# Importance of nonuniform Brillouin zone sampling for *ab initio* Bethe-Salpeter equation calculations of exciton binding energies in crystalline solids

Antonios M. Alvertis,[1, 2, *] Aurélie Champagne,[2, 3] Mauro Del Ben,[4] Felipe H. da Jornada,[5, 6] Diana Y. Qiu,[7] Marina R. Filip,[8] and Jeffrey B. Neaton[2, 3, 9, †]

[1] *KBR, Inc, NASA Ames Research Center, Moffett Field, California 94035, United States*
[2] *Materials Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, United States*
[3] *Department of Physics, University of California Berkeley, Berkeley, United States*
[4] *Applied Mathematics & Computational Research Division,*
*Lawrence Berkeley National Laboratory, Berkeley, California 94720, United States*
[5] *Department of Materials Science and Engineering, Stanford University, Stanford, CA 94305, USA*
[6] *Stanford Institute for Materials and Energy Sciences,*
*SLAC National Accelerator Laboratory, Menlo Park, California 94025, United States*
[7] *Department of Mechanical Engineering and Material Sciences,*
*Yale University, New Haven, Connecticut 06511, United States*
[8] *Department of Physics, University of Oxford, Oxford OX1 3PJ, United Kingdom*
[9] *Kavli Energy NanoScience Institute at Berkeley, Berkeley, United States*

(Dated: November 13, 2023)

Excitons are prevalent in semiconductors and insulators, and their binding energies are critical for optoelectronic applications. The state-of-the-art method for first-principles calculations of excitons in extended systems is the *ab initio* *GW*-Bethe-Salpeter equation (BSE) approach, which can require a fine sampling of reciprocal space to accurately resolve solid-state exciton properties. Here we show, for a wide range of semiconductors and insulators, that the commonly employed approach of uniformly sampling the Brillouin zone can lead to underconverged exciton binding energies, as impractical grid sizes are required to achieve adequate convergence. We further show that nonuniform sampling of the Brillouin zone, focused on the region of reciprocal space where the exciton wavefunction resides, enables efficient rapid numerical convergence of exciton binding energies at a given level of theory. We propose a well-defined convergence procedure, which can be carried out at relatively low computational cost and which in some cases leads to a correction of previous best theoretical estimates by almost a factor of two, qualitatively changing the predicted exciton physics. These results call for the adoption of nonuniform sampling methods for *ab initio* *GW*-BSE calculations, and for revisiting previously computed values for exciton binding energies of many systems.

## I. INTRODUCTION

Excitons are correlated two-particle electron-hole states that predominantly form in semiconductors and insulators. The binding energy of excitons is a critical quantity that determines photocurrent generation in solar cells [1, 2], the possibility of a material forming long-lived excited states for quantum information [3, 4], or the extent to which phonons can screen the attractive Coulomb interaction between the electron and hole [5, 6]. Therefore, the accurate prediction of exciton binding energies from first principles is imperative in the quest for novel semiconductors for diverse optoelectronic applications.

The state-of-the-art method to describe excited state properties in extended systems from first principles is based on many-body perturbation theory within the $GW$ approximation [7, 8] and the Bethe-Salpeter equation approach [9–12]($GW$-BSE), where $G$ is the one-particle Green's function and $W$ is the screened Coulomb interaction. Exciton binding energies computed within the *ab initio* $GW$-BSE framework with current approximations and implementations can be extremely challenging to converge numerically, for two main reasons. Firstly, a very fine sampling of the Brillouin zone (BZ) can be required to resolve essential features of the screened interaction $W$ accurately. This has been identified as an issue in low-dimensional systems and has been addressed elsewhere [13–15]. We therefore focus here on a second convergence challenge, which is that the BSE needs to be solved on ultra-dense **k**-point grids [12, 16, 17] due to the fact that excitons are highly localized in reciprocal space in many known bulk semiconductors of interest. To make such calculations on dense grids possible, so-called "dual-grid" interpolation schemes have been developed, which allow for interpolation between two different uniform **k**-grids across the BZ, a coarse and a fine one [18]. We will refer to these methods as uniform dual grid interpolation (UDGI).

The localized nature of excitons in reciprocal space is consistent with the Wannier-Mott model [19, 20], which describes excitons composed of holes and electrons, with

[*] amalvertis@lbl.gov
[†] jbneaton@lbl.gov

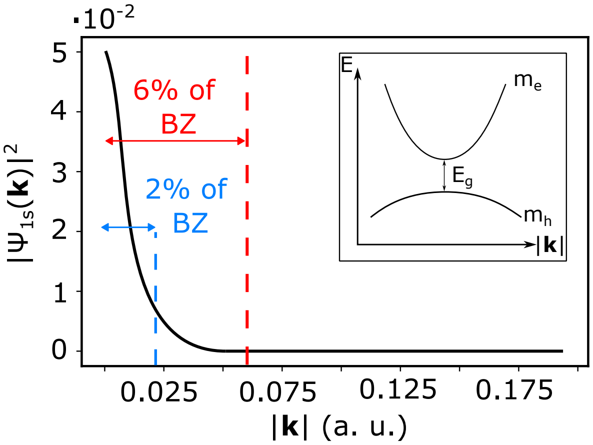

FIG. 1. $1s$ exciton wavefunction in reciprocal space, for GaN in the wurtzite phase, from the Wannier-Mott model (eq. 1). We indicate in blue and red the regions with a cutoff around $\Gamma$ that corresponds to approximately 2% and 6% respectively of the distance to the edge of the first Brillouin zone (BZ). Inset: Schematic band dispersion for a two-band model, including a parabolic valence and conduction band with effective masses $m_h$ and $m_e$, respectively, with a gap of $E_g$ separating the two bands. The GaN material parameters used to compute $|\Psi|^2$ are given in Appendix C, along with the relevant computational details.

parabolic valence and conduction bands and with effective mass $m_h$ and $m_e$ respectively, as visualized in Fig. 1. Within this limit, the reciprocal space wavefunction of the $1s$ exciton can be written as

$$\Psi_{1s}(\mathbf{k}) = \frac{(2a_o)^{3/2}}{\pi} \cdot \frac{1}{(1+a_o^2k^2)^2}, \tag{1}$$

where $a_o = 1/\sqrt{2E_B\mu}$ is the exciton Bohr radius, $E_B$ the exciton binding energy and $\frac{1}{\mu} = \frac{1}{m_e} + \frac{1}{m_h}$ the exciton effective mass. As seen in Fig. 1, when using parameters for GaN, the exciton wavefunction decays rapidly around the zone center. When performing *ab initio* $GW$-BSE calculations using a uniform sampling of the entire BZ, millions of **k**-points are required in order to sufficiently sample the critical region $|\mathbf{k}| \lesssim \frac{2\pi}{a_o}$, with most of the computational effort spent on regions that are not relevant to the exciton physics. Such a strong localization of the exciton wavefunction in reciprocal space is not unique to excitons that are Wannier-Mott-like, and is present in a wide range of bulk systems and beyond, including low-dimensional systems such as transition metal dichalcogenides with excitons localized around the K and K’ valleys [13]. For most materials, the ultra-dense sampling of critical BZ regions required to converge exciton properties is not feasible even when utilizing UDGI. This has resulted in poor numerical convergence of values for the exciton binding energy in some cases, as has been discussed in previous work [17]. Moreover, it has been proposed that convergence of exciton binding energies may be accelerated with UDGI by considering an average screening $W$ in the region near the origin of the BZ [18, 21, 22]; but while this scheme indeed improves the convergence of excitation energies, it has no effect on the exciton binding energies, since it results in a rigid shift of the onset of the exciton continuum [14].

In order to achieve true numerical convergence with respect to **k**-grids, dual grids may be used to interpolate between a coarse uniform grid and a fine nonuniform grid that is designed to include exclusively a patch of the entire BZ, which encompasses the relevant region where the exciton resides. Such nonuniform dual-grid interpolation (NUDGI) approaches allow the BSE to be solved with greatly reduced computational effort, and yields fast and systematic convergence for exciton binding energies [12, 14], as one can afford to effectively increase the **k**-grid density in the critical region without having to sample the entire BZ. Ref. [16] proposed an alternative scheme that does not rely on interpolation between different grids, but instead uses a hybrid **k**-grid across the BZ which is dense in the region of interest and coarse outside it, making it necessary to assign varying weights to the points in the two regions. In Ref. [16] this scheme was used to demonstrate convergence in the exciton binding energies of MgO and InN.

The examples of NUDGI and related strategies when computing exciton binding energies remain rare in the literature, and are mostly focused on systems with very small exciton binding energies, which are known to be challenging to converge, such as GaAs [11], InN [16], and halide perovskites [5, 23]. It is therefore currently unclear to what extent solving the BSE employing NUDGI or alternative sampling schemes is necessary in order to obtain numerically-converged exciton binding energies in general semiconductors of interest, or even whether it is generally possible to obtain accurate values for most systems with the widely used UDGI techniques. Additionally, while convergence with traditional UDGI methods is a matter of increasing the density of the grid used to sample reciprocal space within $GW$-BSE, when using NUDGI methods the size of the patch of the BZ is also a convergence parameter, and there is currently no well-defined procedure for choosing this parameter.

Here we demonstrate that in most semiconductors with Wannier-Mott-like excitons, employing nonuniform sampling when computing exciton binding energies is imperative in order to obtain numerically-converged values. We show that uniformly sampling the BZ when solving the BSE with computationally-feasible grids yields exciton binding energies that are often *greatly* overestimated relative to their converged values, which can lead to qualitatively incorrect predictions of the physics of the exciton, as we discuss for GaN in Section IV. We propose a well-defined procedure for numerical convergence of the $GW$-BSE exciton binding energies at low computational cost, and we show that even when employing NUDGI, errors can arise that, in some cases, lead to fortuitous agreement to experiment. We find that, even for standard semiconductors such as Si and GaN, the lack of convergence of the BZ sampling is the main cause for the discrepancies between reported values of $GW$-BSE exciton binding energies, often leading to differences by more than a factor of three [17, 24]. Our results demonstrate the need for use of nonuniform sampling methods when computing exciton binding energies, and for revisiting reported values in the literature given the underconvergence of binding energies obtained with UDGI. Intriguingly, we find the changes to previously reported values of exciton binding energies through rigorous convergence obtained here can be as significant as corrections associated with dynamical screening of the Coulomb interaction by carrier plasmons [25] and phonons [6], underscoring the need for nonuniform sampling methods for prediction of exciton binding energies.

