Editors' Suggestion

Fully self-consistent finite-temperature *GW* in Gaussian Bloch orbitals for solids

Chia-Nan Yeh,[1] Sergei Iskakov,[1] Dominika Zgid,[2,1] and Emanuel Gull[1]

[1]*Department of Physics, University of Michigan, Ann Arbor, Michigan 48109, USA*
[2]*Department of Chemistry, University of Michigan, Ann Arbor, Michigan 48109, USA*

(Received 15 June 2022; revised 7 October 2022; accepted 16 November 2022; published 5 December 2022)

We present algorithmic and implementation details for the fully self-consistent finite-temperature *GW* method in Gaussian Bloch orbitals for solids. Our implementation is based on the finite-temperature Green's function formalism in which all equations are solved on the imaginary axis, without resorting to analytical continuation during the self-consistency. No quasiparticle approximation is employed and all matrix elements of the self-energy are explicitly evaluated. The method is tested by evaluating the band gaps of selected semiconductors and insulators. We show agreement with other, differently formulated, finite-temperature sc*GW* implementations when finite-size corrections and basis-set errors are taken into account. By migrating computationally intensive calculations to graphics processing units, we obtain scalable results on large supercomputers with nearly optimal performance. Our work demonstrates the applicability of Gaussian orbital based sc*GW* for *ab initio* correlated material simulations and provides a sound starting point for embedding methods built on top of *GW*.

DOI: 10.1103/PhysRevB.106.235104

I. INTRODUCTION

The *GW* method [1] provides direct access to single-particle excitation spectra, unlike ground-state methods such as the density functional theory (DFT) [2]. *Ab initio* simulations of single-particle excitation spectra are essential for direct comparison to experiment such as angle-resolved photoemission spectroscopy (ARPES). *GW* has been widely applied to weakly correlated systems, such as semiconductors [3,4]. Due to the increasing availability of computing power, it is rapidly becoming an alternative to DFT [3,4]. The *GW* method also frequently serves as the first step in embedding frameworks designed to include strong correlations [5–17].

In Hedin's seminal paper [1] a set of exact self-consistent equations was introduced, describing an expansion of the self-energy in terms of the screened interaction. Such an expansion is especially beneficial for metallic systems, where expansions in terms of the bare (i.e., unscreened) interactions may diverge. Hedin's equations are relations between the Green's function, self-energy, vertex function, polarization function, and screened Coulomb interaction. In the so-called *GW* approximation, Hedin's equations are truncated to the first order in the screened Coulomb interaction. The self-consistent solution of these equations is guaranteed to satisfy certain conservation laws and is thermodynamically consistent [1,18,19].

Despite its theoretical simplicity, obtaining results from the *GW* approximation is orders of magnitude more expensive than solving the equations of DFT, and implementations of the *GW* method for materials have not yet reached the maturity of DFT codes, where consistent results for both molecules and solids can be obtained from independent codes with different numerical setups [20–25]. A one-to-one comparison of *GW* implementations is complicated by a dependence on basis sets [e.g., plane waves, linearized augmented plane waves (LAPW), Gaussian orbitals (GTO)], differences in correcting for finite-size effects, differences in methodologies for evaluating quasiparticle peaks and band gaps, and the effect of additional approximations beyond the truncation to first order in the screened interaction. These approximations may consist of a combination of "numerical" approximations (such as the choice of finite basis set in space or frequency, or analytic continuation) and "theoretical" approximations. For molecular systems, multiple comparisons of results from different implementations of the one-shot variant of *GW* (G_0W_0) have been published before [26–29].

Common additional theoretical approximations introduced in practical *GW* calculations include approximations to the self-consistency as well as quasiparticle approximations. The *GW* method is frequently executed non-self-consistently. This approximation is referred to as G_0W_0 [30–40]. Consequently, G_0W_0 results depend on the choice of the starting single-particle Green's function G_0. The selection of a proper starting point requires empirical knowledge of the system of interest. For the band gaps of semiconductors, common choices of G_0 include the DFT Green's function with local density approximation (LDA) and PBE functionals [30–40]. G_0W_0 calculations also employ the quasiparticle approximation, which approximates the frequency dependence of the self-energy by solving the quasiparticle equation in the Kohn-Sham orbital basis. Significant improvement over DFT results has been observed due to well-defined single-particle excitations in *GW*. When compared to experimental data, the somewhat fortuitous agreement has been attributed to an error cancellation between the lack of self-consistency and the absence of vertex corrections in the self-energy and the polarizability [41–44].

Eigenvalue self-consistent *GW* [45–47] and quasiparticle self-consistent *GW* (QS*GW*) [48–51] attempt to eliminate the starting-point dependence with different levels of self-consistency. In QS*GW*, an effective nonlocal static potential is self-consistently determined in the presence of the *GW* self-energy to construct an optimal one-body reference Hamiltonian [49]. While independent of the starting solution, QS*GW* still employs the quasiparticle approximation and obtains the self-energy only at certain frequency points.

The solution of the fully self-consistent *GW* (sc*GW*) approximation became feasible only in recent years, due to numerous numerical advancements, such as the representation of dynamical quantities [52–59] and low-scaling optimizations that exploit the locality of the self-energy [60–62]. Several fully self-consistent *GW* implementations have been reported for molecules [63–65] and periodic systems [44,66–68]. Nevertheless, reaching agreement between different sc*GW* implementations remains challenging [44,68]. Due to the correlated nature of the *GW* approximation, both the core-valence interactions and the interactions between occupied and unoccupied states are included beyond a single-particle picture. Therefore, the quality of the basis sets for unoccupied states and the treatment of core electrons becomes more important than in DFT [69–71].

In this paper, we present a formulation of fully self-consistent finite-temperature *GW* in Gaussian Bloch orbitals for solids. The method is based on the finite-temperature Green's function formalism on the imaginary axis and thus does not require analytical continuation during the self-consistent loop. No quasiparticle approximation is employed. It relies on sparse sampling on the imaginary axis [58] using the intermediate representation [54], Gaussian density fitting for the decomposition of the bare Coulomb integrals [72–74], and Nevanlinna analytical continuation [75] to extract data on the real axis from Green's functions evaluated on the imaginary axis. The implementation makes efficient use of parallel graphics processing unit (GPU) architectures.

Extensions of this work in the presence of strong electron correlations [11–13] or relativistic effects [76] have been discussed previously without presenting details of the *GW* implementation. Here, we focus on implementation details and discuss finite-size effects and finite-size convergence, convergence with respect to the Gaussian basis size, and we benchmark the performance of sc*GW* on GPU architectures. In addition, in the absence of quasiparticle and non-self-consistent approximations, our work provides reference values of sc*GW* band gaps for selected semiconductors and insulators that are compared to another finite-temperature sc*GW* implementation [44], where the numerical setup is entirely different.

