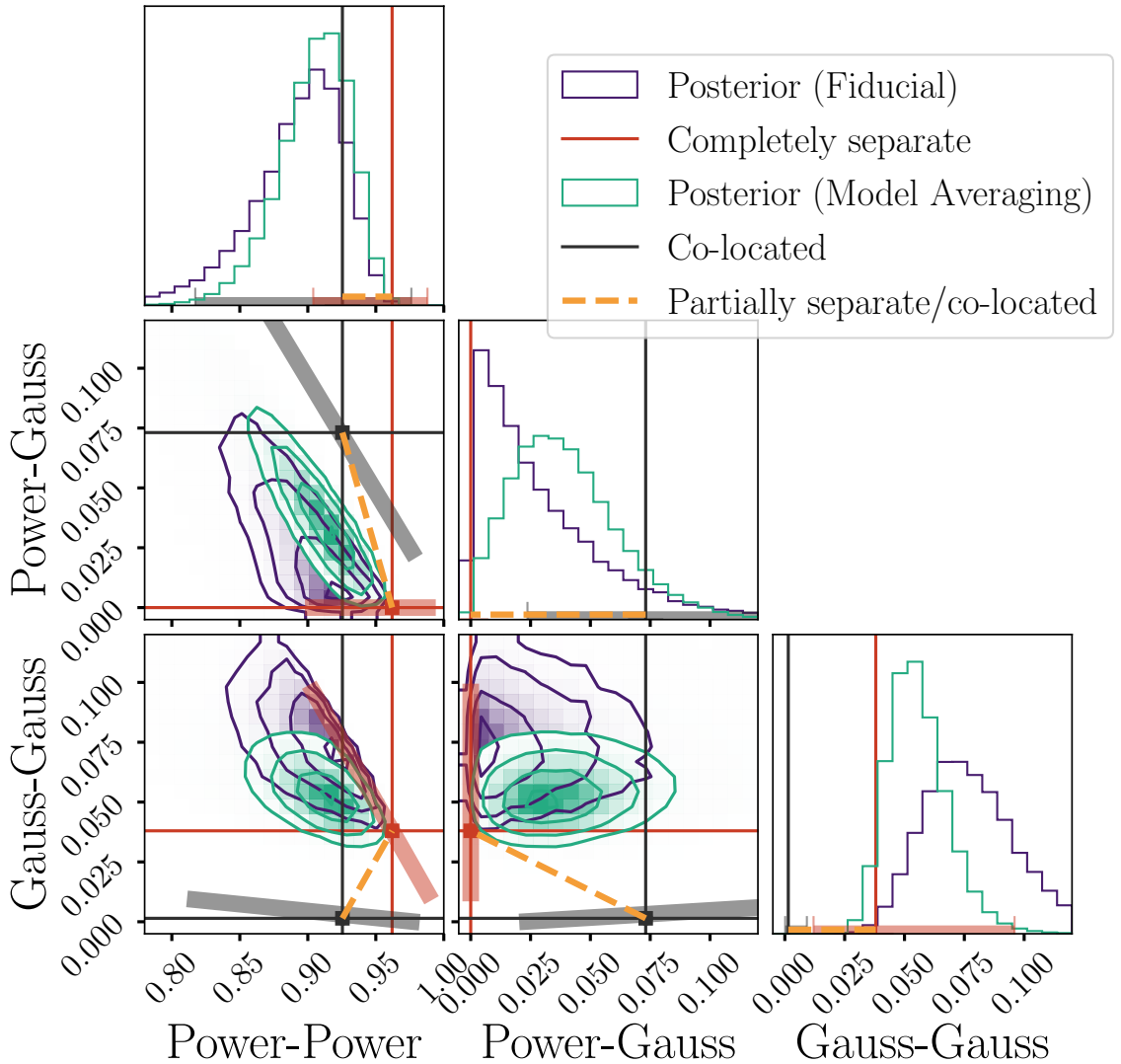


Figure 6.7: Posterior probability for mixing fraction parameters, ψ under different model assumptions (“Fiducial” model in purple and the “Model Averaging” in green). Two extreme hypothetical scenarios are also shown: (1). Red error bars show the case where the Gaussian bump is completely separate from the power-law population. (2). Black error bars show the case where the Gaussian bump and power-law populations are co-located. These hypothetical scenarios are defined using $\lambda_{\text{peak}} = 3.8^{+5.8\%}_{-2.6\%}$ for the Gaussian bump reported in LVK [13]. The orange dashed lines represent the “Partially Separate/Co-located” scenario, showing a situation in which a portion of the Gaussian bump black holes is co-located with the power-law distribution, while the remainder is separate.

To illustrate two extreme situations, we define two hypothetical population model scenarios: “Completely separate” and “Co-located.” In the “Completely separate” scenario, we assume that the power-law and Gaussian populations are entirely separate, resulting in no PG mixing. The relative abundances of the PP and GG populations reflect the fractions of power-law and Gaussian populations in the single-mass distribution, respectively. For the completely separate scenario, we adopt $(\psi_{\text{PP}}, \psi_{\text{PG}}, \psi_{\text{GG}}) = (96.2^{+2.6}_{-5.8}\%, 0.0\%, 3.8^{+5.8}_{-2.6}\%)$. The choice of $\psi_{\text{GG}} = 3.8^{+5.8}_{-2.6}\%$ is based on the relative abundance of the Gaussian bump from the Power-law+Peak model, with the 90% credible intervals reported in Ref. [13] indicating a relative abundance of the Gaussian bump, $\lambda_{\text{peak}} = 3.8^{+5.8}_{-2.6}\%$. Assuming the Gaussian bump population is separate from the power-law, this implies a GG mixing with a relative abundance of approximately $3.8^{+5.8}_{-2.6}\%$ and a zero mixing abundance, $\psi_{\text{PG}} \approx 0\%$. In the “Co-located” scenario, we assume the power-law and Gaussian populations are completely mixed together, resulting in the mixing fraction of PG equal to $2 \times \lambda_{\text{peak}}(1 - \lambda_{\text{peak}})$, giving $(\psi_{\text{PP}}, \psi_{\text{PG}}, \psi_{\text{GG}}) = ((1 - \lambda_{\text{peak}})^2, 2 \times \lambda_{\text{peak}}(1 - \lambda_{\text{peak}}), \lambda_{\text{peak}}^2) \approx (92.5^{+5.1}_{-10.8}\%, 7.3^{+11.7}_{-5.1}\%, 0.1^{+0.8}_{-0.1}\%)$. Interestingly, Figure 6.7 suggests that the posterior distribution from the fiducial model prefers the “Completely separate” scenario over the “Co-located” scenario, although the error bars from each scenario remain substantial.

Figure 6.8 presents the predicted primary/secondary mass functions and mass ratio based on our fiducial inference. For comparison, we also include the Power-law+Peak model with its fiducial parameters obtained from Ref. [13] (also see Table 6.1). The primary mass functions show good agreement, which is expected given that in Section 6.3.3 we have constructed our average mass spectrum to match the Power-law+Peak model. The

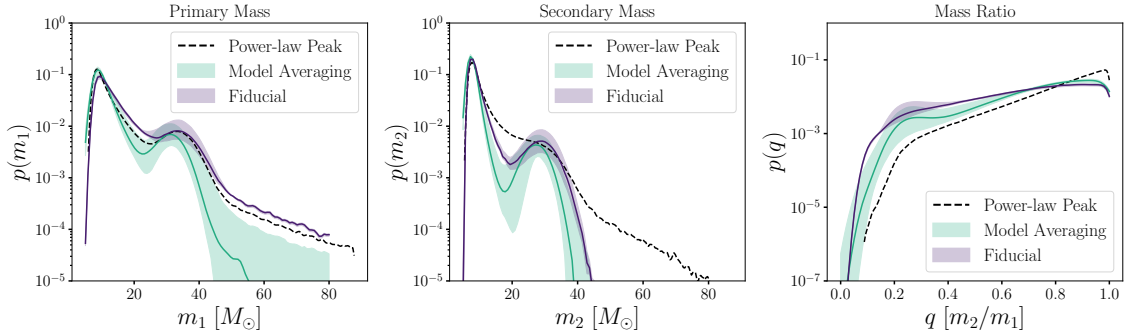


Figure 6.8: Predicted primary/secondary mass and mass ratio functions from the model averaging inference (in light green) and the fiducial inference (in purple). The solid lines represent the MAP and the shaded areas represent the 95% confidence intervals. The light green lines represent predicted functions sampled from the posterior probability of both ψ and λ . The underlying black dashed lines represent the fiducial model of Power-law+Peak from GWTC-3, with the corresponding fiducial parameters detailed in Table 6.1. The 95% confidence interval for the “Fiducial” inference reflects only the posterior uncertainty in ψ and does not include uncertainty regarding λ . As we fix the minimum mass (m_{\min}) and the maximum mass (m_{\max}), the shape uncertainty at the low and high mass ends is not incorporated.

secondary mass functions exhibit some differences. The bump in m_2 is approximately at $\sim 30M_{\odot}$ for both population models. However, the power-law component in the Power-law+Peak is comparatively flatter. This discrepancy might arise from an inherent difference between these two models. The Power-law+Peak models the m_2 via a power-law mass ratio, while our model assumes an average mass spectrum for both m_1 and m_2 . It could mean that the secondary mass spectrum would appear much sharper under our model’s assumptions. Nevertheless, due to the limited dataset size of GW events, inferring the massive end of the mass spectrum remains highly uncertain. Since the primary focus of this paper is on estimating the mixing fraction, we do not emphasize the differences at the tail of the mass spectrum nor trying to infer m_{\max} .

6.5.2 Model averaging

Figure 6.7 compares the posterior of the mixing fractions, $p(\boldsymbol{\psi} \mid \{\mathbf{d}_i\}, \{\text{trig}\}, N_{\text{obs}})$, (“Model Averaging” in green) with the posterior from the fiducial shape parameters (“Fiducial”), $p(\boldsymbol{\psi} \mid \boldsymbol{\lambda}_{\text{fid}}, \{\mathbf{d}_i\}, \{\text{trig}\}, N_{\text{obs}})$. It shows a shift in the posterior mode to approximately the 68 – 95% confidence contour. Additionally, the posterior width for the mixing fractions narrows, suggesting that the fiducial shape values do not provide the best fit to the data and has a lower model evidence. Otherwise, model averaging would result in an increased width of the posterior. This is expected as our population model differs from the Power-law+Peak model, so the fiducial shape parameters do not provide the best fit to the data. The uncertainty in the predicted mass spectrum, as illustrated in Figure 6.8, increases under the “Model Averaging” approach, particularly due to the varying spectral index of the power-law. This increase in uncertainty suggests that, with a flexible power-law model (with a varying spectral index), our mixture model gets a better fit to the data. This better fit is attributed to the fact that both PP and PG models can better explain the observed data, leading to narrower posteriors of the mixing fractions.

The posterior for the mixing fractions, $(\psi_{\text{PP}}, \psi_{\text{PG}}, \psi_{\text{GG}}) = (91.9_{-6.8}^{+3.2}\%, 3.1_{-3.1}^{+5.0}\%, 5.0_{-1.7}^{+3.2}\%)$, indicates that at a 95% confidence level, approximately 3.1% of the binaries in the catalog can be attributed to the mixing of the populations. In Figure 6.7, compared with the “Completely separate” scenario (red error bars) and the “Co-located” scenario (black error bars), we observe that, even with varying shape parameters, the model averaging result still shows a preference for the “Completely separate” scenario between the Gaussian and power-law populations. Nonetheless, there is a notable shift in the mode of PG mixing

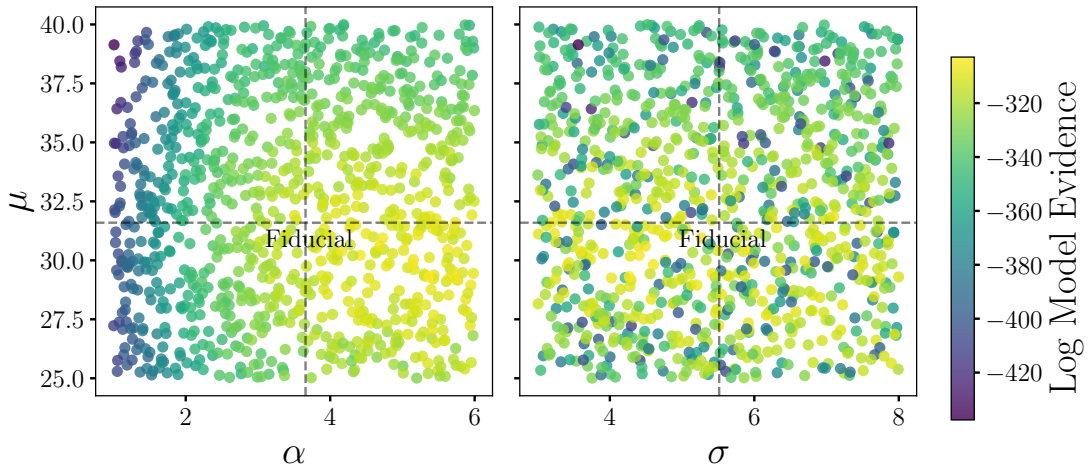


Figure 6.9: The model evidences from 1,000 population inferences utilized in the “Model Averaging”, namely, $p(\boldsymbol{\lambda} \mid \{\mathbf{d}_i\}, \{\text{trig}\}, N_{\text{obs}})$, where we treat each set of $\boldsymbol{\lambda}$ as a model. The colors indicate the model evidence for each of the shape parameters, $(-\alpha, \mu, \sigma)$. There is no obvious correlation between σ and μ , but there is some weak correlation between α and μ with correlation $\simeq -0.3$.

to a slightly higher value, $\psi_{\text{PG}} = 3.1^{+5.0}_{-3.1}\%$, which suggests that the PG mixing posterior now locates itself between the “Completely separate” and “Co-located” scenarios.

Figure 6.9 shows the model evidences for each population model used in the model averaging approach. The peaks of the model evidences, compared to the fiducial values from the average mass spectrum (which is at the center of the prior volume), are shifted towards slightly lower μ and a steeper spectral index (higher $-\alpha$). The shape parameters at the maximum model evidence are $(-\alpha, \mu, \sigma) = (-5.3675, 29.3125, 4.2675)$, which represents a slight shift in the shape parameters compared to the fiducial values. There is a $\simeq -0.3$ correlation between the α and the location of the Gaussian bump, μ . Namely, with a steeper slope of the power-law, the Gaussian bump has to move to a lower mass to compensate.

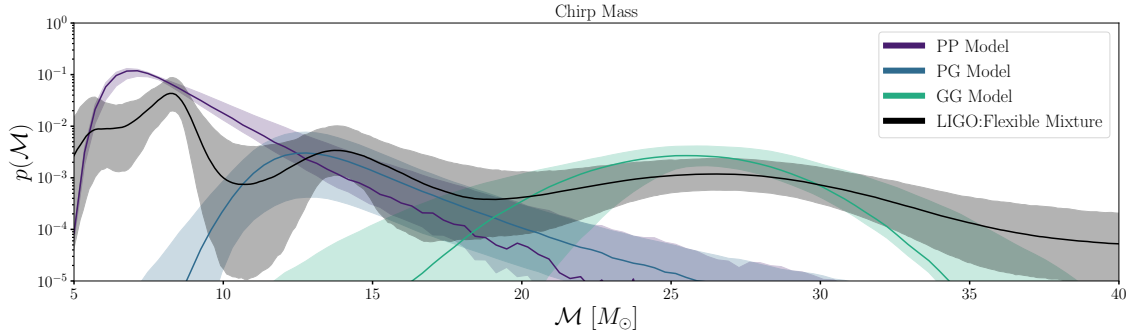


Figure 6.10: The chirp mass spectra predicted by the subpopulations in our population model: PP (purple), PG (blue), and GG (green). The shaded regions represent the 95% confidence intervals, and the subpopulations are normalized to have $\psi_{\text{PP}} + \psi_{\text{PG}} + \psi_{\text{GG}} = 1$. The underlying mass spectrum (black) is from the Flexible Mixture model [14], and utilizes fitting results from Ref. [13].

Figure 6.10 presents the predicted chirp mass spectrum from each subpopulation model and compares these predictions with the Flexible Mixture model from Ref. [13]. The BBH chirp mass spectrum’s three-peak structure is captured by the PP ($\mathcal{M} \sim 8M_{\odot}$), PG ($\mathcal{M} \sim 14M_{\odot}$), and GG ($\mathcal{M} \sim 28M_{\odot}$) subpopulations. We do not vary m_{min} , resulting in an overconfidence in the low-mass of the PP model compared with the predictions of the Flexible Mixture model. However, the relative abundances of both the PG and GG subpopulations align well with the second and third peaks of the chirp mass spectrum.

Our inference results suggest a high fraction of GG mixing and a low fraction of PG mixing ($\psi_{\text{GG}} > \psi_{\text{PG}}$), indicating that the $35M_{\odot}$ Gaussian bump BHs are likely separate from the rest of the population and forming BGBHs. Existing theoretical models for the Gaussian bump thus need to account for the separation of the Gaussian bump black holes from the rest of the black hole population.

6.6 Discussion

6.6.1 High Proportion of Gaussian-Gaussian BBHs.

Figure 6.7 shows a substantial fraction of Gaussian bump black holes merging with other objects from the Gaussian bump (BGBHs), implying that the Gaussian bump population is almost distinct from the power law population. This high fraction of BGBHs presents challenges to existing black hole formation theories. There are at least three explanations which could potentially explain the result: (1). A cluster of stars forming simultaneously in cosmic time, undergoing supernova explosions, and consequently producing dense environments of $\sim 35M_{\odot}$ stellar remnants. (2). Mass segregation [312] in a stellar cluster, causing $\sim 35M_{\odot}$ black holes to gravitate towards the cluster’s core and merge predominantly with similar-mass black holes. (3). A distinct population of $\sim 35M_{\odot}$ black holes with a distinct spatial distribution, or a different set of host halos, from the black holes in the power-law subpopulation.

The prevailing explanation for the $\sim 35M_{\odot}$ Gaussian bump is PPSNe, or the mass accumulation before the pair-instability cutoff at $\gtrsim 40M_{\odot}$. The population of BGBHs challenges our understanding of how PPSNe binaries form, particularly how PPSNe can generate BGBHs without becoming bound to lighter black holes. One possible explanation is the formation of clusters composed exclusively of PPSNe black holes or PPSNe remnants. However, the specific process behind such cluster formation has not been thoroughly explored or discussed in the literature. Another explanation is that these massive BGBHs could come from binary star systems (e.g., see Ref. [313]), where both stars are already very

massive and in the range to produce PPSNe. However, the kicks from supernova explosion of massive stars will likely destroy the binary system and make them unbound.

Next, mass segregation (see Ref. [314]) within stellar clusters might lead to the formation of BGBHs. Heavier black holes sinking to the cluster's center would merge with similar-mass counterparts. This requires an explanation for the production of a high abundance of $\sim 35M_{\odot}$ black holes, potentially through PPSNe or PBHs acting as gravitational centers around which globular clusters form (e.g., Ref. [315]).

Another possibility is Pop III stars [316, 262], forming massive stars at high redshifts that evolve into $\sim 35M_{\odot}$ black hole at the same time (e.g., Ref. [262]). To form BGBHs, this scenario requires these stars to form in clusters and evolve simultaneously into supernovae, thus forming BBHs within the same population. This hypothesis could be further tested through its contributions to cosmic reionization around $z \simeq 6$. However, it's unclear why Pop III stars would preferentially form black holes around $\sim 35M_{\odot}$, but not beyond $40M_{\odot}$, given the absence of pair-instability limitations at such low-metallicity.

Another explanation for a high fraction of BGBHs is PBHs, which are distributed more like the dark matter halo and thus distinct from luminous matter. If such PBHs exist with masses around $\sim 35M_{\odot}$, they would predominantly merge within their group. Gravitational microlensing constrains the fraction of dark matter in the form of PBHs to be less than 10% of the halo [317, 318, 319, 320, 321]. However, the merger rate of PBHs remains highly uncertain (e.g., see Ref.[258, 259]), making it difficult to predict if ψ_{GG} is consistent with PBHs.

6.6.2 Limitation on the Interpretation of the Mixing Fraction Posterior.

A limitation in comparing two hypothetical scenarios using λ_{peak} to estimate mixing fractions is that the Power-law+Peak model only measures the Gaussian bump’s fraction in the primary mass spectrum, not across *all black holes* in the Universe. Thus, using λ_{peak} as a proxy for the bump’s overall abundance may not be directly applicable. However, this approach likely provides a conservative estimate for ψ_{GG} of “Completely separate,” since the primary mass in a BBH is heavier, suggesting ψ_{GG} could be overestimated using λ_{peak} . Our interpretation of Figure 6.7 remains unchanged, the mixing fraction posterior aligning better with the “Completely separate” scenario.

Even if we assume there is no robust estimate of λ_{peak} , by definition, the “Completely separate” scenario would yield $\psi_{\text{PG}} = 0\%$, which is more consistent with our inference results than the “Co-located” scenario, which requires $\psi_{\text{PG}} > \psi_{\text{GG}}$, given the power-law abundance is much higher than the Gaussian bump. To make our $\psi_{\text{GG}} = 5.0^{+3.2\%}_{-1.7\%}$ posterior consistent with the “Co-located” case, the relative abundance of the Gaussian bump would need to be approximately 18 – 28% of all black holes which form BBHs, a significant difference from the λ_{peak} measurement which would be obvious in the inferred primary mass spectrum from GWTC-3. We therefore argue that our inference still suggests that a separate population causes the Gaussian bump in the GWTC-3 catalog.

We assume fixed $(m_{\text{min}}, m_{\text{max}})$, which restricts the explanatory power of the PP and PG models. Secondly, we categorize the black hole population into either a Gaussian bump or a power-law, but the true BBH population might be more complex, containing more than two subpopulations. Another limitation of our mixing approach is that it does not

consider second generation mergers (e.g., [322]), which could be important for the massive end of the mass spectrum. We also did not model the common envelope or isolated channel BBHs. However, we can interpret PP, PG, and GG as isolated channels going through different IMF and metallicity environments, e.g., Ref. [323] can produce PG BBHs (30-10 M_{\odot}) with low-metallicity progenitors with initial $q < 0.5$.

6.7 Conclusion

In this paper, we explore the substructure within the black hole mass spectrum, specifically focusing on the $m_1 \sim 35M_{\odot}$ Gaussian bump in the primary mass spectrum and the $\mathcal{M} \sim 14M_{\odot}$ peak in the BBH chirp mass spectrum. We investigate these substructures through a two-population mixing scenario, examining a power-law and Gaussian population of black holes in the Universe. We define three mixing scenarios: PP binaries, where the power-law population mixes with itself; PG binaries, involving a mix between the power-law population and the Gaussian bump black hole population; and GG binaries, where Gaussian bump black holes merge with themselves. A mixture model was developed to measure the relative abundance of each scenario. The fiducial inference results, aligning with the primary mass spectrum of the Power-law+Peak model without varying the shape parameters of the power-law and Gaussian bump, suggest ($\psi_{\text{PP}} = 92.9^{+2.2}_{-11.1}\%$, $\psi_{\text{PG}} < 8.7\%$, $\psi_{\text{GG}} = 7.1^{+4.8}_{-3.0}\%$). As we vary the shape parameters, including the spectral index of the power-law and the location and width of the bump, our model averaging results indicate ($\psi_{\text{PP}} = 91.9^{+3.2}_{-6.8}\%$, $\psi_{\text{PG}} = 3.1^{+5.0}_{-3.1}\%$, $\psi_{\text{GG}} = 5.0^{+3.2}_{-1.7}\%$). Both sets of results highlight a relatively low PG mixing fraction and a high GG binary mixing fraction, indicating a pref-

ference in the GWTC-3 catalog data for a “Completely separate” scenario. This suggests that $35M_{\odot}$ Gaussian bump black holes are likely separate from the rest of the population.

Our population model’s predicted chirp mass spectrum and the relative abundance of each mixing scenario align well with the Flexible Mixture model results presented in Ref. [13]. The second chirp mass peak at $\mathcal{M} \sim 14M_{\odot}$ closely matches the relative abundance of PG binaries, suggesting partial mixing between the power-law and Gaussian bump populations. Although these populations are likely separated, a fraction mixes, giving rise to the second chirp mass peak.

Most past formation channels explaining the $35M_{\odot}$ Gaussian bump focus on the primary mass spectrum rather than the 2D BBH mass space. For instance, PPSNe are a popular mechanism for the Gaussian bump, facing challenges in explaining BGBH formation without pairing with lighter black holes. One possibility is that such PPSNe Gaussian bump black holes are typically found in star clusters, where mass segregation might facilitate their merger with similar black holes. However, the likelihood of mass segregation and the fraction of Gaussian bump black holes within star clusters remain uncertain. Other formation channels, such as black holes originating from low-metallicity Pop III stars or primordial black holes, could also account for the high fraction of BGBHs, given their separation from other high-metallicity stellar-origin black holes. These channels might explain the separate population of BGBHs, but the precise mechanisms for producing $\sim 35M_{\odot}$ black holes remain unknown and challenging to pinpoint.

We also acknowledge limitations in our population inference, such as the inflexibility of the power-law population model. However, we anticipate that enhancing model

flexibility will likely not significantly alter our current interpretations, due to large error bars and the GWTC-3 catalog's limited size. We suggest that future work on the formation of Gaussian bump black holes should consider the separation of this population, and the potential channels for forming BGBHs.

Facilities: LIGO, Virgo

Software: PyMC5 [324], `scipy` [325], `bilby` [326, 327], `pycbc` [302], `arviz` [328], `corner.py` [329], `matplotlib` [330].

Chapter 7

Conclusions

In this thesis, we have presented three different applications of Bayesian modeling in astrophysics: the Gaussian process finder of the damped Lyman- α absorbers (Chapter 2 and Chapter 3), the multi-fidelity emulator of the matter power spectrum (Chapter 4 and Chapter 5), and the Bayesian hierarchical inference of the binary black hole mass spectrum (Chapter 6). A common theme in these applications is the use of the Bayesian modeling to understand the underlying physics of the data in a model-driven way.

The Gaussian process finder demonstrates a way to probabilistic model the detection of DLAs at a single quasar spectrum level. This hybrid approach combines the atomic physics and the data-driven quasar emission model to infer the DLA parameters. The common difficulty in applying parametric Bayesian modeling to low-level data analysis is the complexity of the likelihood function or the data noise model, which is usually hard to model. The Gaussian process finder provides a way to model the data noise model in a non-parametric way through machine learning.

The multi-fidelity emulator demonstrates a way to make the complex and slow theory model applicable to the fast and efficient data analysis. The common difficulty in applying the theory model to the data analysis is the computational cost of the theory model. In most of the cosmological data analysis, the theory model is usually a simulation code, which is computationally expensive to run. The multi-fidelity emulator provides an efficient way to predict the simulation function at any input parameters by using the data from the lower fidelities. This opens a new way to apply the traditionally slow and expensive simulation codes to various data analysis tasks, such as the parameter inference, the model selection, and the optimization.