The structure of this paper is as follows. Section II reviews the theoretical background of our work. Specifically, Section II A provides an overview of the first-principles description of excitons in solids within the *ab initio* $GW$-BSE formalism, while Section II B describes how one can define a patch of the Brillouin zone in which to sample the exciton properties. In Section III we present our computational results for the exciton binding energy of a range of semiconductors. First we perform a systematic convergence study of the exciton binding energy of the prototypical semiconductors Si and GaN, and demonstrate the importance of employing a nonuniform BZ sampling strategy. Following that, we present our results for a several semiconductors, comparing to previously reported literature values and addressing discrepancies with these prior results due to lack of convergence. Finally, in Section IV we discuss our overall results and their implications for the prediction of exciton properties within the *ab initio* $GW$-BSE framework.

## II. THEORETICAL BACKGROUND

### A. First-principles description of excitons in solids

The Bethe-Salpeter equation (BSE) within the Tamm-Dancoff approximation for zero-momentum excitons in reciprocal space with clamped nuclei is written as [11, 12]

$$(E^{\mathrm{QP}}_{c\mathbf{k}} - E^{\mathrm{QP}}_{v\mathbf{k}})A^S_{cv\mathbf{k}} + \sum_{c'v'\mathbf{k}'} \langle cv\mathbf{k}| K^{eh} |c'v'\mathbf{k}'\rangle A^S_{c'v'\mathbf{k}'} = \Omega^S A^S_{cv\mathbf{k}}, \quad (2)$$

where $E_{c\mathbf{k}}$ and $E_{v\mathbf{k}}$ are the quasiparticle energies of conduction and valence band states, respectively (generally obtained at the $GW$ level). The coefficients $A^S_{cv\mathbf{k}}$ describe the corresponding excited state $S$ with excitation energy $\Omega_S$ as a linear combination of free electron-hole pair wavefunctions, typically obtained from a density functional theory (DFT) calculation. The excited state wavefunction can be written as

$$|S\rangle = \sum_{cv\mathbf{k}} A^S_{cv\mathbf{k}} |cv\mathbf{k}\rangle\,. \quad (3)$$

The kernel $K^{eh}$ in eq. 2 describes the interaction between electrons and holes and consists of direct ($d$) and exchange ($x$) contributions, $K^{eh} = K^d + K^x$. Ignoring the frequency-dependence of the direct term, which is a reasonable approximation if the exciton binding energy is much smaller than the plasma frequency, one may write [12]

$$\langle v c\mathbf{k}| K^d |v'c'\mathbf{k}'\rangle = -\int d\mathbf{r}d\mathbf{r}'\psi^*_c(\mathbf{r})\psi_{c'}(\mathbf{r})W(\mathbf{r},\mathbf{r}',\omega=0)\psi^*_{v'}(\mathbf{r}')\psi_v(\mathbf{r}'), \quad (4)$$

and

$$\langle v c\mathbf{k}| K^x |v'c'\mathbf{k}'\rangle = \int d\mathbf{r}d\mathbf{r}'\psi^*_c(\mathbf{r})\psi_v(\mathbf{r})v(\mathbf{r},\mathbf{r}')\psi^*_{v'}(\mathbf{r}')\psi_{c'}(\mathbf{r}'), \quad (5)$$

with $v$ the bare Coulomb interaction, and

$$W(\mathbf{r},\mathbf{r}',\omega) = \int d\mathbf{r}''\epsilon^{-1}(\mathbf{r},\mathbf{r}'',\omega)v(\mathbf{r}'',\mathbf{r}') \quad (6)$$

the screened Coulomb interaction. Here $\epsilon(\mathbf{r},\mathbf{r}'',\omega)$ is the frequency-dependent, non-local dielectric function. In most applications, and within this work, $\epsilon$ is computed within the random-phase approximation (RPA) [26]. Upon solving the BSE (eq. 2), the exciton binding energy for low-lying resonant exciton $S$ is obtained as

$$E_b = \min_{\mathbf{k}}[E^{\mathrm{QP}}_{c\mathbf{k}} - E^{\mathrm{QP}}_{v\mathbf{k}}] - \Omega_S, \quad (7)$$

*i.e.* as the difference of the minimum direct quasiparticle gap across the BZ and the exciton energy.

In standard *ab initio* BSE calculations of solids, the above kernel matrices are constructed on a coarse grid of **k**-points, usually the same as that used in a preceding $GW$ calculation. However, it is well known that observable quantities such as absorption spectra and exciton binding energies obtained through the solution of the BSE require a very fine grid in order to achieve convergence. Since the calculation of kernel matrix elements on very fine grids can often be computationally prohibitive, dual-grid schemes have been proposed, which generally involve the calculation of DFT wavefunctions on a coarse and a fine **k**-grid, but only require computing the kernel matrix elements on the coarse grid; the BSE Hamiltonian is subsequently interpolated onto the fine grid [18, 21, 27]. Such an interpolation approach has been proposed and implemented in, for example, the BerkeleyGW software package [18], which we employ in this work. The basis for this scheme is a BSE kernel interpolation through a simple expansion of the fine-grid wavefunction in terms of the nearest coarse grid wavefunction as

$$u_{n\mathbf{k}_{fi}} = \sum_{n'} c^{\mathbf{k}_{co}}_{n,n'} u_{n'\mathbf{k}_{co}}, \tag{8}$$

where $u_{n\mathbf{k}}$ is the cell periodic part of the Kohn-Sham wavefunction, $\mathbf{k}_{co}$ the closest coarse-grid point to the fine-grid point $\mathbf{k}_{fi}$, and $n$ the band index. The coefficients $c^{\mathbf{k}_{co}}_{n,n'}$ are obtained as the overlap between coarse- and fine-grid wavefunctions as

$$c^{\mathbf{k}_{co}}_{n,n'} = \int d\mathbf{r} u_{n\mathbf{k}_{fi}}(\mathbf{r}) u^{*}_{n'\mathbf{k}_{co}}(\mathbf{r}). \tag{9}$$

Using these overlap coefficients, one can interpolate the kernel matrix as

$$\langle vc\mathbf{k}_{fi} | K | v'c'\mathbf{k}'_{fi} \rangle = \sum_{n_1,n_2,n_3,n_4} c^{\mathbf{k}_{co}}_{c,n_1} c^{*\mathbf{k}_{co}}_{v,n_2} c^{*\mathbf{k}'_{co}}_{c',n_3} c^{\mathbf{k}'_{co}}_{v',n_4} \langle n_2 n_1 \mathbf{k}_{co} | K | n_4 n_3 \mathbf{k}'_{co} \rangle . \tag{10}$$

The interpolated quantity $K$ can be the exchange kernel, or modified versions of the direct kernel that analytically handle the sharp variations of the matrix elements with respect to the transfer wavevector $\mathbf{q} = \mathbf{k} - \mathbf{k}'$.

In a similar fashion, the conduction and valence $GW$ quasiparticle energies appearing in eq. 2 are interpolated onto the fine grid as

$$E^{\mathrm{QP}}_{n\mathbf{k}_{fi}} = E^{\mathrm{MF}}_{n\mathbf{k}_{fi}} + \left\langle \sum_{n'} |c^{\mathbf{k}_{co}}_{n,n'}|^2 (E^{\mathrm{QP}}_{n'\mathbf{k}_{co}} - E^{\mathrm{MF}}_{n'\mathbf{k}_{co}}) \right\rangle_{\mathbf{k}_{co}} , \tag{11}$$

where the brackets indicate linear interpolation performed using the tetrahedron method, and $E^{\mathrm{MF}}_n$ is the mean-field energy of band $n$, a Kohn-Sham eigenstate from a DFT calculation.

Overall, by interpolating the quasiparticle energies and kernel matrix elements on a fine grid, we may solve the BSE (eq. 2) on this same fine grid, greatly accelerating the convergence of exciton energies $\Omega^S$ and exciton coefficients $A^S_{cv\mathbf{k}}$. Importantly, it is not a requirement for the interpolation of the quasiparticle energies and kernel that the fine grid $\mathbf{k}_{fi}$ be uniform or have any specific characteristics. The fine $\mathbf{k}$-grid may be any general nonuniform grid, such as a patch of the BZ, which, as we demonstrate in the following Section III for a set of representative materials, is particularly important for converging the exciton binding energy. In the following Section II B we discuss different ways of defining patches within the relevant regions of the BZ when computing properties of excitons.

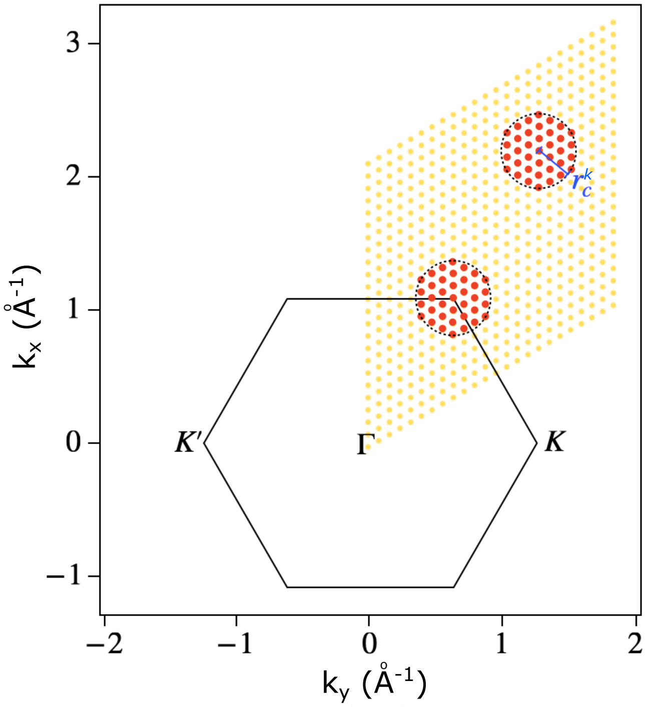

FIG. 2. Schematic of the sampling of the Brillouin zone for monolayer MoS$_2$, with the $\mathbf{k}$-points in a patch of radius $r^k_c = 0.30$ Å$^{-1}$around the K/K' ($\mathbf{k}_o$ = K/K') valleys highlighted in red.

### B. Reciprocal space patches

There are several ways in which one could extract a patch of $\mathbf{k}$-points in the region of the BZ relevant to the exciton under study when starting from a $N_1 \times N_2 \times N_3$ grid that spans the entire BZ. However, the philosophy behind the choice of a patch is always the same: first one must identify the point $\mathbf{k}_o$ of reciprocal space around which the exciton is centered, and then one must decide on the size of the patch, discarding any points outside that region. In this manner, we are left with a truncated fine $\mathbf{k}$-grid of a single density $N_1 \times N_2 \times N_3$, making this approach distinct to methods employing hybrid grids across the entire BZ [16] and removing any need to assign different weights to points within our grid.

