Nonuniform sampling schemes of the Brillouin zone for many-electron perturbation-theory calculations in reduced dimensionality

Felipe H. da Jornada, Diana Y. Qiu, and Steven G. Louie

Phys. Rev. B **95**, 035109 — Published 3 January 2017

DOI: 10.1103/PhysRevB.95.035109

Non-uniform sampling schemes of the Brillouin zone for many-electron perturbation-theory calculations in reduced dimensionality

Felipe H. da Jornada,* Diana Y. Qiu,* and Steven G. Louie†
Department of Physics, University of California at Berkeley, California 94720 and
Materials Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720
(Dated: December 12, 2016)

First principles calculations based on many-electron perturbation theory methods, such as the *ab initio* GW and GW plus Bethe-Salpeter equation (GW-BSE) approach, are reliable ways to predict quasiparticle and optical properties of materials, respectively. However, these methods involve more care in treating the electron-electron interaction and are considerably more computationally demanding when applied to systems with reduced dimensionality, since the electronic confinement leads a slower convergence of sums over the Brillouin zone due to a much more complicated screening environment that manifests in the "head" and "neck" elements of the dielectric matrix. Here, we present two new schemes to sample the Brillouin zone for GW and GW-BSE calculations: the non-uniform neck subsampling method and the clustered sampling interpolation method, which can respectively be used for a family of single-particle problems, such as GW calculations, and for problems involving the scattering of two-particle states, such as when solving the BSE. We tested these methods on several few-layer semiconductors and graphene and show that they perform a much more efficient sampling of the Brillouin zone and yield two to three orders of magnitude reduction in the computer time. These two methods can be readily incorporated into several *ab initio* packages that compute electronic and optical properties through the GW and GW-BSE approaches.

PACS numbers: 73.22.-f, 71.35.-y, 78.67.-n

I. INTRODUCTION

Many-electron perturbation theory methods, especially those based on density-functional theory (DFT) as the starting mean field, are becoming increasingly popular for predicting electronic excited-state properties of novel materials. Some of the most commonly used methods of this family include: the *ab initio* GW approximation, which allows for the computation of quasiparticle (QP) properties of materials[1,2]; the GW plus Bethe-Salpeter equation (GW-BSE) method, which accesses correlated two-particle states such as excitons[3,4]; and the adiabatic-connection fluctuation-dissipation theorem (ACFDT) methods, which allow for accurate computation of the total energy of materials[5–9], among others. These methods are now available in a variety of mature and optimized computer packages[10–13] and have been applied with success to predict electronic and optical properties of a variety of different systems, ranging from bulk 3D semiconductors to systems with reduced dimensionality, such as molecules, graphene, carbon nanotubes, and nanoribbons.

More recently, there has been interest in applying this family of methods to quasi-two-dimensional (quasi-2D) semiconducting materials, which was motivated by the experimental isolation of monolayer transition metal dichalcogenides (TMDs) such as MoS_2 and $MoSe_2$. However, even though the conceptual approximations employed on conventional 3D systems still hold for quasi-2D materials, it has been notoriously harder to perform *ab initio* many-electron perturbation theory calculations on these monolayer TMDs. For example, while one can typically converge GW QP energies on bulk Si with a $4\times4\times4$ k grid, one needs a much finer k grid of $24\times24\times1$ to converge the quasiparticle gap of monolayer MoS_2[14–18]. This is unexpected at first, since: (1) the ground state properties of monolayer MoS_2 calculated with DFT within the local density approximation (LDA) converge on a much coarser $\sim 6\times6\times1$ k grid; and (2) monolayer MoS_2 has a larger bandgap than Si, so one might naively expect that a coarser k grid is enough to converge the electronic properties of monolayer MoS_2.

The difficulty in converging the electronic properties of quasi-2D semiconductors with k-point sampling is an indirect manifestation of unusual features in electron-electron interactions in these systems. In a plane wave basis set, these features are encoded in the dielectric matrix $\varepsilon_{\mathbf{G},\mathbf{G}'}(\mathbf{q})$, which displays a strong, sharply-peaked feature in its $\mathbf{q}$ dispersion in the long wavelength limit not found in typical bulk semiconductors[16–21]. These features in the dielectric screening manifest in a small portion of the Brillouin zone when performing many-electron perturbation theory calculations, and give rise to the very slow convergence with respect to the number of q-vectors included when computing the GW quasiparticle self energy of systems with reduced dimensionality.

In this paper, we address this issue of slow convergence of many-electron perturbation theory calculations with q-point sampling. We introduce two new methods here, the non-uniform neck subsampling (NNS) method, which provides an efficient way to sample the Brillouin zone and capture features of the dielectric matrix due the electronic confinement, which can be readily used in GW and ACFDT calculations; and the clustered sampling interpolation (CSI) method, which is an approximation to efficiently compute matrix elements which arise in

two-body problems, such as in the context of solving the Bethe-Salpeter equation. Specifically for the case of calculating the self-energy and excitonic effects on mono- or few-layer transition metal dichalcogenides, we show that these methods perform a much more efficient sampling of the Brillouin zone and yields orders of magnitude reduction in the computer run time. Moreover, our methods do not assume any analytical form of the dielectric screening[18,19], and can be equally applied to 1D and 2D semiconducting and metallic systems. These two methods can be readily incorporated into several *ab initio* packages that compute electronic and optical properties through many-electron perturbation theory methods.

Our paper is organized as follows: in Section II, we introduce the non-uniform neck subsampling (NNS) method to efficiently calculate sums in the Brillouin zone involving the screened Coulomb interaction. Our main results are in Eqs. 10 and 11, and the NNS is summarized graphically in Fig. 2. We give example of the NNS method by performing calculations on bilayer $MoSe_2$ and graphene. In Section III, we develop the cluster sampling interpolation (CSI) method. The main result of this part is Eq. 24, and the speed up due to the method is presented in Figs. 8 and 9. We conclude in Section IV by summarizing our results.

II. NON-UNIFORM NECK SUBSAMPLING (NNS) METHOD

In this section, we introduce a method to perform non-uniform sampling of the Brillouin zone. Our goal is to efficiently evaluate sums that are common in many-electron perturbation theory calculations with plane-wave basis sets, which involve the screened Coulomb interaction matrix $W_{\mathbf{GG'}}(\mathbf{q},\omega)$. In general, we will be interested in evaluating sums over the Brillouin zone with the form

$$I_{\mathbf{GG'}}(\omega) = \sum_{\mathbf{q}} B_{\mathbf{GG'}}(\mathbf{q},\omega)\, W_{\mathbf{GG'}}(\mathbf{q},\omega)\,, \tag{1}$$

where $\mathbf{q}$ is a transferred momentum, or q-vector, typically defined on a uniform, Γ-centered Monkhorst-Pack grid, $\mathbf{G}$ and $\mathbf{G'}$ are reciprocal lattice vectors, $B_{\mathbf{GG'}}$ is a smooth function, and the screened Coulomb interaction is $W_{\mathbf{GG'}}(\mathbf{q},\omega) = \varepsilon^{-1}_{\mathbf{GG'}}(\mathbf{q},\omega)\, v(\mathbf{q}+\mathbf{G'})$, where $\varepsilon^{-1}(\mathbf{q})$ is the dielectric matrix and $v(\mathbf{q})$ is the bare Coulomb interaction.

We will be particularly interested in evaluating sums related to the electronic self energy, such as the screened-exchange contribution to the GW self energy,

$$\begin{aligned}\Sigma^{\text{sx}}_{n\mathbf{k}}(\omega) &= -\sum_{v\mathbf{GG'}}\left[\sum_{\mathbf{q}} B^{\text{sx}}(\mathbf{q})\, W_{\mathbf{GG'}}(\mathbf{q},\omega - E_{v\mathbf{k}-\mathbf{q}})\right]\\ B^{\text{sx}}(\mathbf{q}) &= \langle u_{n\mathbf{k}}|e^{i\mathbf{G}\cdot\mathbf{r}}|u_{v\mathbf{k}-\mathbf{q}}\rangle\,\langle u_{v\mathbf{k}-\mathbf{q}}|e^{-i\mathbf{G'}\cdot\mathbf{r}}|u_{n\mathbf{k}}\rangle\,,\end{aligned} \tag{2}$$

where v denotes an occupied band and the indices $n, \mathbf{k}, \mathbf{G}$, and $\mathbf{G'}$ are implicit in $B^{\text{sx}}(\mathbf{q})$.

Eq. 1 cannot be applied directly on semiconductors because of the divergence of the Coulomb interaction at $\mathbf{q}$=0. While several treatments have been proposed to deal with this divergence[22,23], we restrict our discussion to one particular stochastic method which is well-suited for systems with reduced dimensionality. We start by taking the continuous limit and then re-discretizing Eq. 1, assuming the matrix elements $B(\mathbf{q})$ to be smooth. This yields

$$I_{\mathbf{GG'}}(\omega) = \sum_{\mathbf{q}} B_{\mathbf{GG'}}(\mathbf{q},\omega)\,\overline{W}_{\mathbf{GG'}}(\mathbf{q},\omega)\,, \tag{3}$$

$$\overline{W}_{\mathbf{GG'}}(\mathbf{q},\omega) = \frac{1}{V_{\mathbf{q}}}\int_{\mathcal{C}_{\mathbf{q}}} \mathrm{d}^D q'\; W_{\mathbf{GG'}}(\mathbf{q}',\omega)\,, \tag{4}$$

where D denotes the number of dimensions on the system. Each integral is performed over the Voronoi cell that surrounds each $\mathbf{q}$ vector, denoted by $\mathcal{C}_{\mathbf{q}}$[24].We denote the volume (or area/length for 2D/1D systems) of $\mathcal{C}_{\mathbf{q}}$ by $V_{\mathbf{q}}$.