The population inference of the binary black hole demonstrates a way to build a hierarchical model to understand a catalog of events. The common difficulty in applying the population model to the data analysis is the selection bias due to the detection efficiency. The population inference provides a way to propagate the measurement uncertainty from each individual event to the population level and simultaneously correcting the selection bias of the catalog.

The above three Bayesian models deal with three different levels of data analysis: the low-level single observation analysis, the catalog-level population analysis, and the theory-level simulation analysis. The Gaussian process finder is a relatively low-level data analysis, where the goal is to build a DLA catalog from inferring the DLA properties from each quasar spectrum. The population inference tells us how to properly propagate a catalog of detections to the physical parameters of the true population. Finally, the emu-

lator approach brings the complex theory model to the data analysis level, allowing us to interpret the data from the first principle.

Science-wise, there are some future directions to extend the works in this thesis. For the Gaussian process finder, the DLA catalog has been proved to be useful in constraining the neutral content of the Universe and studying the systematic bias of the Lyman- α forest cosmology. Though DLAs are well-studied in the terms of cosmology, its lower column density cousins, the Lyman limit systems (LLSs), are less studied. The LLSs are the systems with column density $10^{17.2} \leq N_{\text{HI}} < 10^{20.3} \text{ cm}^{-2}$, and they are the transition between the DLAs and the Lyman- α forest. As the future Lyman- α forest P1D measurements from the DESI survey are coming, the cosmological systematic bias from the LLSs will be more important. Nevertheless, the LLSs are less studied because they are more abundant and harder to detect at the level of low-resolution SDSS spectra. If we can extend the Gaussian process finder to the LLSs and build a robust LLS catalog, we can provide a more complete picture of the cosmological effects of these high column density absorbers in spectroscopic surveys.

For the multi-fidelity emulator, the current work has demonstrated the cost for the emulator training can be largely reduced by using the low-fidelity data. One of the future directions is to extend the emulator to a much higher dimensionality, such as emulators for beyond Λ CDM cosmologies. Another potential direction is to extend the cosmic emulator to the subgrid model emulator in the hydrodynamical simulations (e.g., the subgrid emulator in the FLAMINGO simulations [193]).

For the black hole mixture model, the current framework measures the co-location and separation of the binary black hole populations in a phenomenological way. One of the future directions is to extend the population model to include spins and redshifts, which can provide more information on the formation channels mentioned in Chapter 6. As the number of detected events increases in the future, the information beyond the mass spectrum will be crucial to understand the nature of the $35M_{\odot}$ bump and testing the scenarios such as the mass segregation, the low-metallicity pop-III stars, and the primordial black holes, etc.

I end this thesis with a quote from Laplace, which is also the epigraph of the E.T.

Jaynes' book *Probability Theory: The Logic of Science* [331]:

“Probability theory is nothing but common sense reduced to calculus; it enables us to appreciate with exactness that which accurate minds feel with a sort of instinct for which of themselves they are unable to account.” – Pierre-Simon Laplace, 1819.

Bibliography

- [1] Leah Fauber, Ming-Feng Ho, Simeon Bird, Christian R. Shelton, Roman Garnett, and Ishita Korde. Automated measurement of quasar redshift with a Gaussian process. *MNRAS*, 498(4):5227–5239, November 2020.
- [2] Kimberly K. Boddy, Mariangela Lisanti, Samuel D. McDermott, Nicholas L. Rodd, Christoph Weniger, Yacine Ali-Haïmoud, Malte Buschmann, Ilias Cholis, Djuna Croon, Adrienne L. Erickcek, Vera Gluscevic, Rebecca K. Leane, Siddharth Mishra-Sharma, Julian B. Muñoz, Ethan O. Nadler, Priyamvada Natarajan, Adrian Price-Whelan, Simona Vegetti, and Samuel J. Witte. Snowmass2021 theory frontier white paper: Astrophysical and cosmological probes of dark matter. *Journal of High Energy Astrophysics*, 35:112–138, August 2022.
- [3] Roman Garnett, Shirley Ho, Simeon Bird, and Jeff Schneider. Detecting damped Ly α absorbers with Gaussian processes. *MNRAS*, 472(2):1850–1865, December 2017.
- [4] David Parks, J. Xavier Prochaska, Shawfeng Dong, and Zheng Cai. Deep learning of quasar spectra to discover and characterize damped Ly α systems. *MNRAS*, 476(1):1151–1168, May 2018.
- [5] P. Noterdaeme, P. Petitjean, W. C. Carithers, I. Pâris, A. Font-Ribera, S. Bailey, E. Aubourg, D. Bizyaev, G. Ebelke, H. Finley, J. Ge, E. Malanushenko, V. Malanushenko, J. Miralda-Escudé, A. D. Myers, D. Oravetz, K. Pan, M. M. Pieri, N. P. Ross, D. P. Schneider, A. Simmons, and D. G. York. Column density distribution and cosmological mass density of neutral gas: Sloan Digital Sky Survey-III Data Release 9. *A&A*, 547:L1, November 2012.
- [6] J. Xavier Prochaska and Arthur M. Wolfe. On the (Non)Evolution of H I Gas in Galaxies Over Cosmic Time. *ApJ*, 696(2):1543–1547, May 2009.
- [7] Neil H. M. Crighton, Michael T. Murphy, J. Xavier Prochaska, Gábor Worseck, Marc Rafelski, George D. Becker, Sara L. Ellison, Michele Fumagalli, Sebastian Lopez, Avery Meiksin, and John M. O’Meara. The neutral hydrogen cosmological mass density at $z = 5$. *MNRAS*, 452(1):217–234, September 2015.

- [8] Ming-Feng Ho, Simeon Bird, and Roman Garnett. Detecting multiple DLAs per spectrum in SDSS DR12 with Gaussian processes. *MNRAS*, 496(4):5436–5454, August 2020.
- [9] Trystyn A. M. Berg, Sara L. Ellison, Rubén Sánchez-Ramírez, Sebastián López, Valentina D’Odorico, George D. Becker, Lise Christensen, Guido Cupani, Kelly D. Denney, and Gábor Worsack. Sub-damped Lyman α systems in the XQ-100 survey - I. Identification and contribution to the cosmological H I budget. *MNRAS*, 488(3):4356–4369, September 2019.
- [10] Euclid Collaboration, M. Knabenhans, J. Stadel, D. Potter, J. Dakin, S. Hannestad, T. Tram, S. Marelli, A. Schneider, R. Teyssier, P. Fosalba, S. Andreon, N. Auricchio, C. Baccigalupi, A. Balaguera-Antolínez, M. Baldi, S. Bardelli, P. Battaglia, R. Bender, A. Biviano, C. Bodendorf, E. Bozzo, E. Branchini, M. Brescia, C. Burigana, R. Cabanac, S. Camera, V. Capobianco, A. Cappi, C. Carbone, J. Carretero, C. S. Carvalho, R. Casas, S. Casas, M. Castellano, G. Castignani, S. Cavuoti, R. Cledassou, C. Colodro-Conde, G. Congedo, C. J. Conselice, L. Conversi, Y. Copin, L. Corcione, J. Coupon, H. M. Courtois, A. Da Silva, S. de la Torre, D. Di Ferdinando, C. A. J. Duncan, X. Dupac, G. Fabbian, S. Farrens, P. G. Ferreira, F. Finelli, M. Frailis, E. Franceschi, S. Galeotta, B. Garilli, C. Giocoli, G. Gozaliasl, J. Graciá-Carpio, F. Grupp, L. Guzzo, W. Holmes, F. Hormuth, H. Israel, K. Jahnke, E. Keihanen, S. Kermiche, C. C. Kirkpatrick, B. Kubik, M. Kunz, H. Kurki-Suonio, S. Ligorì, P. B. Lilje, I. Lloro, D. Maino, O. Marggraf, K. Markovic, N. Martinet, F. Marulli, R. Massey, N. Mauri, S. Maurogordato, E. Medinaceli, M. Meneghetti, B. Metcalf, G. Meylan, M. Moresco, B. Morin, L. Moscardini, E. Munari, C. Neissner, S. M. Niemi, C. Padilla, S. Paltani, F. Pasian, L. Patrizzii, V. Pettorino, S. Pires, G. Polenta, M. Poncet, F. Raison, A. Renzi, J. Rhodes, G. Riccio, E. Romelli, M. Roncarelli, R. Saglia, A. G. Sánchez, D. Sapone, P. Schneider, V. Scottez, A. Secroun, S. Serrano, C. Sirignano, G. Sirri, L. Stanco, F. Sureau, P. Tallada Crespí, A. N. Taylor, M. Tenti, I. Tereno, R. Toledo-Moreo, F. Torradeflot, L. Valenziano, J. Valiviita, T. Vassallo, M. Viel, Y. Wang, N. Welikala, L. Whittaker, A. Zacchei, and E. Zucca. Euclid preparation: IX. EuclidEmulator2 - power spectrum emulation with massive neutrinos and self-consistent dark energy perturbations. *MNRAS*, 505(2):2840–2869, August 2021.
- [11] Euclid Collaboration, M. Knabenhans, J. Stadel, D. Potter, J. Dakin, S. Hannestad, T. Tram, S. Marelli, A. Schneider, R. Teyssier, P. Fosalba, S. Andreon, N. Auricchio, C. Baccigalupi, A. Balaguera-Antolínez, M. Baldi, S. Bardelli, P. Battaglia, R. Bender, A. Biviano, C. Bodendorf, E. Bozzo, E. Branchini, M. Brescia, C. Burigana, R. Cabanac, S. Camera, V. Capobianco, A. Cappi, C. Carbone, J. Carretero, C. S. Carvalho, R. Casas, S. Casas, M. Castellano, G. Castignani, S. Cavuoti, R. Cledassou, C. Colodro-Conde, G. Congedo, C. J. Conselice, L. Conversi, Y. Copin, L. Corcione, J. Coupon, H. M. Courtois, A. Da Silva, S. de la Torre, D. Di Ferdinando, C. A. J. Duncan, X. Dupac, G. Fabbian, S. Farrens, P. G. Ferreira, F. Finelli, M. Frailis, E. Franceschi, S. Galeotta, B. Garilli, C. Giocoli, G. Gozaliasl, J. Graciá-Carpio, F. Grupp, L. Guzzo, W. Holmes, F. Hormuth, H. Israel, K. Jahnke, E. Keihanen, S. Kermiche, C. C. Kirkpatrick, B. Kubik, M. Kunz, H. Kurki-Suonio, S. Ligorì,

- P. B. Lilje, I. Lloro, D. Maino, O. Marggraf, K. Markovic, N. Martinet, F. Marulli, R. Massey, N. Mauri, S. Maurogordato, E. Medinaceli, M. Meneghetti, B. Metcalf, G. Meylan, M. Moresco, B. Morin, L. Moscardini, E. Munari, C. Neissner, S. M. Niemi, C. Padilla, S. Paltani, F. Pasian, L. Patrizii, V. Pettorino, S. Pires, G. Polenta, M. Poncet, F. Raison, A. Renzi, J. Rhodes, G. Riccio, E. Romelli, M. Roncarelli, R. Saglia, A. G. Sánchez, D. Sapone, P. Schneider, V. Scottez, A. Secroun, S. Serrano, C. Sirignano, G. Sirri, L. Stanco, F. Sureau, P. Tallada Crespí, A. N. Taylor, M. Tenti, I. Tereno, R. Toledo-Moreo, F. Torradeflot, L. Valenziano, J. Valiviita, T. Vassallo, M. Viel, Y. Wang, N. Welikala, L. Whittaker, A. Zacchei, and E. Zucca. Euclid preparation: IX. EuclidEmulator2 - power spectrum emulation with massive neutrinos and self-consistent dark energy perturbations. *MNRAS*, 505(2):2840–2869, August 2021.
- [12] Shan Ba, William R. Myers, and William A. Brenneman. Optimal sliced latin hypercube designs. *Technometrics*, 57(4):479–487, 2015.
- [13] R. Abbott, T. D. Abbott, F. Acernese, et al. Population of Merging Compact Binaries Inferred Using Gravitational Waves through GWTC-3. *Physical Review X*, 13(1):011048, January 2023.
- [14] Vaibhav Tiwari. VAMANA: modeling binary black hole population with minimal assumptions. *Classical and Quantum Gravity*, 38(15):155007, August 2021.
- [15] Ming-Feng Ho, Simeon Bird, and Roman Garnett. Damped Lyman- α absorbers from Sloan digital sky survey DR16Q with Gaussian processes. *MNRAS*, 507(1):704–719, October 2021.
- [16] Ming-Feng Ho, Simeon Bird, and Christian R. Shelton. Multifidelity emulation for the matter power spectrum using Gaussian processes. *MNRAS*, 509(2):2551–2565, January 2022.
- [17] Ming-Feng Ho, Simeon Bird, Martin A. Fernandez, and Christian R. Shelton. MF-Box: multifidelity and multiscale emulation for the matter power spectrum. *MNRAS*, 526(2):2903–2919, December 2023.
- [18] Andrew Pontzen, Fabio Governato, Max Pettini, C. M. Booth, Greg Stinson, James Wadsley, Alyson Brooks, Thomas Quinn, and Martin Haehnelt. Damped Lyman α systems in galaxy formation simulations. *MNRAS*, 390(4):1349–1371, November 2008.
- [19] Keir K. Rogers, Simeon Bird, Hiranya V. Peiris, Andrew Pontzen, Andreu Font-Ribera, and Boris Leistedt. Simulating the effect of high column density absorbers on the one-dimensional Lyman α forest flux power spectrum. *MNRAS*, 474(3):3032–3042, March 2018.
- [20] Jason X. Prochaska, Stéphane Herbert-Fort, and Arthur M. Wolfe. The SDSS Damped Ly α Survey: Data Release 3. *ApJ*, 635(1):123–142, December 2005.
- [21] Roman Garnett, Shirley Ho, Simeon Bird, and Jeff Schneider. Detecting damped Ly α absorbers with Gaussian processes. *MNRAS*, 472(2):1850–1865, December 2017.

- [22] Solène Chabanier, Thomas Etourneau, Jean-Marc Le Goff, James Rich, Julianna Stermer, Bela Abolfathi, Andreu Font-Ribera, Alma X. Gonzalez-Morales, Axel de la Macorra, Ignasi Pérez-Ràfols, Patrick Petitjean, Matthew M. Pieri, Corentin Ravoux, Graziano Rossi, and Donald P. Schneider. The Completed Sloan Digital Sky Survey IV Extended Baryon Oscillation Spectroscopic Survey: The Damped Ly α Systems Catalog. *ApJS*, 258(1):18, January 2022.
- [23] Ben Wang, Jiaqi Zou, Zheng Cai, J. Xavier Prochaska, Zechang Sun, Jiani Ding, Andreu Font-Ribera, Alma Gonzalez, Hiram K. Herrera-Alcantar, Vid Irsic, Xiaojing Lin, David Brooks, Solène Chabanier, Roger de Belsunce, Nathalie Palanque-Delabrouille, Gregory Tarle, and Zhimin Zhou. Deep Learning of Dark Energy Spectroscopic Instrument Mock Spectra to Find Damped Ly α Systems. *ApJS*, 259(1):28, March 2022.
- [24] Naim Göksel Karaçaylı, Paul Martini, Julien Guy, Corentin Ravoux, Marie Lynn Abdul Karim, Eric Armengaud, Michael Walther, J. Aguilar, S. Ahlen, S. Bailey, J. Bautista, S. F. Beltran, D. Brooks, L. Cabayol-Garcia, S. Chabanier, E. Chaussidon, J. Chaves-Montero, K. Dawson, R. de la Cruz, A. de la Macorra, P. Doel, A. Font-Ribera, J. E. Forero-Romero, S. Gontcho A. Gontcho, A. X. Gonzalez-Morales, C. Gordon, H. K. Herrera-Alcantar, K. Honscheid, V. Iršič, M. Ishak, R. Kehoe, T. Kisner, A. Kremin, M. Landriau, L. Le Guillou, M. E. Levi, Z. Lukić, A. Meisner, R. Miquel, J. Moustakas, E. Mueller, A. Muñoz-Gutiérrez, L. Napolitano, J. Nie, G. Niz, N. Palanque-Delabrouille, W. J. Percival, M. Pieri, C. Poppett, F. Prada, I. Pérez-Ràfols, C. Ramírez-Pérez, G. Rossi, E. Sanchez, H. Seo, F. Sinigaglia, T. Tan, G. Tarlé, B. Wang, B. A. Weaver, C. Yéche, and Z. Zhou. Optimal 1D Ly α forest power spectrum estimation - III. DESI early data. *MNRAS*, 528(3):3941–3963, March 2024.
- [25] Reza Monadi, Ming-Feng Ho, Kathy L. Cooksey, and Simeon Bird. Machine learning uncovers the universe’s hidden gems: A comprehensive catalogue of C IV absorption lines in SDSS DR12. *MNRAS*, 526(3):4557–4574, December 2023.
- [26] Aurel Schneider, Romain Teyssier, Doug Potter, Joachim Stadel, Julian Onions, Darren S. Reed, Robert E. Smith, Volker Springel, Frazer R. Pearce, and Roman Scocimarro. Matter power spectrum and the challenge of percent accuracy. *JCAP*, 2016(4):047, April 2016.
- [27] Katrin Heitmann, David Higdon, Charles Nakhleh, and Salman Habib. Cosmic Calibration. *ApJL*, 646(1):L1–L4, July 2006.
- [28] Katrin Heitmann, David Higdon, Martin White, Salman Habib, Brian J. Williams, Earl Lawrence, and Christian Wagner. The Coyote Universe. II. Cosmological Models and Precision Emulation of the Nonlinear Matter Power Spectrum. *ApJ*, 705(1):156–174, November 2009.
- [29] Katrin Heitmann, Martin White, Christian Wagner, Salman Habib, and David Higdon. The Coyote Universe. I. Precision Determination of the Nonlinear Matter Power Spectrum. *ApJ*, 715(1):104–121, May 2010.

- [30] Thomas J. Santner, Brian J. Williams, and William I. Notz. *The Design and Analysis of Computer Experiments*. Springer series in statistics. Springer, 2003.
- [31] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.
- [32] Giovanni Aricò, Raul E. Angulo, Sergio Contreras, Lurdes Ondaro-Mallea, Marcos Pellejero-Ibañez, and Matteo Zennaro. The BACCO simulation project: a baryonification emulator with neural networks. *MNRAS*, 506(3):4070–4082, September 2021.
- [33] Robin Kooistra, Khee-Gan Lee, and Benjamin Horowitz. Constraining the Fluctuating Gunn-Peterson Approximation using Ly α Forest Tomography at $z = 2$. *ApJ*, 938(2):123, October 2022.
- [34] F. Sinigaglia, F. S. Kitaura, K. Nagamine, Y. Oku, and A. Balaguera-Antolínez. Field-level Lyman- α forest modeling in redshift space via augmented nonlocal Fluctuating Gunn-Peterson Approximation. *A&A*, 682:A21, February 2024.
- [35] Yueying Ni, Shy Genel, Daniel Anglés-Alcázar, Francisco Villaescusa-Navarro, Yongseok Jo, Simeon Bird, Tiziana Di Matteo, Rupert Croft, Nianyi Chen, Natalí S. M. de Santi, Matthew Gebhardt, Helen Shao, Shivam Pandey, Lars Hernquist, and Romeel Dave. The CAMELS Project: Expanding the Galaxy Formation Model Space with New ASTRID and 28-parameter TNG and SIMBA Suites. *ApJ*, 959(2):136, December 2023.
- [36] Francisco Villaescusa-Navarro, Daniel Anglés-Alcázar, Shy Genel, David N. Spergel, Rachel S. Somerville, Romeel Dave, Annalisa Pillepich, Lars Hernquist, Dylan Nelson, Paul Torrey, Desika Narayanan, Yin Li, Oliver Philcox, Valentina La Torre, Ana Maria Delgado, Shirley Ho, Sultan Hassan, Blakesley Burkhart, Digvijay Wadekar, Nicholas Battaglia, Gabriella Contardo, and Greg L. Bryan. The CAMELS Project: Cosmology and Astrophysics with Machine-learning Simulations. *ApJ*, 915(1):71, July 2021.
- [37] Roman Garnett. *Bayesian Optimization*. Cambridge University Press, 2023.
- [38] Dylan Nelson, Volker Springel, Annalisa Pillepich, Vicente Rodriguez-Gomez, Paul Torrey, Shy Genel, Mark Vogelsberger, Ruediger Pakmor, Federico Marinacci, Rainer Weinberger, Luke Kelley, Mark Lovell, Benedikt Diemer, and Lars Hernquist. The IllustrisTNG simulations: public data release. *Computational Astrophysics and Cosmology*, 6(1):2, May 2019.
- [39] Simeon Bird, Yueying Ni, Tiziana Di Matteo, Rupert Croft, Yu Feng, and Nianyi Chen. The ASTRID simulation: galaxy formation and reionization. *MNRAS*, 512(3):3703–3716, May 2022.
- [40] Joop Schaye, Roi Kugel, Matthieu Schaller, John C. Helly, Joey Braspenning, Willem Elbers, Ian G. McCarthy, Marcel P. van Daalen, Bert Vandenbroucke, Carlos S. Frenk,

- Juliana Kwan, Jaime Salcido, Yannick M. Bahé, Josh Borrow, Evgenii Chaikin, Oliver Hahn, Filip Huško, Adrian Jenkins, Cedric G. Lacey, and Folkert S. J. Nobels. The FLAMINGO project: cosmological hydrodynamical simulations for large-scale structure and galaxy cluster surveys. *MNRAS*, 526(4):4978–5020, December 2023.
- [41] MC Kennedy and A O’Hagan. Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1):1–13, 03 2000.
- [42] M. A. Fernandez, Ming-Feng Ho, and Simeon Bird. A Multi-Fidelity Emulator for the Lyman- α Forest Flux Power Spectrum. *arXiv e-prints*, page arXiv:2207.06445, July 2022.
- [43] M. A. Fernandez, Simeon Bird, and Ming-Feng Ho. Cosmological Constraints from the eBOSS Lyman- α Forest using the PRIYA Simulations. *arXiv e-prints*, page arXiv:2309.03943, September 2023.
- [44] Simeon Bird, Martin Fernandez, Ming-Feng Ho, Mahdi Qezlou, Reza Monadi, Yueying Ni, Nianyi Chen, Rupert Croft, and Tiziana Di Matteo. PRIYA: a new suite of Lyman- α forest simulations for cosmology. *JCAP*, 2023(10):037, October 2023.
- [45] Nicholas S. Kern, Adrian Liu, Aaron R. Parsons, Andrei Mesinger, and Bradley Greig. Emulating Simulations of Cosmic Dawn for 21 cm Power Spectrum Constraints on Cosmology, Reionization, and X-Ray Heating. *ApJ*, 848(1):23, October 2017.
- [46] R. E. Smith, J. A. Peacock, A. Jenkins, S. D. M. White, C. S. Frenk, F. R. Pearce, P. A. Thomas, G. Efstathiou, and H. M. P. Couchmann. Stable clustering, the halo model and nonlinear cosmological power spectra. *Mon. Not. Roy. Astron. Soc.*, 341:1311, 2003.
- [47] J. A. Peacock and R. E. Smith. Halo occupation numbers and galaxy bias. *Mon. Not. Roy. Astron. Soc.*, 318:1144, 2000.
- [48] Uros Seljak. Analytic model for galaxy and dark matter clustering. *Mon. Not. Roy. Astron. Soc.*, 318:203, 2000.
- [49] A. M. Wolfe, D. A. Turnshek, H. E. Smith, and R. D. Cohen. Damped Lyman-Alpha Absorption by Disk Galaxies with Large Redshifts. I. The Lick Survey. *ApJS*, 61:249, June 1986.
- [50] Renyue Cen. The Nature of Damped Ly α Systems and Their Hosts in the Standard Cold Dark Matter Universe. *ApJ*, 748(2):121, April 2012.
- [51] Michele Fumagalli, John M. O’Meara, J. Xavier Prochaska, Marc Rafelski, and Nissim Kanekar. Directly imaging damped Ly α galaxies at $z \lesssim 2$ - III. The star formation rates of neutral gas reservoirs at $z \sim 2.7$. *MNRAS*, 446(3):3178–3198, January 2015.
- [52] Jeffrey P. Gardner, Neal Katz, David H. Weinberg, and Lars Hernquist. Testing Cosmological Models against the Abundance of Damped Lyman-Alpha Absorbers. *ApJ*, 486(1):42–47, September 1997.