For example, for bulk systems with 1$s$ Wannier-Mott-like excitons (eq. 1), $\mathbf{k}_o = \Gamma$, since the exciton coefficients peak at $\mathbf{k} = \mathbf{0}$. For transition metal dichalcogenides it has been found that $\mathbf{k}_o$ = K/K' [13]. Generally, if one has no prior knowledge of the system under study, the point $\mathbf{k}_o$ may be determined by identifying where the minimum direct gap occurs in the electronic band structure, or performing an initial BSE calculation on a coarse grid across the entire BZ to identify the relevant region of reciprocal space.

Having determined the exciton center $\mathbf{k}_o$, one may proceed to defining a patch centered around this point. It is also possible to define multiple patches around more than

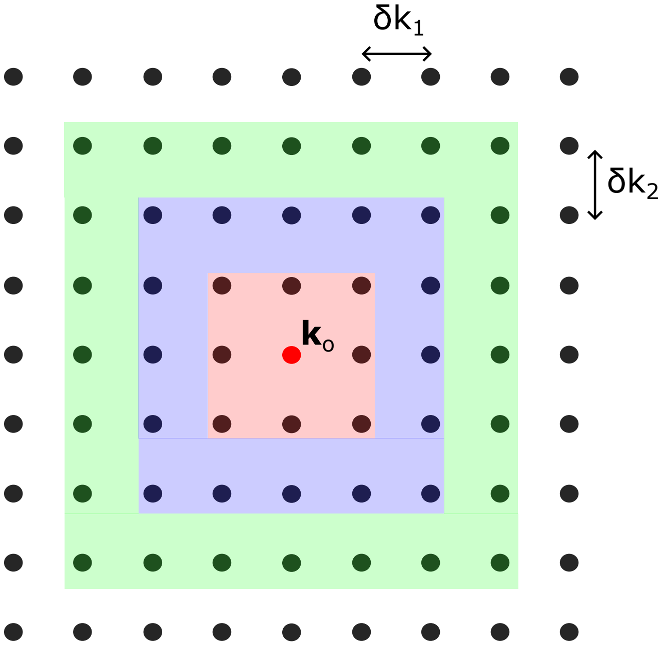

FIG. 3. Schematic of the sampling of the Brillouin zone around a point of interest $\mathbf{k}_o$, using the scheme of eq. 14 based on crystal coordinates. If for example one starts from a $100 \times 100 \times 100$ regular grid and hence $\delta k_1 = \delta k_2 = 0.01$ in crystal coordinates, the region in red here corresponds to a patch with cutoff coordinate $d_c^k = 0.01$, the region in blue to $d_c^k = 0.02$ and so on.

one point of interest in the BZ. One way of capturing the region around a center $\mathbf{k}_o$, is to define a spherical patch of radius $r_c^k$ around that point. The $\mathbf{k}$-points within such a spherical patch satisfy the condition

$$|\mathbf{k} - \mathbf{k}_o| \leq r_c^k, \tag{12}$$

and the value of $r_c^k$ functions as a convergence parameter. An example of a two-dimensional circular patch is visualized in Fig. 2, where a radius of $r_c^k = 0.3$ Å$^{-1}$ has been chosen for a MoS$_2$ monolayer.

An alternative way of generating patches is the following. Let us consider $\mathbf{k}$-points in crystal coordinates ($k_1$,$k_2$,$k_3$), that is, fractions of the primitive reciprocal lattice vectors $\mathbf{b}_1$,$\mathbf{b}_2$, and $\mathbf{b}_3$ as follows

$$\mathbf{k} = k_1\mathbf{b}_1 + k_2\mathbf{b}_2 + k_3\mathbf{b}_3. \tag{13}$$

Here the crystal coordinates ($k_1$,$k_2$,$k_3$) assume values in the range $[-0.5, 0.5]$, and we only retain those points that satisfy the condition

$$-d_c^k \leq k_i - k_{o,i} \leq d_c^k \quad i = 1, 2, 3, \tag{14}$$

where $d_c^k$ a cutoff coordinate. Such a choice of points in reciprocal space is visualized for a two-dimensional example in Fig. 3. While generating a patch based on crystal coordinates has the disadvantage of not allowing one to define a single “cutoff radius” in units of inverse length, it more readily clarifies the percentage of the BZ that is included in the patch along each spatial direction. For example, defining a cutoff coordinate $d_c^k = 0.02$ in crystal coordinates in eq. 14 would suggest that we include 4% of the BZ $[-0.5, 0.5]$ around $\mathbf{k}_o$. Moreover, patches defined in this way are immediately transferable between different systems, as they do not depend on specific material parameters.

Regardless of the method that one chooses to generate a patch in the BZ, the exciton properties will converge to the same answer as long as the relevant region has been adequately sampled by the chosen method. In this work we generate patches in crystal coordinates following eq. 14, with the exception of MoS$_2$, for which we employ circular patches (eq. 12) following previous work [14], and we provide a detailed discussion of its exciton binding energy convergence properties in Appendix A. Moreover, all systems studied in Section III have excitons that are Γ-centered (*i.e.* $\mathbf{k}_o = \mathbf{0}$), with monolayer MoS$_2$ in Appendix A providing an example of a case with $\mathbf{k}_o \neq \mathbf{0}$, reinforcing the relevance of nonuniform sampling methods for non Γ-centered excitons.

## III. RESULTS

We start by presenting the convergence properties of the exciton binding energy for two widely studied semiconductors, Si and GaN, in Section III A. The results emphasize the necessity of using NUDGI or a different nonuniform sampling method when solving the Bethe-Salpeter equation in reciprocal space. In Section III B we present numerically converged exciton binding energies with respect to the BZ sampling for a wider range of prototypical semiconductors and we compare our results, obtained with NUDGI, to literature values as well as to our own calculations employing UDGI. We analyze the convergence behavior of the exciton binding energy with respect to the BZ sampling methods and establish systematic trends. We explain discrepancies between calculated values that have been reported previously in some cases, and we show that these can be attributed to lack of convergence in BZ sampling. More details on all the parameters employed in our DFT and *ab initio* *GW*-BSE calculations are given in Appendix C.

### A. Numerical convergence of exciton binding energies

#### *1. Si*

The first step in converging the exciton binding energy of a material using NUDGI on a patch is to understand the localization behavior of the exciton coefficients $A^S_{cv\mathbf{k}}$ in reciprocal space. This could be achieved for example through an initial solution of the BSE using UDGI, in order to gain a better understanding of the decay of

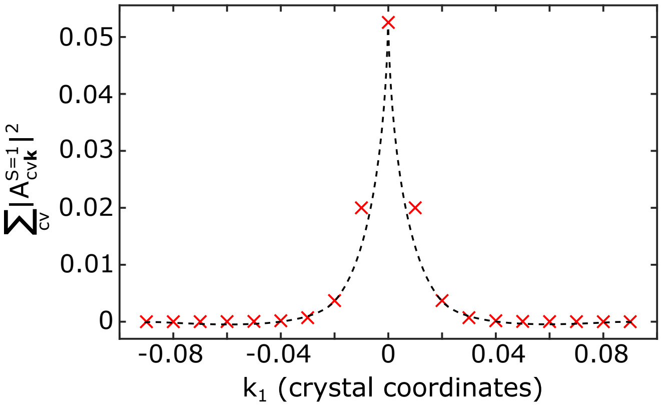

FIG. 4. Exciton coefficients of Si from the solution of the BSE on a patch drawn from a $100 \times 100 \times 100$ regular grid. The decay around Γ of the exciton coefficients is plotted along $k_1$, however is identical to the decay along $k_2, k_3$ given the isotropy of this system. The values of $k_1$, which is along the high symmetry X direction in reciprocal space, are given in crystal coordinates (eq. 13) and the black dashed line serves as a guide to the eye.

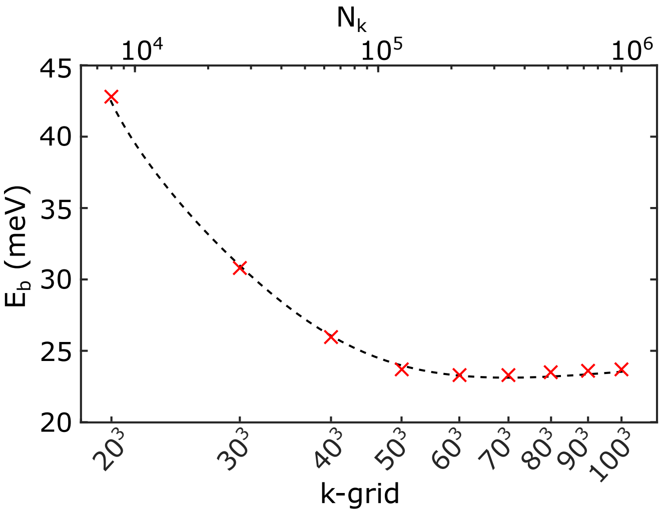

FIG. 5. Convergence of the exciton binding energy of Si with respect to the number of **k**-points, $N_k$, used to solve the BSE, corresponding to grids of $N^3 = N \times N \times N$. Here a patch with a cutoff of $d_c^k = 0.09$ (crystal coordinates) around Γ is employed. The black dashed line serves as a guide to the eye.

the magnitude of the exciton wavefunction relative to its maximum value. For most bulk semiconductors with an exciton binding energy of the order of 10 meV, a few percent of the BZ in each spatial direction is a reasonable guess for the region within which the exciton localizes, as suggested from the Wannier-Mott model, see Fig. 1. Indeed in Fig. 4 we visualize the behavior of the Si exciton coefficients around Γ, as a function of the fractional coordinate $k_1$, as defined in eq. 13. Given the isotropic character of the Si crystal, the behavior is the same along any of the three spatial directions. We see that the exciton wavefunction decays rapidly around Γ.