The paper will proceed as follows. In Sec. II, we introduce the electronic Hamiltonian in the context of Gaussian Bloch orbitals. Section III discusses the self-consistent *GW* equations as well as electronic thermodynamic properties, and Sec. IV describes details of our implementation, including the parallelization scheme employing GPUs. Lastly, Sec. V compares our sc*GW* data to other *GW* implementations on a series of benchmark results. Our conclusions are presented in Sec. VI.

II. ELECTRONIC HAMILTONIAN AND FINITE BASIS SETS

A. Electronic Hamiltonian

We solve a general electronic Hamiltonian $\hat{H} = \hat{H}_0 + \hat{U}$, consisting of a one-electron part $\hat{H}_0$ and two-electron Coulomb interactions $\hat{U}$ (also known as electron repulsion integrals). In a translation-invariant system, $\hat{H}$ in second quantization is

$$\hat{H} = \sum_{\mathbf{k}} \sum_{ij} \sum_{\sigma\sigma'} (H_0)^{\mathbf{k}}_{i\sigma,j\sigma'} \hat{c}^{\mathbf{k}\dagger}_{i\sigma} \hat{c}^{\mathbf{k}}_{j\sigma'} + \frac{1}{2N_k} \sum_{ijkl} \sum_{\mathbf{k}_i\mathbf{k}_j\mathbf{k}_k\mathbf{k}_l} \sum_{\sigma\sigma'} U^{\mathbf{k}_i\mathbf{k}_j\mathbf{k}_k\mathbf{k}_l}_{i\ j\ k\ l} \hat{c}^{\mathbf{k}_i\dagger}_{i\sigma} \hat{c}^{\mathbf{k}_k\dagger}_{k\sigma'} \hat{c}^{\mathbf{k}_l}_{l\sigma'} \hat{c}^{\mathbf{k}+j}_{j\sigma}, \tag{1}$$

where N_k is the number of $\mathbf{k}$ points sampled in the Brillouin zone, and $\hat{c}^{\mathbf{k}\dagger}_{i\sigma}$ ($\hat{c}^{\mathbf{k}}_{i\sigma}$) are the creation (annihilation) operators for electrons in the single-particle spin-orbital basis with crystal momentum $\mathbf{k}$, spin σ, and finite basis index i. The two-electron Coulomb interactions conserve crystal momentum, i.e., $\mathbf{k}_i + \mathbf{k}_k - \mathbf{k}_j - \mathbf{k}_l = \mathbf{G}$, where $\mathbf{G}$ is a reciprocal lattice vector. In general, H_0 exhibits a spin dependence with nonzero off-diagonal spin components. These components appear in the presence of spin-orbit coupling (SOC) and external magnetic fields (see Ref. [76]).

We use bold symbols for matrices in the spin-orbital basis, bold italic symbols for tensors such as the two-electron Coulomb interactions $\boldsymbol{U}$, and regular italic symbols for matrix/tensor elements.

B. Gaussian-type orbitals

We employ Bloch wave functions $g^{\mathbf{k}}_i(\mathbf{r})$ constructed from Gaussian-type orbitals (GTOs) $g^{\mathbf{R}}_i(\mathbf{r})$ as our finite basis set [22–24,77]. The Gaussian Bloch basis $g^{\mathbf{k}}_i(\mathbf{r})$ is expressed as

$$g^{\mathbf{k}}_i(\mathbf{r}) = \sum_{\mathbf{R}} g^{\mathbf{R}}_i(\mathbf{r}) e^{i\mathbf{k}\cdot\mathbf{R}}, \tag{2}$$

where $\mathbf{k}$ is a crystal momentum in the first Brillouin zone of the reciprocal space, and $g^{\mathbf{R}}_i(\mathbf{r})$ is the ith Gaussian atomic orbital centered in unit cell $\mathbf{R}$ [78]. The summation over $\mathbf{R}$ extends to the whole lattice.

Gaussian Bloch waves for different crystal momenta are orthogonal. In orbital space, they define the overlap matrix

$$S^{\mathbf{k}}_{ij} = \int_{\Omega} d\mathbf{r}\, g^{\mathbf{k}*}_i(\mathbf{r}) g^{\mathbf{k}}_j(\mathbf{r}), \tag{3}$$

where Ω denotes the unit-cell volume.

In the nonrelativistic case, the one-electron Hamiltonian H_0 and the two-electron Coulomb integrals U are defined as

$$\begin{aligned}(H_0)^{\mathbf{k}}_{i\sigma,j\sigma'} &= (H_0)^{\mathbf{k}}_{ij}\delta_{\sigma\sigma'}\\ &= \delta_{\sigma\sigma'} \int_{\Omega} d\mathbf{r}\, g^{\mathbf{k}*}_i(\mathbf{r}) \left[-\frac{1}{2}\nabla^2_{\mathbf{r}} + \sum_{\alpha} \frac{Z_\alpha}{|\mathbf{r} - \mathbf{r}_\alpha|} \right] g^{\mathbf{k}}_j(\mathbf{r})\end{aligned} \tag{4}$$

and

$$U^{\mathbf{k}_i\mathbf{k}_j\mathbf{k}_k\mathbf{k}_l}_{i\,j\,k\,l} = \iint d\mathbf{r}\, d\mathbf{r}'\, g_i^{\mathbf{k}_i *}(\mathbf{r}) g_j^{\mathbf{k}_j}(\mathbf{r}) \frac{1}{|\mathbf{r}-\mathbf{r}'|} g_k^{\mathbf{k}_k *}(\mathbf{r}') g_l^{\mathbf{k}_l}(\mathbf{r}'). \tag{5}$$

Reference [76] derives the corresponding expression for a relativistic one-electron Hamiltonian in the exact two-component theory with one-electron approximation (X2C1e).

C. Decomposition of two-electron Coulomb interactions

The two-electron Coulomb interaction, Eq. (5), has a memory requirement of $O(N_k^3 N_{\mathrm{orb}}^4)$ where N_{orb} is the number of atomic orbitals in the unit cell. To reduce the size of this tensor, we use Coulomb potential decompositions. These decompositions can in general be expressed as

$$U^{\mathbf{k}_i\mathbf{k}_j\mathbf{k}_k\mathbf{k}_l}_{i\,j\,k\,l} = \sum_Q V^{\mathbf{k}_i\mathbf{k}_j}_{i\,j}(Q) V^{\mathbf{k}_k\mathbf{k}_l}_{k\,l}(Q), \tag{6}$$

where Q denotes an auxiliary decomposition index and $V^{\mathbf{k}_i\mathbf{k}_j}_{i\,j}(Q)$ is a tensor with two momenta, two orbital indices, and an auxiliary index. Decomposition procedures include the Cholesky decomposition [79] and the density fitting technique [also known as the resolution-of-identity (RI) approximation] [80–82]. In this work, we employ periodic Gaussian density fitting (GDF) with the overlap metric [72–74]. Given an additional set of auxiliary Gaussian orbitals $\chi_Q^{\mathbf{q}}(\mathbf{r})$ as the auxiliary basis, Eq. (6) is computed as