- [53] T. Zafar, C. Péroux, A. Popping, B. Milliard, J. M. Deharveng, and S. Frank. The ESO UVES advanced data products quasar sample. II. Cosmological evolution of the neutral gas mass density. *A&A*, 556:A141, August 2013.
- [54] Martin G. Haehnelt, Matthias Steinmetz, and Michael Rauch. Damped Ly α Absorber at High Redshift: Large Disks or Galactic Building Blocks? *ApJ*, 495(2):647–658, March 1998.
- [55] Jason X. Prochaska and Arthur M. Wolfe. On the Kinematics of the Damped Lyman- α Protogalaxies. *ApJ*, 487(1):73–95, September 1997.
- [56] Simeon Bird, Mark Vogelsberger, Martin Haehnelt, Debora Sijacki, Shy Genel, Paul Torrey, Volker Springel, and Lars Hernquist. Damped Lyman α absorbers as a probe of stellar feedback. *MNRAS*, 445(3):2313–2324, December 2014.
- [57] Simeon Bird, Martin Haehnelt, Marcel Neeleman, Shy Genel, Mark Vogelsberger, and Lars Hernquist. Reproducing the kinematics of damped Lyman α systems. *MNRAS*, 447(2):1834–1846, February 2015.
- [58] Arthur M. Wolfe, Eric Gawiser, and Jason X. Prochaska. Damped Ly α Systems. *ARA&A*, 43(1):861–918, September 2005.
- [59] Anže Slosar, Andreu Font-Ribera, Matthew M. Pieri, James Rich, Jean-Marc Le Goff, Éric Aubourg, Jon Brinkmann, Nicolas Busca, Bill Carithers, Romain Charlassier, Marina Cortês, Rupert Croft, Kyle S. Dawson, Daniel Eisenstein, Jean-Christophe Hamilton, Shirley Ho, Khee-Gan Lee, Robert Lupton, Patrick McDonald, Bumbarija Medolin, Demitri Muna, Jordi Miralda-Escudé, Adam D. Myers, Robert C. Nichol, Nathalie Palanque-Delabrouille, Isabelle Pâris, Patrick Petitjean, Yodovina Piškur, Emmanuel Rollinde, Nicholas P. Ross, David J. Schlegel, Donald P. Schneider, Erin Sheldon, Benjamin A. Weaver, David H. Weinberg, Christophe Yèche, and Donald G. York. The Lyman- α forest in three dimensions: measurements of large scale flux correlations from BOSS 1st-year data. *JCAP*, 2011(9):001, September 2011.
- [60] Donald G. York, J. Adelman, Jr. Anderson, John E., Scott F. Anderson, James Annis, Neta A. Bahcall, J. A. Bakken, Robert Barkhouser, Steven Bastian, Eileen Berman, William N. Boroski, Steve Bracker, Charlie Briegel, John W. Briggs, J. Brinkmann, Robert Brunner, Scott Burles, Larry Carey, Michael A. Carr, Francisco J. Castander, Bing Chen, Patrick L. Colestock, A. J. Connolly, J. H. Crocker, István Csabai, Paul C. Czarapata, John Eric Davis, Mamoru Doi, Tom Dombeck, Daniel Eisenstein, Nancy Ellman, Brian R. Elms, Michael L. Evans, Xiaohui Fan, Glenn R. Federwitz, Larry Fiscelli, Scott Friedman, Joshua A. Frieman, Masataka Fukugita, Bruce Gillespie, James E. Gunn, Vijay K. Gurbani, Ernst de Haas, Merle Haldeman, Frederick H. Harris, J. Hayes, Timothy M. Heckman, G. S. Hennessy, Robert B. Hindsley, Scott Holm, Donald J. Holmgren, Chi-hao Huang, Charles Hull, Don Husby, Shin-Ichi Ichikawa, Takashi Ichikawa, Željko Ivezić, Stephen Kent, Rita S. J. Kim, E. Kinney, Mark Klaene, A. N. Kleinman, S. Kleinman, G. R. Knapp, John Korienek,

Richard G. Kron, Peter Z. Kunszt, D. Q. Lamb, B. Lee, R. French Leger, Siriluk Limmongkol, Carl Lindenmeyer, Daniel C. Long, Craig Loomis, Jon Loveday, Rich Lucinio, Robert H. Lupton, Bryan MacKinnon, Edward J. Mannery, P. M. Mantsch, Bruce Margon, Peregrine McGehee, Timothy A. McKay, Avery Meiksin, Aronne Merelli, David G. Monet, Jeffrey A. Munn, Vijay K. Narayanan, Thomas Nash, Eric Neilsen, Rich Neswold, Heidi Jo Newberg, R. C. Nichol, Tom Nicinski, Mario Nonino, Norio Okada, Sadanori Okamura, Jeremiah P. Ostriker, Russell Owen, A. George Pauls, John Peoples, R. L. Peterson, Donald Petravick, Jeffrey R. Pier, Adrian Pope, Ruth Pordes, Angela Prosapio, Ron Rechenmacher, Thomas R. Quinn, Gordon T. Richards, Michael W. Richmond, Claudio H. Rivetta, Constance M. Rockosi, Kurt Ruthmansdorfer, Dale Sandford, David J. Schlegel, Donald P. Schneider, Maki Sekiguchi, Gary Sergey, Kazuhiro Shimasaku, Walter A. Siegmund, Stephen Smee, J. Allyn Smith, S. Snedden, R. Stone, Chris Stoughton, Michael A. Strauss, Christopher Stubbs, Mark SubbaRao, Alexander S. Szalay, Istvan Szapudi, Gyula P. Szokoly, Anirudda R. Thakar, Christy Tremonti, Douglas L. Tucker, Alan Uomoto, Dan Vanden Berk, Michael S. Vogeley, Patrick Waddell, Shu-i. Wang, Masaru Watanabe, David H. Weinberg, Brian Yanny, Naoki Yasuda, and SDSS Collaboration. The Sloan Digital Sky Survey: Technical Summary. *AJ*, 120(3):1579–1587, September 2000.

- [61] Isabelle Pâris, Patrick Petitjean, Éric Aubourg, Adam D. Myers, Alina Streblyanska, Brad W. Lyke, Scott F. Anderson, Éric Armengaud, Julian Bautista, Michael R. Blanton, Michael Blomqvist, Jonathan Brinkmann, Joel R. Brownstein, William Nielsen Brandt, Étienne Burtin, Kyle Dawson, Sylvain de la Torre, Antonis Georgakakis, Héctor Gil-Marín, Paul J. Green, Patrick B. Hall, Jean-Paul Kneib, Stephanie M. LaMassa, Jean-Marc Le Goff, Chelsea MacLeod, Vivek Mariappan, Ian D. McGreer, Andrea Merloni, Pasquier Noterdaeme, Nathalie Palanque-Delabrouille, Will J. Percival, Ashley J. Ross, Graziano Rossi, Donald P. Schneider, Hee-Jong Seo, Rita Tojeiro, Benjamin A. Weaver, Anne-Marie Weijmans, Christophe Yèche, Pauline Zarrouk, and Gong-Bo Zhao. The Sloan Digital Sky Survey Quasar Catalog: Fourteenth data release. *A&A*, 613:A51, May 2018.
- [62] Simeon Bird, Roman Garnett, and Shirley Ho. Statistical properties of damped Lyman-alpha systems from Sloan Digital Sky Survey DR12. *MNRAS*, 466(2):2111–2122, April 2017.
- [63] W. Carithers. DLA Concordance Catalog. *Published internally to SDSS*, 2012.
- [64] Khee-Gan Lee, Stephen Bailey, Leslie E. Bartsch, William Carithers, Kyle S. Dawson, David Kirkby, Britt Lundgren, Daniel Margala, Nathalie Palanque-Delabrouille, Matthew M. Pieri, David J. Schlegel, David H. Weinberg, Christophe Yèche, Éric Aubourg, Julian Bautista, Dmitry Bizyaev, Michael Blomqvist, Adam S. Bolton, Arnaud Borde, Howard Brewington, Nicolás G. Busca, Rupert A. C. Croft, Timothée Delubac, Garrett Ebelke, Daniel J. Eisenstein, Andreu Font-Ribera, Jian Ge, Jean-Christophe Hamilton, Joseph F. Hennawi, Shirley Ho, Klaus Honscheid, Jean-Marc Le Goff, Elena Malanushenko, Viktor Malanushenko, Jordi Miralda-Escudé, Adam D.

- Myers, Pasquier Noterdaeme, Daniel Oravetz, Kaike Pan, Isabelle Pâris, Patrick Petitjean, James Rich, Emmanuel Rollinde, Nicholas P. Ross, Graziano Rossi, Donald P. Schneider, Audrey Simmons, Stephanie Snedden, Anže Slosar, David N. Spergel, Nao Suzuki, Matteo Viel, and Benjamin A. Weaver. The BOSS Ly α Forest Sample from SDSS Data Release 9. *AJ*, 145(3):69, March 2013.
- [65] Leah Fauber, Ming-Feng Ho, Simeon Bird, Christian R. Shelton, Roman Garnett, and Ishita Korde. Automated measurement of quasar redshift with a Gaussian process. *MNRAS*, 498(4):5227–5239, November 2020.
- [66] T. S. Kim, J. S. Bolton, M. Viel, M. G. Haehnelt, and R. F. Carswell. An improved measurement of the flux distribution of the Ly α forest in QSO absorption spectra: the effect of continuum fitting, metal contamination and noise properties. *MNRAS*, 382(4):1657–1674, December 2007.
- [67] George D. Becker, Paul C. Hewett, Gábor Worseck, and J. Xavier Prochaska. A refined measurement of the mean transmitted flux in the Ly α forest over $2 < z < 5$ using composite quasar spectra. *MNRAS*, 430(3):2067–2081, April 2013.
- [68] Lucien Le Cam. An approximation theorem for the poisson binomial distribution. *Pacific J. Math.*, 10(4):1181–1197, 1960.
- [69] M. Fernandez and S. Williams. Closed-form expression for the poisson-binomial probability density function. *IEEE Transactions on Aerospace and Electronic Systems*, 46(2):803–817, April 2010.
- [70] J. X. Prochaska, G. Worseck, and J. M. O’Meara. A Direct Measurement of the Inter-galactic Medium Opacity to H I Ionizing Photons. *ApJL*, 705:L113–L117, November 2009.
- [71] G. Worseck and J. X. Prochaska. GALEX Far-ultraviolet Color Selection of UV-bright High-redshift Quasars. *ApJ*, 728:23, February 2011.
- [72] M. Fumagalli, J. M. O’Meara, J. X. Prochaska, and G. Worseck. Dissecting the Properties of Optically Thick Hydrogen at the Peak of Cosmic Star Formation History. *ApJ*, 775:78, September 2013.
- [73] R. Sánchez-Ramírez, S. L. Ellison, J. X. Prochaska, T. A. M. Berg, S. López, V. D’Odorico, G. D. Becker, L. Christensen, G. Cupani, K. D. Denney, I. Pâris, G. Worseck, and J. Gorosabel. The evolution of neutral gas in damped Lyman α systems from the XQ-100 survey. *MNRAS*, 456(4):4488–4505, March 2016.
- [74] Alireza Rahmati and Joop Schaye. Predictions for the relation between strong HI absorbers and galaxies at redshift 3. *MNRAS*, 438(1):529–547, February 2014.
- [75] Andreu Font-Ribera, Jordi Miralda-Escudé, Eduard Arnau, Bill Carithers, Khee-Gan Lee, Pasquier Noterdaeme, Isabelle Pâris, Patrick Petitjean, James Rich, Emmanuel Rollinde, Nicholas P. Ross, Donald P. Schneider, Martin White, and Donald G. York.

- The large-scale cross-correlation of Damped Lyman alpha systems with the Lyman alpha forest: first measurements from BOSS. *JCAP*, 2012(11):059, November 2012.
- [76] Ignasi Pérez-Ràfols, Andreu Font-Ribera, Jordi Miralda-Escudé, Michael Blomqvist, Simeon Bird, Nicolás Busca, Hélión du Mas des Bourboux, Lluís Mas-Ribas, Pasquier Noterdaeme, Patrick Petitjean, James Rich, and Donald P. Schneider. The SDSS-DR12 large-scale cross-correlation of damped Lyman alpha systems with the Lyman alpha forest. *MNRAS*, 473(3):3019–3038, January 2018.
- [77] D. Alonso, J. Colosimo, A. Font-Ribera, and A. Slosar. Bias of damped Lyman- α systems from their cross-correlation with CMB lensing. *JCAP*, 2018(4):053, April 2018.
- [78] Xiaojing Lin, Zheng Cai, Yin Li, Alex Krolewski, and Simone Ferraro. Constraining the Halo Mass of Damped Ly α Absorption Systems (DLAs) at $z = 2 - 3.5$ using the Quasar-CMB Lensing Cross-correlation. *arXiv e-prints*, page arXiv:2011.01234, November 2020.
- [79] Rupert A. C. Croft, David H. Weinberg, Neal Katz, and Lars Hernquist. Recovery of the Power Spectrum of Mass Fluctuations from Observations of the Ly α Forest. *ApJ*, 495(1):44–62, March 1998.
- [80] Patrick McDonald, Jordi Miralda-Escudé, Michael Rauch, Wallace L. W. Sargent, Tom A. Barlow, Renyue Cen, and Jeremiah P. Ostriker. The Observed Probability Distribution Function, Power Spectrum, and Correlation Function of the Transmitted Flux in the Ly α Forest. *ApJ*, 543(1):1–23, November 2000.
- [81] M. Viel, M. G. Haehnelt, R. F. Carswell, and T. S. Kim. The effect of (strong) discrete absorption systems on the Lyman α forest flux power spectrum. *MNRAS*, 349(3):L33–L37, April 2004.
- [82] Patrick McDonald, Uroš Seljak, Renyue Cen, David Shih, David H. Weinberg, Scott Burles, Donald P. Schneider, David J. Schlegel, Neta A. Bahcall, John W. Briggs, J. Brinkmann, Masataka Fukugita, Željko Ivezić, Stephen Kent, and Daniel E. Vanden Berk. The Linear Theory Power Spectrum from the Ly α Forest in the Sloan Digital Sky Survey. *ApJ*, 635(2):761–783, December 2005.
- [83] Vid Iršič, Matteo Viel, Trystyn A. M. Berg, Valentina D’Odorico, Martin G. Haehnelt, Stefano Cristiani, Guido Cupani, Tae-Sun Kim, Sebastian López, Sara Ellison, George D. Becker, Lise Christensen, Kelly D. Denney, Gábor Worseck, and James S. Bolton. The Lyman α forest power spectrum from the XQ-100 Legacy Survey. *MNRAS*, 466(4):4332–4345, April 2017.
- [84] Solène Chabanier, Nathalie Palanque-Delabrouille, Christophe Yèche, Jean-Marc Le Goff, Eric Armengaud, Julian Bautista, Michael Blomqvist, Nicolas Busca, Kyle Dawson, Thomas Etourneau, Andreu Font-Ribera, Youngbae Lee, Hélión du Mas des Bourboux, Matthew Pieri, James Rich, Graziano Rossi, Donald Schneider, and Anže

- Slosar. The one-dimensional power spectrum from the SDSS DR14 Ly α forests. *JCAP*, 2019(7):017, July 2019.
- [85] Patrick McDonald, Uroš Seljak, Renyue Cen, Paul Bode, and Jeremiah P. Ostriker. Physical effects on the Ly α forest flux power spectrum: damping wings, ionizing radiation fluctuations and galactic winds. *MNRAS*, 360(4):1471–1482, July 2005.
- [86] Keir K. Rogers, Simeon Bird, Hiranya V. Peiris, Andrew Pontzen, Andreu Font-Ribera, and Boris Leistedt. Simulating the effect of high column density absorbers on the one-dimensional Lyman α forest flux power spectrum. *MNRAS*, 474(3):3032–3042, March 2018.
- [87] Keir K. Rogers, Simeon Bird, Hiranya V. Peiris, Andrew Pontzen, Andreu Font-Ribera, and Boris Leistedt. Correlations in the three-dimensional Lyman-alpha forest contaminated by high column density absorbers. *MNRAS*, 476(3):3716–3728, May 2018.
- [88] Andrei Cuceu, Andreu Font-Ribera, and Benjamin Joachimi. Bayesian methods for fitting Baryon Acoustic Oscillations in the Lyman- α forest. *JCAP*, 2020(7):035, July 2020.
- [89] Kyle S. Dawson, Jean-Paul Kneib, Will J. Percival, Shadab Alam, Franco D. Albareti, Scott F. Anderson, Eric Armengaud, Éric Aubourg, Stephen Bailey, Julian E. Bautista, Andreas A. Berlind, Matthew A. Bershady, Florian Beutler, Dmitry Bizyaev, Michael R. Blanton, Michael Blomqvist, Adam S. Bolton, Jo Bovy, W. N. Brandt, Jon Brinkmann, Joel R. Brownstein, Etienne Burtin, N. G. Busca, Zheng Cai, Chia-Hsun Chuang, Nicolas Clerc, Johan Comparat, Frances Cope, Rupert A. C. Croft, Irene Cruz-Gonzalez, Luiz N. da Costa, Marie-Claude Cousinou, Jeremy Darling, Axel de la Macorra, Sylvain de la Torre, Timothée Delubac, Hélión du Mas des Bourboux, Tom Dwelly, Anne Ealet, Daniel J. Eisenstein, Michael Eracleous, S. Escoffier, Xiaohui Fan, Alexis Finoguenov, Andreu Font-Ribera, Peter Frinchaboy, Patrick Gaulme, Antonis Georgakakis, Paul Green, Hong Guo, Julien Guy, Shirley Ho, Diana Holder, Joe Huehnerhoff, Timothy Hutchinson, Yipeng Jing, Eric Jullo, Vikrant Kamble, Karen Kinemuchi, David Kirkby, Francisco-Shu Kitaura, Mark A. Klaene, Russ R. Laher, Dustin Lang, Pierre Laurent, Jean-Marc Le Goff, Cheng Li, Yu Liang, Marcos Lima, Qiufan Lin, Weipeng Lin, Yen-Ting Lin, Daniel C. Long, Britt Lundgren, Nicholas MacDonald, Marcio Antonio Geimba Maia, Elena Malanushenko, Viktor Malanushenko, Vivek Mariappan, Cameron K. McBride, Ian D. McGreer, Brice Ménard, Andrea Merloni, Andres Meza, Antonio D. Montero-Dorta, Demitri Muna, Adam D. Myers, Kirpal Nandra, Tracy Naugle, Jeffrey A. Newman, Pasquier Noterdaeme, Peter Nugent, Ricardo Ogando, Matthew D. Olmstead, Audrey Oravetz, Daniel J. Oravetz, Nikhil Padmanabhan, Nathalie Palanque-Delabrouille, Kaike Pan, John K. Parejko, Isabelle Pâris, John A. Peacock, Patrick Petitjean, Matthew M. Pieri, Alice Pisani, Francisco Prada, Abhishek Prakash, Anand Raichoor, Beth Reid, James Rich, Jethro Ridl, Sergio Rodriguez-Torres, Aurelio Carnero Rosell, Ashley J. Ross, Graziano Rossi, John Ruan, Mara Salvato, Conor Sayres, Donald P. Schneider,

David J. Schlegel, Uros Seljak, Hee-Jong Seo, Branimir Sesar, Sarah Shandera, Yiping Shu, Anže Slosar, Flavia Sobreira, Alina Streblyanska, Nao Suzuki, Donna Taylor, Charling Tao, Jeremy L. Tinker, Rita Tojeiro, Mariana Vargas-Magaña, Yuting Wang, Benjamin A. Weaver, David H. Weinberg, Martin White, W. M. Wood-Vasey, Christophe Yeche, Zhongxu Zhai, Cheng Zhao, Gong-bo Zhao, Zheng Zheng, Guangtun Ben Zhu, and Hu Zou. The SDSS-IV Extended Baryon Oscillation Spectroscopic Survey: Overview and Early Data. *AJ*, 151(2):44, February 2016.

- [90] Brad W. Lyke, Alexandra N. Higley, J. N. McLane, Danielle P. Schurhammer, Adam D. Myers, Ashley J. Ross, Kyle Dawson, Solène Chabanier, Paul Martini, Nicolás G. Busca, Hélión du Mas des Bourboux, Mara Salvato, Alina Streblyanska, Pauline Zarrouk, Etienne Burtin, Scott F. Anderson, Julian Bautista, Dmitry Bizyaev, W. N. Brandt, Jonathan Brinkmann, Joel R. Brownstein, Johan Comparat, Paul Green, Axel de la Macorra, Andrea Muñoz Gutiérrez, Jiamin Hou, Jeffrey A. Newman, Nathalie Palanque-Delabrouille, Isabelle Pâris, Will J. Percival, Patrick Petitjean, James Rich, Graziano Rossi, Donald P. Schneider, Alexander Smith, M. Vivek, and Benjamin Alan Weaver. The Sloan Digital Sky Survey Quasar Catalog: Sixteenth Data Release. *ApJS*, 250(1):8, September 2020.
- [91] Daniel J. Eisenstein, David H. Weinberg, Eric Agol, Hiroaki Aihara, Carlos Allende Prieto, Scott F. Anderson, James A. Arns, Éric Aubourg, Stephen Bailey, Eduardo Balbinot, Robert Barkhouser, Timothy C. Beers, Andreas A. Berlind, Steven J. Bickerton, Dmitry Bizyaev, Michael R. Blanton, John J. Bochanski, Adam S. Bolton, Casey T. Bosman, Jo Bovy, W. N. Brandt, Ben Breslauer, Howard J. Brewington, J. Brinkmann, Peter J. Brown, Joel R. Brownstein, Dan Burger, Nicolas G. Busca, Heather Campbell, Phillip A. Cargile, William C. Carithers, Joleen K. Carlberg, Michael A. Carr, Liang Chang, Yanmei Chen, Cristina Chiappini, Johan Comparat, Natalia Connolly, Marina Cortes, Rupert A. C. Croft, Katia Cunha, Luiz N. da Costa, James R. A. Davenport, Kyle Dawson, Nathan De Lee, Gustavo F. Porto de Mello, Fernando de Simoni, Janice Dean, Saurav Dhital, Anne Ealet, Garrett L. Ebelke, Edward M. Edmondson, Jacob M. Eiting, Stephanie Escoffier, Massimiliano Esposito, Michael L. Evans, Xiaohui Fan, Bruno Femenía Castellá, Leticia Dutra Ferreira, Greg Fitzgerald, Scott W. Fleming, Andreu Font-Ribera, Eric B. Ford, Peter M. Frinchaboy, Ana Elia García Pérez, B. Scott Gaudi, Jian Ge, Luan Ghezzi, Bruce A. Gillespie, G. Gilmore, Léo Girardi, J. Richard Gott, Andrew Gould, Eva K. Grebel, James E. Gunn, Jean-Christophe Hamilton, Paul Harding, David W. Harris, Suzanne L. Hawley, Frederick R. Hearty, Joseph F. Hennawi, Jonay I. González Hernández, Shirley Ho, David W. Hogg, Jon A. Holtzman, Klaus Honscheid, Naohisa Inada, Inese I. Ivans, Linhua Jiang, Peng Jiang, Jennifer A. Johnson, Cathy Jordan, Wendell P. Jordan, Guinevere Kauffmann, Eyal Kazin, David Kirkby, Mark A. Klaene, G. R. Knapp, Jean-Paul Kneib, C. S. Kochanek, Lars Koesterke, Juna A. Kollmeier, Richard G. Kron, Hubert Lampeitl, Dustin Lang, James E. Lawler, Jean-Marc Le Goff, Brian L. Lee, Young Sun Lee, Jarron M. Leisenring, Yen-Ting Lin, Jian Liu, Daniel C. Long, Craig P. Loomis, Sara Lucatello, Britt Lundgren,

Robert H. Lupton, Bo Ma, Zhibo Ma, Nicholas MacDonald, Claude Mack, Suvrath Mahadevan, Marcio A. G. Maia, Steven R. Majewski, Martin Makler, Elena Malanushenko, Viktor Malanushenko, Rachel Mandelbaum, Claudia Maraston, Daniel Margala, Paul Maseman, Karen L. Masters, Cameron K. McBride, Patrick McDonald, Ian D. McGreer, Richard G. McMahon, Olga Mena Requejo, Brice Ménard, Jordi Miralda-Escudé, Heather L. Morrison, Fergal Mullally, Demitri Muna, Hitoshi Murayama, Adam D. Myers, Tracy Naugle, Angelo Fausti Neto, Duy Cuong Nguyen, Robert C. Nichol, David L. Nidever, Robert W. O'Connell, Ricardo L. C. Ogando, Matthew D. Olmstead, Daniel J. Oravetz, Nikhil Padmanabhan, Martin Paegert, Nathalie Palanque-Delabrouille, Kaike Pan, Parul Pandey, John K. Parejko, Isabelle Pâris, Paulo Pellegrini, Joshua Pepper, Will J. Percival, Patrick Petitjean, Robert Pfaffenberger, Janine Pforr, Stefanie Phleps, Christophe Pichon, Matthew M. Pieri, Francisco Prada, Adrian M. Price-Whelan, M. Jordan Raddick, Beatriz H. F. Ramos, I. Neill Reid, Celine Reyle, James Rich, Gordon T. Richards, George H. Rieke, Marcia J. Rieke, Hans-Walter Rix, Annie C. Robin, Helio J. Rocha-Pinto, Constance M. Rockosi, Natalie A. Roe, Emmanuel Rollinde, Ashley J. Ross, Nicholas P. Ross, Bruno Rossetto, Ariel G. Sánchez, Basilio Santiago, Conor Sayres, Ricardo Schiavon, David J. Schlegel, Katharine J. Schlesinger, Sarah J. Schmidt, Donald P. Schneider, Kris Sellgren, Alaina Sheldon, Erin Sheldon, Matthew Shetrone, Yiping Shu, John D. Silverman, Jennifer Simmerer, Audrey E. Simmons, Thirupathi Sivarani, M. F. Skrutskie, Anže Slosar, Stephen Smee, Verne V. Smith, Stephanie A. Snedden, Keivan G. Stassun, Oliver Steele, Matthias Steinmetz, Mark H. Stockett, Todd Stollberg, Michael A. Strauss, Alexander S. Szalay, Masayuki Tanaka, Aniruddha R. Thakar, Daniel Thomas, Jeremy L. Tinker, Benjamin M. Tofflemire, Rita Tojeiro, Christy A. Tremonti, Mariana Vargas Magaña, Licia Verde, Nicole P. Vogt, David A. Wake, Xiaoke Wan, Ji Wang, Benjamin A. Weaver, Martin White, Simon D. M. White, John C. Wilson, John P. Wisniewski, W. Michael Wood-Vasey, Brian Yanny, Naoki Yasuda, Christophe Yèche, Donald G. York, Erick Young, Gail Zasowski, Idit Zehavi, and Bo Zhao. SDSS-III: Massive Spectroscopic Surveys of the Distant Universe, the Milky Way, and Extra-Solar Planetary Systems. *AJ*, 142(3):72, September 2011.