From Fig. 4, it is reasonable to assume that a patch cutoff coordinate of $d_c^k = 0.09$ is a good first estimate for capturing the relevant part of the BZ when solving the BSE for Si. We now proceed to solve the BSE on patches of this size, which are drawn from grids of varying density. Fig. 5 shows the convergence of the exciton binding energy with respect to the starting **k**-grid for the patch. We see that the converged value of 24 meV is only reached for extremely dense grids of size of at least $50 \times 50 \times 50$. Attempting to use UDGI to solve the BSE on such dense grids would be unfeasible for practical applications, for which grids of $20 \times 20 \times 20$ are commonly considered sufficiently fine. Yet we see that a $20 \times 20 \times 20$ grid for Si leads to a significant overestimation of the exciton binding energy relative to the numerically converged value at this level of theory by almost a factor of 2. Moreover, it is common in the literature to extrapolate plots of the exciton binding energy obtained within UDGI as a function of $1/N_k$ to the limit $N_k \to \infty$ in order to obtain converged $E_b$ values. In Fig. 12 of Appendix B we show for Si that while indeed such an extrapolation results in exciton binding energies that are substantially more converged, one needs grid densities of at least $40^3$ in order to obtain exciton binding energies that are within 1 meV of the converged value. For Si, performing UDGI on $20 \times 20 \times 20$ and $30 \times 30 \times 30$ grids would allow one to obtain an extrapolated $N_k \to \infty$ value of 26 meV for the exciton binding energy, compared to the numerically converged value of approximately 23 meV as shown in Fig. 5. However, our most expensive calculation in Fig. 5 using a patch of cutoff coordinate $d_c^k = 0.09$ drawn from a $100 \times 100 \times 100$ grid, includes only 6,859 **k**-points, which is less than the 8,000 included in a full uniform $20 \times 20 \times 20$ grid. Thus, our converged calculation on a patch drawn from a $100 \times 100 \times 100$ grid has 14% fewer k-points, resulting into a factor 1.6 reduction in computational cost for the diagonalization of the BSE Hamiltonian when compared to the severely underconverged uniform $20 \times 20 \times 20$ calculation. Considering the cost of the $50 \times 50 \times 50$ calculation on a patch, which is converged within 1 meV, we find that it requires only 1% of the computational resources of a full $20 \times 20 \times 20$ calculation. Therefore, sampling a patch of the BZ through NUDGI not only allows us to achieve convergence of the exciton binding energy within 1 meV, which is computationally impractical using UDGI, but it also greatly reduces the computational cost of BSE calculations. Notably, extrapolating to $N_k \to \infty$ with UDGI not only leads to less converged results than NUDGI, but comes at a much higher computational cost.

Having established that a grid of at least $50 \times 50 \times 50$ **k**-points is required to converge the Si exciton binding energy within 1 meV, we return to the issue of choosing a patch cutoff $d_c^k$ that is sufficiently large to capture the relevant part of the BZ. In Fig. 6 we show that increasing the cutoff of a patch drawn from a $50 \times 50 \times 50$ grid around Γ generally increases the exciton binding energy, leading

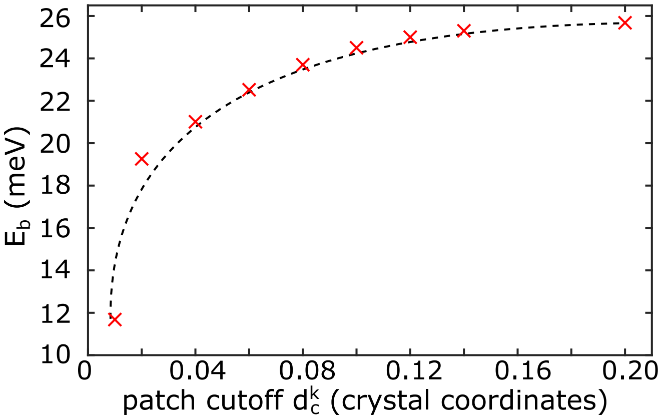

FIG. 6. Convergence of the exciton binding energy of Si with respect to the cutoff coordinate $d_c^k$ of a Γ-centered patch drawn from a $50 \times 50 \times 50$ grid of **k**-points, which is used for interpolation of the BSE kernel. The black dashed line serves as a guide to the eye.

to a converged value of 26 meV for Si. It is therefore not sufficient to converge the density of the grid from which a patch is drawn, but also the patch size.

The convergence procedure developed here is general. One starts from a reasonable first guess of the region of the BZ that has to be included in a BSE calculation, either through the solution of BSE using UDGI, from the physical expectations drawn from the Wannier-Mott model, from the electronic band structure, prior knowledge of the studied system, or other information. Once this initial patch cutoff coordinate has been decided, the grid density is varied within that region until convergence. Then, a separate convergence test is performed for the patch cutoff $d_c^k$, while keeping the density of the grid equal to the one determined in the previous step. Since this procedure is specific to individual exciton states, it cannot be naively transferred to other excited states. In order to obtain converged spectra involving multiple excitons one would have to ensure that the most localized exciton appearing in the spectrum is converged, and that the patch size is sufficient to include the relevant region of every excited state considered. Excitons beyond the 1$s$ state considered here will be even more localized in reciprocal space. While this suggests that the patch size that convergences the 1$s$ exciton will be sufficient to study higher-lying excited states, it is likely that denser grids will be necessary for these states, making the use of NUDGI even more imperative for their study.

### *2. GaN*

Similar to the case of Si, we start the convergence procedure for the exciton binding energy of GaN with an initial guess for a patch that captures the relevant region in the BZ. In Fig. 7 we visualize the decay of the exciton coefficients of this material within a region of $d_c^k = 0.09$ centered at Γ; unlike Si, GaN exhibits different behavior along the $\mathbf{b}_3$ reciprocal lattice vector compared to that along $\mathbf{b}_1$ and $\mathbf{b}_2$ due to its hexagonal symmetry. We see that the exciton wavefunction decays rapidly within this region making a patch cutoff of 0.09 a reasonable starting point for our calculations.

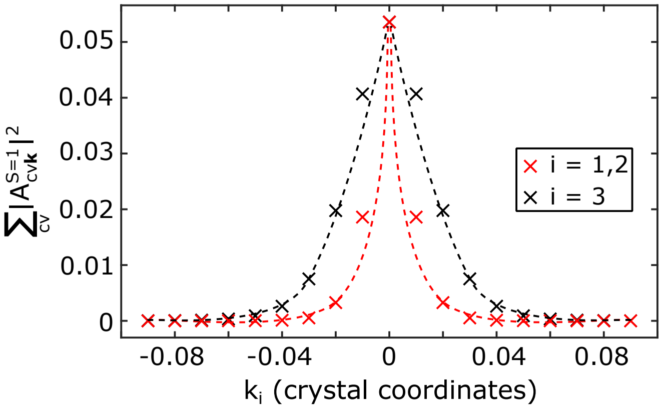

FIG. 7. Exciton coefficients of GaN from the solution of the BSE on a patch drawn from a $100 \times 100 \times 100$ regular grid. The decay around Γ of the exciton coefficients is plotted along $k_1, k_2$ ($k_3$) in red (black) crosses, with the values of $k_i$ given in crystal coordinates (eq. 13). The dashed lines serve as a guide to the eye.

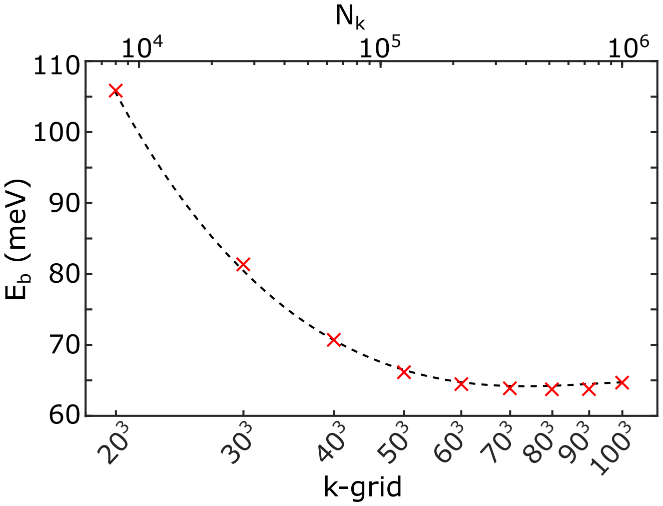

FIG. 8. Convergence of the exciton binding energy of GaN with respect to the number of **k**-points used to solve the BSE, corresponding to grids of $N \times N \times N$, starting from $N = 20$ and in steps of 10 up to $N = 100$. Here a patch with a cutoff of $d_c^k = 0.09$ (crystal coordinates) around Γ is employed. The black dashed line serves as a guide to the eye.

We proceed in Fig. 8 to vary the density of the initial grid with a patch of cutoff $d_c^k = 0.09$, and examine the convergence of the exciton binding energy. Here we find that in order to reach the converged value of 65 meV, a grid of $60 \times 60 \times 60$ or denser is required, which is currently computationally intractable for UDGI BSE calculations. Even if it were possible, the vast majority of the com-

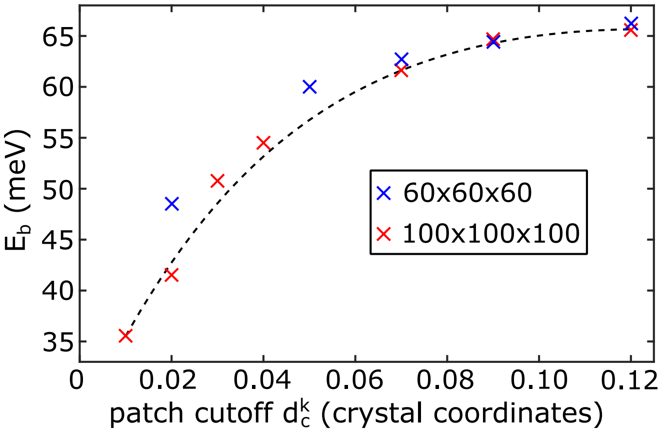

FIG. 9. Convergence of the exciton binding energy of GaN with respect to the cutoff coordinate $d_c^k$ of a Γ-centered patch drawn from $60 \times 60 \times 60$ and $100 \times 100 \times 100$ grid of **k**-points, which is used for interpolation of the BSE kernel. The black dashed line serves as a guide to the eye.

putational workload would be spent on sampling parts of the BZ that are irrelevant to the exciton. Increasing the patch cutoff for a grid of converged density in Fig. 9, shows that a cutoff of 0.09 is sufficient to converge the exciton binding energy of GaN within 1 meV. Finally, we show in Appendix B that for GaN (as with Si), $N_k \to \infty$ extrapolation schemes result in an exciton binding energy within only 3 meV of the converged value once one solves the BSE on grids at least as dense as $40 \times 40 \times 40$. Therefore converging the exciton binding energy using UDGI and an extrapolation is far more computationally demanding than the procedure described above.