$$V^{\mathbf{k}_i\mathbf{k}_j}_{i\,j}(Q) = \sum_{PP'} (\mathbf{J}^{\mathbf{q}})^{1/2}_{QP'} (\mathbf{A}^{\mathbf{q}})^{-1}_{P'P} B^{\mathbf{k}_i\mathbf{k}_j}_{i\,j}(P), \tag{7}$$

where $\mathbf{q} = \mathbf{k}_j - \mathbf{k}_i$, and

$$A^{\mathbf{q}}_{P'P} = \int_\Omega d\mathbf{r}\, \chi_P^{\mathbf{q}*}(\mathbf{r}) \chi_{P'}^{\mathbf{q}}(\mathbf{r}), \tag{8}$$

$$B^{\mathbf{k}_i\mathbf{k}_j}_{i\,j}(P) = \int_\Omega d\mathbf{r}\, \chi_P^{\mathbf{q}*}(\mathbf{r}) g_i^{\mathbf{k}_i *}(\mathbf{r}) g_j^{\mathbf{k}_j}(\mathbf{r}), \tag{9}$$

$$J^{\mathbf{q}}_{PQ} = \iint d\mathbf{r}\, d\mathbf{r}' \frac{\chi_P^{\mathbf{q}*}(\mathbf{r}) \chi_Q^{\mathbf{q}}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}. \tag{10}$$

We choose the even-tempered basis (ETB) with the default progression parameter $\beta_{\mathrm{ETB}} = 2.0$ [83] for $\chi_Q^{\mathbf{q}}(\mathbf{r})$. The typical size of the auxiliary basis $\{\chi_Q^{\mathbf{q}}(\mathbf{r})\}$ is roughly $3 \sim 10$ times size of the GTO basis $\{g_i^{\mathbf{k}}(\mathbf{r})\}$ [84]. Thus, Eq. (7) provides a much more compact representation of two-electron Coulomb interactions than Eq. (6). In addition to lowering memory requirements, the decomposed three-index tensor reduces the scaling of sc*GW* (see Sec. III).

III. SELF-CONSISTENT FINITE-TEMPERATURE *GW*

The central objects of finite-temperature perturbation theory are finite-temperature one-particle Green's functions $G^{\mathbf{k}}_{i\sigma,j\sigma}(\tau)$ and self-energies $\Sigma^{\mathbf{k}}_{i\sigma,j\sigma}(\tau)$. The Green's function $G^{\mathbf{k}}_{i\sigma,j\sigma}(\tau)$ is defined as

$$G^{\mathbf{k}}_{i\sigma,j\sigma}(\tau) = -\frac{1}{Z}\mathrm{Tr}[e^{-(\beta-\tau)(\hat{H}-\mu\hat{N})} \hat{c}^{\mathbf{k}}_{i\sigma} e^{-\tau(\hat{H}-\mu\hat{N})} \hat{c}^{\mathbf{k},\dagger}_{j\sigma}], \tag{11}$$

$$Z = \mathrm{Tr}[e^{-\beta(\hat{H}-\mu\hat{N})}], \tag{12}$$

where Z is the partition function, β the inverse temperature, μ the chemical potential, $\hat{N}$ the particle-number operator, and $\tau \in [0, \beta]$ the imaginary time. Fourier transforms between imaginary-time and Matsubara frequency are defined as

$$G^{\mathbf{k}}_{i\sigma,j\sigma}(i\omega_n) = \int_0^\beta d\tau\, G^{\mathbf{k}}_{i\sigma,j\sigma}(\tau) e^{i\omega_n\tau} \tag{13}$$

and

$$G^{\mathbf{k}}_{i\sigma,j\sigma}(\tau) = \frac{1}{\beta}\sum_n G^{\mathbf{k}}_{i\sigma,j\sigma}(i\omega_n) e^{-i\omega_n\tau}, \tag{14}$$

where $\omega_n = (2n+1)\pi/\beta$, $n \in \mathbb{Z}$, are fermionic Matsubara frequencies. Given an interacting Green's function $\mathbf{G}^{\mathbf{k}}(\tau)$, the correlated density matrix is $\boldsymbol{\gamma}^{\mathbf{k}} = -\mathbf{G}^{\mathbf{k}}(\tau = \beta^-)$ and the total number of electrons N_e is determined as

$$N_e = \frac{1}{N_k}\sum_{\mathbf{k}} \mathrm{tr}[\boldsymbol{\gamma}^{\mathbf{k}} \mathbf{S}^{\mathbf{k}}], \tag{15}$$

where the trace implies a sum over the diagonals in the spin-orbital space.

In the Matsubara frequency domain, the Dyson equation relating self-energies to Green's functions is

$$\begin{aligned}[\mathbf{G}^{\mathbf{k}}(i\omega_n)]^{-1} &= (i\omega_n + \mu)\mathbf{S}^{\mathbf{k}} - \mathbf{H}_0^{\mathbf{k}} - \boldsymbol{\Sigma}^{\mathbf{k}}(i\omega_n)[\boldsymbol{G}] \\ &= [\mathbf{G}_0^{\mathbf{k}}(i\omega_n)]^{-1} - \boldsymbol{\Sigma}^{\mathbf{k}}(i\omega_n)[\boldsymbol{G}],\end{aligned} \tag{16}$$

where $[\mathbf{G}_0^{\mathbf{k}}(i\omega_n)]^{-1} = (i\omega_n + \mu)\mathbf{S}^{\mathbf{k}} - \mathbf{H}_0^{\mathbf{k}}$ is the noninteracting Green's function of the one-electron Hamiltonian $\mathbf{H}_0^{\mathbf{k}}$ and $\boldsymbol{\Sigma}^{\mathbf{k}}[\boldsymbol{G}]$ is the self-energy which is a functional of the full interacting Green's function $\mathbf{G}^{\mathbf{k}}(i\omega_n)$. The inverse is defined as a matrix inversion in spin-orbital space for any given $\mathbf{k}$ and $i\omega_n$.

Self-consistent *GW* yields a particular approximation $(\boldsymbol{\Sigma}^{GW})^{\mathbf{k}}[\boldsymbol{G}]$ of the exact self-energy. Separating the self-energy into its static and dynamical parts,

$$(\boldsymbol{\Sigma}^{GW})^{\mathbf{k}}[\boldsymbol{G}](i\omega_n) = (\boldsymbol{\Sigma}^{GW}_\infty)^{\mathbf{k}}[\boldsymbol{G}] + (\tilde{\boldsymbol{\Sigma}}^{GW})^{\mathbf{k}}[\boldsymbol{G}](i\omega_n), \tag{17}$$

$(\boldsymbol{\Sigma}^{GW}_\infty)^{\mathbf{k}}$ is the static Hartree-Fock (HF) self-energy, and $(\tilde{\boldsymbol{\Sigma}}^{GW})^{\mathbf{k}}(i\omega_n)$ corresponds to the frequency-dependent *GW* self-energy which is obtained via the summation of an infinite series of RPA-like "bubble" diagrams [1]. Together with the Dyson equation [Eq. (16)], the self-energy [Eq. (17)] is solved as a functional of the interacting Green's function $\mathbf{G}^{\mathbf{k}}(i\omega_n)$ iteratively until self-consistency between $\mathbf{G}^{\mathbf{k}}(i\omega_n)$ and $(\boldsymbol{\Sigma}^{GW})^{\mathbf{k}}(i\omega_n)$ is achieved.