- [92] Kyle S. Dawson, David J. Schlegel, Christopher P. Ahn, Scott F. Anderson, Éric Aubourg, Stephen Bailey, Robert H. Barkhouser, Julian E. Bautista, Alessandra Beifiori, Andreas A. Berlind, Vaishali Bhardwaj, Dmitry Bizyaev, Cullen H. Blake, Michael R. Blanton, Michael Blomqvist, Adam S. Bolton, Arnaud Borde, Jo Bovy, W. N. Brandt, Howard Brewington, Jon Brinkmann, Peter J. Brown, Joel R. Brownstein, Kevin Bundy, N. G. Busca, William Carithers, Aurelio R. Carnero, Michael A. Carr, Yanmei Chen, Johan Comparat, Natalia Connolly, Frances Cope, Rupert A. C. Croft, Antonio J. Cuesta, Luiz N. da Costa, James R. A. Davenport, Timothée Delubac, Roland de Putter, Saurav Dhital, Anne Ealet, Garrett L. Ebelke, Daniel J. Eisenstein, S. Escoffier, Xiaohui Fan, N. Filiz Ak, Hayley Finley, Andreu Font-Ribera, R. Génova-Santos, James E. Gunn, Hong Guo, Daryl Haggard, Patrick B. Hall, Jean-Christophe Hamilton, Ben Harris, David W. Harris, Shirley Ho, David W. Hogg, Diana Holder, Klaus Honscheid, Joe Huehnerhoff, Beatrice Jordan, Wendell P. Jor-

dan, Guinevere Kauffmann, Eyal A. Kazin, David Kirkby, Mark A. Klaene, Jean-Paul Kneib, Jean-Marc Le Goff, Khee-Gan Lee, Daniel C. Long, Craig P. Loomis, Britt Lundgren, Robert H. Lupton, Marcio A. G. Maia, Martin Makler, Elena Malanushenko, Viktor Malanushenko, Rachel Mandelbaum, Marc Manera, Claudia Maraston, Daniel Margala, Karen L. Masters, Cameron K. McBride, Patrick McDonald, Ian D. McGreer, Richard G. McMahon, Olga Mena, Jordi Miralda-Escudé, Antonio D. Montero-Dorta, Francesco Montesano, Demitri Muna, Adam D. Myers, Tracy Naugle, Robert C. Nichol, Pasquier Noterdaeme, Sebastián E. Nuza, Matthew D. Olmstead, Audrey Oravetz, Daniel J. Oravetz, Russell Owen, Nikhil Padmanabhan, Nathalie Palanque-Delabrouille, Kaike Pan, John K. Parejko, Isabelle Pâris, Will J. Percival, Ismael Pérez-Fournon, Ignasi Pérez-Ràfols, Patrick Petitjean, Robert Pfaenger, Janine Pforr, Matthew M. Pieri, Francisco Prada, Adrian M. Price-Whelan, M. Jordan Raddick, Rafael Rebolo, James Rich, Gordon T. Richards, Constance M. Rockosi, Natalie A. Roe, Ashley J. Ross, Nicholas P. Ross, Graziano Rossi, J. A. Rubiño-Martín, Lado Samushia, Ariel G. Sánchez, Conor Sayres, Sarah J. Schmidt, Donald P. Schneider, C. G. Scóccola, Hee-Jong Seo, Alaina Shelden, Erin Sheldon, Yue Shen, Yiping Shu, Anže Slosar, Stephen A. Smeed, Stephanie A. Snedden, Fritz Stauffer, Oliver Steele, Michael A. Strauss, Alina Streblyanska, Nao Suzuki, Molly E. C. Swanson, Tomer Tal, Masayuki Tanaka, Daniel Thomas, Jeremy L. Tinker, Rita Tojeiro, Christy A. Tremonti, M. Vargas Magaña, Licia Verde, Matteo Viel, David A. Wake, Mike Watson, Benjamin A. Weaver, David H. Weinberg, Benjamin J. Weiner, Andrew A. West, Martin White, W. M. Wood-Vasey, Christophe Yèche, Idit Zehavi, Gong-Bo Zhao, and Zheng Zheng. The Baryon Oscillation Spectroscopic Survey of SDSS-III. *AJ*, 145(1):10, January 2013.

- [93] Shadab Alam, Franco D. Albareti, Carlos Allende Prieto, F. Anders, Scott F. Anderson, Timothy Anderton, Brett H. Andrews, Eric Armengaud, Éric Aubourg, Stephen Bailey, Sarbani Basu, Julian E. Bautista, Rachael L. Beaton, Timothy C. Beers, Chad F. Bender, Andreas A. Berlind, Florian Beutler, Vaishali Bhardwaj, Jonathan C. Bird, Dmitry Bizyaev, Cullen H. Blake, Michael R. Blanton, Michael Blomqvist, John J. Bochanski, Adam S. Bolton, Jo Bovy, A. Shelden Bradley, W. N. Brandt, D. E. Brauer, J. Brinkmann, Peter J. Brown, Joel R. Brownstein, Angela Burden, Etienne Burtin, Nicolás G. Busca, Zheng Cai, Diego Capozzi, Aurelio Carnero Rosell, Michael A. Carr, Ricardo Carrera, K. C. Chambers, William James Chaplin, Yen-Chi Chen, Cristina Chiappini, S. Drew Chojnowski, Chia-Hsun Chuang, Nicolas Clerc, Johan Comparat, Kevin Covey, Rupert A. C. Croft, Antonio J. Cuesta, Katia Cunha, Luiz N. da Costa, Nicola Da Rio, James R. A. Davenport, Kyle S. Dawson, Nathan De Lee, Timothée Delubac, Rohit Deshpande, Saurav Dhital, Letícia Dutra-Ferreira, Tom Dwelly, Anne Ealet, Garrett L. Ebelke, Edward M. Edmondson, Daniel J. Eisenstein, Tristan Ellsworth, Yvonne Elsworth, Courtney R. Epstein, Michael Eracleous, Stephanie Escoffier, Massimiliano Esposito, Michael L. Evans, Xiaohui Fan, Emma Fernández-Alvar, Diane Feuillet, Nurten Filiz Ak, Hayley Finley, Alexis Finoguenov, Kevin Flaherty, Scott W. Fleming, Andreu Font-Ribera, Jonathan Foster, Peter M. Frinchaboy, J. G. Galbraith-Frew, Rafael A. García, D. A. García-Hernández, Ana E. García Pérez, Patrick Gaulme, Jian Ge, R. Génova-Santos, A. Georgakakis, Luan

Ghezzi, Bruce A. Gillespie, Léo Girardi, Daniel Goddard, Satya Gontcho A. Gontcho, Jonay I. González Hernández, Eva K. Grebel, Paul J. Green, Jan Niklas Grieb, Nolan Grieves, James E. Gunn, Hong Guo, Paul Harding, Sten Hasselquist, Suzanne L. Hawley, Michael Hayden, Fred R. Hearty, Saskia Hekker, Shirley Ho, David W. Hogg, Kelly Holley-Bockelmann, Jon A. Holtzman, Klaus Honscheid, Daniel Huber, Joseph Huehnerhoff, Inese I. Ivans, Linhua Jiang, Jennifer A. Johnson, Karen Kinemuchi, David Kirkby, Francisco Kitaura, Mark A. Klaene, Gillian R. Knapp, Jean-Paul Kneib, Xavier P. Koenig, Charles R. Lam, Ting-Wen Lan, Dustin Lang, Pierre Laurent, Jean-Marc Le Goff, Alexie Leauthaud, Khee-Gan Lee, Young Sun Lee, Timothy C. Licquia, Jian Liu, Daniel C. Long, Martín López-Corredoira, Diego Lorenzo-Oliveira, Sara Lucatello, Britt Lundgren, Robert H. Lupton, III Mack, Claude E., Suvrath Mahadevan, Marcio A. G. Maia, Steven R. Majewski, Elena Malanushenko, Viktor Malanushenko, A. Machado, Marc Manera, Qingqing Mao, Claudia Maraston, Robert C. Marchwinski, Daniel Margala, Sarah L. Martell, Marie Martig, Karen L. Masters, Savita Mathur, Cameron K. McBride, Peregrine M. McGehee, Ian D. McGreer, Richard G. McMahon, Brice Ménard, Marie-Luise Menzel, Andrea Merloni, Szabolcs Mészáros, Adam A. Miller, Jordi Miralda-Escudé, Hironao Miyatake, Antonio D. Montero-Dorta, Surhud More, Eric Morganson, Xan Morice-Atkinson, Heather L. Morrison, Benoit Mosser, Demetri Muna, Adam D. Myers, Kirpal Nandra, Jeffrey A. Newman, Mark Neyrinck, Duy Cuong Nguyen, Robert C. Nichol, David L. Nidever, Pasquier Noterdaeme, Sebastián E. Nuza, Julia E. O’Connell, Robert W. O’Connell, Ross O’Connell, Ricardo L. C. Ogando, Matthew D. Olmstead, Audrey E. Oravetz, Daniel J. Oravetz, Keisuke Osumi, Russell Owen, Deborah L. Padgett, Nikhil Padmanabhan, Martin Paegert, Nathalie Palanque-Delabrouille, Kaike Pan, John K. Parejko, Isabelle Pâris, Changbom Park, Petchara Pattarakijwanich, M. Pellejero-Ibanez, Joshua Pepper, Will J. Percival, Ismael Pérez-Fournon, Ignasi Pérez-Ràfols, Patrick Petitjean, Matthew M. Pieri, Marc H. Pinsonneault, Gustavo F. Porto de Mello, Francisco Prada, Abhishek Prakash, Adrian M. Price-Whelan, Pavlos Protopapas, M. Jordan Raddick, Mubdi Rahman, Beth A. Reid, James Rich, Hans-Walter Rix, Annie C. Robin, Constance M. Rockosi, Thaïse S. Rodrigues, Sergio Rodríguez-Torres, Natalie A. Roe, Ashley J. Ross, Nicholas P. Ross, Graziano Rossi, John J. Ruan, J. A. Rubiño-Martín, Eli S. Rykoff, Salvador Salazar-Albornoz, Mara Salvato, Lado Samushia, Ariel G. Sánchez, Basílio Santiago, Conor Sayres, Ricardo P. Schiavon, David J. Schlegel, Sarah J. Schmidt, Donald P. Schneider, Mathias Schultheis, Axel D. Schwöpe, C. G. Scóccola, Caroline Scott, Kris Sellgren, Hee-Jong Seo, Aldo Serenelli, Neville Shane, Yue Shen, Matthew Shetrone, Yiping Shu, V. Silva Aguirre, Thirupathi Sivarani, M. F. Skrutskie, Anže Slosar, Verne V. Smith, Flávia Sobreira, Diogo Souto, Keivan G. Stassun, Matthias Steinmetz, Dennis Stello, Michael A. Strauss, Alina Streblyanska, Nao Suzuki, Molly E. C. Swanson, Jonathan C. Tan, Jamie Tayar, Ryan C. Terrien, Aniruddha R. Thakar, Daniel Thomas, Neil Thomas, Benjamin A. Thompson, Jeremy L. Tinker, Rita Tojeiro, Nicholas W. Troup, Mariana Vargas-Magaña, Jose A. Vazquez, Licia Verde, Matteo Viel, Nicole P. Vogt, David A. Wake, Ji Wang, Benjamin A. Weaver, David H. Weinberg, Benjamin J. Weiner, Martin White, John C. Wilson, John P. Wisniewski, W. M. Wood-Vasey, Christophe

- Ye'che, Donald G. York, Nadia L. Zakamska, O. Zamora, Gail Zasowski, Idit Zehavi, Gong-Bo Zhao, Zheng Zheng, Xu Zhou, Zhimin Zhou, Hu Zou, and Guangtun Zhu. The Eleventh and Twelfth Data Releases of the Sloan Digital Sky Survey: Final Data from SDSS-III. *ApJS*, 219(1):12, July 2015.
- [94] P. Noterdaeme, P. Petitjean, C. Ledoux, and R. Srianand. Evolution of the cosmological mass density of neutral gas from Sloan Digital Sky Survey II - Data Release 7. *A&A*, 505(3):1087–1098, October 2009.
- [95] Nicolas Busca and Christophe Balland. QuasarNET: Human-level spectral classification and redshifting with Deep Neural Networks. *arXiv e-prints*, page arXiv:1808.09955, August 2018.
- [96] Zhiyuan Guo and Paul Martini. Classification of Broad Absorption Line Quasars with a Convolutional Neural Network. *ApJ*, 879(2):72, July 2019.
- [97] Vikrant Kamble, Kyle Dawson, Hélion du Mas des Bourboux, Julian Bautista, and Donald P. Scheinder. Measurements of Effective Optical Depth in the Ly α Forest from the BOSS DR12 Quasar Sample. *ApJ*, 892(1):70, March 2020.
- [98] Joop Schaye. A Physical Upper Limit on the H I Column Density of Gas Clouds. *ApJL*, 562(1):L95–L98, November 2001.
- [99] Sultan Hassan, Kristian Finlator, Romeel Davé, Christopher W. Churchill, and J. Xavier Prochaska. Testing galaxy formation simulations with damped Lyman- α abundance and metallicity evolution. *MNRAS*, 492(2):2835–2846, February 2020.
- [100] Jens-Kristian Krogager, Johan P. U. Fynbo, Palle Møller, Pasquier Noterdaeme, Kasper E. Heintz, and Max Pettini. The effect of dust bias on the census of neutral gas and metals in the high-redshift Universe due to SDSS-II quasar colour selection. *MNRAS*, 486(3):4377–4397, July 2019.
- [101] T. M. C. Abbott, M. Agüena, A. Alarcon, S. Allam, S. Allen, J. Annis, S. Avila, D. Bacon, K. Bechtol, A. Bermeo, G. M. Bernstein, E. Bertin, S. Bhargava, S. Bocquet, D. Brooks, D. Brout, E. Buckley-Geer, D. L. Burke, A. Carnero Rosell, M. Carrasco Kind, J. Carretero, F. J. Castander, R. Cawthon, C. Chang, X. Chen, A. Choi, M. Costanzi, M. Crocce, L. N. da Costa, T. M. Davis, J. De Vicente, J. DeRose, S. Desai, H. T. Diehl, J. P. Dietrich, S. Dodelson, P. Doel, A. Drlica-Wagner, K. Eckert, T. F. Eifler, J. Elvin-Poole, J. Estrada, S. Everett, A. E. Evrard, A. Farahi, I. Ferrero, B. Flaugher, P. Fosalba, J. Frieman, J. García-Bellido, M. Gatti, E. Gaztanaga, D. W. Gerdes, T. Giannantonio, P. Giles, S. Grandis, D. Gruen, R. A. Gruendl, J. Gschwend, G. Gutierrez, W. G. Hartley, S. R. Hinton, D. L. Hollowood, K. Honscheid, B. Hoyle, D. Huterer, D. J. James, M. Jarvis, T. Jeltema, M. W. G. Johnson, M. D. Johnson, S. Kent, E. Krause, R. Kron, K. Kuehn, N. Kuropatkin, O. Lahav, T. S. Li, C. Lidman, M. Lima, H. Lin, N. MacCrann, M. A. G. Maia, A. Mantz, J. L. Marshall, P. Martini, J. Mayers, P. Melchior, J. Mena-Fernández, F. Menanteau, R. Miquel, J. J. Mohr, R. C. Nichol, B. Nord, R. L. C. Ogando, A. Palmese, F. Paz-Chinchón,

- A. A. Plazas, J. Prat, M. M. Rau, A. K. Romer, A. Roodman, P. Rooney, E. Rozo, E. S. Rykoff, M. Sako, S. Samuroff, C. Sánchez, E. Sanchez, A. Saro, V. Scarpine, M. Schubnell, D. Scolnic, S. Serrano, I. Sevilla-Noarbe, E. Sheldon, J. Allyn. Smith, M. Smith, E. Suchyta, M. E. C. Swanson, G. Tarle, D. Thomas, C. To, M. A. Troxel, D. L. Tucker, T. N. Varga, A. von der Linden, A. R. Walker, R. H. Wechsler, J. Weller, R. D. Wilkinson, H. Wu, B. Yanny, Y. Zhang, Z. Zhang, J. Zuntz, and DES Collaboration. Dark Energy Survey Year 1 Results: Cosmological constraints from cluster abundances and weak lensing. *PhRvD*, 102(2):023509, July 2020.
- [102] J. Anthony Tyson. Large Synoptic Survey Telescope: Overview. In J. Anthony Tyson and Sidney Wolff, editors, *Survey and Other Telescope Technologies and Discoveries*, volume 4836 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, pages 10–20, December 2002.
- [103] Luca Amendola, Stephen Appleby, Anastasios Avgoustidis, David Bacon, Tessa Baker, Marco Baldi, Nicola Bartolo, Alain Blanchard, Camille Bonvin, Stefano Borgani, Enzo Branchini, Clare Burrage, Stefano Camera, Carmelita Carbone, Luciano Casarini, Mark Cropper, Claudia de Rham, Jörg P. Dietrich, Cinzia Di Porto, Ruth Durrer, Anne Ealet, Pedro G. Ferreira, Fabio Finelli, Juan García-Bellido, Tommaso Giannantonio, Luigi Guzzo, Alan Heavens, Lavinia Heisenberg, Catherine Heymans, Henk Hoekstra, Lukas Hollenstein, Rory Holmes, Zhiqi Hwang, Knud Jahnke, Thomas D. Kitching, Tomi Koivisto, Martin Kunz, Giuseppe La Vacca, Eric Linder, Marisa March, Valerio Marra, Carlos Martins, Elisabetta Majerotto, Dida Markovic, David Marsh, Federico Marulli, Richard Massey, Yannick Mellier, Francesco Montanari, David F. Mota, Nelson J. Nunes, Will Percival, Valeria Pettorino, Cristiano Porciani, Claudia Quercellini, Justin Read, Massimiliano Rinaldi, Domenico Sapone, Ignacy Sawicki, Roberto Scaramella, Constantinos Skordis, Fergus Simpson, Andy Taylor, Shaun Thomas, Roberto Trotta, Licia Verde, Filippo Vernizzi, Adrian Vollmer, Yun Wang, Jochen Weller, and Tom Zlosnik. Cosmology and fundamental physics with the Euclid satellite. *Living Reviews in Relativity*, 21(1):2, April 2018.
- [104] DESI Collaboration, Amir Aghamousa, Jessica Aguilar, Steve Ahlen, Shadab Alam, Lori E. Allen, Carlos Allende Prieto, James Annis, Stephen Bailey, Christophe Balleland, Otger Ballester, Charles Baltay, Lucas Beaufore, Chris Bebek, Timothy C. Beers, Eric F. Bell, José Luis Bernal, Robert Besuner, Florian Beutler, Chris Blake, Hannes Bleuler, Michael Blomqvist, Robert Blum, Adam S. Bolton, Cesar Briceno, David Brooks, Joel R. Brownstein, Elizabeth Buckley-Geer, Angela Burden, Etienne Burtin, Nicolas G. Busca, Robert N. Cahn, Yan-Chuan Cai, Laia Cardiel-Sas, Raymond G. Carlberg, Pierre-Henri Carton, Ricard Casas, Francisco J. Castander, Jorge L. Cervantes-Cota, Todd M. Claybaugh, Madeline Close, Carl T. Coker, Shaun Cole, Johan Comparat, Andrew P. Cooper, M. C. Cousinou, Martin Crocce, Jean-Gabriel Cuby, Daniel P. Cunningham, Tamara M. Davis, Kyle S. Dawson, Axel de la Macorra, Juan De Vicente, Timothée Delubac, Mark Derwent, Arjun Dey, Govinda Dhungana, Zhejie Ding, Peter Doel, Yutong T. Duan, Anne Ealet, Jerry Edelstein, Sarah Eftekharzadeh, Daniel J. Eisenstein, Ann Elliott, Stéphanie Escoffier, Matthew Evatt, Parker Fagrelus, Xiaohui Fan, Kevin Fanning, Arya Farahi,

Jay Farihi, Ginevra Favole, Yu Feng, Enrique Fernandez, Joseph R. Findlay, Douglas P. Finkbeiner, Michael J. Fitzpatrick, Brenna Flaughner, Samuel Flender, Andreu Font-Ribera, Jaime E. Forero-Romero, Pablo Fosalba, Carlos S. Frenk, Michele Fumagalli, Boris T. Gaensicke, Giuseppe Gallo, Juan Garcia-Bellido, Enrique Gaztanaga, Nicola Pietro Gentile Fusillo, Terry Gerard, Irena Gershkovich, Tommaso Giannantonio, Denis Gillet, Guillermo Gonzalez-de-Rivera, Violeta Gonzalez-Perez, Shelby Gott, Or Graur, Gaston Gutierrez, Julien Guy, Salman Habib, Henry Heetderks, Ian Heetderks, Katrin Heitmann, Wojciech A. Hellwing, David A. Herrera, Shirley Ho, Stephen Holland, Klaus Honscheid, Eric Huff, Timothy A. Hutchinson, Dragan Huterer, Ho Seong Hwang, Joseph Maria Illa Laguna, Yuzo Ishikawa, Dianna Jacobs, Niall Jeffrey, Patrick Jelinsky, Elise Jennings, Linhua Jiang, Jorge Jimenez, Jennifer Johnson, Richard Joyce, Eric Jullo, Stéphanie Juneau, Sami Kama, Armin Karcher, Sonia Karkar, Robert Kehoe, Noble Kennamer, Stephen Kent, Martin Kilbinger, Alex G. Kim, David Kirkby, Theodore Kisner, Ellie Kitanidis, Jean-Paul Kneib, Sergey Koposov, Eve Kovacs, Kazuya Koyama, Anthony Kremin, Richard Kron, Luzius Kronig, Andrea Kueter-Young, Cedric G. Lacey, Robin Lafever, Ofer Lahav, Andrew Lambert, Michael Lampton, Martin Landriau, Dustin Lang, Tod R. Lauer, Jean-Marc Le Goff, Laurent Le Guillou, Auguste Le Van Suu, Jae Hyeon Lee, Su-Jeong Lee, Daniela Leitner, Michael Lesser, Michael E. Levi, Benjamin L’Huillier, Baojiu Li, Ming Liang, Huan Lin, Eric Linder, Sarah R. Loebman, Zarija Lukić, Jun Ma, Niall MacCrann, Christophe Magneville, Laleh Makarem, Marc Manera, Christopher J. Manser, Robert Marshall, Paul Martini, Richard Massey, Thomas Matheson, Jeremy McCauley, Patrick McDonald, Ian D. McGreer, Aaron Meisner, Nigel Metcalfe, Timothy N. Miller, Ramon Miquel, John Moustakas, Adam Myers, Milind Naik, Jeffrey A. Newman, Robert C. Nichol, Andrina Nicola, Luiz Nicolati da Costa, Jundan Nie, Gustavo Niz, Peder Norberg, Brian Nord, Dara Norman, Peter Nugent, Thomas O’Brien, Minji Oh, Knut A. G. Olsen, Cristobal Padilla, Hamsa Padmanabhan, Nikhil Padmanabhan, Nathalie Palanque-Delabrouille, Antonella Palmese, Daniel Pappalardo, Isabelle Pâris, Changbom Park, Anna Patej, John A. Peacock, Hiranya V. Peiris, Xiyan Peng, Will J. Percival, Sandrine Perruchot, Matthew M. Pieri, Richard Pogge, Jennifer E. Pollack, Claire Poppett, Francisco Prada, Abhishek Prakash, Ronald G. Probst, David Rabinowitz, Anand Raichoor, Chang Hee Ree, Alexandre Refregier, Xavier Regal, Beth Reid, Kevin Reil, Mehdi Rezaie, Constance M. Rockosi, Natalie Roe, Samuel Ronayette, Aaron Roodman, Ashley J. Ross, Nicholas P. Ross, Graziano Rossi, Eduardo Roza, Vanina Ruhlmann-Kleider, Eli S. Rykoff, Cristiano Sabiu, Lado Samushia, Eusebio Sanchez, Javier Sanchez, David J. Schlegel, Michael Schneider, Michael Schubnell, Aurélie Secroun, Uros Seljak, Hee-Jong Seo, Santiago Serrano, Arman Shafieloo, Huanyuan Shan, Ray Sharples, Michael J. Sholl, William V. Shourt, Joseph H. Silber, David R. Silva, Martin M. Sirk, Anze Slosar, Alex Smith, George F. Smoot, Debopam Som, Yong-Seon Song, David Sprayberry, Ryan Staten, Andy Stefanik, Gregory Tarle, Suk Sien Tie, Jeremy L. Tinker, Rita Tojeiro, Francisco Valdes, Octavio Valenzuela, Monica Valluri, Mariana Vargas-Magana, Licia Verde, Alistair R. Walker, Jiali Wang, Yuting Wang, Benjamin A. Weaver, Curtis Weaverdyck, Risa H. Wechsler, David H. Weinberg, Mar-

- tin White, Qian Yang, Christophe Yeche, Tianmeng Zhang, Gong-Bo Zhao, Yi Zheng, Xu Zhou, Zhimin Zhou, Yaling Zhu, Hu Zou, and Ying Zu. The DESI Experiment Part I: Science, Targeting, and Survey Design. *arXiv e-prints*, page arXiv:1611.00036, October 2016.
- [105] D. Spergel, N. Gehrels, J. Breckinridge, M. Donahue, A. Dressler, B. S. Gaudi, T. Greene, O. Guyon, C. Hirata, J. Kalirai, N. J. Kasdin, W. Moos, S. Perlmutter, M. Postman, B. Rauscher, J. Rhodes, Y. Wang, D. Weinberg, J. Centrella, W. Traub, C. Baltay, J. Colbert, D. Bennett, A. Kiessling, B. Macintosh, J. Merten, M. Mortonson, M. Penny, E. Rozo, D. Savransky, K. Stapelfeldt, Y. Zu, C. Baker, E. Cheng, D. Content, J. Dooley, M. Foote, R. Goullioud, K. Grady, C. Jackson, J. Kruk, M. Levine, M. Melton, C. Peddie, J. Ruffa, and S. Shaklan. Wide-Field InfraRed Survey Telescope-Astrophysics Focused Telescope Assets WFIRST-AFTA Final Report. *arXiv e-prints*, page arXiv:1305.5422, May 2013.
- [106] Robert R. Caldwell and Marc Kamionkowski. The Physics of Cosmic Acceleration. *Annual Review of Nuclear and Particle Science*, 59(1):397–429, November 2009.
- [107] Jonathan L. Feng. Dark Matter Candidates from Particle Physics and Methods of Detection. *ARA&A*, 48:495–545, September 2010.
- [108] Yvonne Y. Y. Wong. Neutrino Mass in Cosmology: Status and Prospects. *Annual Review of Nuclear and Particle Science*, 61(1):69–98, November 2011.
- [109] R. W. Hockney and J. W. Eastwood. *Computer simulation using particles*. 1988.
- [110] Josh Barnes and Piet Hut. A hierarchical $O(N \log N)$ force-calculation algorithm. *Nature*, 324(6096):446–449, December 1986.
- [111] H. M. P. Couchman, P. A. Thomas, and F. R. Pearce. Hydra: an Adaptive-Mesh Implementation of P 3M-SPH. *ApJ*, 452:797, October 1995.
- [112] L. Greengard and V. Rokhlin. A Fast Algorithm for Particle Simulations. *Journal of Computational Physics*, 73(2):325–348, December 1987.
- [113] Walter Dehnen. A Hierarchical $O(N^2)$ Force Calculation Algorithm. *Journal of Computational Physics*, 179(1):27–42, June 2002.
- [114] Salman Habib, Katrin Heitmann, David Higdon, Charles Nakhleh, and Brian Williams. Cosmic calibration: Constraints from the matter power spectrum and the cosmic microwave background. *PhRvD*, 76(8):083503, October 2007.
- [115] Earl Lawrence, Katrin Heitmann, Martin White, David Higdon, Christian Wagner, Salman Habib, and Brian Williams. The Coyote Universe. III. Simulation Suite and Precision Emulator for the Nonlinear Matter Power Spectrum. *ApJ*, 713(2):1322–1331, April 2010.
- [116] Katrin Heitmann, Earl Lawrence, Juliana Kwan, Salman Habib, and David Higdon. The Coyote Universe Extended: Precision Emulation of the Matter Power Spectrum. *ApJ*, 780(1):111, January 2014.