To evaluate the integral in Eq. 4, it is possible to employ the exact analytic behavior of $\varepsilon^{-1}_{\mathbf{GG'}}(\mathbf{q}{\to}0)$ and use a Monte Carlo average scheme to more efficiently sample Brillouin zones with arbitrary shapes. This stochastic approach is used in a few GW packages[10,11]. For example, for bulk 3D semiconductors, $\varepsilon^{-1}_{\mathbf{GG'}}(\mathbf{q}{\to}0)$ is smooth and approaches a constant as $\mathbf{q}{\to}0$, so the integral in Eq. 4 can be evaluated for 3D semiconductors as

$$\begin{aligned}\overline{W}^{\text{MC}}_{\mathbf{GG'}}(\mathbf{q}{\neq}0,\omega) &= \varepsilon^{-1}_{\mathbf{GG'}}(\mathbf{q},\omega)\; v^{\text{MC}}(\mathbf{q},\mathbf{G'}),\\ \overline{W}^{\text{MC}}_{\mathbf{GG'}}(\mathbf{q}{=}0,\omega) &= \varepsilon^{-1}_{\mathbf{GG'}}(\mathbf{q}_0,\omega)\; v^{\text{MC}}(0,\mathbf{G'}),\\ v^{\text{MC}}(\mathbf{q},\mathbf{G}) &\coloneqq \frac{1}{N_{\text{MC}}}\sum_{\mathbf{q}_{\text{MC}}\in\mathcal{C}_{\mathbf{q}}} v(\mathbf{q}_{\text{MC}}+\mathbf{G}),\end{aligned} \tag{5}$$

where a small but finite vector $\mathbf{q}_0$ is employed to compute the long wavelength limit of the dielectric matrix, and a value of $N_{\text{MC}} \sim 10^6$ Monte Carlo samples $\{\mathbf{q}_{\text{MC}} \in \mathcal{C}_{\mathbf{q}}\}$ is typically enough to converge the sum to within a few meVs.

To extend the above evaluation to systems with reduced dimensionality, one needs to resort to a supercell approach, where a large vacuum region is included in the confined direction to separate periodic images. In these cases, GW calculations typically converge very slowly with the supercell size due to the long-range nature of the Coulomb interaction between replicas of charged excitations on neighboring cells. A common solution to overcome this drawback is to truncate the Coulomb potential along the confined direction to prevent spurious interactions between periodic images. For instance, for a quasi-2D crystal, one can truncate the Coulomb interaction in real space as

$$v_{\text{trunc}}(\mathbf{r}) = \frac{e^2}{|\mathbf{r}|}\theta(L_z/2 - z), \tag{6}$$

where L_z is the length of the supercell along the confined direction[19]. Such a potential has the following form in

reciprocal space,

$$v_{\rm trunc}(\mathbf{q},\mathbf{G}) = \frac{4\pi e^2}{|\mathbf{q}+\mathbf{G}|^2}\left[1 - e^{-qL_z/2}\cos\frac{G_z L_z}{2}\right], \quad (7)$$

where the $\mathbf{q}$ vectors are only sampled in the extended 2D plane, but all the reciprocal lattice vectors, $\mathbf{G}$, including those with components along confined directions, are sampled. With such a truncated Coulomb potential, converged calculations typically depend very weakly on the length of the supercell in the confined direction[17,19].

Once we truncate the Coulomb potential[19], Eqs. 3 and 5 can also be applied to quasi-2D semiconductors. In fact, it is possible to converge the *absolute* Fock exchange energy on bilayer $MoSe_2$ to within 70 meV on a $6\times6\times1$ q grid, which shows that the matrix elements $B^{\rm sx}(\mathbf{q})$ in Eq. 2 are smooth functions even for systems with reduced dimensionality, and that Monte Carlo sampling methods can effectively capture the fast variations in the Coulomb interaction. However, these stochastic methods, as usually employed, become much less efficient to evaluate the total GW self energy for quasi-2D semiconductors. Still for the case of bilayer $MoSe_2$, a $24\times24\times1$ q grid is necessary to converge the GW self energy to within 50 meV, even if we use a more sophisticated analytic expression for the inverse dielectric matrix, $\varepsilon^{-1}_{00}(\mathbf{q})$.

The slow convergence of Eq. 3 on systems with reduced dimensionality is a sign that the analytic models typically employed for the dielectric matrix are no longer accurate in the range of $\mathbf{q}$- and $\mathbf{G}$-vectors we are interested in. This is mainly due to two factors: first, the dielectric matrices of these systems have many features as a function of $\mathbf{q}$ which are hard to model analytically[16,17,20,21]; and second, these Monte Carlo averages should be performed not only for the head element ($\mathbf{G}$=$\mathbf{G}'$=0), but also for a series of reciprocal lattice vectors $\mathbf{G}_\perp, \mathbf{G}'_\perp$ in the confined direction (e.g., along the direction of the normal vector for a 2D material), which we denote to as the *neck* elements of the matrix.

The physical motivation for focusing on these neck matrix elements of the dielectric matrix, $\varepsilon_{\mathbf{G}_\perp\mathbf{G}'_\perp}(\mathbf{q})$, is that the $\mathbf{G}_\perp$ vectors in the confined direction become continuous as the simulation supercell grows in the confined direction, and so they become almost as important as the $\mathbf{G}$=0 vector. For example, in our calculation on bilayer $MoSe_2$, the magnitude of the smallest, nonzero reciprocal lattice vector $\mathbf{G}_\perp$ corresponding to the out-of-plane direction is 5% of that of the in-plane, primitive reciprocal lattice vector. Consequently, not only will the head element of the screened Coulomb potential $W_{\mathbf{00}}(\mathbf{q}) = \varepsilon^{-1}_{00}(\mathbf{q})v(\mathbf{q})$ be large, but a series of neck elements $W_{\mathbf{G}_\perp,\mathbf{G}'_\perp}(\mathbf{q})$ will also be large, as long as $|\mathbf{G}_\perp|$ and $|\mathbf{G}'_\perp|$ are smaller than, or of the same order of magnitude as the q-vectors in $\mathcal{C}_{\mathbf{q}=\mathbf{0}}$.

We illustrate the sharp features in the inverse dielectric matrix of bilayer $MoSe_2$ by plotting in Fig. 1 some selected components of $1/(\varepsilon_{\mathbf{G}_\perp,\mathbf{G}_\perp})$ and $\varepsilon^{-1}_{\mathbf{G}_\perp,\mathbf{G}_\perp}$, with $\mathbf{G}_\perp = (0,0,G_z)$. It becomes evident that the inverse dielectric matrix has completely different $\mathbf{q}\to0$ behavior

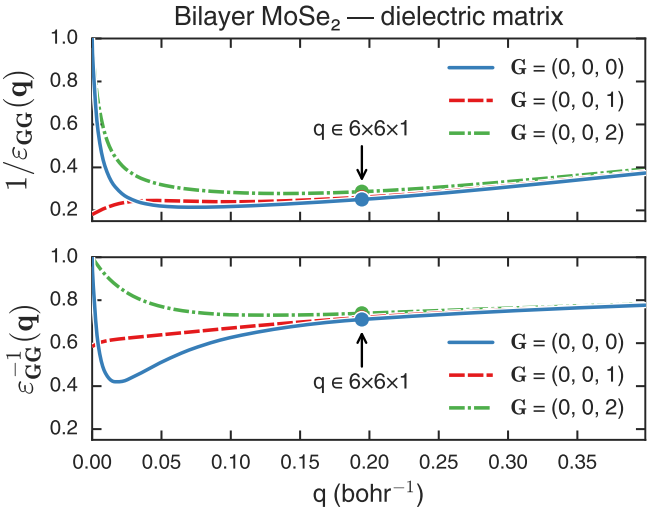

FIG. 1. (Color online) Selected matrix elements of the dielectric matrix of bilayer $MoSe_2$: $1/\varepsilon^{-1}_{\mathbf{GG}}(\mathbf{q})$ (top panel) and $\varepsilon^{-1}_{\mathbf{GG}}(\mathbf{q})$ (bottom panel). The difference between the two curves is related to local field effects. Dots represent calculations performed with smallest non-zero $\mathbf{q}$ point from an uniform $6\times6\times1$ q grid.

depending on whether G_z is odd or even. More importantly, even for a simple system such as bilayer $MoSe_2$, local fields play a very important role, as different $\mathbf{G}_\perp$ components of the inverse dielectric function $\varepsilon^{-1}_{\mathbf{G}_\perp,\mathbf{G}_\perp}$ disperse in a qualitatively different way from the matrix elements $1/(\varepsilon_{\mathbf{G}_\perp,\mathbf{G}_\perp})$. Because these differences are not uniform among different $\mathbf{G}_\perp$ components, this shows that local fields mix different components of the neck of the inverse dielectric function in a non-obvious way, in the process of inverting the dielectric matrix.

Since local fields are important in systems with reduced dimensionality, an accurate analytical model for the neck of the inverse dielectric matrix $\varepsilon^{-1}_{\mathbf{G},\mathbf{G}'}$ requires us to cary out the inversion of the dielectric matrix and explicitly include a series of off-diagonal matrix elements. So, while it is important to capture the q-dispersion of $\varepsilon^{-1}_{\mathbf{G}_\perp\mathbf{G}'_\perp}(\mathbf{q})$ in order to evaluate the sums in the Brillouin zone, it seems unlikely that there is a compact and reliable analytic expression for ε^{-1} for the range of the $\mathbf{G}_\perp$- and q-vectors we are interested in, especially one that is valid for a wide range of complex materials.

Even with no analytical expression for $\varepsilon^{-1}_{\mathbf{G}_\perp\mathbf{G}'_\perp}(\mathbf{q})$, we can still speedup the convergence of the sum in Eq. 4 dramatically if we sample more efficiently the inverse dielectric matrix in the region where both the Coulomb interaction is larger and where $\varepsilon^{-1}(\mathbf{q})$ varies the most in $\mathcal{C}_{\mathbf{0}}$. We propose to explicitly capture these variations by breaking up the integral in Eq. 4 into one radial and one angular part, where the radial part is divided into N_s annuli, each one having a thickness Δ_s. We also approximate the radial integral with a discrete sum over N_s points q_s, which we refer to as subsampling points, and

write

$$\overline{W}^{\rm sub}_{\mathbf{G}_\perp\mathbf{G}'_\perp}(\mathbf{q}{=}0,\omega) \equiv \sum_{s=1}^{N_s} w_s\, \varepsilon^{-1}_{\mathbf{G}_\perp\mathbf{G}'_\perp}(\mathbf{q}_s\,\omega)\, v(\mathbf{q}_s+\mathbf{G}'_\perp), \tag{8}$$

$$w_s \approx \frac{1}{N_{\rm MC}} \sum_{\mathbf{q}_{\rm MC}} \theta(|\mathbf{q}_{\rm MC}|-a_s)\theta(a_{s+1}-|\mathbf{q}_{\rm MC}|), \tag{9}$$

where is w_s the weight associated with each subsampling point, and a_s is just a shorthand for the inner radius of each annulus, i.e., $a_s \equiv \sum_{i=1}^{s-1}\Delta_s$.