A. Hartree-Fock self-energy

The Hartree-Fock self-energy is static and can be further divided into a Hartree term ($\boldsymbol{J}$) and an exchange term ($\boldsymbol{K}$):

$$\left(\Sigma^{GW}_\infty\right)^{\mathbf{k}}_{i\sigma,j\sigma} = J^{\mathbf{k}}_{i\sigma,j\sigma} + K^{\mathbf{k}}_{i\sigma,j\sigma}. \tag{18}$$

Individually, each term can be expressed in terms of the GDF Coulomb tensor $V^{\mathbf{k}\mathbf{k}'}_{ij}(Q)$ and the density matrix $\gamma^{\mathbf{k}'}_{b\sigma,a\sigma'}$ as

$$J^{\mathbf{k}}_{i\sigma,j\sigma} = \frac{1}{N_k}\sum_{\mathbf{k}'}\sum_{\sigma_1}\sum_{ab} \gamma^{\mathbf{k}'}_{a\sigma_1,b\sigma_1} U^{\mathbf{k}\mathbf{k}\mathbf{k}'\mathbf{k}'}_{i\,j\,b\,a} \tag{19}$$

$$= \frac{1}{N_k}\sum_{\mathbf{k}'}\sum_{\sigma_1}\sum_{ab}\sum_Q V^{\mathbf{k}\mathbf{k}}_{ij}(Q) \gamma^{\mathbf{k}'}_{a\sigma_1,b\sigma_1} V^{\mathbf{k}'\mathbf{k}'}_{b\,a}(Q) \tag{20}$$

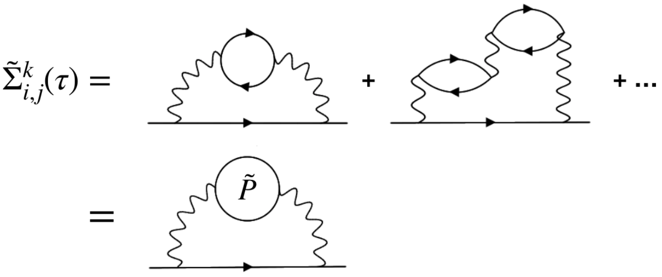

FIG. 1. Self-energy diagrams for the dynamical part of the GW self-energy. Lines with arrows denote interacting Green's function $G^{\mathbf{k}}_{i\sigma,j\sigma}$, wiggly lines denote the interaction $U^{\mathbf{k},\mathbf{k}-\mathbf{q},\mathbf{k}',\mathbf{k}'+\mathbf{q}}_{i,\ j,\ k,\ l}$. $\tilde{\boldsymbol{P}}$ represents the sum of all "bubble" diagrams.

and

$$K^{\mathbf{k}}_{i\sigma,j\sigma'} = -\frac{1}{N_k}\sum_{\mathbf{k}'}\sum_{ab}\gamma^{\mathbf{k}'}_{a\sigma,b\sigma'}U^{\mathbf{k}\mathbf{k}'\mathbf{k}'\mathbf{k}}_{i\,a\,b\,j} \tag{21}$$

$$= -\frac{1}{N_k}\sum_{\mathbf{k}'}\sum_{ab}\sum_{Q}V^{\mathbf{k}\mathbf{k}'}_{i\,a}(Q)\gamma^{\mathbf{k}'}_{a\sigma,b\sigma'}V^{\mathbf{k}'\mathbf{k}}_{b\,j}(Q). \tag{22}$$

B. Dynamical part of the *GW* self-energy

In GW, the dynamical part of the self-energy is approximated as the sum of an infinite series of RPA-like "bubble" diagrams [1] as shown in Fig. 1. On the imaginary-time axis, $(\tilde{\boldsymbol{\Sigma}}^{GW})^{\mathbf{k}}(\tau)$ reads as

$$(\tilde{\Sigma}^{GW})^{\mathbf{k}}_{i\sigma,j\sigma}(\tau) = -\frac{1}{N_k}\sum_{\mathbf{q}}\sum_{ab}G^{\mathbf{k}-\mathbf{q}}_{a\sigma,b\sigma}(\tau)\tilde{W}^{\mathbf{k},\mathbf{k}-\mathbf{q},\mathbf{k}-\mathbf{q},\mathbf{k}}_{i\ \ a\ \ \ b\ \ \ j}(\tau), \tag{23}$$

where $\tilde{\boldsymbol{W}}$ is the effective screened interaction tensor, defined as the difference between the full dynamically screened interaction $\boldsymbol{W}$ and the bare interaction $\boldsymbol{U}$, i.e., $\tilde{\boldsymbol{W}} = \boldsymbol{W} - \boldsymbol{U}$. In the GW approximation, the screened interaction $\boldsymbol{W}$ is expressed as [1]

$$W^{\mathbf{k}_1\mathbf{k}_2\mathbf{k}_3\mathbf{k}_4}_{i\ j\ k\ l}(i\Omega_n)=U^{\mathbf{k}_1\mathbf{k}_2\mathbf{k}_3\mathbf{k}_4}_{i\ j\ k\ l}+\frac{1}{N_k}\sum_{\mathbf{k}_5\mathbf{k}_6\mathbf{k}_7\mathbf{k}_8}\sum_{abcd}U^{\mathbf{k}_1\mathbf{k}_2\mathbf{k}_5\mathbf{k}_6}_{i\ j\ a\ b}\Pi^{\mathbf{k}_5\mathbf{k}_6\mathbf{k}_7\mathbf{k}_8}_{a\ b\ c\ d}$$
$$\times\ (i\Omega_n)W^{\mathbf{k}_7\mathbf{k}_8\mathbf{k}_3\mathbf{k}_4}_{c\ d\ k\ l}(i\Omega_n), \tag{24}$$

where $\boldsymbol{\Pi}$ is the noninteracting polarization function

$$\Pi^{\mathbf{k}_1\mathbf{k}_2\mathbf{k}_3\mathbf{k}_4}_{a\ b\ c\ d}(\tau) = \sum_{\sigma}G^{\mathbf{k}_1}_{d\sigma,a\sigma}(\tau)G^{\mathbf{k}_2}_{b\sigma,c\sigma}(-\tau)\delta_{\mathbf{k}_1\mathbf{k}_4}\delta_{\mathbf{k}_2\mathbf{k}_3}. \tag{25}$$