- [117] Katrin Heitmann, Derek Bingham, Earl Lawrence, Steven Bergner, Salman Habib, David Higdon, Adrian Pope, Rahul Biswas, Hal Finkel, Nicholas Frontiere, and Suman Bhattacharya. The Mira-Titan Universe: Precision Predictions for Dark Energy Surveys. *ApJ*, 820(2):108, April 2016.
- [118] Earl Lawrence, Katrin Heitmann, Juliana Kwan, Amol Upadhye, Derek Bingham, Salman Habib, David Higdon, Adrian Pope, Hal Finkel, and Nicholas Frontiere. The Mira-Titan Universe. II. Matter Power Spectrum Emulation. *ApJ*, 847(1):50, September 2017.
- [119] Shankar Agarwal, Filipe B. Abdalla, Hume A. Feldman, Ofer Lahav, and Shaun A. Thomas. PkANN - II. A non-linear matter power spectrum interpolator developed using artificial neural networks. *MNRAS*, 439(2):2102–2121, April 2014.
- [120] Sebastian Bocquet, Katrin Heitmann, Salman Habib, Earl Lawrence, Thomas Uram, Nicholas Frontiere, Adrian Pope, and Hal Finkel. The Mira-Titan Universe. III. Emulation of the Halo Mass Function. *ApJ*, 901(1):5, September 2020.
- [121] Juliana Kwan, Suman Bhattacharya, Katrin Heitmann, and Salman Habib. Cosmic Emulation: The Concentration-Mass Relation for Λ CDM Universes. *ApJ*, 768(2):123, May 2013.
- [122] Juliana Kwan, Katrin Heitmann, Salman Habib, Nikhil Padmanabhan, Earl Lawrence, Hal Finkel, Nicholas Frontiere, and Adrian Pope. Cosmic Emulation: Fast Predictions for the Galaxy Power Spectrum. *ApJ*, 810(1):35, September 2015.
- [123] Zhongxu Zhai, Jeremy L. Tinker, Matthew R. Becker, Joseph DeRose, Yao-Yuan Mao, Thomas McClintock, Sean McLaughlin, Eduardo Rozo, and Risa H. Wechsler. The Aemulus Project. III. Emulation of the Galaxy Correlation Function. *ApJ*, 874(1):95, March 2019.
- [124] Thomas McClintock, Eduardo Rozo, Matthew R. Becker, Joseph DeRose, Yao-Yuan Mao, Sean McLaughlin, Jeremy L. Tinker, Risa H. Wechsler, and Zhongxu Zhai. The Aemulus Project. II. Emulating the Halo Mass Function. *ApJ*, 872(1):53, February 2019.
- [125] Jia Liu, Andrea Petri, Zoltán Haiman, Lam Hui, Jan M. Kratochvil, and Morgan May. Cosmology constraints from the weak lensing peak counts and the power spectrum in CFHTLenS data. *PhRvD*, 91(6):063507, March 2015.
- [126] J. Harnois-Déraps, B. Giblin, and B. Joachimi. Cosmic shear covariance matrix in Λ CDM: Cosmology matters. *A&A*, 631:A160, November 2019.
- [127] Christopher T. Davies, Marius Cautun, Benjamin Giblin, Baojiu Li, Joachim Harnois-Déraps, and Yan-Chuan Cai. Constraining cosmology with weak lensing voids. *arXiv e-prints*, page arXiv:2010.11954, October 2020.

- [128] Simeon Bird, Keir K. Rogers, Hiranya V. Peiris, Licia Verde, Andreu Font-Ribera, and Andrew Pontzen. An emulator for the Lyman- α forest. *JCAP*, 2019(2):050, February 2019.
- [129] Benjamin Giblin, Matteo Cataneo, Ben Moews, and Catherine Heymans. On the road to per cent accuracy - II. Calibration of the non-linear matter power spectrum for arbitrary cosmologies. *MNRAS*, 490(4):4826–4840, December 2019.
- [130] Christian Pedersen, Andreu Font-Ribera, Keir K. Rogers, Patrick McDonald, Hiranya V. Peiris, Andrew Pontzen, and Anže Slosar. An emulator for the Lyman- α forest in beyond- Λ CDM cosmologies. *JCAP*, 2021(5):033, May 2021.
- [131] Nesar Ramachandra, Georgios Valogiannis, Mustapha Ishak, Katrin Heitmann, and LSST Dark Energy Science Collaboration. Matter power spectrum emulator for $f(R)$ modified gravity cosmologies. *PhRvD*, 103(12):123525, June 2021.
- [132] Edwin V Bonilla, Kian Chai, and Christopher Williams. Multi-task gaussian process prediction. volume 20. NIPS, 2008.
- [133] Zarija Lukić, Casey W. Stark, Peter Nugent, Martin White, Avery A. Meiksin, and Ann Almgren. The Lyman α forest in optically thin hydrodynamical simulations. *MNRAS*, 446(4):3697–3724, February 2015.
- [134] Nicolas Chartier, Benjamin Wandelt, Yashar Akrami, and Francisco Villaescusa-Navarro. CARPool: fast, accurate computation of large-scale structure statistics by pairing costly and cheap cosmological simulations. *arXiv e-prints*, page arXiv:2009.08970, September 2020.
- [135] L. F. Richardson. The Approximate Arithmetical Solution by Finite Differences of Physical Problems Involving Differential Equations, with an Application to the Stresses in a Masonry Dam. *Philosophical Transactions of the Royal Society of London Series A*, 210:307–357, January 1911.
- [136] Doogesh Kodi Ramanah, Tom Charnock, Francisco Villaescusa-Navarro, and Benjamin D. Wandelt. Super-resolution emulator of cosmological simulations using deep physical models. *MNRAS*, 495(4):4227–4236, July 2020.
- [137] Yin Li, Yueying Ni, Rupert A. C. Croft, Tiziana Di Matteo, Simeon Bird, and Yu Feng. AI-assisted super-resolution cosmological simulations. *arXiv e-prints*, page arXiv:2010.06608, October 2020.
- [138] Keir K. Rogers, Hiranya V. Peiris, Andrew Pontzen, Simeon Bird, Licia Verde, and Andreu Font-Ribera. Bayesian emulator optimisation for cosmology: application to the Lyman-alpha forest. *JCAP*, 2019(2):031, February 2019.
- [139] Florent Leclercq. Bayesian optimization for likelihood-free cosmological inference. *PhRvD*, 98(6):063511, September 2018.

- [140] Timur Takhtaganov, Zarija Lukić, Juliane Müller, and Dmitriy Morozov. Cosmic Inference: Constraining Parameters with Observations and a Highly Limited Number of Simulations. *ApJ*, 906(2):74, January 2021.
- [141] Marcos Pellejero-Ibañez, Raul E. Angulo, Giovanni Aricó, Matteo Zennaro, Sergio Contreras, and Jens Stücker. Cosmological parameter estimation via iterative emulation of likelihoods. *MNRAS*, 499(4):5257–5268, December 2020.
- [142] D. Huang, T. T. Allen, W. I. Notz, and R. A. Miller. Sequential kriging optimization using multiple-fidelity evaluations. *Struct. Multidisc. Optim.*, 32:369–382, November 2006.
- [143] Alexander I.J Forrester, Andràs Sóbester, and Andy J. Keane. Multi-fidelity optimization via surrogate modelling. *Proc. R. Soc. A.*, 463:3251–3269, October 2007.
- [144] Remi Lam, Douglas Allaire, and Karen E Willcox. Multifidelity Optimization using Statistical Surrogate Modeling for Non-Hierarchical Information Sources. *56th AIAA/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*, January 2015.
- [145] Matthias Poloczek, Jialei Wang, and Peter I. Frazier. Multi-Information Source Optimization. *arXiv e-prints*, page arXiv:1603.00389, March 2016.
- [146] Mark McLeod, Michael A. Osborne, and Stephen J. Roberts. Practical Bayesian Optimization for Variable Cost Objectives. *arXiv e-prints*, page arXiv:1703.04335, March 2017.
- [147] Peter I. Frazier. A Tutorial on Bayesian Optimization. *arXiv e-prints*, page arXiv:1807.02811, July 2018.
- [148] Volker Springel and Lars Hernquist. Cosmological smoothed particle hydrodynamics simulations: a hybrid multiphase model for star formation. *MNRAS*, 339(2):289–311, February 2003.
- [149] Yu Feng, Simeon Bird, Lauren Anderson, Andreu Font-Ribera, and Chris Pedersen. Mp-gadget/mp-gadget: A tag for getting a doi, October 2018.
- [150] Marcel P. van Daalen, Joop Schaye, C. M. Booth, and Claudio Dalla Vecchia. The effects of galaxy formation on the matter power spectrum: a challenge for precision cosmology. *MNRAS*, 415(4):3649–3665, August 2011.
- [151] Martin White. Baryons and weak lensing power spectra. *Astroparticle Physics*, 22(2):211–217, November 2004.
- [152] Aurel Schneider, Nicola Stoira, Alexandre Refregier, Andreas J. Weiss, Mischa Knabenhans, Joachim Stadel, and Romain Teyssier. Baryonic effects for weak lensing. Part I. Power spectrum and covariance matrix. *JCAP*, 2020(4):019, April 2020.

- [153] P. Perdikaris, M. Raissi, A. Damianou, N. D. Lawrence, and G. E. Karniadakis. Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling. *Proc. R. Soc. A.*, 473(20160751), 2017.
- [154] Kurt Cutajar, Mark Pullin, Andreas Damianou, Neil Lawrence, and Javier González. Deep Gaussian Processes for Multi-fidelity Modeling. *arXiv e-prints*, page arXiv:1903.07320, March 2019.
- [155] Andrei Paleyes, Mark Pullin, Maren Mahsereci, Neil Lawrence, and Javier González. Emulation of physical processes with emukit. In *Second Workshop on Machine Learning and the Physical Sciences, NIPS*, 2019.
- [156] Yu Feng, Simeon Bird, Lauren Anderson, Andreu Font-Ribera, and Chris Pedersen. Mp-gadget/mp-gadget: A tag for getting a doi, October 2018.
- [157] Julien Lesgourgues. The Cosmic Linear Anisotropy Solving System (CLASS) I: Overview. *arXiv e-prints*, page arXiv:1104.2932, April 2011.
- [158] Y. B. Zel'Dovich. Reprint of 1970A&A.....5...84Z. Gravitational instability: an approximate theory for large density perturbations. *A&A*, 500:13–18, March 1970.
- [159] Raul E. Angulo and Andrew Pontzen. Cosmological N-body simulations with suppressed variance. *MNRAS*, 462(1):L1–L5, October 2016.
- [160] G. Hinshaw, D. Larson, E. Komatsu, D. N. Spergel, C. L. Bennett, J. Dunkley, M. R. Nolta, M. Halpern, R. S. Hill, N. Odegard, L. Page, K. M. Smith, J. L. Weiland, B. Gold, N. Jarosik, A. Kogut, M. Limon, S. S. Meyer, G. S. Tucker, E. Wollack, and E. L. Wright. Nine-year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Cosmological Parameter Results. *ApJS*, 208(2):19, October 2013.
- [161] Benjamin Peherstorfer, Karen Willcox, and Max Gunzburger. Survey of multifidelity methods in uncertainty propagation, inference, and optimization. *SIAM Review*, 60(3):550–591, 2018.
- [162] Andreas Damianou and Neil Lawrence. Deep gaussian processes. In Carlos M. Carvalho and Pradeep Ravikumar, editors, *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pages 207–215, Scottsdale, Arizona, USA, 29 Apr–01 May 2013. PMLR.
- [163] Simon D. M. White. Formation and Evolution of Galaxies: Les Houches Lectures. *arXiv e-prints*, pages astro-ph/9410043, Oct 1994.
- [164] Giovanni Aricò, Raul E. Angulo, Sergio Contreras, Lurdes Ondaro-Mallea, Marcos Pellejero-Ibañez, and Matteo Zennaro. The BACCO Simulation Project: A baryonification emulator with Neural Networks. *arXiv e-prints*, page arXiv:2011.15018, November 2020.

- [165] GPpy. GPpy: A gaussian process framework in python. <http://github.com/SheffieldML/GPy>, since 2012.
- [166] Carla Currin, Toby Mitchell, Max Morris, and Don Ylvisaker. Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. *Journal of the American Statistical Association*, 86(416):953–963, 1991.
- [167] A. O’Hagan. Bayesian analysis of computer code outputs: A tutorial. *Reliability Engineering & System Safety*, 91(10):1290–1300, 2006. The Fourth International Conference on Sensitivity Analysis of Model Output (SAMO 2004).
- [168] T. Auld, M. Bridges, M. P. Hobson, and S. F. Gull. Fast cosmological parameter estimation using neural networks. *MNRAS*, 376(1):L11–L15, March 2007.
- [169] T. Auld, M. Bridges, and M. P. Hobson. COSMONET: fast cosmological parameter estimation in non-flat models using neural networks. *MNRAS*, 387(4):1575–1582, July 2008.
- [170] Giovanni Aricò, Raul E. Angulo, and Matteo Zennaro. Accelerating large-scale-structure data analyses by emulating boltzmann solvers and lagrangian perturbation theory [version 2; peer review: 2 approved]. *Open Research Europe*, 1(152), 2022.
- [171] Alessio Spurio Mancini, Davide Piras, Justin Alsing, Benjamin Joachimi, and Michael P. Hobson. COSMOPower: emulating cosmological power spectra for accelerated Bayesian inference from next-generation surveys. *MNRAS*, 511(2):1771–1788, April 2022.
- [172] Andreas Nygaard, Emil Brinch Holm, Steen Hannestad, and Thomas Tram. CONNECT: A neural network based framework for emulating cosmological observables and cosmological parameter inference. *arXiv e-prints*, page arXiv:2205.15726, May 2022.
- [173] Sven Günther, Julien Lesgourgues, Georgios Samaras, Nils Schöneberg, Florian Stadtmann, Christian Fidler, and Jesús Torrado. CosmicNet II: Emulating extended cosmologies with efficient and accurate neural networks. *arXiv e-prints*, page arXiv:2207.05707, July 2022.
- [174] Jonas El Gammal, Nils Schöneberg, Jesús Torrado, and Christian Fidler. Fast and robust Bayesian Inference using Gaussian Processes with GPpy. *arXiv e-prints*, page arXiv:2211.02045, November 2022.
- [175] Euclid Collaboration, Mischa Knabenhans, Joachim Stadel, Stefano Marelli, Doug Potter, Romain Teyssier, Laurent Legrand, Aurel Schneider, Bruno Sudret, Linda Blot, Saeeda Awan, Carlo Burigana, Carla Sofia Carvalho, Hannu Kurki-Suonio, and Gabriele Sirri. Euclid preparation: II. The EUCLIDEMULATOR - a tool to compute the cosmology dependence of the nonlinear matter power spectrum. *MNRAS*, 484(4):5509–5529, April 2019.

- [176] Sambit K. Giri and Aurel Schneider. Emulation of baryonic effects on the matter power spectrum and constraints from galaxy cluster data. *JCAP*, 2021(12):046, December 2021.
- [177] Christian Arnold, Baojiu Li, Benjamin Giblin, Joachim Harnois-Déraps, and Yan-Chuan Cai. FORGE: the f(R)-gravity cosmic emulator project - I. Introduction and matter power spectrum emulator. *MNRAS*, 515(3):4161–4175, September 2022.
- [178] Joachim Harnois-Déraps, Cesar Hernandez-Aguayo, Carolina Cuesta-Lazaro, Christian Arnold, Baojiu Li, Christopher T. Davies, and Yan-Chuan Cai. MGLENS: Modified gravity weak lensing simulations for emulation-based cosmological inference. *MNRAS*, 525(4):6336–6358, November 2023.
- [179] Christopher T. Davies, Marius Cautun, Benjamin Giblin, Baojiu Li, Joachim Harnois-Déraps, and Yan-Chuan Cai. Constraining cosmology with weak lensing voids. *MNRAS*, 507(2):2267–2282, October 2021.
- [180] Benjamin Giblin, Yan-Chuan Cai, and Joachim Harnois-Déraps. Enhancing cosmic shear with the multiscale lensing probability density function. *MNRAS*, 520(2):1721–1737, April 2023.
- [181] Takahiro Nishimichi, Masahiro Takada, Ryuichi Takahashi, Ken Osato, Masato Shirasaki, Taira Oogi, Hironao Miyatake, Masamune Oguri, Ryoma Murata, Yosuke Kobayashi, and Naoki Yoshida. Dark Quest. I. Fast and Accurate Emulation of Halo Clustering Statistics and Its Application to Galaxy Clustering. *ApJ*, 884(1):29, October 2019.
- [182] Sebastian Bocquet, Katrin Heitmann, Salman Habib, Earl Lawrence, Thomas Uram, Nicholas Frontiere, Adrian Pope, and Hal Finkel. The Mira-Titan Universe. III. Emulation of the Halo Mass Function. *ApJ*, 901(1):5, September 2020.
- [183] Aviad Cohen, Anastasia Fialkov, Rennan Barkana, and Raul A. Monsalve. Emulating the global 21-cm signal from Cosmic Dawn and Reionization. *MNRAS*, 495(4):4845–4859, July 2020.
- [184] H. T. J. Bevins, W. J. Handley, A. Fialkov, E. de Lera Acedo, and K. Javid. GLOB-ALEMU: a novel and robust approach for emulating the sky-averaged 21-cm signal from the cosmic dawn and epoch of reionization. *MNRAS*, 508(2):2923–2936, December 2021.
- [185] Christian H. Bye, Stephen K. N. Portillo, and Anastasia Fialkov. 21cmVAE: A Very Accurate Emulator of the 21 cm Global Signal. *ApJ*, 930(1):79, May 2022.
- [186] Keir K. Rogers and Hiranya V. Peiris. General framework for cosmological dark matter bounds using N -body simulations. *PhRvD*, 103(4):043526, February 2021.
- [187] Keir K. Rogers and Hiranya V. Peiris. Strong Bound on Canonical Ultralight Axion Dark Matter from the Lyman-Alpha Forest. *PhRvL*, 126(7):071302, February 2021.

- [188] L. Cabayol-Garcia, J. Chaves-Montero, A. Font-Ribera, and C. Pedersen. A neural network emulator for the Lyman- α forest 1D flux power spectrum. *MNRAS*, 525(3):3499–3515, November 2023.
- [189] D. Zürcher, J. Fluri, R. Sgier, T. Kacprzak, M. Gatti, C. Doux, L. Whiteway, A. Réfrégier, C. Chang, N. Jeffrey, B. Jain, P. Lemos, D. Bacon, A. Alarcon, A. Amon, K. Bechtol, M. Becker, G. Bernstein, A. Campos, R. Chen, A. Choi, C. Davis, J. Derose, S. Dodelson, F. Elsner, J. Elvin-Poole, S. Everett, A. Ferte, D. Gruen, I. Harrison, D. Huterer, M. Jarvis, P. F. Leget, N. Maccrann, J. Mccullough, J. Muir, J. Myles, A. Navarro Alsina, S. Pandey, J. Prat, M. Raveri, R. P. Rollins, A. Roodman, C. Sanchez, L. F. Secco, E. Sheldon, T. Shin, M. Troxel, I. Tutusaus, B. Yin, M. Agüena, S. Allam, F. Andrade-Oliveira, J. Annis, E. Bertin, D. Brooks, D. Burke, A. Carnero Rosell, M. Carrasco Kind, J. Carretero, F. Castander, R. Cawthon, C. Conselice, M. Costanzi, L. da Costa, M. E. da Silva Pereira, T. Davis, J. De Vicente, S. Desai, H. T. Diehl, J. Dietrich, P. Doel, K. Eckert, A. Evrard, I. Ferrero, B. Flaugher, P. Fosalba, D. Friedel, J. Frieman, J. Garcia-Bellido, E. Gaztanaga, D. Gerdes, T. Giannantonio, R. Gruendl, J. Gschwend, G. Gutierrez, S. Hinton, D. L. Hollowood, K. Honscheid, B. Hoyle, D. James, K. Kuehn, N. Kuropatkin, O. Lahav, C. Lidman, M. Lima, M. Maia, J. Marshall, P. Melchior, F. Menanteau, R. Miquel, R. Morgan, A. Palmese, F. Paz-Chinchon, A. Pieres, A. Plazas Malagón, K. Reil, M. Rodriguez Monroy, K. Romer, E. Sanchez, V. Scarpine, M. Schubnell, S. Serrano, I. Sevilla, M. Smith, E. Suchyta, G. Tarle, D. Thomas, C. To, T. N. Varga, J. Weller, R. Wilkinson, and DES Collaboration. Dark energy survey year 3 results: Cosmology with peaks using an emulator approach. *MNRAS*, 511(2):2075–2104, April 2022.
- [190] Richard Neveux, Etienne Burtin, Vanina Ruhlmann-Kleider, Arnaud de Mattia, Agne Semenaite, Kyle S. Dawson, Axel de la Macorra, Will J. Percival, Graziano Rossi, Donald P. Schneider, and Gong-Bo Zhao. Combined full shape analysis of BOSS galaxies and eBOSS quasars using an iterative emulator. *arXiv e-prints*, page arXiv:2201.04679, January 2022.
- [191] Yongseok Jo, Shy Genel, Benjamin Wandelt, Rachel S. Somerville, Francisco Villaescusa-Navarro, Greg L. Bryan, Daniel Anglés-Alcázar, Daniel Foreman-Mackey, Dylan Nelson, and Ji-hoon Kim. Calibrating Cosmological Simulations with Implicit Likelihood Inference Using Galaxy Growth Observables. *ApJ*, 944(1):67, February 2023.
- [192] Jaime Salcido, Ian G. McCarthy, Juliana Kwan, Amol Upadhye, and Andreea S. Font. SP(k) - a hydrodynamical simulation-based model for the impact of baryon physics on the non-linear matter power spectrum. *MNRAS*, 523(2):2247–2262, August 2023.
- [193] Roi Kugel, Joop Schaye, Matthieu Schaller, John C. Helly, Joey Braspenning, Willem Elbers, Carlos S. Frenk, Ian G. McCarthy, Juliana Kwan, Jaime Salcido, Marcel P. van Daalen, Bert Vandenbroucke, Yannick M. Bahé, Josh Borrow, Evgenii Chaikin, Filip Huško, Adrian Jenkins, Cedric G. Lacey, Folkert S. J. Nobels, and Ian Vernon. FLAMINGO: Calibrating large cosmological hydrodynamical simulations with machine learning. *arXiv e-prints*, page arXiv:2306.05492, June 2023.

- [194] James G. Rogers, Clàudia Janó Muñoz, James E. Owen, and Richard A. Booth. Exoplanet atmosphere evolution: emulation with random forests. *arXiv e-prints*, page arXiv:2110.15162, October 2021.
- [195] Damon H. T. Cheung, Kaze W. K. Wong, Otto A. Hannuksela, Tjonnie G. F. Li, and Shirley Ho. Testing the robustness of simulation-based gravitational-wave population inference. *arXiv e-prints*, page arXiv:2112.06707, December 2021.
- [196] Justin Alsing, Hiranya Peiris, Joel Leja, ChangHoon Hahn, Rita Tojeiro, Daniel Mortlock, Boris Leistedt, Benjamin D. Johnson, and Charlie Conroy. SPECULATOR: Emulating Stellar Population Synthesis for Fast and Accurate Galaxy Spectra and Photometry. *ApJS*, 249(1):5, July 2020.
- [197] Yi Ji, Simon Mak, Derek Soeder, J-F Paquet, and Steffen A. Bass. A graphical multi-fidelity Gaussian process model, with application to emulation of expensive computer simulations. *arXiv e-prints*, page arXiv:2108.00306, July 2021.
- [198] Yi Ji, Henry Shaowu Yuchi, Derek Soeder, J. F. Paquet, Steffen A. Bass, V. Roshan Joseph, C. F. Jeff Wu, and Simon Mak. Multi-Stage Multi-Fidelity Gaussian Process Modeling, with Application to Heavy-Ion Collisions. *arXiv e-prints*, page arXiv:2209.13748, September 2022.
- [199] J. Holdship, S. Viti, T. J. Haworth, and J. D. Ilee. Chemulator: Fast, accurate thermochemistry for dynamical models through emulation. *A&A*, 653:A76, September 2021.
- [200] Ian Vernon, Junli Liu, Michael Goldstein, James Rowe, Jen Topping, and Keith Lindsey. Bayesian uncertainty analysis for complex systems biology models: emulation, global parameter searches and evaluation of gene functions. *BMC Systems Biology*, 12(1):1, 2018.
- [201] Kelly R. Moran, Katrin Heitmann, Earl Lawrence, Salman Habib, Derek Bingham, Amol Upadhye, Juliana Kwan, David Higdon, and Richard Payne. The Mira-Titan Universe - IV. High-precision power spectrum emulation. *MNRAS*, 520(3):3443–3458, April 2023.
- [202] Francisco Villaescusa-Navarro, Daniel Anglés-Alcázar, Shy Genel, David N. Spergel, Rachel S. Somerville, Romeel Dave, Annalisa Pillepich, Lars Hernquist, Dylan Nelson, Paul Torrey, Desika Narayanan, Yin Li, Oliver Philcox, Valentina La Torre, Ana Maria Delgado, Shirley Ho, Sultan Hassan, Blakesley Burkhart, Digvijay Wadekar, Nicholas Battaglia, Gabriella Contardo, and Greg L. Bryan. The CAMELS Project: Cosmology and Astrophysics with Machine-learning Simulations. *ApJ*, 915(1):71, July 2021.
- [203] Nicolas Chartier, Benjamin Wandelt, Yashar Akrami, and Francisco Villaescusa-Navarro. CARPool: fast, accurate computation of large-scale structure statistics by pairing costly and cheap cosmological simulations. *MNRAS*, 503(2):1897–1914, May 2021.

- [204] Nicolas Chartier and Benjamin D. Wandelt. Bayesian control variates for optimal covariance estimation with pairs of simulations and surrogates. *MNRAS*, 515(1):1296–1315, September 2022.
- [205] J. Andrés Christen and Colin Fox. Markov chain monte carlo using an approximation. *Journal of Computational and Graphical Statistics*, 14(4):795–810, 2005.
- [206] Mikkel B. Lykkegaard, Grigorios Mingas, Robert Scheichl, Colin Fox, and Tim J. Dodwell. Multilevel Delayed Acceptance MCMC with an Adaptive Error Model in PyMC3. *arXiv e-prints*, page arXiv:2012.05668, December 2020.
- [207] Yin Li, Yueying Ni, Rupert A. C. Croft, Tiziana Di Matteo, Simeon Bird, and Yu Feng. AI-assisted superresolution cosmological simulations. *Proceedings of the National Academy of Science*, 118(19):e2022038118, May 2021.
- [208] Yueying Ni, Yin Li, Patrick Lachance, Rupert A. C. Croft, Tiziana Di Matteo, Simeon Bird, and Yu Feng. AI-assisted superresolution cosmological simulations - II. Halo substructures, velocities, and higher order statistics. *MNRAS*, 507(1):1021–1033, October 2021.
- [209] Alex Cole, Benjamin K. Miller, Samuel J. Witte, Maxwell X. Cai, Meiert W. Grootes, Francesco Nattino, and Christoph Weniger. Fast and credible likelihood-free cosmology with truncated marginal neural ratio estimation. *JCAP*, 2022(9):004, September 2022.
- [210] Supranta S. Boruah, Tim Eifler, Vivian Miranda, and P. M. Sai Krishanth. Accelerating cosmological inference with Gaussian processes and neural networks - an application to LSST Y1 weak lensing and galaxy clustering. *MNRAS*, 518(4):4818–4831, February 2023.
- [211] Holger Wendland. *Scattered Data Approximation*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2004.
- [212] Yueying Ni, Tiziana Di Matteo, Simeon Bird, Rupert Croft, Yu Feng, Nianyi Chen, Michael Tremmel, Colin DeGraf, and Yin Li. The ASTRID simulation: the evolution of supermassive black holes. *MNRAS*, 513(1):670–692, June 2022.
- [213] Volker Springel, Rüdiger Pakmor, Oliver Zier, and Martin Reinecke. Simulating cosmic structure formation with the GADGET-4 code. *MNRAS*, 506(2):2871–2949, September 2021.
- [214] Nick Hand, Yu Feng, Florian Beutler, Yin Li, Chirag Modi, Uroš Seljak, and Zachary Slepian. nboddykit: An Open-source, Massively Parallel Toolkit for Large-scale Structure. *AJ*, 156(4):160, October 2018.
- [215] Peter Z. G. Qian. Sliced latin hypercube designs. *Journal of the American Statistical Association*, 107(497):393–399, 2012.