While the approximation in Eq. 8 works best with isotropic materials, we stress that most of the variation of $\varepsilon^{-1}(\mathbf{q})$ only depends on $|\mathbf{q}|$, and we can always choose a direction for each subsampling point q_s that yields the same value for the inverse dielectric function as the angle-averaged inverse dielectric function (at least for one particular pair of G-vectors).

For a given N_s number of subsampled points, we have the freedom to define two quantities: the N_s subsampling points q_s where the dielectric matrix has to be explicitly computed, and the N_s annulus thicknesses Δ_s. As we derive in the Appendix from simple assumptions of the qualitative behavior of the inverse dielectric matrix, the optimal position of the subsampling point is halfway between between the inner and outer radius of each annulus, $q_s = a_s + \frac{\Delta_s}{2}$, where $a_{s+1} = a_s + \Delta_s$. On the other hand, while the choice of optimal thicknesses is system dependent, a practical solution is to use a polynomial sampling of degree p, $\Delta_s = \Delta_1 \times s^p$ (which corresponds to polynomial sampling of degree $p+1$ for the subsampling points). We tested different samplings by calculating the quasiparticle bandgap of bilayer $MoSe_2$ with a set of q-vectors defined on a regular $6\times6\times1$ Monkhorst-Pack grid plus a set of $N_s = 10$ subsampling $\mathbf{q}$ points. We tested thicknesses generated with a polynomial of degree $p = 0$, $p = 1$ and $p = 2$ and found little difference in the resulting energies, with the *absolute* quasiparticle energies differing by less than 10 meV for states near the Fermi energy between calculations generated with $p = 1$ and $p = 2$, and by less than 20 meV between calculations generated with $p = 0$ and $p = 1$. However, subsampling points generated with a quadratic grid ($p = 1$) capture the dip feature in Fig. 1 better than subsampling points generated with a linear grid ($p = 0$), so, for simplicity, we use $p = 1$ from here on when performing subsequent calculations.

In principle, the extra cost associated with the sampling technique would be the ratio of the number of subsampling points to the number of q-vectors on the regular grid of q-vectors. However, since the fast variations in the dielectric matrix are confined to a small number of $\mathbf{G}_\perp$-vectors, we only need to calculate the dielectric matrices for the subsampling contribution in Eq. 8 for a small number of neck $\mathbf{G}_\perp$-vectors. We choose these vectors on the condition that $|\mathbf{G}|^2 \le E^{\rm sub}_{\rm cut} \equiv |\mathbf{G}^{\rm min}_\parallel|^2$, where $\mathbf{G}^{\rm min}_\parallel$ is the smallest reciprocal lattice vector in a periodic direction. The cutoff $E^{\rm sub}_{\rm cut}$ used for the subsampling points is typically much smaller than the cutoff $E_{\rm cut}$ needed for the full dielectric matrix in GW calculations, so the extra cost associated with the NNS scheme is small. For instance, on bilayer $MoSe_2$, the number of G-vectors up to $E_{\rm cut} = 35$ Ry and $E^{\rm sub}_{\rm cut} \approx 1.36$ Ry is 11667 and 37, respectively. We arrive at the final expression to evaluate the sum in Eq. 1 within the NNS method,

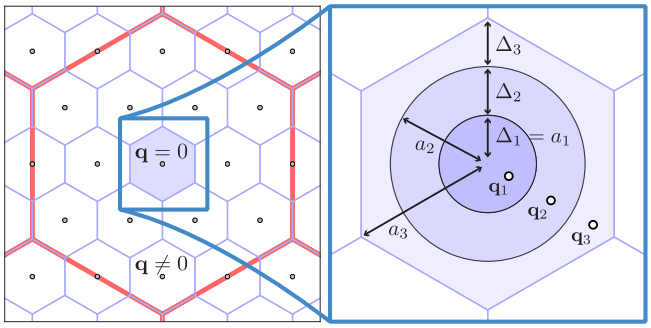

FIG. 2. (Color online) Graphical representation of the NNS scheme. Left panel: $\mathbf{q}$ points involved in the sum in Eq. 10. Each smaller hexagon represents the Voronoi cell $\mathcal{C}_\mathbf{q}$ that encloses each $\mathbf{q}$ point. The thicker red line denotes the Brillouin zone edge, and each dot on the left panel represents one point where we calculate both the matrix elements B and the screened Coulomb interaction W. Right panel: special treatment for $\mathbf{q}$=0 point of Eq. 10. Each dot in this panel represents a $\mathbf{q}$ point where we compute the screened Coulomb interaction. In this example, we use $N_s = 3$ subsampled points.

$$I_{\mathbf{G}\mathbf{G}'}(\omega) = \sum_{\mathbf{q}} B_{\mathbf{G}\mathbf{G}'}(\mathbf{q},\omega)\overline{W}^{\rm sub}_{\mathbf{G}\mathbf{G}'}(\mathbf{q},\omega) \tag{10}$$

$$\overline{W}^{\rm sub}_{\mathbf{G}\mathbf{G}'}(\mathbf{q}{\neq}0,\omega) = W_{\mathbf{G}\mathbf{G}'}(\mathbf{q},\omega)$$
$$\overline{W}^{\rm sub}_{\mathbf{G}\mathbf{G}'}(\mathbf{q}{=}0,\omega) = \begin{cases} W_{\mathbf{G}\mathbf{G}'}(\mathbf{q}_0,\omega) & \text{for } |\mathbf{G}|^2 \text{ and } |\mathbf{G}'|^2 > E^{\rm sub}_{\rm cut} \\ \sum_{s=1}^{N_s} w_s W_{\mathbf{G}\mathbf{G}'}(\mathbf{q}_s,\omega) & \text{otherwise.} \end{cases} \tag{11}$$

where $\mathbf{q}_0$ is an arbitrarily small but non-zero vector. Eqs. 10 and 11, together with the definition of the sub-

sampling weight in Eq. 9 form the basis of the NNS method.

A graphical representation of this discretization procedure is given in Fig. 2 for a quasi-2D system with hexagonal symmetry. In the left panel, we show the $\mathbf{q}$ points involved in the sum in Eq. 10 (a $4\times4\times1$ q grid is used in this example). The thicker red line denotes the Brillouin zone edge, and each dot on the left panel represents one point where we calculate both the matrix elements B and the screened Coulomb interaction W. In the right panel, we show the special treatment for the Voronoi region associated with the $\mathbf{q}$=0 point from Eq. 10. Each dot in the right panel represents a $\mathbf{q}$ point where we compute the screened Coulomb interaction. In this example, we use $N_s = 3$ subsampled points.

When the inverse dielectric matrix is anisotropic, i.e., $\varepsilon^{-1}(\mathbf{q}\to0)$ depends on the direction of $\mathbf{q}$, there is an additional complication when employing either the uniform sampling (Eqs. 5) or the NNS scheme (Eqs. 10 and 11). Still, the angular dependence on the screened Coulomb interaction is typically much less important than the radial dependence, and a simple model can effectively capture most of the anisotropy in $\varepsilon^{-1}(\mathbf{q}\to0)$ without additional computational cost.

For quasi-2D systems and in the long wavelength limit, one can show that the head of the inverse longitudinal dielectric matrix can be expressed as

$$\varepsilon^{-1}_{00}(\mathbf{q}\to0) = 1 - q\ \hat{q}\cdot\underline{\underline{\alpha}}\cdot\hat{q}, \tag{12}$$

where $\underline{\underline{\alpha}}$ is a 2×2 Hermitian tensor. The eigenvectors $\{\mathbf{u}_i\}_i$ of $\underline{\underline{\alpha}}$ give the principal axes of polarization for the head of the inverse longitudinal dielectric function, and it can be determined from either symmetry considerations of the crystal or from explicit calculations of $\varepsilon^{-1}_{00}(\mathbf{q}_0)$ along 3 different directions of $\mathbf{q}_0$.

Our goal is to find an optimal direction $\hat{q}_0$ for the subsampled $\mathbf{q}$ vectors such that, for $|\mathbf{q}'| = |\mathbf{q}_0| \to 0$,

$$\frac{1}{2\pi}\int \mathrm{d}\theta'\ \varepsilon^{-1}_{00}(\mathbf{q}') = \varepsilon^{-1}_{00}(\mathbf{q}_0), \tag{13}$$

which results in a vector $\mathbf{q}_0$ that is parallel to the average of the eigenvectors of $\underline{\underline{\alpha}}$, $\hat{q}_0 = \frac{1}{2}(\hat{u}_1 + \hat{u}_2)$.

For the angular average of $\varepsilon^{-1}(\mathbf{q})$ in Eq. 8 to be accurately represented by a single evaluation of the inverse dielectric matrix along an average direction, one should also choose a Voronoi region that is as isotropic as possible. So, one should keep the ratio b_i/N_{k_i} approximately constant for all extended directions i, where each b_i and N_{k_i} is a primitive, reciprocal lattice constant and the number of $\mathbf{k}$ points along each direction, respectively. With this geometrical setup, the NNS scheme can be readily employed on systems with anisotropic screening response in the long wavelength limit as long as the NNS is performed in the direction that averages the screening response. This direction can be obtained by either 3 computationally inexpensive evaluations of the head of the inverse dielectric function along different directions, or by symmetry considerations.

We now turn to applying the NNS method as defined in Eqs. 10 and 11 for some systems of interest. We will first discuss the convergence on semiconducting systems having both isotropic and anisotropic dielectric response – bilayer $MoSe_2$ and monolayer black phosprohous –, and on graphene.