Due to the size of the interaction tensor $\boldsymbol{U}$, solving Eqs. (24) and (25) directly is not practical in large simulations. As illustrated in Fig. 2, the decomposition of Coulomb integrals allows us to express $\tilde{\boldsymbol{W}}$ as

$$\tilde{W}^{\mathbf{k},\mathbf{k}-\mathbf{q},\mathbf{k}-\mathbf{q},\mathbf{k}}_{i\ \ j\ \ \ \ k\ \ \ \ l}(i\Omega_n)$$
$$= \sum_{Q,Q'}V^{\mathbf{k},\mathbf{k}-\mathbf{q}}_{i\ \ j}(Q)\big\{\tilde{P}^{\mathbf{q}}_{0,QQ'}(i\Omega_n) + \big[\tilde{P}^{\mathbf{q}}_{0}(i\Omega_n)\big]^2_{QQ'} + \cdots\big\}$$
$$\times V^{\mathbf{k}-\mathbf{q},\mathbf{k}}_{k\ \ \ \ l}(Q') \tag{26}$$

$$= \sum_{Q,Q'}V^{\mathbf{k},\mathbf{k}-\mathbf{q}}_{i\ \ j}(Q)\tilde{P}^{\mathbf{q}}_{QQ'}(i\Omega_n)V^{\mathbf{k}-\mathbf{q},\mathbf{k}}_{k\ \ \ \ l}(Q'), \tag{27}$$

where $\Omega_n = 2n\pi/\beta$ $(n\in\mathbb{Z})$ are bosonic Matsubara frequencies. The noninteracting auxiliary function $\tilde{\mathbf{P}}^{\mathbf{q}}_0(i\Omega_n)$ is defined as

$$\tilde{P}^{\mathbf{q}}_{0,QQ'}(i\Omega_n) = \int_0^{\beta}d\tau\,\tilde{P}^{\mathbf{q}}_{0,QQ'}(\tau)e^{i\Omega_n\tau}, \tag{28a}$$

$$\tilde{P}^{\mathbf{q}}_{0,QQ'}(\tau) = \frac{-1}{N_k}\sum_{\mathbf{k}}\sum_{\sigma\sigma'}\sum_{abcd}V^{\mathbf{k},\mathbf{k}+\mathbf{q}}_{d\ \ a}(Q)$$
$$\times G^{\mathbf{k}}_{c\sigma',d\sigma}(-\tau)G^{\mathbf{k}+\mathbf{q}}_{a\sigma\ b\sigma'}(\tau)V^{\mathbf{k}+\mathbf{q},\mathbf{k}}_{b\ \ \ \ c}(Q'), \tag{28b}$$

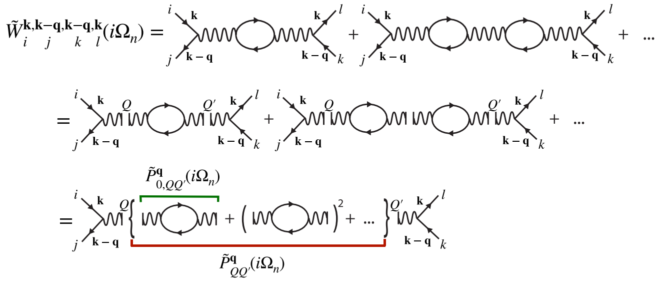
$\tilde{W}^{\mathbf{k},\mathbf{k}-\mathbf{q},\mathbf{k}-\mathbf{q},\mathbf{k}}_{i\ \ j\ \ \ \ k\ \ \ \ l}(i\Omega_n) =$
i
k
j
k − q
k
l
k − q
k
+
+ ...
=
Q
Q′
+
+ ...
=
$\tilde{P}^{\mathbf{q}}_{0,QQ'}(i\Omega_n)$
Q
{
+
2
+ ...
}
Q′
$\tilde{P}^{\mathbf{q}}_{QQ'}(i\Omega_n)$

FIG. 2. Diagrammatic expression of $\tilde{\boldsymbol{W}}$.

and the renormalized auxiliary function $\tilde{\mathbf{P}}^{\mathbf{q}}(i\Omega_n)$ is computed using the geometric series

$$\tilde{\mathbf{P}}^{\mathbf{q}}(i\Omega_n) = \sum_{m=1}^{\infty}\big[\tilde{\mathbf{P}}^{\mathbf{q}}_0(i\Omega_n)\big]^m = \big[\mathbf{I} - \tilde{\mathbf{P}}^{\mathbf{q}}_0(i\Omega_n)\big]^{-1}\tilde{\mathbf{P}}^{\mathbf{q}}_0(i\Omega_n), \tag{29}$$

where the inverse denotes a matrix inversion in the auxiliary orbital space and $\mathbf{I}$ is a unitary matrix. Transforming $\tilde{\mathbf{P}}^{\mathbf{q}}(i\Omega_n)$ from the Matsubara frequency to the imaginary-time domain

$$\tilde{P}^{\mathbf{q}}_{QQ'}(\tau) = \frac{1}{\beta}\sum_{n}\tilde{P}^{\mathbf{q}}_{QQ'}(i\Omega_n)e^{-i\Omega_n\tau}, \tag{30}$$

and then inserting it into Eq. (23), we arrive at

$$(\tilde{\Sigma}^{GW})^{\mathbf{k}}_{i\sigma,j\sigma}(\tau) = \frac{-1}{N_k}\sum_{\mathbf{q}}\sum_{ab}\sum_{QQ'}G^{\mathbf{k}-\mathbf{q}}_{a\sigma,b\sigma}(\tau)$$
$$\times\ V^{\mathbf{k},\mathbf{k}-\mathbf{q}}_{i\ \ a}(Q)\tilde{P}^{\mathbf{q}}_{QQ'}(\tau)V^{\mathbf{k}-\mathbf{q},\mathbf{k}}_{b\ \ \ \ j}(Q'). \tag{31}$$

The dynamical self-energy, expressed in such a way, is then evaluated directly on the imaginary-time axis. Self-consistent iterations on the imaginary axis are straightforward and stable. Further approximations to the placement of poles, "quasiparticle" approximations, or approximations to the off-diagonal self-energy structure are not needed. However, the evaluation of real-frequency spectra and band gaps requires an analytical continuation to real frequencies. Recently developed complex analysis techniques [75,85] can be employed to perform this step accurately.