### B. Comparison to literature values and BSE on a regular reciprocal-space grid

We now present numerically-converged results with respect to the BZ sampling for exciton binding energies for a range of semiconductors and insulators of interest using a standard *ab initio* $GW$-BSE approach (see Appendix C for details). The studied systems are given in Table I, along with their relevant structural parameters obtained from the Materials Project database [28]. Table II summarizes the converged results for the exciton binding energies $E_{B,\text{patch}}$ of these systems when employing a patch drawn from a $100 \times 100 \times 100$ grid, with a cutoff of 0.12 in crystal coordinates, which is sufficient to converge all values within 1 meV. We compare to the exciton binding energy obtained from a BSE calculation on a regular grid across the entire BZ, $E_{B,\text{regular}}$, with the grid size for each case given in the table in parentheses. Naturally, the latter grids are necessarily much coarser than $100 \times 100 \times 100$ due to the large number of **k**-points to be considered in the region outside the critical region which is included in the patched sampling strategy. We also compare our results to exciton binding energies reported in the literature and obtained from a BSE calculation using a regular grid across the entire BZ, as well as to experimental values.

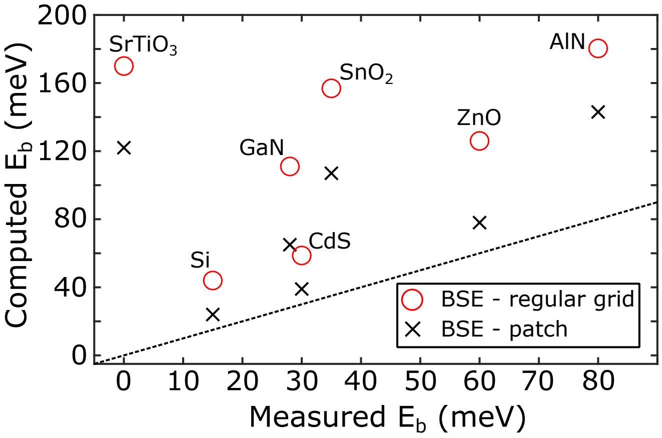

FIG. 10. Exciton binding energies computed with the *ab initio* Bethe-Salpeter equation (BSE) approach (see Section II) on a regular grid (red circles) or a patch (black crosses), and compared to experimental values. Perfect agreement is indicated by the $y = x$ line (black dashed line). For SrTiO$_3$, there is no reported experimental value to the best of our knowledge, and we have set the value to zero. The regular grid BSE values are the literature values summarized in Table II. For ZnO we could not find literature reports of its exciton binding energy within $GW$-BSE, we therefore include our own value on a $24 \times 24 \times 12$ grid across the entire BZ. For experimental values, we pick the largest of the reported values for each system as summarized in Table II.

From Table II, using NUDGI and solving the BSE on a patch results in dramatically reduced values for the exciton binding energy in every studied system. In some cases such as Si, the reduction is almost by a factor of two. This level of numerical convergence is only reachable with the finer grids obtainable with NUDGI, which shows a tendency to substantially reduce the exciton binding energies, as seen in Figs. 5 and 8. We note in passing that the underconverged exciton binding energies computed in this work with UDGI are in fairly good agreement with literature results when using similar grid sizes, which further validates our approach. Moreover, while computed exciton binding energies in all cases overestimate the experimental values, agreement to experiment is substantially improved once rigorous convergence through the nonuniform sampling of the BZ is ensured. This is shown in Fig. 10 for all materials studied. We have excluded MgO from Fig. 10, in order to improve visibility of the data points, due to the large exciton binding energy of this system.

We note that for some of the systems studied here, BSE calculations have previously been computed on a patch within the BZ. In Table III we compare our converged exciton binding energies for AlN, GaN, Si and ZnO to the values reported in Ref. [24], which employed a NUDGI strategy. The values of Ref. [24] are much lower than our

| Material | Structure | $a$ (Å) | $c/a$ | Space Group | Identifier |
|---|---|---|---|---|---|
| AlN | Wurtzite | 3.128 | 1.604 | P6$_3$mc | mp-661 |
| CdS | Zincblende | 4.200 | 1 | F$\bar{4}$3m | mp-2469 |
| GaN | Wurtzite | 3.215 | 1.630 | P6$_3$mc | mp-804 |
| MgO | Halite, Rock Salt | 3.010 | 1 | Fm$\bar{3}$m | mp-1265 |
| Si | Diamond | 3.849 | 1 | Fd$\bar{3}$m | mp-149 |
| SnO$_2$ | Rutile | 4.765 | 0.673 | P4$_2$/mnm | mp-856 |
| SrTiO$_3$ | Cubic Perovskite | 3.852 | 1 | Pm$\bar{3}$m | mp-5229 |
| ZnO | Wurtzite | 3.237 | 1.614 | P6$_3$mc | mp-2133 |

TABLE I. Studied materials, their structure, lattice parameters, space group, and identifier in the Materials Project database [28]. We performed geometry optimization for the atomic positions of these systems using DFT within the PBE exchange-correlation functional, keeping their lattice parameters fixed, with the exception of SrTiO$_3$, for which we used the local density approximation (LDA) and optimized both the atomic positions and its lattice parameter (the LDA has been discussed in the literature to yield more accurate results for structural properties of SrTiO$_3$ compared to PBE [29]).

| Material | $E_{B,\text{patch}}$ | $E_{B,\text{uniform}}$ (this work) | $E_{B,\text{uniform}}$ (literature) | $E_{B,\text{exp.}}$ |
|---|---|---|---|---|
| AlN | 147 | 184 ($24 \times 24 \times 12$) | 181 [17] ($24 \times 24 \times 12$) | 48 [30], 80 [31] |
| CdS | 39 | 65 ($28 \times 28 \times 28$) | 59 [17] ($24 \times 24 \times 24$) | 28 [32], 30 [33] |
| GaN | 65 | 111 ($24 \times 24 \times 12$) | 110 [17] ($24 \times 24 \times 12$) | 20 [34], 28 [35] |
| MgO | 323 | 360 ($24 \times 24 \times 24$) | 370 [5] ($24 \times 24 \times 24$) | 80 [36], 145 [37] |
| Si | 25 | 44 ($20 \times 20 \times 20$) | 42 [17] ($28 \times 28 \times 28$) | 15 [38] |
| SnO$_2$ | 107 | 124 ($18 \times 18 \times 27$) | 157 [39] ($4 \times 4 \times 6$) | 33 [40], 35 [41] |
| SrTiO$_3$ | 122 | 148 ($18 \times 18 \times 18$) | 170 [42] ($20 \times 20 \times 20$) | — |
| ZnO | 78 | 125 ($24 \times 24 \times 12$) | — | 60 [43], 63 [44] |

TABLE II. Exciton binding energies computed within nonuniform dual-grid interpolation (NUDGI) for *GW*-BSE using a patch of cutoff $d_c^k = 0.12$ (in crystal coordinates) drawn from a $100 \times 100 \times 100$ regular grid ($E_{B,\text{patch}}$), uniform dual-grid interpolation (UDGI, $E_{B,\text{uniform}}$), computed within this work and also reported in the literature, with the associated grid given in parentheses in every case, and reported experimental values ($E_{B,\text{exp.}}$). All values are given in meV.

computed values and fortuitously, in much closer agreement to experiment, since the BZ is undersampled, and the studied region too small to yield convergence. As shown in Figs. 6 and 9, using a small patch cutoff can lead to significant *underestimation* of the exciton binding energy. Indeed we expect, after rigorously converging the exciton binding energies, to find an *overestimation* compared to experimental values. This is in fact consistent with screening effects coming from different sources which have not been considered here, such as for example the screening of excitons by phonons [5, 6] and free charge carriers [25]. These effects tend to reduce the exciton binding energy and thus result in closer agreement to experimental values. We therefore conclude that fortuitous agreement with experiment in past work for the exciton binding energies can result from the cancellation of two errors: lack of convergence of the BSE exciton binding energy obtained within NUDGI with respect to the patch cutoff, and not accounting for additional physics such as temperature-dependent dynamical screening of excitons.

## IV. DISCUSSION AND CONCLUSIONS

In this work we demonstrate that the calculation of exciton binding energies within the *ab initio* *GW*-BSE framework may only realistically be converged through a nonuniform BZ sampling strategy, employing a patch around the region where the exciton localizes. Converging BSE calculations with respect to the sampling of reciprocal space using traditional uniform BZ sampling is inefficient, requires extremely fine grids, and leads to prohibitive computational cost even for simple materials. We demonstrate this conclusion by studying the convergence behavior of the BSE exciton binding energy over a wide range of commonly studied semiconductors and insulators. As a result, most values reported in the literature for bulk systems, which rely on sampling the entire BZ using a uniform grid, are underconverged with respect to the employed **k**-grid, leading in some cases to a significant overestimation of exciton binding energies compared to their converged values by up to 40%. This calls for revisiting certain *GW*-BSE predictions of exciton binding energies reported in the literature, and more generally, the use of methods that rely on sampling reciprocal space to obtain exciton properties.

We have presented the convergence behavior of exciton binding energies when nonuniformly sampling the BZ, establishing a scheme for converging this quantity systematically, with the density of the **k**-grid from which we draw patches, and the size of these patches, as the main convergence parameters. Compared to previously underconverged results, nonuniform sampling of the BZ corrects previously reported exciton binding energies by

| Material | $E_{B,\text{patch}}$ - this work | $E_{B,\text{patch}}$ - Ref. [24] | $E_{B,\text{exp.}}$ |
|---|---|---|---|
| AlN | 147 | 70 | 48 [30], 80 [31] |
| GaN | 65 | 30 | 20 [34], 28 [35] |
| Si | 25 | 15 | 15 [38] |
| ZnO | 78 | 60 | 60 [43], 63 [44] |

TABLE III. Exciton binding energies computed within $GW$-BSE employing nonuniform dual-grid interpolation (NUDGI), using a patch of cutoff $d_c^k = 0.12$ (in crystal coordinates) drawn from a $100 \times 100 \times 100$ regular grid ($E_{B,\text{patch}}$), values reported in Ref. [24], and reported experimental values ($E_{B,\text{exp.}}$). All values are given in meV. Values without citations are computed within this work.

an amount that is at least as significant as corrections associated with additional physics, such as temperature-dependent dynamical screening through phonons [5, 6] and free charge carriers [25]. Rigorous convergence of the exciton binding energy with respect to the **k**-grid through nonuniform sampling methods is therefore a critical prerequisite for any calculation that describes such effects.