A. NNS method applied to semiconductors

The application of the NNS method is straightforward for systems with isotropic in-plane dielectric response (i.e., such that $\varepsilon^{-1}_{\mathbf{G}\,\mathbf{G}'}(\mathbf{q}\to0)$ does not depend on the direction of $\mathbf{q}$). In Fig. 3 (a), we compare the convergence of the quasiparticle gap of bilayer $MoSe_2$ with the number of $\mathbf{q}$ vectors on a regular grid, using both a conventional sampling of the Brillouin zone with the Monte Carlo sampling method and the new NNS method. Our calculation was performed with in a supercell arrangement with $L_z = 53\,\text{Å}$. We chose bilayer $MoSe_2$ instead of monolayer $MoSe_2$ because the gap is indirect in the bilayer structure and converges slightly slower with the number of $\mathbf{q}$ vectors. Whereas one would require a sampling of the q-vectors on a $36\times36\times1$ grid to converge the quasiparticle gap of bilayer $MoSe_2$ to within 50 meV using a uniform sampling of the Brillouin zone – and even finer grids to check that the answer is indeed converged – we can achieve a much better convergence by sampling the $\mathbf{q}$ vectors on a $6\times6\times1$ regular grid with an additional set of $N_s = 10$ subsampled $\mathbf{q}$ points, which effectively samples features that would only be captured on a regular $1143\times1143\times1$ q grid. We also show the error[25] of the ionization potential (IP) of bilayer $MoSe_2$ as a function of the q grid in Fig. 3 (b). The convergence of the IP with q grid is very similar to the convergence of the QP bandband, which is a result of the overall larger screened-exchange contribution to the GW self-energy for valence states compared to conduction states.

We also compare the convergence of QP gap the as a function of the number of subsampling points in Fig. 3 (c), where it is evident that the NNS method converges very fast with the number of subsampling points. Indeed, the quasiparticle gap of bilayer $MoSe_2$ changes by just 3 meV if we vary N_s from 8 to 15. The extra cost associated with the NNS scheme is also very small in this system, as shown in Table I. Therefore, the NNS method allows one to converge the quasiparticle gap of bilayer $MoSe_2$ in an efficient way, providing savings of about 2 orders of magnitude in the CPU time compared to a traditional uniform sampling of the Brilluoin zone.

Next, we illustrate the convergence for materials with anisotropic dielectric response by studying monolayer black phosphorous, which is another prototypical quasi-2D semiconductor which exhibits large optical anisotropy, linear optical dichroism, and strong many-body interactions. By the symmetry of the crystal, the

TABLE I. Comparison of the uniform sampling of the Brillouin zone with the non-uniform neck subsampling method for bilayer $MoSe_2$ with $N_s = 10$ subsampling points. We compare the indirect $\Gamma \to \Lambda$ gap, CPU usage, and the number of q points in the irreducible portion of the Brillouin zone. For the NNS method, we report the effective q grid spanned by the smaller subsampling point. The two calculations with denser q grids were performed with a smaller cutoff and extrapolated according to the process described on the text.

q grid	Uniform sampling			Non-uniform neck subsampling (NNS) method			
	$\Gamma \to \Lambda$ (eV)	CPU usage (core–hour)	# of q points in irr. BZ	$\Gamma \to \Lambda$ (eV)	CPU usage (core–hour)	# of q points in irr. BZ	Effective q grid
$6 \times 6 \times 1$	3.31	96	7	1.75	157	16	$1143 \times 1143 \times 1$
$12 \times 12 \times 1$	2.14	930	19	1.76	1373	28	$2286 \times 2286 \times 1$
$24 \times 24 \times 1^*$	1.85	3620	61	1.76	5130	70	$4573 \times 4573 \times 1$
$36 \times 36 \times 1^*$	1.80	12280	127	1.76	15390	136	$6859 \times 6859 \times 1$

principal axes of $1/\varepsilon^{-1}_{00}(\mathbf{q})$, i.e., the eigenvectors of $\underline{\underline{\alpha}}$, have to lie along high-symmetry lines. If we setup the lattice such that (100) and (010) correspond to the armchair and zigzag directions of the the black phosphorous monolayer, respectively, we find that the dielectric response is indeed anisotropic in this material, with the eigenvalues of $\underline{\underline{\alpha}}$ being different along the two directions: $\alpha_{(100)} \approx 52.6$ 1/bohr and $\alpha_{(010)} \approx 72.1$ 1/bohr. The optimal direction to compute the dielectric response is $\hat{q}_0 = \frac{1}{2}[(100) + (010)]$, which does not coincide with the (110) direction because the in-plane lattice constants are not the same.

In Fig. 4, we show the convergence of the GW quasiparticle gap on monolayer black phosphorous as a function of the q grid, computed in a supercell arrangement with $L_z = 20$ Å. Just as in the case of bilayer $MoSe_2$, we observe that the convergence is much faster with the proposed NNS scheme, where we obtain a quasiparticle gap converged within 50 meV employing $\mathbf{q}$ vectors on a grid as coarse as $7 \times 5 \times 1$ with additional $N_s = 10$ subsampled $\mathbf{q}$ points. In addition, we also show that the converge is remarkably fast if we perform the NNS along the optimal direction, $\hat{q}_0 = \frac{1}{2}[(100) + (010)]$. Still, regardless of the direction, the NNS scheme is much more efficient to converge the quasiparticle gap than using a uniform sampling of the Brillouin zone.

B. NNS method applied to quasi-2D metals and quasi-metals

The NNS scheme can also be efficiently applied on systems other than 2D semiconductors. Before we proceed, we must carefully distinguish $\mathbf{k}$ points used to compute the dielectric matrix from the set of transfer momenta $\mathbf{q}$ where we evaluate the dielectric matrix. While the NNS scheme deals with the sampling of q-vectors, we have used so far a uniform k grid when we compute $\varepsilon^{-1}_{\mathbf{G}\mathbf{G}'}(\mathbf{q})$ for each particular $\mathbf{q}$. However, there is a known additional difficulty when calculating the dielectric matrix for metallic systems because the k grid must be fine enough to sample intraband transitions. This problem can be mitigated by also sampling the $\mathbf{k}$ points in the Brillouin zone in a non-uniform fashion. We write the polarizability matrix as

$$\chi^0_{\mathbf{G}\mathbf{G}'}(\mathbf{q},\omega) = \frac{g_s}{V_{\text{cell}}} \sum_{nn'\mathbf{k}} w_{\mathbf{k}} \left[f(E_{n'\mathbf{k}+\mathbf{q}}) - f(E_{n\mathbf{k}})\right] \frac{M^*_{\mathbf{G}}(n',n,\mathbf{k},\mathbf{q}) M_{\mathbf{G}'}(n',n,\mathbf{k},\mathbf{q})}{\omega - (E_{n\mathbf{k}} - E_{n'\mathbf{k}+\mathbf{q}}) + i0^+ \,\text{sgn}(E_{n\mathbf{k}} - E_{n'\mathbf{k}+\mathbf{q}})} \tag{14}$$

$$M_{\mathbf{G}}(n',n,\mathbf{k},\mathbf{q}) = \langle u_{n'\mathbf{k}+\mathbf{q}} | e^{i\mathbf{G}\cdot\mathbf{r}} | u_{n\mathbf{k}} \rangle \,, \tag{15}$$

where V_{cell}, g_s, f, and $w_{\mathbf{k}}$ denote the unit cell volume, the spin degeneracy of the calculation, the Fermi occupation factor, and the weight of each $\mathbf{k}$ point, with $\sum_{\mathbf{k}} w_{\mathbf{w}} = 1$, respectively.

Here, we propose to associate different weights, $w_{\mathbf{k}}$, with each $\mathbf{k}$ point, proportionally to the volume $V_{\mathbf{k}}$ of the Voronoi cell $\mathcal{C}_{\mathbf{k}}$ that surrounds each $\mathbf{k}$ point. The weights $w_{\mathbf{k}}$ can be determined uniquely by the Voronoi tessellation of the Brillouin zone, and we use the Voro++ package in BerkeleyGW[10] to efficiently compute the Voronoi tessellation including periodic boundary conditions. This allows one to use non-uniform k grids to evaluate the sum in Eq. 14 and more efficiently capture complicated regions of the Brillouin zone, such as those associated with intraband transitions[26].

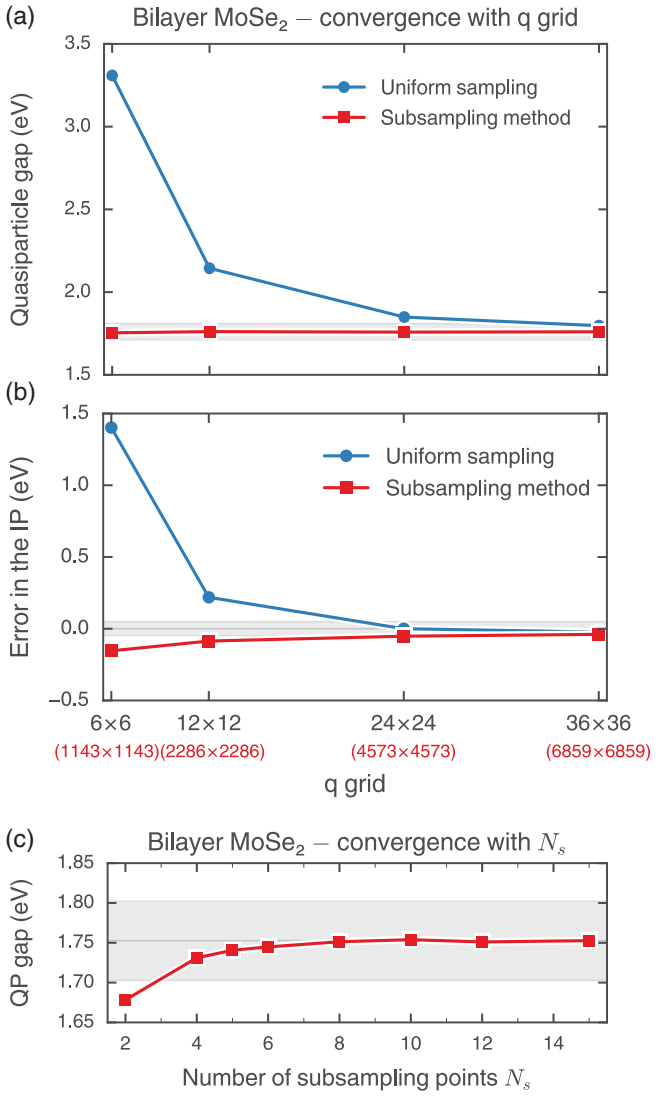

FIG. 3. (Color online) (a) Convergence of the quasiparticle gap of bilayer $MoSe_2$ as function of the q grid using a uniform sampling of the Brillouin zone (Monte Carlo averaging scheme) and the proposed NNS method, with $N_s = 10$. The gray shaded region corresponds to an interval of ± 50 meV compared to our most converged value. (b) Error in the ionization potential (IP) of bilayer $MoSe_2$ as a function of the q grid. The values in parenthesis represent the effective q grid captured by the smallest subsampling point. (c) Convergence of the quasiparticle gap as a function of N_s for a $6 \times 6 \times 1$ q grid.