C. Thermodynamic properties

For a conserving approximation, the grand potential Ω is defined in terms of a Green's function, a self-energy, and the

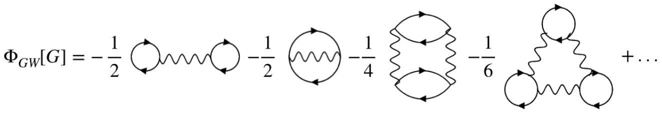

FIG. 3. Diagrammatic expansion of Φ function in the *GW* approximation.

corresponding Φ functional as [86]

$$\Omega[\boldsymbol{G}] = \Phi[\boldsymbol{G}] - \mathrm{Tr}\{\ln[-\boldsymbol{G}^{-1}\} - \mathrm{Tr}\big\{\boldsymbol{\Sigma}[\boldsymbol{G}]\boldsymbol{G}\big\}, \qquad (32)$$

where the symbol Tr{. . .} includes summations over crystal momentum ($\frac{1}{N_k}\sum_{\mathbf{k}}$), Matsubara frequency ($\frac{1}{\beta}\sum_n$), and spin-orbital index (i, σ). The Luttinger-Ward functional $\Phi[\boldsymbol{G}]$ is a functional of $\boldsymbol{G}$ expressed as

$$\Phi[\boldsymbol{G}] = \sum_{m=1}^{\infty} \frac{1}{2m}\mathrm{Tr}\big\{\boldsymbol{\Sigma}^{(m)}[\boldsymbol{G}]\boldsymbol{G}\big\}, \qquad (33)$$

where $\boldsymbol{\Sigma}^{(m)}[\boldsymbol{G}]$ is the mth-order skeleton diagram of the self-energy $\boldsymbol{\Sigma}$. Within the *GW* approximation, the Φ functional, as shown in Fig. 3, is expressed as

$$\Phi_{GW}[G] = -\frac{1}{2}\mathrm{Tr}\big\{\boldsymbol{\Sigma}^{GW}_{\infty}\boldsymbol{\gamma}\big\} - \frac{1}{4}\mathrm{Tr}\{(\boldsymbol{U\Pi})^2\} + \cdots \qquad (34)$$

$$= \Phi^{GW}_{\infty} + \tilde{\Phi}_{GW}, \qquad (35)$$

where $\boldsymbol{U}$ are the bare Coulomb integrals [Eq. (5)] and $\boldsymbol{\Pi}$ is the noninteracting polarization function [Eq. (25)]. We define the first term in Eq. (35) as the contribution of the static *GW* self-energy Φ^{GW}_{∞} and attribute the rest coming from the dynamical *GW* self-energy diagrams as $\tilde{\Phi}_{GW}$.

Using the decomposed Coulomb integrals [Eq. (6)], the first term in $\tilde{\Phi}_{GW}$ is rewritten as

$$\begin{aligned}\frac{1}{4}\mathrm{Tr}\{(\boldsymbol{U\Pi})^2\} &= \frac{1}{4N_k}\sum_{\mathbf{q}}\frac{1}{\beta}\sum_n\sum_{QQ'}\tilde{P}^{\mathbf{q}}_{0,QQ'}(i\Omega_n)\tilde{P}^{\mathbf{q}}_{0,Q'Q}(i\Omega_n)\\ &= \frac{1}{4N_k}\sum_{\mathbf{q}}\frac{1}{\beta}\sum_n \mathrm{tr}\big\{\tilde{\mathbf{P}}^{\mathbf{q}}_0(i\Omega_n)^2\big\}\\ &= \frac{1}{4}\mathrm{Tr}\{(\tilde{\boldsymbol{P}}_0)^2\},\end{aligned} \qquad (36)$$

where $\tilde{\boldsymbol{P}}_0$ is the noninteracting auxiliary function defined in Eq. (28) and tr{. . .} represents the trace of a matrix in the auxiliary Gaussian orbital space. Similarly, $\tilde{\Phi}_{GW}$ can be rewritten in terms of the noninteracting auxiliary function $\tilde{\boldsymbol{P}}_0$ as

$$\begin{aligned}\tilde{\Phi}_{GW} &= -\frac{1}{2}\Big([\mathrm{Tr}\{\tilde{\boldsymbol{P}}_0\} + \frac{1}{2}\mathrm{Tr}\{(\tilde{\boldsymbol{P}}_0)^2\} + \cdots] - \mathrm{Tr}\{\tilde{\boldsymbol{P}}_0\}\Big)\\ &= \frac{1}{2N_k}\sum_{\mathbf{q}}\frac{1}{\beta}\sum_n \mathrm{tr}\{\ln[\mathbf{I} - \tilde{\mathbf{P}}^{\mathbf{q}}_0(i\Omega_n)] + \tilde{\mathbf{P}}^{\mathbf{q}}_0(i\Omega_n)\}.\end{aligned} \qquad (37)$$

Inserting Eqs. (35) and (37) into Eq. (32), the *GW* grand potential is defined. Other thermodynamic quantities such as entropies, specific heats, total energies, and free energies are then evaluated using standard thermodynamic expressions [86–90].

In Φ-derivable self-consistent methods such as sc*GW* and the self-consistent second-order Green's function perturbation theory (GF2) [90,91], thermodynamic quantities are independent of the integration path and the choice of method [18,19]. For example, different approaches for obtaining the total energy from the single-particle Green's function, such as using the Galitskii-Migdal formula or thermodynamic integration, all lead to the same result when the self-energy is Φ derivable [18,19,89]. Φ-derivable approximations also guarantee conservation of particle number, momentum, and energy [19].

D. Treatment of the integrable divergence

The two-electron Coulomb integral (6) has a singularity at $\mathbf{q} \to \mathbf{0}$ due to the long-range contribution of the Coulomb kernel $\sim 1/|\mathbf{q}+\mathbf{G}|^2$ at $\mathbf{G} = \mathbf{0}$. In calculations of the GDF Coulomb tensors $V^{\mathbf{k}_1\mathbf{k}_2}_{i\ j}(Q)$ [Eq. (7)], the $\mathbf{G} = \mathbf{0}$ contribution is manually excluded at $\mathbf{q} = \mathbf{0}$ to regularize the divergence.

In the Coulomb potential $\boldsymbol{J}$ [Eq. (20)], the singularity can be safely excluded since the divergence is canceled exactly by the electron-nucleus Coulomb potential counterpart. When an infinite number of $\mathbf{k}$ points is sampled in the first Brillouin zone, the singularities are in fact integrable in both the HF exchange potential [Eq. (22)] and the dynamical *GW* self-energy [Eq. (23)], resulting in finite contributions [37,41,92–96]. However, in practice, for any finite-size $\mathbf{k}$ mesh, both Eqs. (22) and (23) exhibit a divergence. This can be understood by rewriting the effective screened interaction $\tilde{W}$ in Eq. (23) in terms of the plane-wave basis ($\mathbf{G}$),

$$\begin{aligned}\tilde{W}^{\mathbf{k},\mathbf{k}-\mathbf{q},\mathbf{k}-\mathbf{q},\mathbf{k}}_{i\ \ a\ \ \ b\ \ \ j}(\tau) &= \frac{1}{\Omega}\sum_{\mathbf{G}\mathbf{G}'}\rho^{\mathbf{k}-\mathbf{q}\mathbf{k}*}_{a\ \ i}(\mathbf{G})\frac{\sqrt{4\pi}}{|\mathbf{q}+\mathbf{G}|}\\ &\quad\times(\epsilon^{\mathbf{q},-1}_{\mathbf{G}\mathbf{G}'}(\tau) - \delta_{\mathbf{G}\mathbf{G}'})\frac{\sqrt{4\pi}}{|\mathbf{q}+\mathbf{G}'|}\rho^{\mathbf{k}-\mathbf{q}\mathbf{k}}_{b\ \ j}(\mathbf{G}'),\end{aligned} \qquad (38)$$

where $\epsilon^{\mathbf{q}}_{\mathbf{G}\mathbf{G}'}(\tau)$ is the effective dielectric function in the plane-wave basis, and $\rho^{\mathbf{k}_1\mathbf{k}_2}_{i\ j}(\mathbf{G})$ is the Fourier transform of a GTO pair density function $\rho^{\mathbf{k}_1\mathbf{k}_2}_{i\ j}(\mathbf{r}) = g^{\mathbf{k}_1*}_i(\mathbf{r})g^{\mathbf{k}_2}_j(\mathbf{r})$,

$$\rho^{\mathbf{k}_1\mathbf{k}_2}_{i\ j}(\mathbf{G}) = \int_{\Omega} d\mathbf{r}\,\rho^{\mathbf{k}_1\mathbf{k}_2}_{i\ j}(\mathbf{r})e^{-i(\mathbf{k}_2-\mathbf{k}_1+\mathbf{G})\mathbf{r}}. \qquad (39)$$