- [216] John Salvatier, Thomas V. Wiecki, and Christopher Fonnesbeck. Probabilistic programming in python using PyMC3. *PeerJ Computer Science*, 2:e55, apr 2016.
- [217] Alexander H. Nitz, Sumit Kumar, Yi-Fan Wang, Shilpa Kastha, Shichao Wu, Marlin Schäfer, Rahul Dhurkunde, and Collin D. Capano. 4-OGC: Catalog of Gravitational Waves from Compact Binary Mergers. *ApJ*, 946(2):59, April 2023.
- [218] The LIGO Scientific Collaboration, the Virgo Collaboration, the KAGRA Collaboration, R. Abbott, H. Abe, et al. Tests of General Relativity with GWTC-3. *arXiv e-prints*, page arXiv:2112.06861, December 2021.
- [219] The LIGO Scientific Collaboration, the Virgo Collaboration, the KAGRA Collaboration, R. Abbott, H. Abe, et al. Constraints on the cosmic expansion history from GWTC-3. *arXiv e-prints*, page arXiv:2111.03604, November 2021.
- [220] The LIGO Scientific Collaboration, the Virgo Collaboration, the KAGRA Collaboration, R. Abbott, T. D. Abbott, et al. GWTC-3: Compact Binary Coalescences Observed by LIGO and Virgo During the Second Part of the Third Observing Run. *arXiv e-prints*, page arXiv:2111.03606, November 2021.
- [221] F. Acernese, M. Agathos, K. Agatsuma, et al. Advanced Virgo: a second-generation interferometric gravitational wave detector. *Classical and Quantum Gravity*, 32(2):024001, January 2015.
- [222] T. Akutsu, M. Ando, K. Arai, et al. Overview of KAGRA: KAGRA science. *Progress of Theoretical and Experimental Physics*, 2021(5):05A103, May 2021.
- [223] B. P. Abbott, R. Abbott, T. D. Abbott, LIGO Scientific Collaboration, and Virgo Collaboration. Binary Black Hole Population Properties Inferred from the First and Second Observing Runs of Advanced LIGO and Advanced Virgo. *ApJL*, 882(2):L24, September 2019.
- [224] Tejaswi Venumadhav, Barak Zackay, Javier Roulet, Liang Dai, and Matias Zaldarriaga. New search pipeline for compact binary mergers: Results for binary black holes in the first observing run of Advanced LIGO. *PhRvD*, 100(2):023011, July 2019.
- [225] R. Abbott, T. D. Abbott, S. Abraham, LIGO Scientific Collaboration, and Virgo Collaboration. Population Properties of Compact Objects from the Second LIGO-Virgo Gravitational-Wave Transient Catalog. *ApJL*, 913(1):L7, May 2021.
- [226] Digvijay Wadekar, Javier Roulet, Tejaswi Venumadhav, Ajit Kumar Mehta, Barak Zackay, Jonathan Mushkin, Seth Olsen, and Matias Zaldarriaga. New black hole mergers in the LIGO-Virgo O3 data from a gravitational wave search including higher-order harmonics. *arXiv e-prints*, page arXiv:2312.06631, December 2023.
- [227] Michael Zevin, Chris Pankow, Carl L. Rodriguez, Laura Sampson, Eve Chase, Vassiliki Kalogera, and Frederic A. Rasio. Constraining Formation Models of Binary Black Holes with Gravitational-wave Observations. *ApJ*, 846(1):82, September 2017.

- [228] Michael Zevin, Simone S. Bavera, Christopher P. L. Berry, Vicky Kalogera, Tassos Fragos, Pablo Marchant, Carl L. Rodriguez, Fabio Antonini, Daniel E. Holz, and Chris Pankow. One Channel to Rule Them All? Constraining the Origins of Binary Black Holes Using Multiple Formation Pathways. *Astrophys. J.*, 910(2):152, 2021.
- [229] Bruce Edelman, Zoheyr Doctor, Jaxen Godfrey, and Ben Farr. Ain't No Mountain High Enough: Semiparametric Modeling of LIGO-Virgo's Binary Black Hole Mass Distribution. *ApJ*, 924(2):101, January 2022.
- [230] Simon Stevenson, Frank Ohme, and Stephen Fairhurst. Distinguishing Compact Binary Population Synthesis Models Using Gravitational Wave Observations of Coalescing Binary Black Holes. *ApJ*, 810(1):58, September 2015.
- [231] Feryal Özel, Dimitrios Psaltis, Ramesh Narayan, and Jeffrey E. McClintock. The Black Hole Mass Distribution in the Galaxy. *ApJ*, 725(2):1918–1927, December 2010.
- [232] Will M. Farr, Niharika Sravan, Andrew Cantrell, Laura Kreidberg, Charles D. Bailyn, Ilya Mandel, and Vicky Kalogera. The Mass Distribution of Stellar-mass Black Holes. *ApJ*, 741(2):103, November 2011.
- [233] Maya Fishbach, Reed Essick, and Daniel E. Holz. Does Matter Matter? Using the Mass Distribution to Distinguish Neutron Stars and Black Holes. *ApJL*, 899(1):L8, August 2020.
- [234] Amanda Farah, Maya Fishbach, Reed Essick, Daniel E. Holz, and Shanika Galaudage. Bridging the Gap: Categorizing Gravitational-wave Events at the Transition between Neutron Stars and Black Holes. *ApJ*, 931(2):108, June 2022.
- [235] Ang Li, Zhiqiang Miao, Sophia Han, and Bing Zhang. Constraints on the Maximum Mass of Neutron Stars with a Quark Core from GW170817 and NICER PSR J0030+0451 Data. *ApJ*, 913(1):27, May 2021.
- [236] Rachel A. Patton, Tuguldur Sukhbold, and J. J. Eldridge. Comparing compact object distributions from mass- and presupernova core structure-based prescriptions. *MNRAS*, 511(1):903–913, March 2022.
- [237] Jared C. Siegel, Ilia Kiato, Vicky Kalogera, Christopher P. L. Berry, Thomas J. Maccarone, Katelyn Breivik, Jeff J. Andrews, Simone S. Bavera, Aaron Dotter, Tassos Fragos, Konstantinos Kovalakas, Devina Misra, Kyle A. Rocha, Philipp M. Srivastava, Meng Sun, Zepei Xing, and Emmanouil Zapartas. Investigating the Lower Mass Gap with Low-mass X-Ray Binary Population Synthesis. *ApJ*, 954(2):212, September 2023.
- [238] Chris L. Fryer, Krzysztof Belczynski, Grzegorz Wiktorowicz, Michal Dominik, Vicky Kalogera, and Daniel E. Holz. Compact Remnant Mass Function: Dependence on the Explosion Mechanism and Metallicity. *ApJ*, 749(1):91, April 2012.
- [239] Michael Zevin, Mario Spera, Christopher P. L. Berry, and Vicky Kalogera. Exploring the Lower Mass Gap and Unequal Mass Regime in Compact Binary Evolution. *ApJL*, 899(1):L1, August 2020.

- [240] Ilya Mandel and Bernhard Müller. Simple recipes for compact remnant masses and natal kicks. *MNRAS*, 499(3):3214–3221, December 2020.
- [241] L. A. C. van Son, S. E. de Mink, M. Renzo, S. Justham, E. Zapartas, K. Breivik, T. Callister, W. M. Farr, and C. Conroy. No Peaks without Valleys: The Stable Mass Transfer Channel for Gravitational-wave Sources in Light of the Neutron Star-Black Hole Mass Gap. *ApJ*, 940(2):184, December 2022.
- [242] Fabian R. N. Schneider, Philipp Podsiadlowski, and Eva Laplace. Bimodal Black Hole Mass Distribution and Chirp Masses of Binary Black Hole Mergers. *ApJL*, 950(2):L9, June 2023.
- [243] William A. Fowler and F. Hoyle. Neutrino Processes and Pair Formation in Massive Stars and Supernovae. *ApJS*, 9:201, December 1964.
- [244] Z. Barkat, G. Rakavy, and N. Sack. Dynamics of Supernova Explosion Resulting from Pair Formation. *PhRvL*, 18(10):379–381, March 1967.
- [245] A. Heger and S. E. Woosley. The Nucleosynthetic Signature of Population III. *ApJ*, 567(1):532–543, March 2002.
- [246] A. Heger, C. L. Fryer, S. E. Woosley, N. Langer, and D. H. Hartmann. How Massive Single Stars End Their Life. *ApJ*, 591(1):288–300, July 2003.
- [247] Stan. E. Woosley and Alexander Heger. The Deaths of Very Massive Stars. In Jorick S. Vink, editor, *Very Massive Stars in the Local Universe*, volume 412 of *Astrophysics and Space Science Library*, page 199, January 2015.
- [248] K. Belczynski, A. Heger, W. Gladysz, A. J. Ruiter, S. Woosley, G. Wiktorowicz, H. Y. Chen, T. Bulik, R. O’Shaughnessy, D. E. Holz, C. L. Fryer, and E. Berti. The effect of pair-instability mass loss on black-hole mergers. *A&A*, 594:A97, October 2016.
- [249] Colm Talbot and Eric Thrane. Measuring the Binary Black Hole Mass Spectrum with an Astrophysically Motivated Parameterization. *ApJ*, 856(2):173, April 2018.
- [250] Pablo Marchant, Mathieu Renzo, Robert Farmer, Kaliroe M. W. Pappas, Ronald E. Taam, Selma E. de Mink, and Vassiliki Kalogera. Pulsational Pair-instability Supernovae in Very Close Binaries. *ApJ*, 882(1):36, September 2019.
- [251] S. E. Woosley. The Evolution of Massive Helium Stars, Including Mass Loss. *ApJ*, 878(1):49, June 2019.
- [252] M. Renzo, R. Farmer, S. Justham, Y. Götberg, S. E. de Mink, E. Zapartas, P. Marchant, and N. Smith. Predictions for the hydrogen-free ejecta of pulsational pair-instability supernovae. *A&A*, 640:A56, August 2020.
- [253] R. Farmer, M. Renzo, S. E. de Mink, P. Marchant, and S. Justham. Mind the Gap: The Location of the Lower Edge of the Pair-instability Supernova Black Hole Mass Gap. *ApJ*, 887(1):53, December 2019.

- [254] D. D. Hendriks, L. A. C. van Son, M. Renzo, R. G. Izzard, and R. Farmer. Pulsational pair-instability supernovae in gravitational-wave and electromagnetic transients. *MNRAS*, 526(3):4130–4147, December 2023.
- [255] M. M. Briel, H. F. Stevance, and J. J. Eldridge. Understanding the high-mass binary black hole population from stable mass transfer and super-Eddington accretion in BPASS. *MNRAS*, 520(4):5724–5745, April 2023.
- [256] G. F. Chapline. Cosmological effects of primordial black holes. *Nature*, 253(5489):251–252, January 1975.
- [257] B. J. Carr. The primordial black hole mass spectrum. *ApJ*, 201:1–19, October 1975.
- [258] Simeon Bird, Ilias Cholis, Julian B. Muñoz, Yacine Ali-Haïmoud, Marc Kamionkowski, Ely D. Kovetz, Alvise Raccanelli, and Adam G. Riess. Did LIGO Detect Dark Matter? *PhRvL*, 116(20):201301, May 2016.
- [259] Misao Sasaki, Teruaki Suyama, Takahiro Tanaka, and Shuichiro Yokoyama. Primordial black holes—perspectives in gravitational wave astronomy. *Classical and Quantum Gravity*, 35(6):063001, March 2018.
- [260] Fabio Antonini, Silvia Toonen, and Adrian S. Hamers. Binary Black Hole Mergers from Field Triples: Properties, Rates, and the Impact of Stellar Evolution. *ApJ*, 841(2):77, June 2017.
- [261] Kedron Silsbee and Scott Tremaine. Lidov-Kozai Cycles with Gravitational Radiation: Merging Black Holes in Isolated Triple Systems. *ApJ*, 836(1):39, February 2017.
- [262] Kohei Inayoshi, Kazumi Kashiyama, Eli Visbal, and Zoltán Haiman. Gravitational wave background from Population III binary black holes consistent with cosmic reionization. *MNRAS*, 461(3):2722–2727, September 2016.
- [263] Vaibhav Tiwari and Stephen Fairhurst. The Emergence of Structure in the Binary Black Hole Mass Distribution. *ApJL*, 913(2):L19, June 2021.
- [264] Carl L. Rodriguez, Sourav Chatterjee, and Frederic A. Rasio. Binary black hole mergers from globular clusters: Masses, merger rates, and the impact of stellar evolution. *PhRvD*, 93(8):084029, April 2016.
- [265] Pau Amaro-Seoane and Xian Chen. Relativistic mergers of black hole binaries have large, similar masses, low spins and are circular. *MNRAS*, 458(3):3075–3082, May 2016.
- [266] Amanda M. Farah, Maya Fishbach, and Daniel E. Holz. Two of a Kind: Comparing big and small black holes in binaries with gravitational waves. *arXiv e-prints*, page arXiv:2308.05102, August 2023.
- [267] Erez Michaely and Hagai B. Perets. Gravitational-wave Sources from Mergers of Binary Black Holes Catalyzed by Flyby Interactions in the Field. *ApJL*, 887(2):L36, December 2019.

- [268] Michal Dominik, Emanuele Berti, Richard O’Shaughnessy, Ilya Mandel, Krzysztof Belczynski, Christopher Fryer, Daniel E. Holz, Tomasz Bulik, and Francesco Panarale. Double Compact Objects III: Gravitational-wave Detection Rates. *ApJ*, 806(2):263, June 2015.
- [269] Nicola Giacobbo, Michela Mapelli, and Mario Spera. Merging black hole binaries: the effects of progenitor’s metallicity, mass-loss rate and Eddington factor. *MNRAS*, 474(3):2959–2974, March 2018.
- [270] Mario Spera, Michela Mapelli, Nicola Giacobbo, Alessandro A. Trani, Alessandro Bressan, and Guglielmo Costa. Merging black hole binaries with the SEVN code. *MNRAS*, 485(1):889–907, May 2019.
- [271] Michael Y. Grudić, Stella S. R. Offner, Dávid Guszejnov, Claude-André Faucher-Giguère, and Philip F. Hopkins. Does God play dice with star clusters? *arXiv e-prints*, page arXiv:2307.00052, June 2023.
- [272] Michal Dominik, Krzysztof Belczynski, Christopher Fryer, Daniel E. Holz, Emanuele Berti, Tomasz Bulik, Ilya Mandel, and Richard O’Shaughnessy. Double Compact Objects. I. The Significance of the Common Envelope on Merger Rates. *ApJ*, 759(1):52, November 2012.
- [273] Simon Stevenson, Alejandro Vigna-Gómez, Ilya Mandel, Jim W. Barrett, Coenraad J. Neijssel, David Perkins, and Selma E. de Mink. Formation of the first three gravitational-wave observations through isolated binary evolution. *Nature Communications*, 8:14906, April 2017.
- [274] E. Laplace, S. Justham, M. Renzo, Y. Götzberg, R. Farmer, D. Vartanyan, and S. E. de Mink. Different to the core: The pre-supernova structures of massive single and binary-stripped stars. *A&A*, 656:A58, December 2021.
- [275] A. Olejak and K. Belczynski. The Implications of High Black Hole Spins for the Origin of Binary Black Hole Mergers. *ApJL*, 921(1):L2, November 2021.
- [276] Floor S. Broekgaarden, Simon Stevenson, and Eric Thrane. Signatures of Mass Ratio Reversal in Gravitational Waves from Merging Binary Black Holes. *ApJ*, 938(1):45, October 2022.
- [277] Michael Zevin and Simone S. Bavera. Suspicious Siblings: The Distribution of Mass and Spin across Component Black Holes in Isolated Binary Evolution. *ApJ*, 933(1):86, July 2022.
- [278] Ely D. Kovetz, Ilias Cholis, Patrick C. Breysse, and Marc Kamionkowski. Black hole mass function from gravitational wave measurements. *PhRvD*, 95(10):103010, May 2017.
- [279] Maya Fishbach and Daniel E. Holz. Where Are LIGO’s Big Black Holes? *ApJL*, 851(2):L25, December 2017.

- [280] Thomas A. Callister and Will M. Farr. A Parameter-Free Tour of the Binary Black Hole Population. *arXiv e-prints*, page arXiv:2302.07289, February 2023.
- [281] Jaxen Godfrey, Bruce Edelman, and Ben Farr. Cosmic Cousins: Identification of a Subpopulation of Binary Black Holes Consistent with Isolated Binary Evolution. *arXiv e-prints*, page arXiv:2304.01288, April 2023.
- [282] Maya Fishbach and Daniel E. Holz. Picky Partners: The Pairing of Component Masses in Binary Black Hole Mergers. *ApJL*, 891(1):L27, March 2020.
- [283] Kaze Wong and Miles Cranmer. Automated discovery of interpretable gravitational-wave population models. In *Machine Learning for Astrophysics*, page 25, July 2022.
- [284] Amanda M. Farah, Bruce Edelman, Michael Zevin, Maya Fishbach, Jose María Ezquiaga, Ben Farr, and Daniel E. Holz. Things That Might Go Bump in the Night: Assessing Structure in the Binary Black Hole Mass Spectrum. *ApJ*, 955(2):107, October 2023.
- [285] Vaibhav Tiwari. What’s in a binary black hole’s mass parameter? *MNRAS*, 527(1):298–306, January 2024.
- [286] The LIGO Scientific Collaboration, the Virgo Collaboration, R. Abbott, T. D. Abbott, et al. GWTC-2.1: Deep Extended Catalog of Compact Binary Coalescences Observed by LIGO and Virgo During the First Half of the Third Observing Run. *arXiv e-prints*, page arXiv:2108.01045, August 2021.
- [287] Geraint Pratten, Cecilio García-Quirós, Marta Colleoni, Antoni Ramos-Buades, Héctor Estellés, Maite Mateu-Lucena, Rafel Jaume, Maria Haney, David Keitel, Jonathan E. Thompson, and Sascha Husa. Computationally efficient models for the dominant and subdominant harmonic modes of precessing binary black holes. *PhRvD*, 103(10):104056, May 2021.
- [288] Serguei Ossokine, Alessandra Buonanno, Sylvain Marsat, Roberto Cotesta, Stanislav Babak, Tim Dietrich, Roland Haas, Ian Hinder, Harald P. Pfeiffer, Michael Pürrer, Charles J. Woodford, Michael Boyle, Lawrence E. Kidder, Mark A. Scheel, and Béla Szilágyi. Multipolar effective-one-body waveforms for precessing binary black holes: Construction and validation. *PhRvD*, 102(4):044055, August 2020.
- [289] Thomas J. Loredo. Accounting for source uncertainties in analyses of astronomical survey data. *AIP Conference Proceedings*, 735(1):195–206, 2004.
- [290] Salvatore Vitale, Davide Gerosa, Will M. Farr, and Stephen R. Taylor. *Inferring the properties of a population of compact binaries in presence of selection effects*. 7 2020.
- [291] Stephen R. Taylor and Davide Gerosa. Mining Gravitational-wave Catalogs To Understand Binary Stellar Evolution: A New Hierarchical Bayesian Framework. *Phys. Rev. D*, 98(8):083017, 2018.

- [292] Ilya Mandel, Will M. Farr, and Jonathan R. Gair. Extracting distribution parameters from multiple uncertain observations with selection biases. *Mon. Not. Roy. Astron. Soc.*, 486(1):1086–1093, 2019.
- [293] Maya Fishbach, Daniel E. Holz, and Will M. Farr. Does the Black Hole Merger Rate Evolve with Redshift? *Astrophys. J. Lett.*, 863(2):L41, 2018.
- [294] Scott Ellis Perkins, Peter McGill, William Dawson, Natasha S. Abrams, Casey Y. Lam, Ming-Feng Ho, Jessica R. Lu, Simeon Bird, Kerianne Pruett, Nathan Golovich, and George Chapline. Disentangling the Black Hole Mass Spectrum with Photometric Microlensing Surveys. *arXiv e-prints*, page arXiv:2310.03943, October 2023.
- [295] Scott E. Perkins, Nicolás Yunes, and Emanuele Berti. Probing fundamental physics with gravitational waves: The next generation. *PhRvD*, 103(4):044024, February 2021.
- [296] Lee Samuel Finn. Binary inspiral, gravitational radiation, and cosmology. *Phys. Rev. D*, 53:2878–2894, 1996.
- [297] Lee Samuel Finn and David F. Chernoff. Observing binary inspiral in gravitational radiation: One interferometer. *Phys. Rev. D*, 47:2198–2219, 1993.
- [298] Michal Dominik, Emanuele Berti, Richard O’Shaughnessy, Ilya Mandel, Krzysztof Belczynski, Christopher Fryer, Daniel E. Holz, Tomasz Bulik, and Francesco Pannarale. Double Compact Objects III: Gravitational Wave Detection Rates. *Astrophys. J.*, 806(2):263, 2015.
- [299] Vaibhav Tiwari, Stephen Fairhurst, and Mark Hannam. Constraining Black Hole Spins with Gravitational-wave Observations. *ApJ*, 868(2):140, December 2018.
- [300] M. Campanelli, C. O. Lousto, and Y. Zlochower. Spinning-black-hole binaries: The orbital hang-up. *Phys. Rev. D*, 74:041501, Aug 2006.
- [301] LIGO Scientific Collaboration, Virgo Collaboration, and KAGRA Collaboration. LVK Algorithm Library - LALSuite. Free software (GPL), 2018.
- [302] C. M. Biwer, Collin D. Capano, Soumi De, Miriam Cabero, Duncan A. Brown, Alexander H. Nitz, and V. Raymond. PyCBC Inference: A Python-based Parameter Estimation Toolkit for Compact Binary Coalescence Signal. *PASP*, 131(996):024503, February 2019.
- [303] Lee Samuel Finn and David F. Chernoff. Observing binary inspiral in gravitational radiation: One interferometer. *PhRvD*, 47(6):2198–2219, March 1993.
- [304] Lee Samuel Finn. Binary inspiral, gravitational radiation, and cosmology. *PhRvD*, 53(6):2878–2894, March 1996.
- [305] Sebastian Khan, Sascha Husa, Mark Hannam, Frank Ohme, Michael Pürrer, Xisco Jiménez Forteza, and Alejandro Bohé. Frequency-domain gravitational waves from nonprecessing black-hole binaries. II. A phenomenological model for the advanced detector era. *PhRvD*, 93(4):044007, February 2016.

- [306] Sascha Husa, Sebastian Khan, Mark Hannam, Michael Pürrer, Frank Ohme, Xisco Jiménez Forteza, and Alejandro Bohé. Frequency-domain gravitational waves from nonprecessing black-hole binaries. I. New numerical waveforms and anatomy of the signal. *PhRvD*, 93(4):044006, February 2016.
- [307] Michael Pürrer, Sebastian Khan, Frank Ohme, Ofek Birnholtz, and Lionel London. IMRPhenomD: Phenomenological waveform model. Astrophysics Source Code Library, record ascl:2307.019, July 2023.
- [308] Davide Gerosa, Sizheng Ma, Kaze W. K. Wong, Emanuele Berti, Richard O’Shaughnessy, Yanbei Chen, and Krzysztof Belczynski. Multiband gravitational-wave event rates and stellar physics. *Phys. Rev. D*, 99(10):103004, 2019.
- [309] B. P. Abbott et al. Supplement: The Rate of Binary Black Hole Mergers Inferred from Advanced LIGO Observations Surrounding GW150914. *Astrophys. J. Suppl.*, 227(2):14, 2016.
- [310] B. P. Abbott et al. The Rate of Binary Black Hole Mergers Inferred from Advanced LIGO Observations Surrounding GW150914. *Astrophys. J. Lett.*, 833(1):L1, 2016.
- [311] David W. Hogg, Adam D. Myers, and Jo Bovy. Inferring the Eccentricity Distribution. *ApJ*, 725(2):2166–2175, December 2010.
- [312] James Binney and Scott Tremaine. *Galactic Dynamics: Second Edition*. 2008.
- [313] Krzysztof Belczynski, Alessandra Buonanno, Matteo Cantiello, Chris L. Fryer, Daniel E. Holz, Ilya Mandel, M. Coleman Miller, and Marek Walczak. The Formation and Gravitational-wave Detection of Massive Stellar Black Hole Binaries. *ApJ*, 789(2):120, July 2014.
- [314] Clovis Hopman and Tal Alexander. The Effect of Mass Segregation on Gravitational Wave Sources near Massive Black Holes. *ApJL*, 645(2):L133–L136, July 2006.
- [315] Thomas Lacroix and Joseph Silk. Intermediate-mass Black Holes and Dark Matter at the Galactic Center. *ApJL*, 853(1):L16, January 2018.
- [316] Tomoya Kinugawa, Kohei Inayoshi, Kenta Hotokezaka, Daisuke Nakauchi, and Takashi Nakamura. Possible indirect confirmation of the existence of Pop III massive stars by gravitational wave. *MNRAS*, 442(4):2963–2992, August 2014.
- [317] C. Alcock, R. A. Allsman, D. R. Alves, T. S. Axelrod, A. C. Becker, D. P. Bennett, K. H. Cook, N. Dalal, A. J. Drake, K. C. Freeman, M. Geha, K. Griest, M. J. Lehner, S. L. Marshall, D. Minniti, C. A. Nelson, B. A. Peterson, P. Popowski, M. R. Pratt, P. J. Quinn, C. W. Stubbs, W. Sutherland, A. B. Tomaney, T. Vandehei, and D. L. Welch. MACHO Project Limits on Black Hole Dark Matter in the 1-30 M_{solar} Range. *ApJL*, 550(2):L169–L172, April 2001.