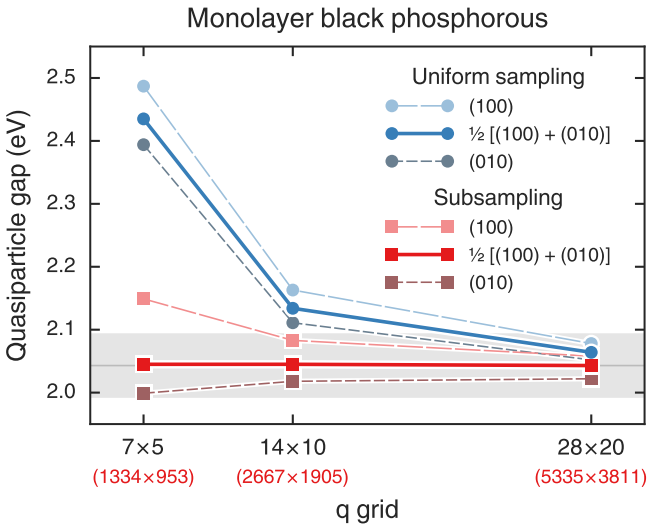

FIG. 4. (Color online) Convergence of the quasiparticle gap of monolayer black phosphorous as function of the q grid using a uniform sampling of the Brillouin zone (Monte Carlo averaging scheme) and the NNS method. For both methods, we compare the converge rate for different directions of the small $\mathbf{q}$ vector(s), where (100) and (010) follow the armchair and zigzag directions, respectively. The gray shaded region corresponds to an interval of ± 50 meV compared to our most converged value, and corresponds to the convergence threshold one is typically interested. The values in parenthesis represent the effective q grid captured by the subsampling points.

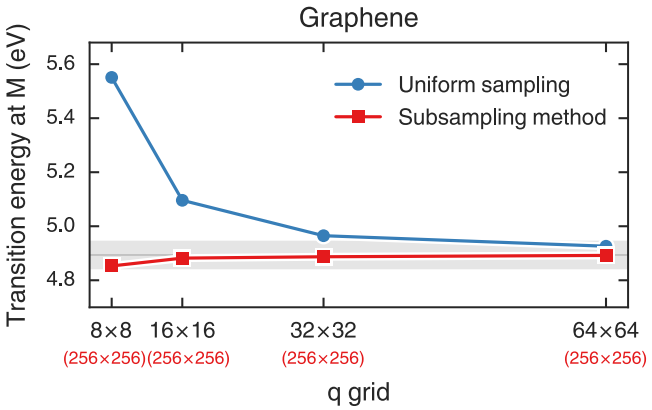

FIG. 5. (Color online) Convergence of the quasiparticle interband transition energy at the M point of the Brillouin zone of graphene as function of the q grid. The two curves and the shaded region are the same as in Fig. 3. The values in parenthesis represent the effective q grid captured by the subsampling points, which was kept fixed in this calculation.

With this new method, we can now employ the NNS scheme on graphene, another prototypical 2D material where the k-point sampling is also complicated. We setup our supercell with a distance $L_z = 17$ Å between repeated graphene layers. We employ the new non-uniform k-point sampling scheme of the dielectric matrix by including $\mathbf{k}$ points from a coarse $8 \times 8 \times 1$ grid if $\mathbf{k}$ is far away from the Dirac points at the K and K' points of the Brillouin zone but include more $\mathbf{k}$ points commensurate with a much finer $512 \times 512 \times 1$ grid near the Dirac points. When we lay out the radial subsampling $\mathbf{q}$ points, we employ a constant thicknesses $\Delta_s = \Delta_1$, as we can then reuse more information when we construct the polarizability matrix for different subsampling $\mathbf{q}_s$ vectors. In Fig. 5, we compare the convergence of the quasiparticle interband transition energy at the M point of graphene

with the number of q-vectors on the regular grid, for the two methods. Once again, the curve obtained with the proposed NNS method converges much faster with the number of q-vectors, and a regular $8 \times 8 \times 1$ grid for the q-vectors is enough to converge the quasiparticle transition energy at the M point to within 50 meV.

Therefore, the NNS method can indeed be applied on a variety of systems with reduced dimensionality and with different screening properties. Although the method was tested here for GW calculations, which show order-of-magnitude speedups in the computer runtime compared to a regular sampling of the Brillouin zone, it can be applied to other kinds of many-electron perturbation theory calculations as well.

III. CLUSTERED SAMPLING INTERPOLATION (CSI) METHOD

In the previous section, we introduced the NNS method to compute sums associated with one-electron integrals, such as those needed to calculate the GW quasiparticle self energy. In this section, we introduce the clustered sampling interpolation (CSI) method to efficiently sample the Brillouin zone for problems involving two-particle correlated states, such as those given by the Bethe-Salpeter equation (BSE). These problems are characterized by associated Hamiltonians, typically written in the occupation representation, with matrix elements describing scattering amplitudes from one two-particle state at a $\mathbf{k}$ point to another state at a different $\mathbf{k}$ point.

When solving the BSE to obtain the optical absorption spectrum, it is a well-known problem that interaction matrix elements need to be constructed on a very fine k grid because excitons are correlated states with wavefunction that has fine structures in k-space. For example, even for bulk semiconductors such as GaAs, very fine uniform grids containing over a million $\mathbf{k}$ points are necessary to resolve the exciton energies and wavefunctions[27]. In many cases, the bottleneck for solving the BSE is in computing the interaction matrix elements, and in the past, schemes based on interpolation between two different k grids, which we refer to as "dual-grid" methods, have been employed to make these calculations feasible[4,27]. We will review here the scheme implemented in the current released version of BerkeleyGW[10], describe its shortcomings when applying it on systems with reduced dimensionality, and propose an extension for the scheme to mitigate these shortcomings.

We are interested in evaluating the two-particle matrix elements that are in the Bethe-Salpeter equation, which is of the form

$$(E_{c\mathbf{k}+\mathbf{Q}} - E_{v\mathbf{k}})A^S_{vc\mathbf{k}} + \sum_{v'c'\mathbf{k}'} K^{\mathrm{eh}}_{vc;v'c'}(\mathbf{k}, \mathbf{q}{=}\mathbf{k}'{-}\mathbf{k})A^S_{v'c'\mathbf{k}'} = \Omega^S A^S_{vc\mathbf{k}}. \quad (16)$$

Here, S indexes the exciton states; $\mathbf{Q}$ is the center-of-mass momentum of the electron-hole pair; $A^S_{vc\mathbf{k}}$ is the amplitude of a free electron-hole pair consisting of an electron in $|c\mathbf{k}+\mathbf{Q}\rangle$ and one missing from $|v\mathbf{k}\rangle$; Ω^S is the exciton excitation energy; $E_{c\mathbf{k}+\mathbf{Q}}$ and $E_{v\mathbf{k}}$ are the quasiparticle energies, and K^{eh} is the electron-hole interaction kernel. The kernel contains contributions from a direct term and an exchange term. The direct term is

$$K^{\mathrm{d}}_{vc;v'c'}(\mathbf{k}, \mathbf{q}{=}\mathbf{k}'{-}\mathbf{k}) = -\sum_{\mathbf{G}\mathbf{G}'} M^*_{\mathbf{G}}(c, c', \mathbf{k}, \mathbf{q}) W_{\mathbf{G},\mathbf{G}'}(\mathbf{q})\, M_{\mathbf{G}'}(v, v', \mathbf{k}, \mathbf{q}), \quad (17)$$

where the matrix elements M are given by Eq. 15, and the exchange term is

$$K^{\mathrm{x}}_{vc;v'c'}(\mathbf{k}, \mathbf{q}{=}\mathbf{k}'{-}\mathbf{k}) = \sum_{\mathbf{G}} M^*_{\mathbf{G}}(c, v, \mathbf{k}, \mathbf{Q}) v_{\mathbf{G}}(\mathbf{Q})\, M_{\mathbf{G}}(c', v', \mathbf{k}', \mathbf{Q}). \quad (18)$$

In the BerkeleyGW code package, the original formulation of the dual-grid interpolation method employs two sets of $\mathbf{k}$ points: one set of $\mathbf{k}_{\mathrm{co}}$ $\mathbf{k}$ points defined on a coarse grid and a set of $\mathbf{k}_{\mathrm{fi}}$ $\mathbf{k}$ points defined on a fine grid. The direct (K^{d}) and exchange (K^{x}) matrix elements in the BSE kernel are explicitly calculated on $\mathbf{k}_{\mathrm{co}}$. Then, an interpolation is performed by expanding each fine-grid Bloch state in terms of the closest coarse-grid Bloch state,

$$|u_{n\mathbf{k}_{\mathrm{fi}}}\rangle = \sum_m C^{\mathbf{k}_{\mathrm{co}}}_{nm} |u_{m\mathbf{k}_{\mathrm{co}}}\rangle \quad (19)$$

$$C^{\mathbf{k}_{\mathrm{co}}}_{nm} = \int \mathrm{d}^3 r\; u_{n\mathbf{k}_{\mathrm{fi}}}(\mathbf{r})\, u^*_{m\mathbf{k}_{\mathrm{co}}}(\mathbf{r}), \quad (20)$$

which allows the kernel matrix elements to be approximated as

$$K^{\mathrm{d/x}}_{mn;m'n'}(\mathbf{k}_{\mathrm{fi}}, \mathbf{q}_{\mathrm{fi}}{=}\mathbf{k}'_{\mathrm{fi}}{-}\mathbf{k}_{\mathrm{fi}}) \approx \sum_{\substack{n_1 n_2 \\ m_1 m_2}} C^{\mathbf{k}_{\mathrm{co}}}_{nn_1} C^{\mathbf{k}_{\mathrm{co}}*}_{mm_1} \times C^{\mathbf{k}'_{\mathrm{co}}*}_{n'n_2} C^{\mathbf{k}'_{\mathrm{co}}}_{m'm_2}\, K^{\mathrm{d/x}}_{mn;m'n'}(\mathbf{k}_{\mathrm{fi}}, \mathbf{q}_{\mathrm{co}}{=}\mathbf{k}'_{\mathrm{co}}{-}\mathbf{k}_{\mathrm{co}}). \quad (21)$$