Inserting Eq. (38) into (23), we obtain

$$\begin{aligned}(\tilde{\Sigma}^{GW})^{\mathbf{k}}_{i\sigma,j\sigma}(\tau) &= \frac{-1}{N_k\Omega}\sum_{\mathbf{q}}\sum_{\mathbf{G}\mathbf{G}'}\sum_{ab}G^{\mathbf{k}-\mathbf{q}}_{a\sigma,b\sigma}(\tau)\\ &\quad\times\rho^{\mathbf{k}-\mathbf{q}\mathbf{k}*}_{a\ \ i}(\mathbf{G})\frac{\sqrt{4\pi}}{|\mathbf{q}+\mathbf{G}|}(\epsilon^{\mathbf{q},-1}_{\mathbf{G}\mathbf{G}'}(\tau) - \delta_{\mathbf{G}\mathbf{G}'})\\ &\quad\times\frac{\sqrt{4\pi}}{|\mathbf{q}+\mathbf{G}'|}\rho^{\mathbf{k}-\mathbf{q}\mathbf{k}}_{b\ \ j}(\mathbf{G}').\end{aligned} \qquad (40)$$

At $\mathbf{q} \to \mathbf{0}$, the screened interaction in the plane-wave basis diverges when $\mathbf{G} = \mathbf{0}$ or $\mathbf{G}' = \mathbf{0}$. A similar divergence in the HF exchange potential can be understood by replacing $(\epsilon^{\mathbf{q},-1}_{\mathbf{G}\mathbf{G}'}(\tau) - \delta_{\mathbf{G}\mathbf{G}'})$ with $\delta_{\mathbf{G}\mathbf{G}'}$. The explicit exclusion of the $\mathbf{G} = \mathbf{0}$ contribution in the two-electron Coulomb integrals avoids these divergences. At the same time, it will result in a slow convergence with respect to the number of $\mathbf{k}$ points sampled, which impedes a rapid convergence to the thermodynamic limit (TDL) in practical calculations. A finite-size correction is therefore necessary to accelerate convergence to the TDL.

Several strategies of correcting these finite-size effects have been proposed [37,41,92–96]. We follow the procedure of Gygi and Baldereschi [92] in which an auxiliary function is subtracted and added back on the right-hand side of Eqs. (22) and (23). The singularity is first removed by subtracting an auxiliary function that exhibits the same divergence $\sim 1/\mathbf{q}^2$ as $\mathbf{q} \to \mathbf{0}$. Therefore, the resulting smooth integrand can be evaluated accurately by a summation over a finite number of $\mathbf{k}$ points, and the singularity is transferred to the added term expressed by the auxiliary function. The key point is that this added term can be analytically integrated. In principle, the choice of auxiliary function is arbitrary since convergence will be achieved upon increasing the number of $\mathbf{k}$ points, irrespective of the correction. However, a proper choice of auxiliary function will accelerate the convergence with respect to the number of $\mathbf{k}$ points.

In this work, the auxiliary function used to correct the HF exchange potential in Ref. [94] is adopted for both Eqs. (22) and (23). For the dynamical *GW* self-energy, only the leading-order correction at $\mathbf{G} = \mathbf{G}' = \mathbf{0}$ is included, which is the so-called head correction

$$(\Delta^{GW})^{\mathbf{k}}_{i\sigma,j\sigma'}(\tau) = -\chi \sum_{ab} G^{\mathbf{k}-\mathbf{q}}_{a\sigma,b\sigma'}(\tau)\rho^{\mathbf{k}-\mathbf{q}\mathbf{k}*}_{a\ \ i}(\mathbf{G}) \times \left[\epsilon^{\mathbf{q},-1}_{\mathbf{G}\mathbf{G}'}(-\tau) - \delta_{\mathbf{G}\mathbf{G}'}\right]\rho^{\mathbf{k}-\mathbf{q}\mathbf{k}}_{b\ \ j}(\mathbf{G}')\Big|_{\mathbf{q}=\mathbf{G}=\mathbf{G}'=\mathbf{0}} \tag{41}$$

$$= -\chi\left[\epsilon^{\mathbf{0},-1}_{\mathbf{0}\mathbf{0}}(-\tau) - 1\right]\sum_{ab} S^{\mathbf{k}}_{ia} G^{\mathbf{k}}_{a\sigma,b\sigma'}(\tau) S^{\mathbf{k}}_{bj}, \tag{42}$$

where χ is the supercell Madelung constant [94]. The head correction is dynamical and requires the knowledge of the dielectric constant in the long-wavelength limit. However, a direct evaluation of $\epsilon^{\mathbf{q}=\mathbf{0}}_{\mathbf{G}=\mathbf{0},\mathbf{G}'=\mathbf{0}}$ is not possible due to the singularity of the bare Coulomb interaction [44,68]. Instead, we fit $\epsilon^{\mathbf{q}}_{\mathbf{G}=\mathbf{0},\mathbf{G}'=\mathbf{0}}$ using a least-square fit with a finite number of $\mathbf{q}$ points around the Γ point, and then extrapolate to $\mathbf{q} = \mathbf{0}$. Lastly, a static finite-size correction for the HF exchange potential reads as

$$(\Delta^{\mathrm{HF}})^{\mathbf{k}}_{i\sigma,j\sigma'} = -\chi \sum_{ab} S^{\mathbf{k}}_{ia}\gamma^{\mathbf{k}}_{a\sigma,b\sigma'}S^{\mathbf{k}}_{bj}. \tag{43}$$

Additional details of the derivation of Eqs. (42) and (43) are shown in Appendix A.