- [318] P. Tisserand, L. Le Guillou, C. Afonso, J. N. Albert, J. Andersen, R. Ansari, É. Aubourg, P. Bareyre, J. P. Beaulieu, X. Charlot, C. Coutures, R. Ferlet, P. Fouqué, J. F. Glicenstein, B. Goldman, A. Gould, D. Graff, M. Gros, J. Haissinski, C. Hamadache, J. de Kat, T. Lasserre, É. Lesquoy, C. Loup, C. Magneville, J. B. Marquette, É. Maurice, A. Maury, A. Milsztajn, M. Moniez, N. Palanque-Delabrouille, O. Perdereau, Y. R. Rahal, J. Rich, M. Spiro, A. Vidal-Madjar, L. Vigroux, S. Zylberajch, and EROS-2 Collaboration. Limits on the Macho content of the Galactic Halo from the EROS-2 Survey of the Magellanic Clouds. *A&A*, 469(2):387–404, July 2007.
- [319] L. Wyrzykowski, J. Skowron, S. Kozłowski, A. Udalski, M. K. Szymański, M. Kubiak, G. Pietrzyński, I. Soszyński, O. Szewczyk, K. Ulaczyk, R. Poleski, and P. Tisserand. The OGLE view of microlensing towards the Magellanic Clouds - IV. OGLE-III SMC data and final conclusions on MACHOs. *MNRAS*, 416(4):2949–2961, October 2011.
- [320] T. Blaineau, M. Moniez, C. Afonso, J. N. Albert, R. Ansari, E. Aubourg, C. Coutures, J. F. Glicenstein, B. Goldman, C. Hamadache, T. Lasserre, L. Le Guillou, E. Lesquoy, C. Magneville, J. B. Marquette, N. Palanque-Delabrouille, O. Perdereau, J. Rich, M. Spiro, and P. Tisserand. New limits from microlensing on Galactic black holes in the mass range $10 M_{\odot} \leq M \leq 1000 M_{\odot}$. *A&A*, 664:A106, August 2022.
- [321] Simeon Bird, Andrea Albert, Will Dawson, Yacine Ali-Haïmoud, Adam Coogan, Alex Drlica-Wagner, Qi Feng, Derek Inman, Keisuke Inomata, Ely Kovetz, Alexander Kusenko, Benjamin V. Lehmann, Julian B. Muñoz, Rajeev Singh, Volodymyr Takhistov, and Yu-Dai Tsai. Snowmass2021 Cosmic Frontier White Paper: Primordial black hole dark matter. *Physics of the Dark Universe*, 41:101231, August 2023.
- [322] Davide Gerosa and Maya Fishbach. Hierarchical mergers of stellar-mass black holes and their gravitational-wave signatures. *Nature Astron.*, 5(8):749–760, 2021.
- [323] A. Olejak, M. Fishbach, K. Belczynski, D. E. Holz, J. P. Lasota, M. C. Miller, and T. Bulik. The Origin of Inequality: Isolated Formation of a $30+10 M_{\odot}$ Binary Black Hole Merger. *ApJL*, 901(2):L39, October 2020.
- [324] Abril-Pla Oriol, Andreani Virgile, Carroll Colin, Dong Larry, Fonnesbeck Christopher J., Kochurov Maxim, Kumar Ravin, Lao Jupeng, Luhmann Christian C., Martin Osvaldo A., Osthege Michael, Vieira Ricardo, Wiecki Thomas, and Zinkov Robert. Pymc: A modern and comprehensive probabilistic programming framework in python. *PeerJ Computer Science*, 9:e1516, 2023.
- [325] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, Ilhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and

- SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.
- [326] Gregory Ashton, Moritz Hübner, Paul D. Lasky, Colm Talbot, Kendall Ackley, Sylvia Biscoveanu, Qi Chu, Atul Divakarla, Paul J. Easter, Boris Goncharov, Francisco Hernandez Vivanco, Jan Harms, Marcus E. Lower, Grant D. Meadors, Denyz Melchor, Ethan Payne, Matthew D. Pitkin, Jade Powell, Nikhil Sarin, Rory J. E. Smith, and Eric Thrane. BILBY: A User-friendly Bayesian Inference Library for Gravitational-wave Astronomy. *ApJS*, 241(2):27, April 2019.
- [327] I. M. Romero-Shaw, C. Talbot, S. Biscoveanu, V. D’Emilio, G. Ashton, C. P. L. Berry, S. Coughlin, S. Galaudage, C. Hoy, M. Hübner, K. S. Phukon, M. Pitkin, M. Rizzo, N. Sarin, R. Smith, S. Stevenson, A. Vajpeyi, M. Arène, K. Athar, S. Banagiri, N. Bose, M. Carney, K. Chatziioannou, J. A. Clark, M. Colleoni, R. Cotesta, B. Edelman, H. Estellés, C. García-Quirós, Abhirup Ghosh, R. Green, C. J. Haster, S. Husa, D. Keitel, A. X. Kim, F. Hernandez-Vivanco, I. Magaña Hernandez, C. Karathanasis, P. D. Lasky, N. De Lillo, M. E. Lower, D. Macleod, M. Mateu-Lucena, A. Miller, M. Millhouse, S. Morisaki, S. H. Oh, S. Ossokine, E. Payne, J. Powell, G. Pratten, M. Pürerer, A. Ramos-Buades, V. Raymond, E. Thrane, J. Veitch, D. Williams, M. J. Williams, and L. Xiao. Bayesian inference for compact binary coalescences with BILBY: validation and application to the first LIGO-Virgo gravitational-wave transient catalogue. *MNRAS*, 499(3):3295–3319, December 2020.
- [328] Ravin Kumar, Colin Carroll, Ari Hartikainen, and Osvaldo Martin. Arviz a unified library for exploratory analysis of bayesian models in python. *Journal of Open Source Software*, 4(33):1143, 2019.
- [329] Daniel Foreman-Mackey. corner.py: Scatterplot matrices in python. *The Journal of Open Source Software*, 1(2):24, jun 2016.
- [330] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.
- [331] E. T. Jaynes. *Probability Theory: The Logic of Science*. CUP, 2003.
- [332] Jens-Kristian Krogager, Palle Møller, Lise B. Christensen, Pasquier Noterdaeme, Johan P. U. Fynbo, and Wolfram Freudling. High-redshift damped Ly α absorbing galaxy model reproducing the N H I - Z distribution. *MNRAS*, 495(3):3014–3021, May 2020.
- [333] Tom Theuns. Connecting cosmological accretion to strong Ly α absorbers. *MNRAS*, November 2020.
- [334] Michele Fumagalli, John M. O’Meara, J. Xavier Prochaska, and Gabor Worseck. Dissecting the Properties of Optically Thick Hydrogen at the Peak of Cosmic Star Formation History. *ApJ*, 775(1):78, September 2013.

- [335] Robert R. Gibson, Linhua Jiang, W. N. Brandt, Patrick B. Hall, Yue Shen, Jianfeng Wu, Scott F. Anderson, Donald P. Schneider, Daniel Vanden Berk, S. C. Gallagher, Xiaohui Fan, and Donald G. York. A Catalog of Broad Absorption Line Quasars in Sloan Digital Sky Survey Data Release 5. *ApJ*, 692(1):758–777, February 2009.
- [336] I. Pâris, P. Petitjean, É. Aubourg, S. Bailey, N. P. Ross, A. D. Myers, M. A. Strauss, S. F. Anderson, E. Arnau, J. Bautista, D. Bizyaev, A. S. Bolton, J. Bovy, W. N. Brandt, H. Brewington, J. R. Browstein, N. Busca, D. Capellupo, W. Carithers, R. A. C. Croft, K. Dawson, T. Delubac, G. Ebelke, D. J. Eisenstein, P. Engelke, X. Fan, N. Filiz Ak, H. Finley, A. Font-Ribera, J. Ge, R. R. Gibson, P. B. Hall, F. Hamann, J. F. Hennawi, S. Ho, D. W. Hogg, Ž. Ivezić, L. Jiang, A. E. Kimball, D. Kirkby, J. A. Kirkpatrick, K. G. Lee, J. M. Le Goff, B. Lundgren, C. L. MacLeod, E. Malanushenko, V. Malanushenko, C. Maraston, I. D. McGreer, R. G. McMahon, J. Miralda-Escudé, D. Muna, P. Noterdaeme, D. Oravetz, N. Palanque-Delabrouille, K. Pan, I. Perez-Fournon, M. M. Pieri, G. T. Richards, E. Rollinde, E. S. Sheldon, D. J. Schlegel, D. P. Schneider, A. Slosar, A. Shelden, Y. Shen, A. Simmons, S. Snedden, N. Suzuki, J. Tinker, M. Viel, B. A. Weaver, D. H. Weinberg, M. White, W. M. Wood-Vasey, and C. Yèche. The Sloan Digital Sky Survey quasar catalog: ninth data release. *A&A*, 548:A66, December 2012.
- [337] Isabelle Pâris, Patrick Petitjean, Éric Aubourg, Adam D. Myers, Alina Streblyanska, Brad W. Lyke, Scott F. Anderson, Éric Armengaud, Julian Bautista, Michael R. Blanton, Michael Blomqvist, Jonathan Brinkmann, Joel R. Brownstein, William Nielsen Brandt, Étienne Burtin, Kyle Dawson, Sylvain de la Torre, Antonis Georgakakis, Héctor Gil-Marín, Paul J. Green, Patrick B. Hall, Jean-Paul Kneib, Stephanie M. LaMassa, Jean-Marc Le Goff, Chelsea MacLeod, Vivek Mariappan, Ian D. McGreer, Andrea Merloni, Pasquier Noterdaeme, Nathalie Palanque-Delabrouille, Will J. Percival, Ashley J. Ross, Graziano Rossi, Donald P. Schneider, Hee-Jong Seo, Rita Tojeiro, Benjamin A. Weaver, Anne-Marie Weijmans, Christophe Yèche, Pauline Zarrouk, and Gong-Bo Zhao. The Sloan Digital Sky Survey Quasar Catalog: Fourteenth data release. *A&A*, 613:A51, May 2018.
- [338] I. Pâris, P. Petitjean, É. Aubourg, S. Bailey, N. P. Ross, A. D. Myers, M. A. Strauss, S. F. Anderson, E. Arnau, J. Bautista, D. Bizyaev, A. S. Bolton, J. Bovy, W. N. Brandt, H. Brewington, J. R. Browstein, N. Busca, D. Capellupo, W. Carithers, R. A. C. Croft, K. Dawson, T. Delubac, G. Ebelke, D. J. Eisenstein, P. Engelke, X. Fan, N. Filiz Ak, H. Finley, A. Font-Ribera, J. Ge, R. R. Gibson, P. B. Hall, F. Hamann, J. F. Hennawi, S. Ho, D. W. Hogg, Ž. Ivezić, L. Jiang, A. E. Kimball, D. Kirkby, J. A. Kirkpatrick, K. G. Lee, J. M. Le Goff, B. Lundgren, C. L. MacLeod, E. Malanushenko, V. Malanushenko, C. Maraston, I. D. McGreer, R. G. McMahon, J. Miralda-Escudé, D. Muna, P. Noterdaeme, D. Oravetz, N. Palanque-Delabrouille, K. Pan, I. Perez-Fournon, M. M. Pieri, G. T. Richards, E. Rollinde, E. S. Sheldon, D. J. Schlegel, D. P. Schneider, A. Slosar, A. Shelden, Y. Shen, A. Simmons, S. Snedden, N. Suzuki, J. Tinker, M. Viel, B. A. Weaver, D. H. Weinberg, M. White, W. M. Wood-Vasey, and C. Yèche. The Sloan Digital Sky Survey quasar catalog: ninth data release. *A&A*, 548:A66, December 2012.

- [339] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.
- [340] Joshua V. Dillon, Ian Langmore, Dustin Tran, Eugene Brevdo, Srinivas Vasudevan, Dave Moore, Brian Patton, Alex Alemi, Matt Hoffman, and Rif A. Saurous. Tensor-Flow Distributions. *arXiv e-prints*, page arXiv:1711.10604, November 2017.
- [341] Gustavo Malkomes and Roman Garnett. Automating bayesian optimization with bayesian optimization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS’18*, page 5988–5997, Red Hook, NY, USA, 2018. Curran Associates Inc.
- [342] Zong-Min Wu and Robert Schaback. Local error estimates for radial basis function interpolation of scattered data. *IMA Journal of Numerical Analysis*, 13(1):13–27, 01 1993.
- [343] Yosuke Kobayashi, Takahiro Nishimichi, Masahiro Takada, and Hironao Miyatake. Full-shape cosmology analysis of the SDSS-III BOSS galaxy power spectrum using an emulator-based halo model: A 5% determination of σ_8 . *PhRvD*, 105(8):083517, April 2022.
- [344] Salman Habib, Adrian Pope, Hal Finkel, Nicholas Frontiere, Katrin Heitmann, David Daniel, Patricia Fasel, Vitali Morozov, George Zagaris, Tom Peterka, Venkatram Vishwanath, Zarija Lukić, Saba Sehrish, and Wei-keng Liao. HACC: Simulating sky surveys on state-of-the-art supercomputing architectures. *NewA*, 42:49–65, January 2016.
- [345] Andreas Damianou and Neil D. Lawrence. Deep Gaussian processes. In Carlos M. Carvalho and Pradeep Ravikumar, editors, *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pages 207–215, Scottsdale, Arizona, USA, 29 Apr–01 May 2013. PMLR.
- [346] Rouzbeh Allahverdi, Robert Brandenberger, Francis-Yan Cyr-Racine, and Anupam Mazumdar. Reheating in Inflationary Cosmology: Theory and Applications. *Annual Review of Nuclear and Particle Science*, 60:27–51, November 2010.
- [347] Erik Holmberg. On the Clustering Tendencies among the Nebulae. II. a Study of Encounters Between Laboratory Models of Stellar Systems by a New Integration Procedure. *ApJ*, 94:385, November 1941.
- [348] Aurel Schneider and Romain Teyssier. A new method to quantify the effects of baryons on the matter power spectrum. *JCAP*, 2015(12):049, December 2015.

- [349] Douglas Potter, Joachim Stadel, and Romain Teyssier. PKDGRAV3: beyond trillion particle cosmological simulations for the next era of galaxy surveys. *Computational Astrophysics and Cosmology*, 4(1):2, May 2017.
- [350] Loic Le Gratiet and Josselin Garnier. Recursive co-kriging model for design of computer experiments with multiple levels of fidelity. *International Journal for Uncertainty Quantification*, 4(5):365–386, 2014.
- [351] Hugh Salimbeni and Marc Deisenroth. Doubly Stochastic Variational Inference for Deep Gaussian Processes. *arXiv e-prints*, page arXiv:1705.08933, May 2017.
- [352] Planck Collaboration, P. A. R. Ade, N. Aghanim, M. Arnaud, M. Ashdown, J. Aumont, C. Baccigalupi, A. J. Banday, R. B. Barreiro, J. G. Bartlett, N. Bartolo, E. Battaner, R. Battye, K. Benabed, A. Benoît, A. Benoit-Lévy, J. P. Bernard, M. Bersanelli, P. Bielewicz, J. J. Bock, A. Bonaldi, L. Bonavera, J. R. Bond, J. Borrill, F. R. Bouchet, F. Boulanger, M. Bucher, C. Burigana, R. C. Butler, E. Calabrese, J. F. Cardoso, A. Catalano, A. Challinor, A. Chamballu, R. R. Chary, H. C. Chiang, J. Chluba, P. R. Christensen, S. Church, D. L. Clements, S. Colombi, L. P. L. Colombo, C. Combet, A. Coulais, B. P. Crill, A. Curto, F. Cuttaia, L. Danese, R. D. Davies, R. J. Davis, P. de Bernardis, A. de Rosa, G. de Zotti, J. Delabrouille, F. X. Désert, E. Di Valentino, C. Dickinson, J. M. Diego, K. Dolag, H. Dole, S. Donzelli, O. Doré, M. Douspis, A. Ducout, J. Dunkley, X. Dupac, G. Efstathiou, F. Elsner, T. A. Enßlin, H. K. Eriksen, M. Farhang, J. Fergusson, F. Finelli, O. Forni, M. Frailis, A. A. Fraisse, E. Franceschi, A. Frejsel, S. Galeotta, S. Galli, K. Ganga, C. Gauthier, M. Gerbino, T. Ghosh, M. Giard, Y. Giraud-Héraud, E. Giusarma, E. Gjerløw, J. González-Nuevo, K. M. Górski, S. Gratton, A. Gregorio, A. Gruppuso, J. E. Gudmundsson, J. Hamann, F. K. Hansen, D. Hanson, D. L. Harrison, G. Helou, S. Henrot-Versillé, C. Hernández-Monteagudo, D. Herranz, S. R. Hildebrandt, E. Hivon, M. Hobson, W. A. Holmes, A. Hornstrup, W. Hovest, Z. Huang, K. M. Huffenberger, G. Hurier, A. H. Jaffe, T. R. Jaffe, W. C. Jones, M. Juvela, E. Keihänen, R. Keskitalo, T. S. Kisner, R. Kneissl, J. Knoche, L. Knox, M. Kunz, H. Kurki-Suonio, G. Lagache, A. Lähteenmäki, J. M. Lamarre, A. Lasenby, M. Lattanzi, C. R. Lawrence, J. P. Leahy, R. Leonardi, J. Lesgourgues, F. Levrier, A. Lewis, M. Liguori, P. B. Lilje, M. Linden-Vørnle, M. López-Cañiego, P. M. Lubin, J. F. Macías-Pérez, G. Maggio, D. Maino, N. Mandolesi, A. Mangilli, A. Marchini, M. Maris, P. G. Martin, M. Martinelli, E. Martínez-González, S. Masi, S. Matarrese, P. McGehee, P. R. Meinhold, A. Melchiorri, J. B. Melin, L. Mendes, A. Mennella, M. Migliaccio, M. Millea, S. Mitra, M. A. Miville-Deschênes, A. Moneti, L. Montier, G. Morgante, D. Mortlock, A. Moss, D. Munshi, J. A. Murphy, P. Naselsky, F. Nati, P. Natoli, C. B. Netterfield, H. U. Nørgaard-Nielsen, F. Noviello, D. Novikov, I. Novikov, C. A. Oxborrow, F. Paci, L. Pagano, F. Pajot, R. Paladini, D. Paoletti, B. Partridge, F. Pasian, G. Patanchon, T. J. Pearson, O. Perdereau, L. Perotto, F. Perrotta, V. Pettorino, F. Piacentini, M. Piat, E. Pierpaoli, D. Pietrobon, S. Plaszczynski, E. Pointecouteau, G. Polenta, L. Popa, G. W. Pratt, G. Prézeau, S. Prunet, J. L. Puget, J. P. Rachen, W. T. Reach, R. Rebolo, M. Reinecke, M. Remazeilles, C. Renault, A. Renzi, I. Ristorcelli, G. Rocha, C. Rosset, M. Rossetti, G. Roudier, B. Rouillé d’Orfeuil, M. Rowan-Robinson, J. A.

- Rubiño-Martín, B. Rusholme, N. Said, V. Salvatelli, L. Salvati, M. Sandri, D. Santos, M. Savelainen, G. Savini, D. Scott, M. D. Seiffert, P. Serra, E. P. S. Shellard, L. D. Spencer, M. Spinelli, V. Stolyarov, R. Stompor, R. Sudiwala, R. Sunyaev, D. Sutton, A. S. Suur-Uski, J. F. Sygnet, J. A. Tauber, L. Terenzi, L. Toffolatti, M. Tomasi, M. Tristram, T. Trombetti, M. Tucci, J. Tuovinen, M. Türler, G. Umama, L. Valenziano, J. Valiviita, F. Van Tent, P. Vielva, F. Villa, L. A. Wade, B. D. Wandelt, I. K. Wehus, M. White, S. D. M. White, A. Wilkinson, D. Yvon, A. Zacchei, and A. Zonca. Planck 2015 results. XIII. Cosmological parameters. *A&A*, 594:A13, September 2016.
- [353] Ming-Feng Ho, Scott E. Perkins, Simeon Bird, William Dawson, Nathan Golovich, Jessica R. Lu, and Peter McGill. Evidence for two non-homogeneous black hole populations in LIGO GWTC-3. (*unpublished companion paper*).
- [354] S. E. Woosley. Pulsational Pair-instability Supernovae. *ApJ*, 836(2):244, February 2017.
- [355] Amanda M. Farah, Bruce Edelman, Michael Zevin, Maya Fishbach, Jose María Ezquiaga, Ben Farr, and Daniel E. Holz. Things That Might Go Bump in the Night: Assessing Structure in the Binary Black Hole Mass Spectrum. *Astrophys. J.*, 955(2):107, 2023.
- [356] Will M. Farr, Simon Stevenson, M. Coleman Miller, Ilya Mandel, Ben Farr, and Alberto Vecchio. Distinguishing Spin-Aligned and Isotropic Black Hole Populations With Gravitational Waves. *Nature*, 548:426, 2017.

.1 Sample posteriors for $\mathcal{M}_{\text{DLA}(2)}$

The calculation of the Poisson-Binomial process in Eq. 2.70 computes the probability of N DLAs within a given column density or redshift bin on the sample posteriors $p_{\text{DLA}}^i(\theta) = p(\{\mathcal{M}_{\text{DLA}}\} \mid \mathbf{y}_i, \boldsymbol{\lambda}_i, \boldsymbol{\nu}_i, \theta, z_{\text{QSO}i})$, where i represents the index of the quasar sample. However, with more than 1 DLA, we will not just have parameters in two-dimensions $\theta = (\log_{10} N_{\text{HI}}, z_{\text{DLA}})$ but also parameters from the second or third DLAs $\{\theta_j\}_{j=1}^k = \{(\log_{10} N_{\text{HI}j}, z_{\text{DLA}j})\}_{j=1}^k$, with k DLAs. It is thus not straightforward to see how we can calculate the Poisson-Binomial process on the sample posteriors with more than 1 DLA.

Here we provide a procedure to calculate the sample posteriors of the second DLA given the parameters of the first DLA. The sample posteriors of the second DLA could be written as:

$$\begin{aligned}
& p(\text{2nd DLA at } \theta = (\log_{10} N_{\text{HI}}, z_{\text{DLA}})) \\
&= \int_{\text{1st DLA} \in \theta'} p(\text{1st DLA at } \theta' \text{ and 2nd DLA at } \theta) d\theta' \\
&= \int_{\theta'} p(\theta, \theta' \mid \mathcal{M}_{\text{DLA}(2)}, \mathcal{D}) d\theta'
\end{aligned} \tag{.1}$$

where we marginalize the first DLA at parameters $\theta' = (\log_{10} N_{\text{HI}}', z'_{\text{DLA}})$ with a given 2nd DLA parameter $\theta = (\log_{10} N_{\text{HI}}, z_{\text{DLA}})$. We can furthermore write the joint posterior density into a likelihood density using Bayes rule:

$$\begin{aligned}
& p(\theta, \theta' \mid \mathcal{M}_{\text{DLA}(2)}, \mathcal{D}) \\
& \propto p(\mathcal{D} \mid \theta, \theta', \mathcal{M}_{\text{DLA}(2)}) p(\theta', \theta \mid \mathcal{M}_{\text{DLA}(2)}) \\
& = p(\mathcal{D} \mid \theta, \theta', \mathcal{M}_{\text{DLA}(2)}) p(\theta \mid \mathcal{M}_{\text{DLA}(1)}) \\
& \quad p(\theta' \mid \mathcal{M}_{\text{DLA}(1)}, \mathcal{D})
\end{aligned} \tag{.2}$$

where the joint prior density $p(\theta', \theta \mid \mathcal{M}_{\text{DLA}(2)})$ could be written as a product of a non-informative prior and an informed prior.

The posterior density of the second DLA could thus be expressed as a discrete sum over θ' at the informed prior density:

$$\begin{aligned}
& p(\text{2nd DLA at } \theta) \\
& \propto \int_{\theta'} p(\mathcal{D} \mid \theta, \theta', \mathcal{M}_{\text{DLA}(2)}) p(\theta \mid \mathcal{M}_{\text{DLA}(1)}) \\
& \quad p(\theta' \mid \mathcal{M}_{\text{DLA}(1)}, \mathcal{D}) d\theta' \\
& \simeq \frac{1}{N} \sum_{i=1}^N p(\mathcal{D} \mid \theta, \theta'_i, \mathcal{M}_{\text{DLA}(2)}) p(\theta \mid \mathcal{M}_{\text{DLA}(1)}),
\end{aligned} \tag{.3}$$

where

$$\theta'_i \sim p(\theta' \mid \mathcal{M}_{\text{DLA}(1)}, \mathcal{D}). \quad (.4)$$

However, for each θ , we only have one θ' . We thus can simplify the discrete sum as:

$$\begin{aligned} p(\text{2nd DLA at } \theta) \\ \propto p(\mathcal{D} \mid \theta, \theta'_i, \mathcal{M}_{\text{DLA}(2)})p(\theta \mid \mathcal{M}_{\text{DLA}(1)}), \end{aligned} \quad (.5)$$

where the non-informative prior $p(\theta \mid \mathcal{M}_{\text{DLA}(1)})$ expresses the way we sample θ for $p(\text{2nd DLA at } \theta)$.