Notice in Eq. 17 and Eq. 18 that K^{d} depends sensitively on the relative reciprocal vector $\mathbf{q}$, whereas K^{x} depends only on the center-of-mass momentum $\mathbf{Q}$, which is a constant. The bare Coulomb interaction $v(\mathbf{q})$ diverges as $\mathbf{q} \to 0$. We therefore expect that K^{d} will change very rapidly at small $\mathbf{q}$, and consequently, any direct interpolation of K^{d} must converge very slowly. To avoid this problem, as currently implemented in BerkeleyGW, K^{d} is decomposed into its head (K^{h}), wing (K^{w}) and body (K^{b}) contributions, each of which has different limiting behavior for the Coulomb interaction as $\mathbf{q} \to 0$. The head contains the terms where $\mathbf{G}{=}\mathbf{G}'{=}0$ and diverges as $\frac{1}{q^2}$ in 3D and $\frac{1}{q}$ in 2D, for semiconductors and insulators. The wing contains the sum over terms where $\mathbf{G}{=}0 \neq \mathbf{G}'$ or $\mathbf{G}'{=}0 \neq \mathbf{G}$ and diverges as $\frac{1}{q}$ in 3D and goes to a constant in 2D, for semiconductors and insulators. The body contains the sum over terms where $\mathbf{G} \neq 0$ and $\mathbf{G}' \neq 0$

and goes to a constant value in the limit of small q. With this understanding, for the 2D and 3D cases, the direct term can be written as

$$K^{\rm d}_{mn;m'n'}(\mathbf{k},\mathbf{q}) = \frac{a_{mn;m'n'}(\mathbf{k},\mathbf{q})}{q^{d-1}} + \frac{b_{mn;m'n'}(\mathbf{k},\mathbf{q})}{q^{d-2}} + c_{mn;m'n'}(\mathbf{k},\mathbf{q}), \quad (22)$$

where d is the effective dimension, and

$$\begin{aligned} a_{mn;m'n'}(\mathbf{k},\mathbf{q}) &= q^{d-1} \times K^{\rm h}_{mn;m'n'}(\mathbf{k},\mathbf{q}), \\ b_{mn;m'n'}(\mathbf{k},\mathbf{q}) &= q^{d-2} \times K^{\rm w}_{mn;m'n'}(\mathbf{k},\mathbf{q}), \\ c_{mn;m'n'}(\mathbf{k},\mathbf{q}) &= K^{\rm b}_{mn;m'n'}(\mathbf{k},\mathbf{q}). \end{aligned} \quad (23)$$

The functions a, b, and c, where the divergence in the Coulomb interaction is removed, are interpolated in the dual-grid scheme and then used to construct K^d.

The interpolation procedure described above works efficiently for 3D metals and semiconductors, where a, b, and c are smooth functions of $\mathbf{q}$ because the inverse dielectric matrix is also a smooth function of $\mathbf{q}$. However, as we previously discussed, $\varepsilon^{-1}_{\mathbf{GG}'}(\mathbf{q})$ displays sharp features in systems with reduced dimensionality as $\mathbf{q}\to 0$, and thus, the head, wing and body components of the matrix elements, which depend on ε^{-1}, may also vary considerably with $\mathbf{q}$ even when the divergence in the Coulomb interaction is removed. We illustrate this behavior by plotting the head and wing components of the matrix elements associated with the direct Coulomb term of the BSE (Eq. 17) for silicon and monolayer MoS_2 on Fig. 6. In bulk silicon, we multiply the head matrix elements by q^2 and the wing matrix elements by $\mathbf{q}$ to remove the divergence due to the bare Coulomb interaction. Then, the matrix elements are smooth functions of $\mathbf{q}$. In 2D, however, the non-smooth behavior cannot be removed by multiplying any simple factor. We remove the divergence due to the bare Coulomb interaction by multiplying the head matrix element by q, but even after removing the Coulomb divergence, both the head and wing components still have a sharp features at small $\mathbf{q}$. These features are a consequence of the sharp feature in the inverse dielectric matrix (Fig. 1).

To capture these sharp features in 2D, it is important to explicitly calculate $K^{\rm h/w/b}_{mn;m'n'}(\mathbf{k},\mathbf{q})$ for a variety of small $\mathbf{q}$. Consequently, a dual-grid scheme as described above necessarily converges very slowly with respect to sampling of the coarse grid, which must be fine enough to resolve the sharp feature in $K^{\rm h/w/b}$, and quickly becomes prohibitively expensive, since the cost of calculating the matrix elements scales with the number of coarse $\mathbf{k}$ points squared.

In contrast to their sharply varying q-dependence, however, the head, wing, and body matrix elements do not depend much on $\mathbf{k}$. This is because the k-dependence comes in solely in the matrix elements $M(m,n,\mathbf{k},\mathbf{q})$, which are typically smooth functions of $\mathbf{k}$, since the periodic part of the Bloch functions are smoothly varying quantities. This is illustrated for the case of monolayer MoS_2 in Fig. 7, where the contribution to the BSE from the direct screened Coulomb interaction, $K^{\rm d}(\mathbf{k},\mathbf{k}'{=}\mathbf{k}{+}\mathbf{q})$, displays a very small spread over a wide range of values for $\mathbf{k}$. Thus, in order to capture all of the screening effects, we need to minimally sample a large number of finely-spaced $\mathbf{q}$ transitions from a set of $\mathbf{k}$ points that can be relatively coarse.

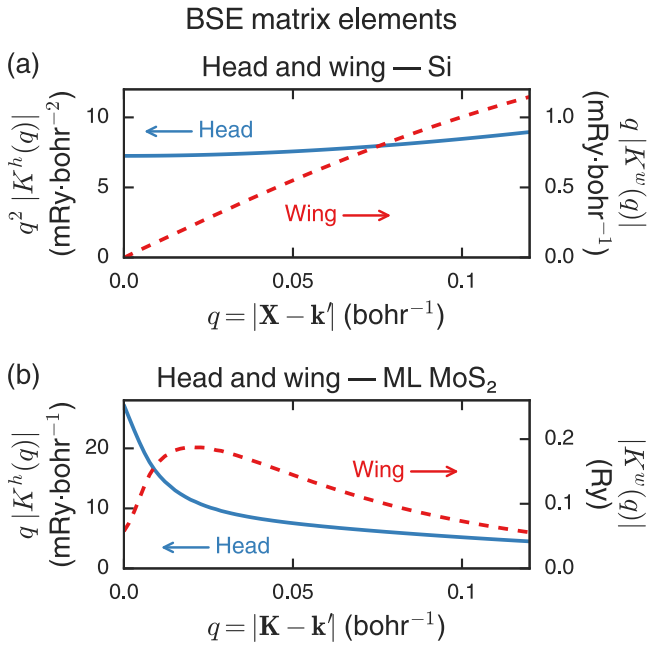

FIG. 6. (Color online) Head and wing components of the BSE matrix elements for bulk Si (top) and monolayer MoS_2 (bottom).

In order to explicitly capture the small-$\mathbf{q}$ behavior, we develop an extension on the dual-grid interpolation scheme. In addition to calculating BSE matrix elements on a coarse k grid, we also explicitly calculate BSE matrix elements for scattering from each $\mathbf{k}$ point, $\mathbf{k}_{\rm co}$, on the coarse grid to an arbitrary cluster of $\mathbf{k}$ points, $\mathbf{k}_{\rm cl}$, close to each $\mathbf{k}_{\rm co}$. We will refer to this scheme as clustered sampling interpolation (CSI). For simplicity, we will focus on the case of isotropic materials, where $K^{\rm h/w/b}_{mn;m'n'}(\mathbf{k},\mathbf{q})$ depends only on $|\mathbf{q}|$. In this case, the cluster of points can be chosen to lie along a radial line extending from each $\mathbf{k}_{\rm co}$. The generalization to anisotropic materials is straightforward with a small computational overhead.

We interpolate the BSE matrix elements from the coarse grid to the fine grid using a conditional scheme. If the distance between two points on the fine grid, $|\mathbf{k}'_{\rm fi}{-}\mathbf{k}_{\rm fi}|$, is greater than the smallest distance between two points on the coarse grid $\Delta_{\rm co}$, the interpolation is identical to the original dual-grid scheme. If the distance between two points on the fine grid, $|\mathbf{k}'_{\rm fi}-\mathbf{k}_{\rm fi}|$, is less than $\Delta_{\rm co}$, we expand the Bloch state at $\mathbf{k}_{\rm fi}$ over the Bloch states at the closest coarse point, $\mathbf{k}_{\rm co}$, and we expand the Bloch state at $\mathbf{k}'_{\rm fi}$ over the Bloch states at a cluster point, $\mathbf{k}_{\rm cl}$, for which $K^{\rm h/w/b}_{mn;m'n'}(\mathbf{k}_{\rm co},\mathbf{q}{=}\mathbf{k}_{\rm cl}-\mathbf{k}_{\rm co})$ has already been cal-