Rigorous convergence of exciton binding energies can lead to large quantitative changes to their values, and it can also lead to qualitative differences in the predicted physics of an exciton. For example, GaN has a longitudinal optical (LO) phonon of frequency $\omega_{LO} = 84\,\text{meV}$ [5]. The converged exciton binding energy of this system is $E_b = 65\,\text{meV}$ as we found in Section III, which means that $\omega_{LO} > E_b$. A direct consequence of this is that absorption of a single LO phonon by the exciton can lead to its dissociation into a free electron-hole pair, which has been predicted to occur on ultra-fast timescales [6]. On the other hand, as we see in Table II, employing a regular grid that spans the entire BZ for this system yields an exciton binding energy of roughly 110 meV, which would suggest $\omega_{LO} < E_b$, suggesting that absorbing a single phonon is *not* sufficient to dissociate the exciton.

In contrast to the semiconducting compounds studied in the present work, we note that there are also examples of systems where excitons are delocalized in reciprocal space. In those cases there is little benefit to using NUDGI, and UDGI can already provide accurate values for exciton binding energies. Molecular crystals, for example, host Frenkel-like excitons that are relatively localized in real space, and hence highly delocalized in reciprocal space [45, 46]. Halide double perovskites are another class of systems that can host excitons that are delocalized in reciprocal space [47].

We also emphasize that our conclusions on the importance of nonuniform sampling towards obtaining converged exciton properties are not limited to excitons that are Γ-centered. In Section II B we describe the generation of patches centered around arbitrary points in the BZ, which we illustrate in Appendix A for the two-dimensional $MoS_2$ system with excitons centered around the $K$ and $K'$ valleys of the BZ. The methodology described in this work is distinct from the clustered sampling interpolation (CSI) technique, which improves the kernel interpolation procedure and allows to converge the exciton properties of two-dimensional materials with respect to the coarse **k**-grid [15], as also elaborated on in Appendix A.

Overall, our results suggest that a nonuniform sampling of the BZ is critical to obtain numerically converged exciton binding energies within *ab initio* $GW$-BSE and related frameworks used to compute exciton properties in reciprocal space. Additionally, such calculations typically come at a much lower computational cost compared to traditional uniform sampling methods that sample the entire BZ. Nonuniform dual grid interpolation for $GW$-BSE calculations has been implemented and is freely available within the BerkeleyGW software package [18]. We hope that our work will raise visibility for the need to carefully converge exciton binding energies and contribute towards the wide adoption of nonuniform sampling methods of the exciton properties of materials, leading to a more comprehensive understanding of the optoelectronic properties of complex materials.

## V. ACKNOWLEDGMENTS

This work was primarily supported by the Center for Computational Study of Excited-State Phenomena in Energy Materials (C2SEPEM) as part of the Computational Materials Sciences Program, which provided advanced codes, at the Lawrence Berkeley National Laboratory, funded by the U.S. Department of Energy, Office of Science, Basic Energy Sciences, Materials Sciences and Engineering Division, under Contract No. DE-AC02-05CH11231. Computational resources were provided by the National Energy Research Scientific Computing Center (NERSC). F.H.J. acknowledges support by the National Science Foundation (NSF) CAREER award through Grant No. DMR-2238328. D.Y.Q. acknowledges support from the Department of Energy Early Career Award Number DE-SC0021965. M.R.F. acknowledges support from the UK Engineering and Physical Sciences Research Council (EPSRC), Grant EP/V010840/1 and EP/X038777/1.

### Appendix A: Exciton binding energy of $MoS_2$ and comparison to the clustered sampling interpolation method

As discussed above for a set of semiconductors and insulators, achieving converged exciton binding energies

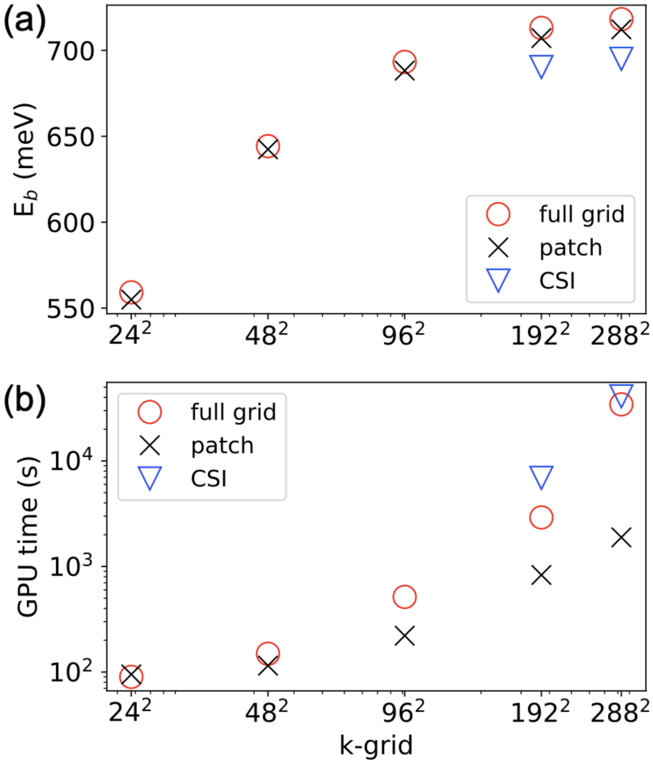

FIG. 11. For a monolayer MoS$_2$: (a) exciton binding energies computed within BSE using the UDGI scheme with a coarse $24 \times 24 \times 1$ grid and a uniform fine grid (red circles) or a $0.3\,\text{Å}^{-1}$ patch around K/K' (black crosses), and compared to the CSI method (blue triangles); and (b) associated GPU time in seconds.

and wavefunctions often requires revisiting the UDGI method implemented in BerkeleyGW [11, 12, 18], such that the fine **k**-grid is no longer a uniform grid across the entire BZ, but rather a dense patch around the Γ-point. This interpolation method is effective for 3D systems as it involves an explicit calculation of the direct and exchange matrix elements in the BSE kernel on a relatively coarse **k**-grid. In contrast, for quasi-2D systems, the sharp features in the inverse dielectric matrix as $\mathbf{q} \to 0$, where $\mathbf{q} = \mathbf{k} - \mathbf{k}'$, the **k**-point difference or momentum transfer, lead to strong variations with **q** in the kernel matrix elements, hence requiring explicit calculation of the interaction matrix elements for several small **q** [15]. The clustered sampling interpolation (CSI) method was proposed to address the challenges in capturing small-**q** features, where the BSE matrix elements are calculated explicitly on a coarse **k**-grid and a cluster of nearby **k** points for each **k** point on the coarse grid [13–15], with the **k**-points in the clusters drawn from a fine grid.

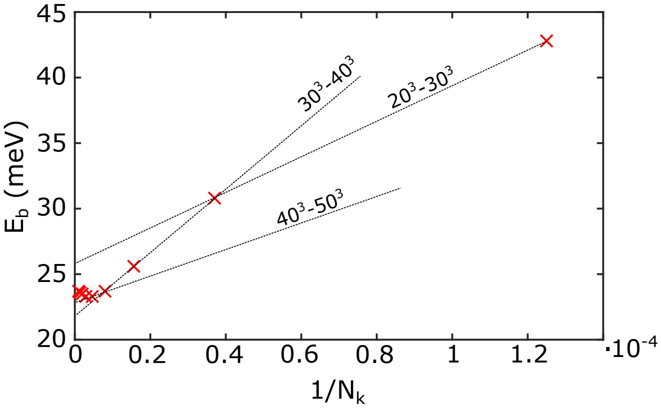

FIG. 12. Convergence of the exciton binding energy of Si as a function of the inverse number of **k**-points included in the sampling of the BZ. The dashed lines extrapolate pairs of points on grids $N^3$ to the $N_k \to \infty$ limit. For example, the line noted as $20^3 - 30^3$ extrapolates to $N_k \to \infty$ by using the $E_b$ values obtained through calculations on $20^3$ and $30^3$ grids.

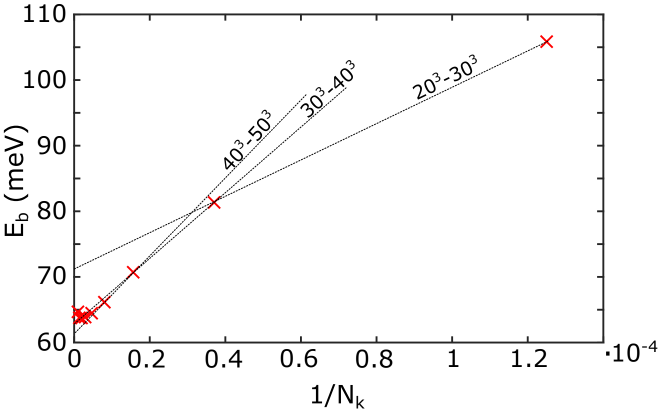

FIG. 13. Convergence of the exciton binding energy of GaN as a function of the inverse number of **k**-points included in the sampling of the BZ. The dashed lines extrapolate pairs of points on grids $N^3$ to the $N_k \to \infty$ limit. For example, the line noted as $20^3 - 30^3$ extrapolates to $N_k \to \infty$ by using the $E_b$ values obtained through calculations on $20^3$ and $30^3$ grids.

Here, we investigate the performance of a nonuniform sampling of the BZ both through the NUDGI (patched sampling) strategy employed in the main manuscript, and the CSI method, when converging the exciton binding energy for the quasi-2D MoS$_2$ system. Starting from a DFT calculation using the PBE functional [48] with a wavefunction cutoff energy of 140 Ry, we compute the self-energy correction using the Godby-Needs plasmon pole model [49], the semiconductor screening for the treatment of the $\mathbf{q} \to 0$ limit, and the slab truncation. We consider a dielectric cutoff of 20 Ry, and include 2000 bands in the Coulomb-hole summation. This gives a QP band gap of 2.81 eV, in good agreement with previous reports [14]. We do not include spin-orbit coupling effects, which are known to split the VBM and CBM at the K/K' valleys. For the BSE calculation, we compute the interaction matrix elements of the kernel on a coarse $24 \times 24 \times 1$ **k**-grid for 4 conduction and 4 valence bands, then interpolate on fine **k**-grids with various densities, including 1 valence and 2 conduction bands.