To get the normalized posterior density for the 2nd DLA, we can directly normalize on the joint likelihood density:

$$\begin{aligned} p(\text{2nd DLA at } \theta) \\ &= \frac{p(\mathcal{D} \mid \{\theta, \theta'\}_j, \mathcal{M}_{\text{DLA}(2)})}{\sum_{j=1}^N p(\mathcal{D} \mid \{\theta, \theta'\}_j, \mathcal{M}_{\text{DLA}(2)})} \\ &= \frac{p(\mathcal{D} \mid \{\theta, \theta'\}_j, \mathcal{M}_{\text{DLA}(2)})}{N^2 \frac{1}{N^2} \sum_{j=1}^N p(\mathcal{D} \mid \{\theta, \theta'\}_j, \mathcal{M}_{\text{DLA}(2)})} \\ &= \frac{1}{N^2} \frac{p(\mathcal{D} \mid \{\theta, \theta'\}_j, \mathcal{M}_{\text{DLA}(2)})}{p(\mathcal{D} \mid \mathcal{M}_{\text{DLA}(2)})} \end{aligned} \quad (.6)$$

We thus can compute the posterior density for the first DLA and second DLA at a given θ :

$$\begin{aligned} p(\text{1 or 2 DLAs at } \theta) \\ &= \Pr(\mathcal{M}_{\text{DLA}(1)})p(\text{1st DLA at } \theta) \\ &\quad + \Pr(\mathcal{M}_{\text{DLA}(2)})p(\text{2nd DLA at } \theta) \end{aligned} \quad (.7)$$

$\log_{10} N_{\text{HI}}$	$f(N_{\text{HI}}) (10^{-21})$	68% limits (10^{-21})	95% limits (10^{-21})
20.0 – 20.1	0.413	0.405 – 0.422	0.398 – 0.430
20.1 – 20.2	0.241	0.235 – 0.247	0.229 – 0.252
20.2 – 20.3	0.175	0.171 – 0.180	0.167 – 0.184
20.3 – 20.4	0.136	0.132 – 0.139	0.129 – 0.142
20.4 – 20.5	0.101	$[9.88 – 10.41] \times 10^{-2}$	$[9.63 – 10.67] \times 10^{-2}$
20.5 – 20.6	7.60×10^{-2}	$[7.40 – 7.80] \times 10^{-2}$	$[7.21 – 8.00] \times 10^{-2}$
20.6 – 20.7	5.20×10^{-2}	$[5.06 – 5.36] \times 10^{-2}$	$[4.91 – 5.50] \times 10^{-2}$
20.7 – 20.8	3.84×10^{-2}	$[3.73 – 3.96] \times 10^{-2}$	$[3.63 – 4.07] \times 10^{-2}$
20.8 – 20.9	2.52×10^{-2}	$[2.44 – 2.60] \times 10^{-2}$	$[2.37 – 2.69] \times 10^{-2}$
20.9 – 21.0	1.67×10^{-2}	$[1.61 – 1.72] \times 10^{-2}$	$[1.56 – 1.78] \times 10^{-2}$
21.0 – 21.1	1.03×10^{-2}	$[9.94 – 10.75] \times 10^{-3}$	$[9.57 – 11.15] \times 10^{-3}$
21.1 – 21.2	7.21×10^{-3}	$[6.92 – 7.49] \times 10^{-3}$	$[6.65 – 7.78] \times 10^{-3}$
21.2 – 21.3	4.17×10^{-3}	$[3.99 – 4.37] \times 10^{-3}$	$[3.81 – 4.56] \times 10^{-3}$
21.3 – 21.4	2.87×10^{-3}	$[2.74 – 3.01] \times 10^{-3}$	$[2.62 – 3.14] \times 10^{-3}$
21.4 – 21.5	1.49×10^{-3}	$[1.41 – 1.58] \times 10^{-3}$	$[1.33 – 1.68] \times 10^{-3}$
21.5 – 21.6	8.71×10^{-4}	$[8.23 – 9.31] \times 10^{-4}$	$[7.76 – 9.79] \times 10^{-4}$
21.6 – 21.7	4.03×10^{-4}	$[3.74 – 4.41] \times 10^{-4}$	$[3.46 – 4.69] \times 10^{-4}$
21.7 – 21.8	2.15×10^{-4}	$[1.96 – 2.37] \times 10^{-4}$	$[1.77 – 2.60] \times 10^{-4}$
21.8 – 21.9	1.41×10^{-4}	$[1.29 – 1.56] \times 10^{-4}$	$[1.17 – 1.70] \times 10^{-4}$
21.9 – 22.0	4.75×10^{-5}	$[4.04 – 5.70] \times 10^{-5}$	$[3.56 – 6.41] \times 10^{-5}$
22.0 – 22.1	2.08×10^{-5}	$[1.70 – 2.64] \times 10^{-5}$	$[1.32 – 3.21] \times 10^{-5}$
22.1 – 22.2	8.99×10^{-6}	$[7.49 – 13.49] \times 10^{-6}$	$[6.00 – 16.49] \times 10^{-6}$
22.2 – 22.3	4.76×10^{-6}	$[3.57 – 7.14] \times 10^{-6}$	$[2.38 – 9.52] \times 10^{-6}$
22.3 – 22.4	1.89×10^{-6}	$[9.46 – 37.83] \times 10^{-7}$	$[9.46 – 56.74] \times 10^{-7}$
22.4 – 22.5	3.00×10^{-6}	$[2.25 – 3.76] \times 10^{-6}$	$[1.50 – 4.51] \times 10^{-6}$
22.5 – 22.6	5.97×10^{-7}	$[5.97 – 17.90] \times 10^{-7}$	$[5.97 – 29.83] \times 10^{-7}$
22.6 – 22.7	0	$0 – 9.48 \times 10^{-7}$	$0 – 9.48 \times 10^{-7}$
22.7 – 22.8	3.76×10^{-7}	$[3.76 – 11.29] \times 10^{-7}$	$[3.76 – 11.29] \times 10^{-7}$
22.8 – 22.9	0	$0 – 2.99 \times 10^{-7}$	$0 – 5.98 \times 10^{-7}$
22.9 – 23.0	0	$0 – 2.38 \times 10^{-7}$	$0 – 4.75 \times 10^{-7}$

Table .1: Average column density distribution function for all DLAs with $2 < z < 5$. The table is generated by using $p(\{\mathcal{M}_{\text{DLA}}\} | \mathbf{y}, \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\text{QSO}})$. See also Figure 2.14.

z	dN/dX	68% limits	95% limits
2.00 – 2.17	0.0309	0.0302 – 0.0317	0.0294 – 0.0325
2.17 – 2.33	0.0448	0.0438 – 0.0458	0.0428 – 0.0468
2.33 – 2.50	0.0497	0.0485 – 0.0510	0.0474 – 0.0521
2.50 – 2.67	0.0528	0.0514 – 0.0542	0.0501 – 0.0556
2.67 – 2.83	0.0676	0.0658 – 0.0694	0.0641 – 0.0711
2.83 – 3.00	0.0721	0.0700 – 0.0743	0.0680 – 0.0764
3.00 – 3.17	0.0760	0.0734 – 0.0788	0.0709 – 0.0814
3.17 – 3.33	0.0846	0.0811 – 0.0885	0.0779 – 0.0919
3.33 – 3.50	0.0824	0.0785 – 0.0868	0.0747 – 0.0910
3.50 – 3.67	0.0835	0.0786 – 0.0888	0.0737 – 0.0937
3.67 – 3.83	0.0738	0.0671 – 0.0806	0.0618 – 0.0873
3.83 – 4.00	0.0594	0.0512 – 0.0675	0.0454 – 0.0757
4.00 – 4.17	0.0665	0.0570 – 0.0797	0.0475 – 0.0892
4.17 – 4.33	0.1033	0.0893 – 0.1228	0.0726 – 0.1396
4.33 – 4.50	0.0966	0.0805 – 0.1208	0.0644 – 0.1409
4.50 – 4.67	0.1137	0.0885 – 0.1453	0.0632 – 0.1706
4.67 – 4.83	0.1131	0.0754 – 0.1634	0.0503 – 0.2011
4.83 – 5.00	0	0 – 0.057	0 – 0.085

Table .2: Table of dN/dX values from our multi-DLA catalogue for $2 < z < 5$. The table is generated by using $p(\{\mathcal{M}_{\text{DLA}}\} | \mathbf{y}, \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\text{QSO}})$. See also Figure 2.15.

z	$\Omega_{\text{DLA}}(10^{-3})$	68% limits	95% limits
2.00 – 2.17	0.385	0.371 – 0.400	0.358 – 0.416
2.17 – 2.33	0.532	0.516 – 0.550	0.501 – 0.568
2.33 – 2.50	0.645	0.620 – 0.679	0.596 – 0.720
2.50 – 2.67	0.653	0.624 – 0.689	0.598 – 0.728
2.67 – 2.83	0.786	0.759 – 0.814	0.732 – 0.841
2.83 – 3.00	0.792	0.764 – 0.822	0.737 – 0.850
3.00 – 3.17	0.910	0.865 – 0.972	0.826 – 1.046
3.17 – 3.33	1.051	1.002 – 1.101	0.957 – 1.154
3.33 – 3.50	0.958	0.891 – 1.031	0.829 – 1.106
3.50 – 3.67	1.297	1.220 – 1.380	1.147 – 1.455
3.67 – 3.83	1.303	1.222 – 1.391	1.144 – 1.486
3.83 – 4.00	0.891	0.742 – 1.052	0.639 – 1.205
4.00 – 4.17	0.993	0.746 – 1.245	0.564 – 1.488
4.17 – 4.33	1.519	1.341 – 1.718	1.168 – 1.923
4.33 – 4.50	1.085	0.880 – 1.325	0.702 – 1.695
4.50 – 4.67	1.741	1.282 – 2.224	0.666 – 2.851
4.67 – 4.83	1.239	0.826 – 1.712	0.484 – 2.222
4.83 – 5.00	0	0 – 0.213	0 – 0.492

Table .3: Table of Ω_{DLA} values. The table is generated by using $p(\{\mathcal{M}_{\text{DLA}}\} | \mathbf{y}, \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\text{QSO}})$. See also Figure 2.16.

Table .4: Table of dN/dX values, integrated over all putative absorbers with $N_{\text{HI}} > 10^{20.3}$ in our catalogue.

z	dN/dX	68% limits	95% limits
2.00 – 2.17	0.0337	0.0330 – 0.0345	0.0323 – 0.0352
2.17 – 2.33	0.0429	0.0421 – 0.0438	0.0413 – 0.0446
2.33 – 2.50	0.0462	0.0452 – 0.0472	0.0443 – 0.0481
2.50 – 2.67	0.0493	0.0482 – 0.0505	0.0471 – 0.0516
2.67 – 2.83	0.0620	0.0606 – 0.0634	0.0592 – 0.0649
2.83 – 3.00	0.0660	0.0643 – 0.0678	0.0627 – 0.0695
3.00 – 3.17	0.0704	0.0683 – 0.0726	0.0663 – 0.0747
3.17 – 3.33	0.0745	0.0719 – 0.0774	0.0695 – 0.0800
3.33 – 3.50	0.0763	0.0729 – 0.0800	0.0696 – 0.0833
3.50 – 3.67	0.0777	0.0735 – 0.0821	0.0697 – 0.0862
3.67 – 3.83	0.0632	0.0586 – 0.0688	0.0539 – 0.0735
3.83 – 4.00	0.0648	0.0585 – 0.0720	0.0522 – 0.0792
4.00 – 4.17	0.0581	0.0507 – 0.0670	0.0447 – 0.0745
4.17 – 4.33	0.0709	0.0620 – 0.0842	0.0532 – 0.0953
4.33 – 4.50	0.1024	0.0896 – 0.1216	0.0736 – 0.1376
4.50 – 4.67	0.0827	0.0689 – 0.1057	0.0552 – 0.1241
4.67 – 4.83	0.1041	0.0818 – 0.1413	0.0669 – 0.1636
4.83 – 5.00	0.0676	0.0507 – 0.1184	0.0169 – 0.1522

.2 Tables for CDDF, dN/dX , and Ω_{DLA}

.3 Tables of the measurements for DLAs in SDSS DR16

.4 Likelihood Function of Average Mass Spectrum Fitting

In Section 6.3.3, we discuss how we obtain the fiducial parameters for our mixture model through fitting the average mass spectrum of the Power-law+Peak model. In this appendix, we describe the detailed procedures of this fitting.

We first forward sample the (m_1, m_2) pairs using the Power-law+Peak model (with the fiducial parameters in Table 6.1), consisting of a m_1 function in Eq. 6.9 and a power-law

Table .5: Ω_{DLA} values, integrated over all putative absorbers with $N_{\text{HI}} > 10^{20.3}$ in our catalogue.

z	$\Omega_{\text{DLA}}(10^{-3})$	68% limits	95% limits
2.00 – 2.17	0.582	0.550 – 0.619	0.520 – 0.659
2.17 – 2.33	0.610	0.576 – 0.651	0.548 – 0.694
2.33 – 2.50	0.691	0.664 – 0.722	0.638 – 0.755
2.50 – 2.67	0.647	0.621 – 0.676	0.596 – 0.706
2.67 – 2.83	0.770	0.738 – 0.809	0.711 – 0.855
2.83 – 3.00	0.747	0.723 – 0.773	0.701 – 0.799
3.00 – 3.17	0.789	0.758 – 0.829	0.729 – 0.896
3.17 – 3.33	0.850	0.810 – 0.909	0.773 – 1.042
3.33 – 3.50	0.908	0.855 – 0.962	0.792 – 1.019
3.50 – 3.67	1.019	0.953 – 1.087	0.866 – 1.166
3.67 – 3.83	0.664	0.604 – 0.731	0.550 – 0.806
3.83 – 4.00	0.887	0.781 – 1.000	0.683 – 1.112
4.00 – 4.17	0.562	0.508 – 0.622	0.457 – 0.684
4.17 – 4.33	1.061	0.843 – 1.337	0.708 – 1.675
4.33 – 4.50	1.507	1.252 – 1.810	1.038 – 2.182
4.50 – 4.67	0.595	0.473 – 0.737	0.373 – 0.892
4.67 – 4.83	0.913	0.657 – 1.208	0.465 – 1.498
4.83 – 5.00	1.221	0.449 – 1.995	0.127 – 2.449

q function in Eq. 6.10. Then we concatenate a series of (m_1, m_2) pairs to a 1-D array of black hole masses, assuming primary and secondary masses are arbitrary labels.

With a 1-D array of the forward sampled black hole masses, we apply a KDE to obtain the probability density function of this 1-D array, $p_{\text{KDE}}(m)$. We want to know how much the shape parameters, the spectral index and the location and standard deviation of the Gaussian bump change after we concatenate the (m_1, m_2) into a 1-D array. Then, we assume the average mass spectrum ($p_{\text{ave}}(m)$) follows the power-law + peak structure and fit it to $p_{\text{KDE}}(m)$:

$$\begin{aligned}
 p_{\text{ave}}(m \mid -\alpha, \delta_m, m_{\text{max}}, m_{\text{min}}, \mu, \sigma, \lambda_p) = & \\
 (1 - \lambda_p)\mathcal{B}(m \mid -\alpha, m_{\text{max}}, m_{\text{min}}, \delta_m) & \quad (.8) \\
 + \lambda_p\mathcal{G}(m \mid \mu, \sigma, m_{\text{min}}, \delta_m). &
 \end{aligned}$$

Table .6: The column density distribution function integrated over all spectral lengths within $2 < z < 5$.

$\log_{10} N_{\text{HI}}$	$f(N_{\text{HI}}) (10^{-21})$	68% limits (10^{-21})	95% limits (10^{-21})
20.0 – 20.1	0.371	0.365 – 0.378	0.358 – 0.385
20.1 – 20.2	0.235	0.230 – 0.240	0.225 – 0.244
20.2 – 20.3	0.170	0.166 – 0.173	0.162 – 0.177
20.3 – 20.4	0.128	0.125 – 0.131	0.122 – 0.134
20.4 – 20.5	9.58×10^{-2}	$[9.36 - 9.80] \times 10^{-2}$	$[9.15 - 10.02] \times 10^{-2}$
20.5 – 20.6	7.16×10^{-2}	$[6.99 - 7.33] \times 10^{-2}$	$[6.83 - 7.50] \times 10^{-2}$
20.6 – 20.7	5.09×10^{-2}	$[4.97 - 5.23] \times 10^{-2}$	$[4.85 - 5.35] \times 10^{-2}$
20.7 – 20.8	3.56×10^{-2}	$[3.47 - 3.66] \times 10^{-2}$	$[3.38 - 3.75] \times 10^{-2}$
20.8 – 20.9	2.45×10^{-2}	$[2.38 - 2.52] \times 10^{-2}$	$[2.31 - 2.59] \times 10^{-2}$
20.9 – 21.0	1.64×10^{-2}	$[1.59 - 1.69] \times 10^{-2}$	$[1.55 - 1.74] \times 10^{-2}$
21.0 – 21.1	1.06×10^{-2}	$[1.02 - 1.09] \times 10^{-2}$	$[9.92 - 11.29] \times 10^{-3}$
21.1 – 21.2	6.96×10^{-3}	$[6.72 - 7.22] \times 10^{-3}$	$[6.48 - 7.47] \times 10^{-3}$
21.2 – 21.3	4.58×10^{-3}	$[4.41 - 4.77] \times 10^{-3}$	$[4.25 - 4.94] \times 10^{-3}$
21.3 – 21.4	2.66×10^{-3}	$[2.55 - 2.79] \times 10^{-3}$	$[2.43 - 2.91] \times 10^{-3}$
21.4 – 21.5	1.51×10^{-3}	$[1.44 - 1.60] \times 10^{-3}$	$[1.36 - 1.68] \times 10^{-3}$
21.5 – 21.6	9.95×10^{-4}	$[9.43 - 10.56] \times 10^{-4}$	$[8.91 - 11.08] \times 10^{-4}$
21.6 – 21.7	4.82×10^{-4}	$[4.52 - 5.23] \times 10^{-4}$	$[4.18 - 5.57] \times 10^{-4}$
21.7 – 21.8	2.60×10^{-4}	$[2.39 - 2.84] \times 10^{-4}$	$[2.18 - 3.08] \times 10^{-4}$
21.8 – 21.9	1.50×10^{-4}	$[1.35 - 1.69] \times 10^{-4}$	$[1.23 - 1.83] \times 10^{-4}$
21.9 – 22.0	7.73×10^{-5}	$[6.98 - 8.86] \times 10^{-5}$	$[6.03 - 9.81] \times 10^{-5}$
22.0 – 22.1	4.34×10^{-5}	$[3.74 - 5.09] \times 10^{-5}$	$[3.30 - 5.69] \times 10^{-5}$
22.1 – 22.2	1.43×10^{-5}	$[1.19 - 2.02] \times 10^{-5}$	$[8.33 - 23.80] \times 10^{-6}$
22.2 – 22.3	1.13×10^{-5}	$[8.51 - 15.12] \times 10^{-6}$	$[6.62 - 17.96] \times 10^{-6}$
22.3 – 22.4	3.75×10^{-6}	$[3.00 - 6.01] \times 10^{-6}$	$[1.50 - 8.26] \times 10^{-6}$
22.4 – 22.5	2.39×10^{-6}	$[1.79 - 4.77] \times 10^{-6}$	$[5.96 - 59.63] \times 10^{-7}$
22.5 – 22.6	1.42×10^{-6}	$[9.47 - 28.42] \times 10^{-7}$	$[4.74 - 37.90] \times 10^{-7}$
22.6 – 22.7	7.53×10^{-7}	$[3.76 - 15.05] \times 10^{-7}$	$0 - 2.26 \times 10^{-6}$
22.7 – 22.8	5.98×10^{-7}	$[2.99 - 11.96] \times 10^{-7}$	$0 - 1.79 \times 10^{-6}$
22.8 – 22.9	7.12×10^{-7}	$[4.75 - 14.24] \times 10^{-7}$	$[2.37 - 16.62] \times 10^{-7}$
22.9 – 23.0	5.66×10^{-7}	$[1.89 - 9.43] \times 10^{-7}$	$0 - 1.32 \times 10^{-6}$

Here, m refers to the black hole mass regardless of the labels of primary/secondary m_1 and m_2 . The likelihood function for finding the best-fit shape parameters is:

$$\begin{aligned} \log \mathcal{L}(\boldsymbol{\theta}) = & \\ & -\frac{1}{2} \sum \left[\log(2\pi\sigma^2) + \frac{(p_{\text{KDE}}(m) - p_{\text{ave}}(m | \boldsymbol{\theta}))^2}{\sigma^2} \right]. \end{aligned} \quad (.9)$$

Here, we fit the parameters of $\boldsymbol{\theta} = (-\alpha, \delta_m, \mu, \sigma, \lambda_p)$ but fix $(m_{\text{max}}, m_{\text{min}})$ to the input values to the forward sampling of the Power-law+Peak model. The prior for each shape parameter is listed in Table .7. The high mass end of the $p_{\text{KDE}}(m)$ has numerical noise due to a lack of Monte Carlo samples to reconstruct the correct Power-law+Peak through a KDE. To avoid this artifact affecting the fit, we let σ scale as the Poisson error

$$\sigma = \sigma_0 \sqrt{\frac{p_{\text{KDE}}(m)}{N}}, \quad (.10)$$

where N is the total number of the Monte Carlo samples used to build the KDE probability density function. Ideally, with a larger number of samples, the numerical uncertainty is smaller. We assume a broad prior for the scaling constant σ_0 for this numerical uncertainty,

$$\sigma_0 \sim \text{LogNormal}(\mu = 0, \sigma = 1). \quad (.11)$$

The fitting gives $\sigma_0 = 15_{-2}^{+2}$ with $2\text{-}\sigma$ error.

.5 Fiducial Inference with a Different Power Spectral Density

In this work, we utilize the analytical Power Spectral Density (PSD), `AdvMidHighSensitivityP12` from `PyCBC` [302] and the `IMRPhenomD` [305, 306, 307] waveform model, to calculate the

Table .7: The prior for shape parameters in the average mass spectrum fitting.

Parameter	Description	Prior
δ_m	The δm for the low-mass mass spectrum smoothing	U(0, 10)
m_{\max}	Maximum mass bound for the power-law model	-
m_{\min}	Minimum mass bound for both power-law and Gaussian models	-
$-\alpha$	Spectral index of the power-law	U(1, 6)
μ	Mean of the Gaussian model	U(20, 50)
σ	Standard deviation of the Gaussian model	U(1, 6)
λ_p	Mixing fraction of the Gaussian model	U(0, 0.5)

detection efficiency of our population model across different subpopulations. The detection probability, $p_{\text{det}}(\theta)$, is computed following the approach from Ref. [295]. We employ a pre-marginalized version that excludes detector-dependent variables, focusing instead on primary mass, secondary mass, and luminosity distance, where $\theta = (m_1, m_2, L)$. We set $\rho_{\text{threshold}} = 8$, meaning that we consider a trigger to be a detection if the SNR is above 8.

Population models aim to establish physically motivated priors for black hole properties. Our work specifically targets the modeling of the black hole mass spectrum. For each subpopulation model, we sample black hole masses (m_a, m_b) and convert these values to (m_1, m_2) . These Monte Carlo samples are utilized to determine detection efficiency. Given that our model does not account for luminosity distance, we introduce a prior on the luminosity distance, $p(L) \propto L^2$, within a range of (5, 5000) Mpc, ensuring it is uniform across the survey volume.

In principle, the most robust approach for calculating detection efficiency requires marginalizing over the PSDs from various survey operational periods. Throughout this study, we have utilized an analytical PSD, `AdvMidHighSensitivityP1200087`. Therefore, our goal here is to assess the sensitivity of our inference results to different PSDs, ensuring

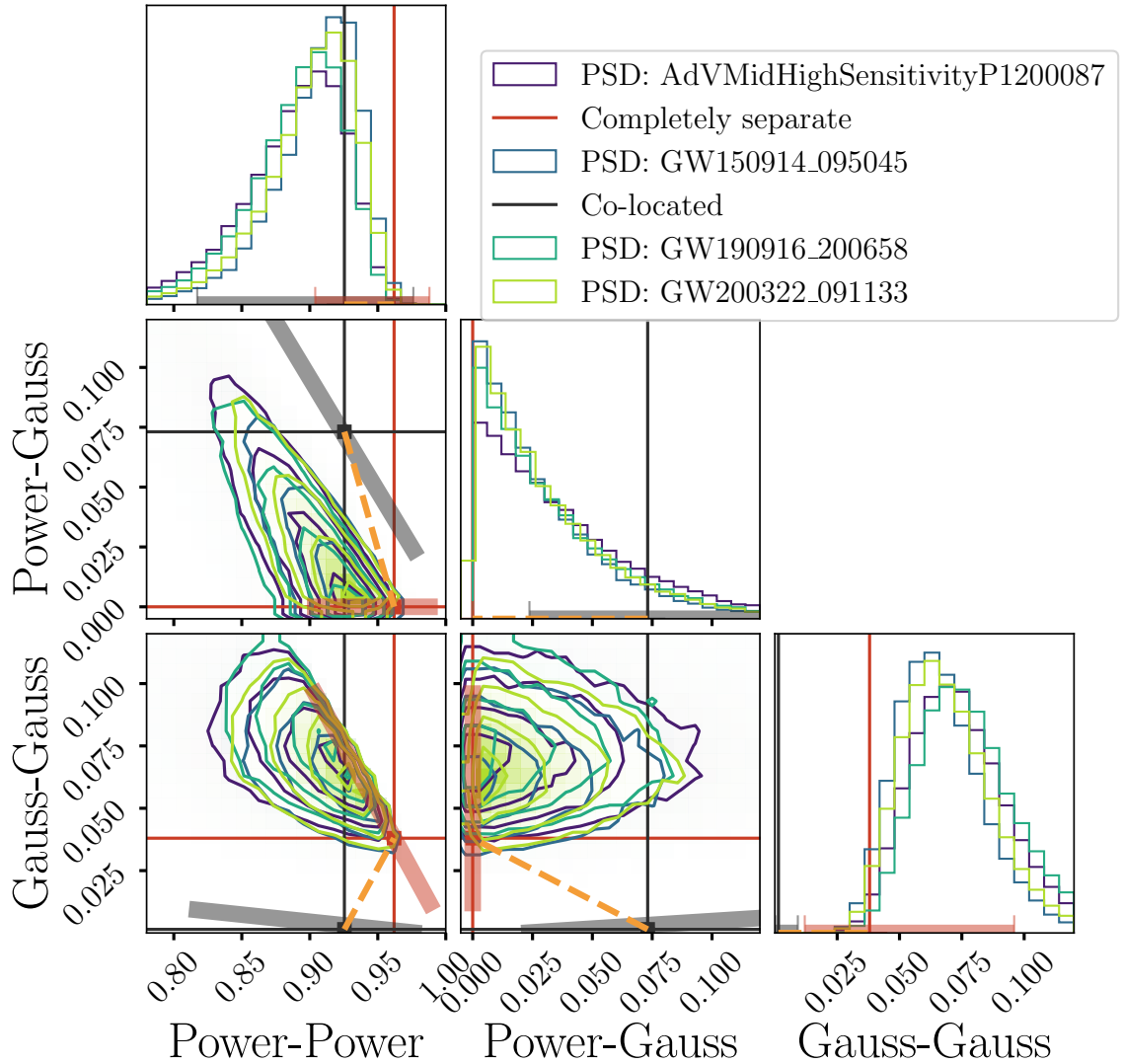


Figure .1: Fiducial MCMC chains using various Power Spectral Densities (PSDs) in the computation of the detection probability, p_{det} . Four different PSDs are utilized: the analytical PSD (AdVMidHighSensitivityP1200087), the PSD from the GWTC-1 event (GW150914_095045), the PSD from the GWTC-2 event (GW190916_200658), and the PSD associated with the GWTC-3 event (GW200322_091133). There is no evident shift in the posterior with different PSDs.

that our conclusions are not significantly impacted by the choice of PSD and that our PSD approximation is sufficient.

Figure .1 presents the “Fiducial” inference results regarding the mixing fractions using various PSDs, including those from GWTC-1 (GW150914_095045), GWTC-2 (GW190916_200658), and GWTC-3 (GW200322_091133). We observe minimal shifts in the mode of the posterior distribution (less than 1-sigma), with the exception that the width of the posterior for the AdVMidHighSensitivityP1200087 PSD is slightly broader than that of the others. This suggests that the impact of using different PSDs on the main results (Figure 6.7) presented in this paper is negligible.