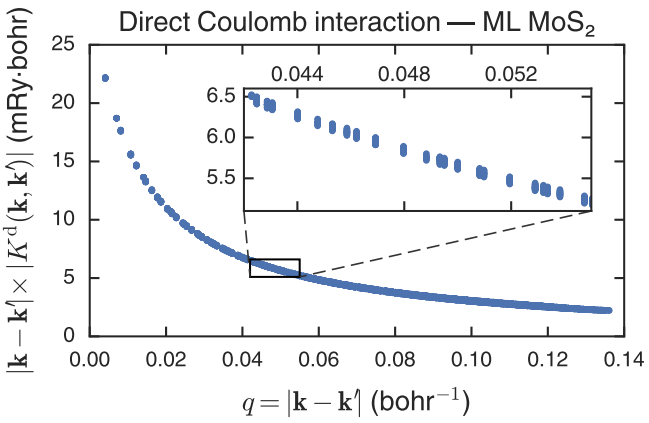

FIG. 7. (Color online) Matrix element for the BSE involving the direct screened Coulomb interaction, $K^{\rm d}(\mathbf{k},\mathbf{k}')$, calculated for an electron and hole states at the conduction band and valence band, respectively, for monolayer MoS_2, where $\mathbf{k}$ and $\mathbf{k}'$ are any points on a $300\times300\times1$ k grid that lies within 0.02 bohr^{-1} of the K point. Note that $K^{\rm d}(\mathbf{k},\mathbf{k}')$ depends very weakly on either $\mathbf{k}$ or $\mathbf{k}'$ individually if $\mathbf{q}{=}\mathbf{k}'-\mathbf{k}$ is kept constant. The inset zooms in on a small region of the plot and illustrates the small spread in values for different $\mathbf{k}$ and $\mathbf{k}'$ pairs with the same q.

culated, while preserving as closely as possible the length of the transfer vector so that $|\mathbf{q}| = |\mathbf{k}_{\rm co}-\mathbf{k}_{\rm cl}| \approx |\mathbf{k}'_{\rm fi}-\mathbf{k}_{\rm fi}|$. Then,

$$K^{\rm h/w/b}_{mn;m'n'}(\mathbf{k}_{\rm fi},\mathbf{q}_{\rm fi}{=}\mathbf{k}'_{\rm fi}{-}\mathbf{k}_{\rm fi}) \approx \sum_{\substack{n_1n_2\\m_1m_2}} C^{\mathbf{k}_{\rm co}}_{nn_1}C^{\mathbf{k}_{\rm co}*}_{mm_1} \times C^{\mathbf{k}_{\rm cl}*}_{n'n_2}C^{\mathbf{k}_{\rm cl}}_{m'm_2}\; K^{\rm h/w/b}_{mn;m'n'}(\mathbf{k}_{\rm fi},\mathbf{q}_{\rm co}{=}\mathbf{k}_{\rm cl}{-}\mathbf{k}_{\rm co}), \tag{24}$$

where

$$C^{\mathbf{k}'_{\rm cl}}_{nm} = \int {\rm d}^3r\; u_{n\mathbf{k}'_{\rm fi}}(\mathbf{r})\, u^*_{m\mathbf{k}_{\rm cl}}(\mathbf{r}). \tag{25}$$

We now apply clustered sampling interpolation to a system of interest, monolayer MoS_2. The calculation is performed in a supercell setup with $L_z = 25\mathring{A}$. Fig. 8 shows how the binding energy of the lowest energy 1s and 2p excitons in monolayer MoS_2 converges with respect to an explicit calculation on a single uniform grid (the single-grid scheme) and with respect to the coarse k grid when using either dual-grid interpolation or clustered sampling interpolation. For both the dual-grid interpolation and CSI, the coarse grid is interpolated to a $300\times300\times1$ fine grid. The binding energy is defined, following Ref.[17], as the difference between the electron-hole continuum and the exciton excitation energy, which is independent of the numerical treatment of the divergence at $W(\mathbf{q}{=}0)$. From Fig. 8, it is clear that the clustered sampling interpolation converges much more quickly than the dual-grid interpolation, requiring only a $18\times18\times1$ coarse grid to converge the binding energy to within 0.1 eV. In contrast, the dual-grid scheme does not converge until the coarse grid sampling is increased beyond $48\times48\times1$. Moreover, we see that while in most cases the dual-grid interpolation is still an improvement on the uniform gird, the convergence fluctuates. This erratic convergence occurs because different uniform k grids sample different regions of the sharp feature in the screening. The different convergence rates are even more dramatic for higher energy states, such as the 2p state (Fig. 8), whose complex nodal structure is even more sensitive to the spatially varying screening at small q.

In general, calculating the BSE matrix elements scales with the total number of $\mathbf{k}$ points squared. Thus, in the dual-grid scheme, the computational cost scales with the number of $\mathbf{k}$ points on the coarse grid squared, $N^2_{\mathbf{k}_{\rm co}}$. Clustered sampling interpolation has an additional cost associated with calculating the matrix elements involving transitions between the coarse $\mathbf{k}$ points and cluster points. This additional term scales as $N_{\mathbf{k}_{\rm co}} \times N_{\mathbf{k}_{\rm cl}}$, where $N_{\mathbf{k}_{\rm cl}}$ is the number of $\mathbf{k}$ points in each cluster. For isotropic systems $N_{\mathbf{k}_{\rm cl}}$ is typically much smaller than $N_{\mathbf{k}_{\rm co}}$, since the sampling is only along one dimension. Thus, the additional cost of clustered sampling interpolation is small compared with the cost of calculating the matrix elements on the coarse grid. The total cost of calculating the matrix elements scales as $N^2_{\mathbf{k}_{\rm co}} + N_{\mathbf{k}_{\rm co}} \times N_{\mathbf{k}_{\rm cl}}$.

Table II shows the CPU time required to calculate the BSE kernel, $K^{\rm eh}$, for MoS_2 in the dual-grid and CSI schemes as a function of the coarse grid and interpolated to a $300\times300\times1$ fine grid. On identical coarse grids, there is a small computational overhead in the CSI scheme, on the order of 10 core-hours, which scales linearly with $N_{\mathbf{k}_{\rm co}}$. However, the binding energy of the 1s exciton state converges in the CSI scheme on an $18\times18\times1$ coarse grid, whereas it is still unconverged on a $48\times48\times1$ coarse grid in the dual-grid scheme. Thus, for MoS_2, CSI results in a speed-up of *at least* one order of magnitude. To directly calculate the BSE matrix elements on a $300\times300\times1$ as reported in Fig. 8, we solve the BSE on a patch, which only includes $\mathbf{k}$ points within $0.2\mathring{A}^{-1}$ of the K point in the Brillouin zone. This allows us to obtain the binding energy of the lowest energy excitons within 20meV of the calculation on the full Brillouin zone but is insufficient to obtain the entire optical spectrum. Since we know that calculating the BSE matrix elements scales as $N^2_{\mathbf{k}_{\rm co}}$, we estimate that directly calculating the BSE matrix on a $300\times300\times1$ k grid would take approximately 15 million core-hours, compared with 228 core hours with CSI scheme.

The proposed clustered sampling scheme assumes an isotropic system, where the BSE matrix elements depend only on the magnitude of q. However, in practice, it also results in improved convergence for anisotropic materials such as few-layer black phosphorus. Fig. 9 shows the performance of the CSI scheme for monolayer black phosphorus, when the clustered points are sampled along the (100) direction, (110) direction and (010) direction. The calculation is done in a supercell setup with $L_z = 20\mathring{A}$. Here, the convergence of the binding energy for the different interpolation schemes is referenced to a single grid

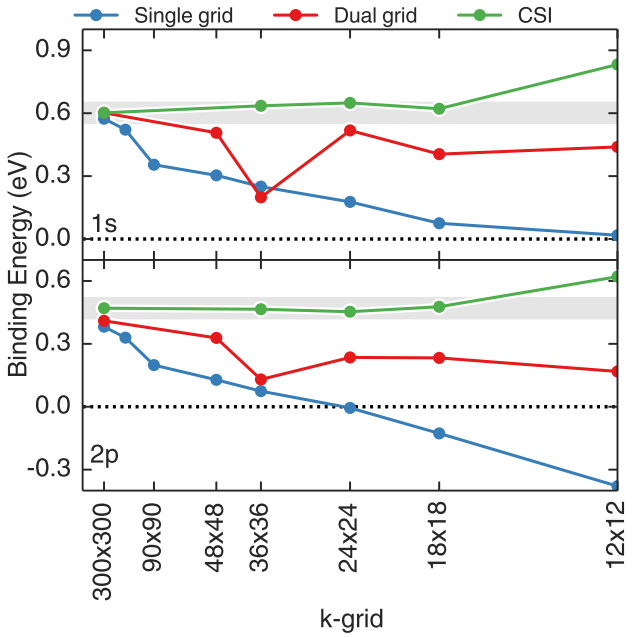

FIG. 8. (Color online) Convergence of the binding energy of the 1s(top) and 2p(bottom) states of the lowest energy series of excitons in MoS_2 using an explicit calculation without interpolation (single grid), a dual-grid method and the proposed CSI method. The x-axis represents the k-point grid used in the single grid method and the coarse grid used in both the dual grid and CSI methods. The gray shaded region corresponds to an interval of ± 50 meV compared to the converged value, which is the convergence threshold one is typically interested in.

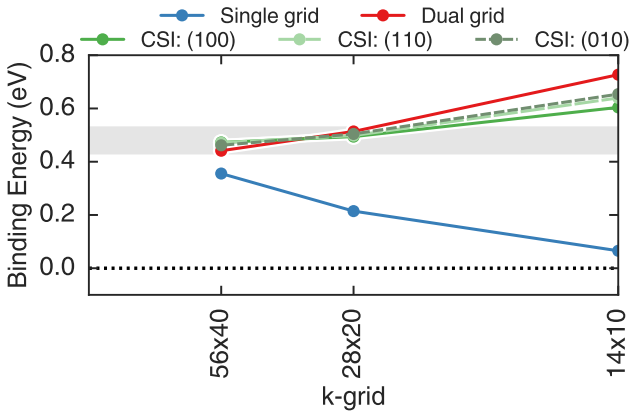

FIG. 9. (Color online) Convergence of the binding energy of the 1s(top) state of the lowest energy series of excitons in monolayer black phosphorus using an explicit calculation without interpolation (single grid), a dual-grid method and the proposed CSI method with sampling along the 100, 110, and 010 directions. The x-axis represents the k-point grid used in the single grid method and the coarse grid used in both the dual grid and CSI methods. The gray shaded region corresponds to an interval of ± 50 meV compared to the converged value, which is the convergence threshold one is typically interested in.

calculation with $160 \times 160 \times 1$ k-points performed in a patch of radius 0.2 Å around the Γ point in the Brillouin zone. The reference converged binding energy is 0.47 eV. For the same coarse and fine k grids, the CSI scheme always converges more quickly than both the single and dual grid scheme, with convergence being the fastest when the clustered points are sampled along the (100) direction, which is the more highly-dispersive armchair direction in black phosphorus. Once again, the convergence of the dual grid scheme is erratic due to the spatially varying screening. While some k grids give similar results to the CSI scheme, as the k grid is increased to $56 \times 40 \times 1$ the dual grid binding energy still undershoots the converged value.