For NUDGI, we generate two circular patches (eq. 12) and center them around the K/K' **k**-points (see Fig. 2), consistent with the direct gap at K/K'. For the convergence with respect to the patch radius, we find that a radius of 0.3 (0.2) $\text{Å}^{-1}$ for each patch enables convergence of the binding energy of the first exciton within 5 (20) meV of the value obtained on a uniform fine grid of the same density. For a constant patch radius of 0.3 $\text{Å}^{-1}$, we consider various densities for the fine **k**-grid; the computed exciton binding energy is shown in Figure 11(a) as a function of the density of the fine grid. An exciton binding energy of 712 meV is obtained with a grid of $288 \times 288 \times 1$ **k**-points. Using a grid of $192 \times 192 \times 1$ provides a binding energy converged within 5 meV with respect to that value. As mentioned above, the coarse $24 \times 24 \times 1$ **k**-grid used in the UDGI scheme (with or without the patch) might not be sufficient to capture the sharp features for $\mathbf{q} \to 0$. We verified this by using the CSI method and show the computed exciton binding energies in Fig. 11(a) for two fine grids. An exciton binding energy of 695 meV is obtained with a $288 \times 288 \times 1$ fine grid. This value is $\sim 17$ meV smaller than the exciton binding energy obtained with the UDGI method on a fine grid of the same density. Our converged value here somewhat overestimates previously reported exciton binding energies for $MoS_2$ [14, 15], due to differences in the computational approach. Specifically, we use a PBE starting point instead of LDA; we do not account for spin-orbit coupling effects, and we use the one-shot $G_oW_o$ approach, instead of self-consistent $GW$. Moreover, here we employ the Godby-Needs plasmon pole model for the frequency dependence of the dielectric function, and a dielectric cutoff energy of 20 Ry instead of 35 Ry. These factors lead to minor variations in the quasiparticle gap, the band curvature, and ultimately the exciton binding energies. Nevertheless, it is clear from Fig. 11(b) that the use of a patch to sample the BZ greatly reduces the computational load (GPU time and memory - not shown), while still obtaining exciton binding energies in reasonable agreement to those computed with the CSI method. We emphasize that for two-dimensional materials, the CSI method ensures rigorous convergence of exciton properties with respect to the coarse **k**-grid, accurately resolving features of the BSE matrix elements for small **q**. The accuracy of exciton binding energies based on dual-grid interpolation for 2D materials can sensitively depend on the coarse **k**-grid that we interpolate from [15], and for underconverged cases, the interpretation of the binding energy can be sensitive to the sampling of the dielectric function around $\mathbf{q} = \mathbf{0}$. Here, we define the binding energy as the difference between the quasiparticle band gap and the exciton energy, eq. 7. However, the screened Coulomb interaction is averaged in the region of $\mathbf{q} = \mathbf{0}$ [18, 21, 22], leading to a shift in the onset of the continuum in the spectrum of exciton energies. This means that for underconverged cases, the exciton continuum does not correspond to the quasiparticle band gap, leading to a potential overestimation of the exciton binding energy [14]. Nevertheless it is clear that even for quasi-2D systems, NUDGI provides a significant speed-up compared to performing UDGI across the entire BZ, consistent with our conclusions for bulk semiconductors.

## Appendix B: Extrapolation of the exciton binding energy to the $N_k \to \infty$ limit

Fig. 12 and 13 visualize the exciton binding energy of Si and GaN respectively, as a function of the inverse number of **k**-points included in the fine grid for the BSE calculations. Such plots allow us to extrapolate the exciton binding energy to the $N_k \to \infty$ limit, which provides a slightly accelerated rate of convergence compared to performing calculations on denser grids, as discussed in the main text. In both cases here, a patch cutoff coordinate of $d_c^k = 0.09$ (crystal coordinates) has been used, in order to accelerate calculations.

## Appendix C: Computational details for DFT and $GW$-BSE calculations

For all studied materials with the exception of $SrTiO_3$, we first optimize the atomic positions starting from structures drawn from the Materials Project database [28], leaving the lattice parameters fixed. Structural parameters are summarized in Table I. For this we employ DFT within the Quantum Espresso software package [50], and we use the generalized gradient approximation (GGA) at the Perdew, Burke and Ernzerhof (PBE) level [48]. For $SrTiO_3$ we use the local density approximation (LDA) [51] and optimize both the atomic positions and lattice parameters of this system.

Using the DFT-PBE Kohn-Sham wavefunctions (LDA for $SrTiO_3$) as a starting point, we perform $GW$ calculations within the BerkeleyGW code [18], choosing calculation parameters in a way as to converge the quasiparticle band gaps within 0.1 eV, following Refs. [5, 6, 42] and using the Hybertsen-Louie generalized plasmon pole model [8] to compute the dielectric function at finite frequencies in most cases, with the exceptions of MgO, ZnO and $SnO_2$, for which we employ the Godby-Needs model [49, 52]. We find that the following parameters lead to converged $GW$ calculations: AlN (400 bands, 32 Ry polarizability cutoff, $6 \times 6 \times 6$ half-shifted k-grid), CdS (500 bands, 40 Ry polarizability cutoff, $6 \times 6 \times 6$ half-shifted k-grid), GaN (400 bands, 40 Ry polarizability cutoff, $4 \times 4 \times 4$ half-shifted k-grid), MgO (600 bands, 50 Ry polarizability cutoff, $6 \times 6 \times 6$ Γ-centered k-grid), Si (400 bands, 30 Ry polarizability cutoff, $6 \times 6 \times 6$ half-shifted k-grid), $SrTiO_3$ (1000 bands, 14 Ry polarizability cutoff, $6 \times 6 \times 6$ half-shifted k-grid), $SnO_2$ (1024 bands, 48 Ry polarizability cutoff, $6 \times 6 \times 9$ half-shifted k-grid), ZnO (1026 bands, 50 Ry polarizability cutoff, $8 \times 8 \times 5$ half-shifted k-grid), where a shifted grid is used in most cases to achieve better convergence of the dielectric function that is used in the self-energy part of the $GW$ calculation [53, 54], which we find to be critical for obtaining converged exciton binding energies. The $GW$ self-energy is always computed on a $\Gamma$-centered $\mathbf{k}$-grid.

The electronic BSE kernel is computed on the same $\mathbf{k}$-grid as the $GW$ eigenvalues. We computed the kernel for the following number of bands: AlN (4 valence and 4 conduction bands), CdS (4 valence and 4 conduction bands), GaN (4 valence and 4 conduction bands), MgO (8 valence and 8 conduction band), Si (4 valence and 10 conduction bands), $SrTiO_3$ (9 valence and 3 conduction bands), $SnO_2$ (4 valence and 4 conduction bands), ZnO (4 valence and 4 conduction bands). For all cases, we use nonuniform dual grid interpolation to interpolate the kernel onto a patch drawn from a range of fine grids (as outlined in detail in the main text) and typically on 3 valence and 1 conduction bands, with the exceptions of $SrTiO_3$ where the interpolated kernel is computed on 9 valence and 3 conduction bands, and Si where the interpolated kernel is computed on 4 valence and 3 conduction bands.

For GaN, we compute the exciton coefficients within the Wannier-Mott model, eq. 1. To compute $A_{\mathbf{k}}$, we use the converged exciton binding energy of 65 meV from $GW$-BSE. Moreover, we compute the effective masses for the top/bottom valence and conduction bands respectively using the finite difference formula $\frac{1}{m^*} = \frac{E(\delta\mathbf{k})+E(-\delta\mathbf{k})-2E(\Gamma)}{\delta\mathbf{k}^2}$, taking $\delta\mathbf{k}$ to be equal to 0.01 (in crystal coordinates) along each spatial direction, and the energies $E$ are computed at the $GW$ level, using DFT-PBE wavefunctions as a starting point. Finally, we average over the three spatial directions and for the hole and electron effective masses we obtain $m_h = 1.013$ and $m_e = 0.152$ respectively.

[1] G. Grancini, M. Maiuri, D. Fazzi, A. Petrozza, H. J. Egelhaaf, D. Brida, G. Cerullo, and G. Lanzani, Hot exciton dissociation in polymer solar cells, Nature Materials **12**, 29 (2013).

[2] T. J. Savenije, C. S. Ponseca, L. Kunneman, M. Abdellah, K. Zheng, Y. Tian, Q. Zhu, S. E. Canton, I. G. Scheblykin, T. Pullerits, A. Yartsev, and V. Sundström, Thermally activated exciton dissociation and recombination control the carrier dynamics in organometal halide perovskite, Journal of Physical Chemistry Letters **5**, 2189 (2014).

[3] S. Deshpande, T. Frost, A. Hazari, and P. Bhattacharya, Electrically pumped single-photon emission at room temperature from a single InGaN/GaN quantum dot, Applied Physics Letters **105**, 10.1063/1.4897640 (2014).

[4] A. Karasahin, R. M. Pettit, N. Von Den Driesch, M. M. Jansen, A. Pawlis, and E. Waks, Single quantum emitters with spin ground states based on Cl bound excitons in ZnSe, Physical Review A **106**, 2 (2022), 2203.05748.

[5] M. R. Filip, J. B. Haber, and J. B. Neaton, Phonon Screening of Excitons in Semiconductors: Halide Perovskites and beyond, Physical Review Letters **127**, 67401 (2021), 2106.08697.

[6] A. M. Alvertis, J. B. Haber, Z. Li, C. Coveney, S. G. Louie, M. R. Filip, and J. B. Neaton, In preparation.

[7] L. Hedin, New Method for Calculating the One-Particle Green's Function with Application to the Electron-Gas Problem, Physical Review **139**, 796 (1965).

[8] M. S. Hybertsen and S. G. Louie, Electron correlation in semiconductors and insulators: Band gaps and quasiparticle energies, Physical Review B **34**, 5390 (1986).

[9] S. Albrecht, L. Reining, R. Del Sole, and G. Onida, Ab initio calculation of excitonic effects in the optical spectra of semiconductors, Phys. Rev. Lett. **80**, 4510 (1998).

[10] L. X. Benedict, E. L. Shirley, and R. B. Bohn, Optical absorption of insulators and the electron-hole interaction: An ab initio calculation, Physical Review Letters **80**, 4514 (1998).

[11] M. Rohlfing and S. G. Louie, Electron-hole excitations in semiconductors and insulators, Physical Review Letters **81**, 2312 (1998).

[12] M. Rohlfing and S. G. Louie, Electron-hole excitations and optical spectra from first principles, Physical Review B - Condensed Matter and Materials Physics **62**, 4927 (2000), 0406203v3 [arXiv:cond-mat].