TABLE II. Comparison of computational time required to calculate the BSE kernel matrix elements K^{eh} with different Brillouin Zone interpolation schemes for MoS_2 on different coarse k grids, k_{co}. The binding energy of the 1s state, E_b^{1s}, is given after interpolating from the coarse k grid to a $300 \times 300 \times 1$ fine grid.

	Dual Grid		CSI	
k_{co} grid	CPU Usage (core–hour)	E_b^{1s} (eV)	CPU Usage (core–hour)	E_b^{1s} (eV)
$12 \times 12 \times 1$	39	0.44	59	0.83
$18 \times 18 \times 1$	196	0.41	223	0.62
$24 \times 24 \times 1$	579	0.52	613	0.65
$36 \times 36 \times 1$	2948	0.20	3017	0.64
$48 \times 48 \times 1$	8857	0.51	8979	0.63

IV. CONCLUSION

In summary, we address the problem that many-electron perturbation theory calculations, such as those performed in GW and GW-BSE theories, on low-dimensional systems converge very slowly with respect to sampling of the Brillouin zone due to sharp features in the spatial variations in screening, which manifest as sharp features in the q-dependence of the dielectric matrix and cannot be described by a simple analytic model due to the complexity of the out-of-plane local fields. Thus, we present two new schemes to sample the Brillouin zone in a computationally efficient way for low-dimensional systems. The first scheme, which we refer to as the non-uniform neck subsampling (NNS) method, allows for efficient sampling of single-particle problems, such as GW and ACFDT. In the NNS method, an additional radial sampling is performed in the Voronoi cell that surrounds each $\mathbf{k}$ point with appropriately chosen weights. The second scheme, clustered sampling interpolation (CSI), addresses two-particle scattering problems,

such as in solving for the solution of the BSE. In CSI, we explicitly calculate two-particle scattering matrix elements in small, uniformly-spaced clusters of $\mathbf{k}$ points and use these clusters to interpolate to a uniform fine grid. Both schemes result in typical speedups of about two orders of magnitude in the computer run-time and can be easily incorporated into several *ab initio* packages that compute electronic and optical properties employing many-body-perturbation theory methods.

This work was supported by the Center for Computational Study of Excited State Phenomena in Energy Materials at the Lawrence Berkeley National Laboratory, which is funded by the U. S. Department of Energy, Office of Basic Energy Sciences under Contract No. DE-AC02-05CH11231. D. Y. Q. acknowledges support from the NSF Graduate Research Fellowship Grant No. DGE 1106400. Computational resources have been provided from the National Energy Research Scientific Computing Center (NERSC), a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

V. APPENDIX

For a given N_s number of subsampled points, we have the freedom to define two quantities: the N_s annulus thicknesses Δ_s that define the intervals for the radial integrals and, thus, the weights w_s; and the N_s subsampling points q_s where the dielectric matrix has to be explicitly computed. In order to make the discretization in Eq. 8 practical, it is necessary to choose appropriate subsampled points q_s and/or the thicknesses Δ_s that approximates the integral in an efficient way. We introduce a scheme to use a crude approximation of the screened Coulomb interaction to provide constraints on either q_s or Δ_s.

For semiconductors, the inverse dielectric matrix approaches a finite constant as $\mathbf{q}\to 0$, so we can write the head of the (truncated) screened Coulomb interaction as $W(q) \propto 1/q^2$, $1/q$ or $\log(q)$, depending on whether we have a 3D, 2D or 1D system, respectively. A good choice of Δ_s and q_s is such that the screened Coulomb interaction evaluated at the subsampled point represents the average value of the W according to the analytic limit,

$$W(\mathbf{q}_s,\omega{=}0)\int_{a_s}^{a_{s+1}} \mathrm{d}^D q' = \int_{a_s}^{a_{s+1}} \mathrm{d}^D q'\, W(\mathbf{q}',\omega{=}0), \quad (26)$$

which, together with the constraint of a_{N_s} from $\mathcal{C}_{\mathbf{q}=\mathbf{0}}$, provide N_s+1 constraints and allows us to obtain the optimal subsampling points q_s given N_s-1 thicknesses Δ_s for the radial integration.

We summarize the relationship between Δ_s and q_s obtained from Eq. 26 for 3D, 2D and 1D semiconducting systems in Table III. For systems other than semiconductors, the screened Coulomb interaction has different analytic behaviors for $\mathbf{q}\to 0$, so other optimal subsampling points could be determined. Fortunately, for metallic systems of any dimensionality, the condition in Eq. 26 is fulfilled for any choice of q_s, and for quasi-2D systems with linear energy dispersion, such as graphene, the optimal subsampling point is still given by $q_s = a_s + \frac{\Delta_s}{2}$. So, we use the relationship between q_s and Δ_s as defined in Table III for all types of systems we consider.

TABLE III. Optimal choice of the subsampling point q_s in terms of the thickness Δ_s of each radial interval for 3D, 2D and 1D semiconductors. The inner radius of each interval is denoted by $a_s \equiv \sum_{i=1}^{s-1}\Delta_s$. We also compare the first optimal point q_1 for different dimensionality.

D	q_s	q_1
3	$\sqrt{a_s^2 + a_s\Delta_s + \Delta_s^2/3}$	$0.577\Delta_1$
2	$a_s + \frac{\Delta_s}{2}$	$0.500\Delta_1$
1	$(a_s+\Delta_s)(1+\Delta_s/a_s)^{a_s/\Delta_s}/e$	$0.368\Delta_1$

* These two authors contributed equally.
† Email: sglouie@berkeley.edu

[1] L. Hedin, Phys. Rev. **139**, A796 (1965).
[2] M. S. Hybertsen and S. G. Louie, Phys. Rev. B **34**, 5390 (1986).
[3] G. Strinati, Riv. Nuovo Cimento **11**, 1 (1988).
[4] M. Rohlfing and S. G. Louie, Phys. Rev. B **62**, 4927 (2000).
[5] M. Fuchs and X. Gonze, Physical Review B **65**, 23 (2002).
[6] A. Marini, P. García-González, and A. Rubio, Physical Review Letters **96**, 2 (2006).
[7] D. Lu, Y. Li, D. Rocca, and G. Galli, Physical Review Letters **102**, 1 (2009).
[8] J. Toulouse, I. C. Gerber, G. Jansen, A. Savin, and J. G. Ángyán, Physical Review Letters **102**, 1 (2009).
[9] S. Lebègue, J. Harl, T. Gould, J. G. Ángyán, G. Kresse, and J. F. Dobson, Physical Review Letters **105**, 1 (2010).
[10] J. Deslippe, G. Samsonidze, D. Strubbe, M. Jain, M. L. Cohen, and S. G. Louie, Comput. Phys. Commun. **183**, 1269 (2012).
[11] A. Marini, H. Conor, M. Gruning, and D. Varsano, Comp. Phys. Comm. **180**, 1392 (2009).
[12] X. Gonze, B. Amadon, P.-M. Anglade, J.-M. Beuken, F. Bottin, P. Boulanger, F. Bruneval, D. Caliste, R. Caracas, M. Ct;, T. Deutsch, L. Genovese, P. Ghosez, M. Giantomassi, S. Goedecker, D. Hamann, P. Hermet, F. Jollet, G. Jomard, S. Leroux, M. Mancini, S. Mazevet, M. Oliveira, G. Onida, Y. Pouillon, T. Rangel, G.-M. Rignanese, D. Sangalli, R. Shaltaf, M. Torrent, M. Verstraete, G. Zerah, and J. Zwanziger, Computer Physics Communications **180**, 2582 (2009).
[13] J. J. Mortensen, L. B. Hansen, and K. W. Jacobsen, Phys.

Rev. B **71**, 035109 (2005).
[14] D. Y. Qiu, F. H. da Jornada, and S. G. Louie, Phys. Rev. Lett. **111**, 216805 (2013), erratum, *ibid.* **115**, 119901 (2015).
[15] A. Molina-Sanchez, D. Sangalli, K. Hummer, A. Marini, and L. Wirtz, Phys. Rev. B **88**, 045412 (2013).
[16] F. Hüser, T. Olsen, and K. Thygesen, Phys. Rev. B **88**, 245309 (2013).
[17] D. Y. Qiu, F. H. da Jornada, and S. G. Louie, Phys. Rev. B **93**, 235435 (2016).
[18] F. A. Rasmussen, P. S. Schmidt, K. T. Winther, and K. S. Thygesen, Phys. Rev. B **94**, 155406 (2016).
[19] S. Ismail-Beigi, Phys. Rev. B **73**, 233103 (2006).
[20] P. Cudazzo, I. V. Tokatly, and A. Rubio, Phys. Rev. B **84**, 085406 (2011).
[21] A. Chernikov, T. C. Berkelbach, H. M. Hill, A. Rigosi, Y. Li, O. B. Aslan, D. R. Reichman, M. S. Hybertsen, and T. F. Heinz, Phys. Rev. Lett. **113**, 076802 (2014).
[22] F. Gygi and A. Baldereschi, Phys. Rev. B **34**, 4405 (1986).
[23] P. Carrier, S. Rohra, and A. Görling, Phys. Rev. B **75**, 205126 (2007).
[24] A Voronoi cell can be thought of as a generalization of the Wigner-Seitz cell for non-uniform grids; if the q-vectors are sampled uniformly on a $n_q \times n_q \times n_q$ grid, as is usually the case for 3D systems, then $\mathcal{C}_{\mathbf{q}}$ is simply the Brillouin zone scaled down isotropically by n_q.
[25] We report the error of the IP instead of the absolute IP because the convergence of the absolute IP is much more sensitive to other numerical parameters, such as the cutoff of the dielectric matrix and number of bands.
[26] C. H. Rycroft, Chaos **19**, 041111 (2009).
[27] M. Rohlfing and S. G. Louie, Phys. Rev. Lett. **81**, 2312 (1998).