[13] D. Y. Qiu, F. H. Da Jornada, and S. G. Louie, Optical spectrum of MoS2: Many-body effects and diversity of exciton states, Physical Review Letters **111**, 1 (2013), 1311.0963.

[14] D. Y. Qiu, F. H. Da Jornada, and S. G. Louie, Screening and many-body effects in two-dimensional crystals: Monolayer MoS2, Physical Review B **93**, 1 (2016).

[15] F. H. Da Jornada, D. Y. Qiu, and S. G. Louie, Nonuniform sampling schemes of the Brillouin zone for many-electron perturbation-theory calculations in reduced dimensionality, Physical Review B **95**, 1 (2017), 1610.06641.

[16] F. Fuchs, C. Rödl, A. Schleife, and F. Bechstedt, Efficient O (N2) approach to solve the Bethe-Salpeter equation for excitonic bound states, Physical Review B - Condensed Matter and Materials Physics **78**, 1 (2008).

[17] J. Sun, J. Yang, and C. A. Ullrich, Low-cost alternatives to the Bethe-Salpeter equation: Towards simple hybrid functionals for excitonic effects in solids, Physical Review Research **2**, 1 (2020).

[18] J. Deslippe, G. Samsonidze, D. A. Strubbe, M. Jain, M. L. Cohen, and S. G. Louie, BerkeleyGW: A massively parallel computer package for the calculation of the quasiparticle and optical properties of materials and nanostructures, Computer Physics Communications **183**, 1269 (2012), 1111.4429.

[19] G. H. Wannier, The structure of electronic excitation levels in insulating crystals, Physical Review **52**, 191 (1937).

[20] N. F. Mott, Conduction in polar crystals. II. The conduction band and ultra-violet absorption of alkali-halide crystals, Trans. Faraday Soc. **34**, 500 (1938).

[21] D. Sangalli, A. Ferretti, H. Miranda, C. Attaccalite, I. Marri, E. Cannuccia, P. Melo, M. Marsili, F. Paleari, A. Marrazzo, G. Prandini, P. Bonfà, M. O. Atambo, F. Affinito, M. Palummo, A. Molina-Sánchez, C. Hogan,

M. Grüning, D. Varsano, and A. Marini, Many-body perturbation theory calculations using the yambo code, Journal of Physics Condensed Matter **31**, 10.1088/1361-648X/ab15d0 (2019), 1902.03837.

[22] T. Rangel, M. Del Ben, D. Varsano, G. Antonius, F. Bruneval, F. H. da Jornada, M. J. van Setten, O. K. Orhan, D. D. O'Regan, A. Canning, A. Ferretti, A. Marini, G. M. Rignanese, J. Deslippe, S. G. Louie, and J. B. Neaton, Reproducibility in G0W0 calculations for solids, Computer Physics Communications **255**, 107242 (2020), 1903.06865.

[23] Y. Chen, S. G. Motti, R. D. Oliver, A. D. Wright, H. J. Snaith, M. B. Johnston, L. M. Herz, and M. R. Filip, Optoelectronic Properties of Mixed Iodide-Bromide Perovskites from First-Principles Computational Modeling and Experiment, Journal of Physical Chemistry Letters , 4184 (2022).

[24] M. Dvorak, S. H. Wei, and Z. Wu, Origin of the variation of exciton binding energy in semiconductors, Physical Review Letters **110**, 1 (2013).

[25] A. Champagne, J. B. Haber, S. Pokawanvit, D. Y. Qiu, S. Biswas, H. A. Atwater, F. H. da Jornada, and J. B. Neaton, Quasiparticle and optical properties of carrier-doped monolayer MoTe2 from first principles., Submitted (Nano Letters) 10.1021/acs.nanolett.3c00386 (2023).

[26] D. M. Ceperley and B. J. Alder, Ground state of the electron gas by a stochastic method, Physical Review Letters **45**, 566 (1980).

[27] Y. Gillet, M. Giantomassi, and X. Gonze, Efficient on-the-fly interpolation technique for Bethe-Salpeter calculations of optical spectra, Computer Physics Communications **203**, 83 (2016), 1602.01863.

[28] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, Commentary: The materials project: A materials genome approach to accelerating materials innovation, APL Materials **1**, 10.1063/1.4812323 (2013).

[29] Y. Zhang, J. Sun, J. P. Perdew, and X. Wu, Comparative first-principles studies of prototypical ferroelectric materials by LDA, GGA, and SCAN meta-GGA, Physical Review B **96**, 1 (2017).

[30] R. A. Leute, M. Feneberg, R. Sauer, K. Thonke, S. B. Thapa, F. Scholz, Y. Taniyasu, and M. Kasu, Photoluminescence of highly excited AlN: Biexcitons and exciton-exciton scattering, Applied Physics Letters **95**, 10.1063/1.3186044 (2009).

[31] J. Li, K. B. Nam, M. L. Nakarmi, J. Y. Lin, H. X. Jiang, P. Carrier, and S. H. Wei, Band structure and fundamental optical transitions in wurtzite AlN, Applied Physics Letters **83**, 5163 (2003).

[32] M. A. Jakobson, V. D. Kagan, R. P. Seisyan, and E. V. Goncharova, Optical properties of "pure" CdS and metal-insulator-semiconductor structures on CdS at electrical operation, Journal of Crystal Growth **138**, 225 (1994).

[33] J. Voigt, F. Spiegelberg, and M. Senoner, Band parameters of CdS and CdSe single crystals determined from optical exciton spectra, Physica Status Solidi (B) **91**, 189 (1979).

[34] J. F. Muth, J. H. Lee, I. K. Shmagin, R. M. Kolbas, H. C. Casey, B. P. Keller, U. K. Mishra, and S. P. DenBaars, Absorption coefficient, energy gap, exciton binding energy, and recombination lifetime of GaN obtained from transmission measurements, Applied Physics Letters **71**, 2572 (1997).

[35] K. Reimann, M. Steube, D. Fröhlich, and S. J. Clarke, Exciton binding energies and band gaps in GaN bulk crystals, Journal of Crystal Growth **189-190**, 652 (1998).

[36] W. C. Walker, D. M. Roessler, and E. Loh, Phonon-induced splitting of exciton lines in MgO and BeO, Physical Review Letters **20**, 847 (1968).

[37] R. C. Whited, C. J. Flaten, and W. C. Walker, Exciton thermoreflectance of MgO and CaO, Solid State Communications **13**, 1903 (1973).

[38] M. A. Green, Improved value for the silicon free exciton binding energy, AIP Advances **3**, 10.1063/1.4828730 (2013).

[39] A. Miglio, R. Saniz, D. Waroquiers, M. Stankovski, M. Giantomassi, G. Hautier, G. M. Rignanese, and X. Gonze, Computed electronic and optical properties of SnO2 under compressive stress, Optical Materials **38**, 161 (2014).

[40] K. Reimann and M. Steube, Experimental determination of the electronic band structure of SnO2, Solid State Communications **105**, 649 (1998).

[41] C. Schweitzer, K. Reimann, and M. Steube, Two-photon spectroscopy of SnO2 under hydrostatic pressure, Solid State Communications **110**, 697 (1999).

[42] S. E. Reyes-Lillo, T. Rangel, F. Bruneval, and J. B. Neaton, Effects of quantum confinement on excited state properties of SrTiO3 from ab initio many-body perturbation theory, Physical Review B **94**, 1 (2016), 1605.01818.

[43] D. C. Reynolds, C. W. Litton, D. C. Look, J. E. Hoelscher, B. Claflin, T. C. Collins, J. Nause, and B. Nemeth, High-quality, melt-grown ZnO single crystals, Journal of Applied Physics **95**, 4802 (2004).

[44] A. Mang, K. Reimann, and S. Rübenacke, Band gaps, crystal-field splitting, spin-orbit coupling, and exciton binding energies in ZnO under hydrostatic pressure, Solid State Communications **94**, 251 (1995).

[45] A. M. Alvertis, R. Pandya, L. A. Muscarella, N. Sawhney, M. Nguyen, B. Ehrler, A. Rao, R. H. Friend, A. W. Chin, and B. Monserrat, Impact of exciton delocalization on exciton-vibration interactions in organic semiconductors, Physical Review B - Condensed Matter and Materials Physics **102**, 081122(R) (2020), 2006.03604.

[46] A. M. Alvertis, J. B. Haber, E. A. Engel, S. Sharifzadeh, and J. B. Neaton, Phonon-Induced Localization of Excitons in Molecular Crystals from First Principles, Physical Review Letters **130**, 86401 (2023), 2301.11944.

[47] R.-I. Biega, Y. Chen, M. R. Filip, and L. Leppert, Chemical Mapping of Excitons in Halide Double Perovskites, Nano Letters 10.1021/acs.nanolett.3c02285 (2023), 2306.11352.

[48] J. P. Perdew, K. Burke, and M. Ernzerhof, Generalized Gradient Approximation Made Simple, Physical Review Letters **77**, 3865 (1996).

[49] R. W. Godby and R. J. Needs, Metal-insulator transition in Kohn-Sham theory and quasiparticle theory, Physical Review Letters **62**, 1169 (1989).

[50] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, A. Dal Corso, S. Fabris, G. Fratesi, S. de Gironcoli, R. Gebauer, U. Gerstmann, C. Gougoussis, A. Kokalj, M. Lazzeri, L. Martin-Samos, N. Marzari, F. Mauri, R. Mazzarello, S. Paolini, A. Pasquarello, L. Paulatto, C. Sbraccia, S. Scandolo, G. Sclauzero, A. P. Seitsonen, A. Smogunov, P. Umari, and R. M. Wentzcovitch, QUANTUM ESPRESSO: a modular and open-

source software project for quantum simulations of materials, Journal of Physics: Condensed Matter **21**, 395502 (2009).

[51] R. O. Jones and O. Gunnarsson, The density functional formalism, its applications and prospects, Reviews of Modern Physics **61**, 689 (1989).

[52] A. Oschlies, R. W. Godby, and R. J. Needs, GW self-energy calculations of carrier-induced band-gap narrowing in n-type silicon, Physical Review B **51**, 1527 (1995).

[53] L. X. Benedict, E. L. Shirley, and R. B. Bohn, Theory of optical absorption in diamond, Si, Ge, and GaAs, Physical Review B - Condensed Matter and Materials Physics **57**, R9385 (1998).

[54] D. Rocca, Y. Ping, R. Gebauer, and G. Galli, Solution of the Bethe-Salpeter equation without empty electronic states: Application to the absorption spectra of bulk systems, Physical Review B - Condensed Matter and Materials Physics **85**, 1 (2012).