IV. IMPLEMENTATION DETAILS

A. sc*GW* workflow

The unit cell and crystal structure, the GTO basis set, the auxiliary basis set, the temperature, and the $\mathbf{k}$ mesh fully define the electronic structure problem and allow to precompute the one-electron Hamiltonian $(H_0)^{\mathbf{k}}_{i\sigma,j\sigma'}$, overlap matrices $S^{\mathbf{k}}_{ij}$, and density-fitted Coulomb interactions $V^{\mathbf{k}\mathbf{k}'}_{ij}(Q)$. In this work, all of the tensors $(H_0)^{\mathbf{k}}_{i\sigma,j\sigma'}$, $S^{\mathbf{k}}_{ij}$, and $V^{\mathbf{k}\mathbf{k}'}_{ij}(Q)$ are precomputed using the PYSCF package [23] and stored on disk. Figure 4 shows the workflow of our sc*GW* implementation. Steps 1 and 2 are functionals of $\mathbf{G}$ while steps 3 and 4 are functionals of $\mathbf{\Sigma}^{GW}$. Starting from an initial guess for the Green's function (obtained, for instance, from Hartree-Fock or DFT), the static Hartree-Fock self-energy $(\mathbf{\Sigma}^{GW}_{\infty})^{\mathbf{k}}$ is computed with Eqs. (20) and (22). This Green's function is also used to compute the dynamical part of the *GW* self-energy $(\tilde{\mathbf{\Sigma}}^{GW})^{\mathbf{k}}(\tau)$ through Eq. (31). The dynamical self-energy is then Fourier transformed to Matsubara space. Next, using the newly computed $(\mathbf{\Sigma}^{GW}_{\infty})^{\mathbf{k}}$ and $(\tilde{\mathbf{\Sigma}}^{GW})^{\mathbf{k}}(i\omega_n)$, the chemical potential is adjusted in the presence of the new *GW* self-energy such that the total number of electrons corresponds to a charge-neutral system. (We employ a threshold value of 10^{-9}.) Lastly, the Green's function is updated via the Dyson equation and serves as an input to the next iteration.

This self-consistent iteration is repeated until convergence in Eqs. (16) and (17) is reached. We validate convergence in the Green's function and the self-energy as well as in the energy and in the chemical potential. The additional computational overhead compared to the one-shot G_0W_0 variant is proportional to the number of iterations required to reach a convergence. Typically, 10 ∼ 100 iterations are needed.

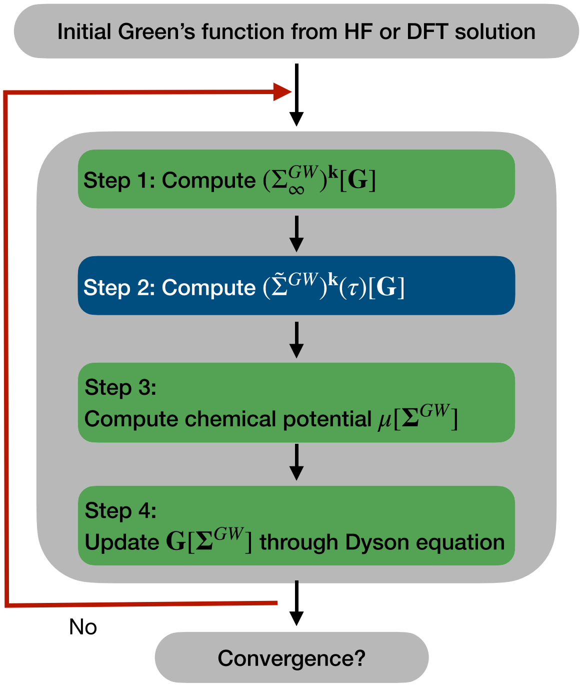
Initial Green's function from HF or DFT solution
Step 1: Compute $(\Sigma^{GW}_{\infty})^{\mathbf{k}}[\mathbf{G}]$
Step 2: Compute $(\tilde{\Sigma}^{GW})^{\mathbf{k}}(\tau)[\mathbf{G}]$
Step 3: Compute chemical potential $\mu[\mathbf{\Sigma}^{GW}]$
Step 4: Update $\mathbf{G}[\mathbf{\Sigma}^{GW}]$ through Dyson equation
No
Convergence?

FIG. 4. Workflow of sc*GW*. Green boxes represent code segments that can be accelerated using MPI while the blue box contains computationally intensive parts implemented in a CPU-GPU hybrid architecture.

B. Convergence acceleration

The number of iterations to reach a self-consistent solution directly affects the efficiency of a sc*GW* implementation. Two aspects will affect the convergence behavior: the initial guess of the Green's function and the iterative solver.

A "good" initial starting point will result in a fast and stable convergence. Aside from the initial guess based on a bare noninteracting Hamiltonian $(H_0)^{\mathbf{k}}_{i\sigma,j\sigma'}$, starting points such as the effective Hartree-Fock or DFT one-electron Green's functions are commonly used. For systems where multiple metastable states are present, an inadequate starting point may guide sc*GW* to converge to a metastable state [97].

Iterative numerical methods that facilitate and stabilize the convergence of self-consistent equations have been an active field of research. Common iterative solvers such as the direct inversion in the iterative subspace (DIIS) [98–101] and the Newton method have been widely applied in DFT and other electronic structure methods for convergence of frequency-independent one-particle quantities [98–100,102,103]. In contrast, applications to Green's-function-based methods are so far limited. In this work, we adopt a DIIS algorithm customized for finite-temperature solution of the Dyson equation [104].

C. Representation on the imaginary axes

The representation of dynamical quantities critically affects the computational cost and the memory requirements. The compact representation of the frequency dependence of one- and two-body quantities is an active field of research [52,54–59]. In this work, all dynamical quantities, including fermionic and bosonic functions, are expanded into the intermediate representation (IR) [54], generated using the IRBASIS open-source software package [105], with sparse sampling on both the imaginary-time and Matsubara frequency axes [58]. The intermediate representation is controlled by a dimensionless parameter λ that should be chosen to be larger than $\beta\tilde{\omega}$ where β is the inverse temperature and $\tilde{\omega}$ is the bandwidth of the system. The typical number of imaginary-time coefficients retained is $10 \sim 200$.

D. Spectral function $A(\omega)$

Once the sc*GW* single-particle Green's function has been computed, the $\mathbf{k}$-resolved spectral function $A^{\mathbf{k}}(\omega)$ can be extracted by inverting the relation

$$G^{\mathbf{k}}_{i\sigma,j\sigma'}(i\omega_n) = \int \frac{A^{\mathbf{k}}_{i\sigma,j\sigma'}(\omega)}{i\omega_n - \omega} d\omega. \tag{44}$$

Due to the nonorthogonality of the GTO basis set, a basis transformation of the Green's functions to an orthonormal basis is necessary in order for $A^{\mathbf{k}}_{i\sigma,i\sigma}(\omega)$ to be normalized to one and strictly positive. In this work, we choose symmetrized atomic orbitals (SAO) [106] constructed from Gaussian Bloch orbitals. Expressing the Green's function in SAO, we continue all diagonal terms of the Green's function from the Matsubara frequency domain to the real frequency axis using Nevanlinna analytical continuation [75] to obtain the $\mathbf{k}$-resolved orbital-dependent spectral functions $A^{\mathbf{k}}_{i\sigma,i\sigma}(\omega)$, and the $\mathbf{k}$-resolved spectral functions $A^{\mathbf{k}}(\omega) = \sum_{i\sigma} A^{\mathbf{k}}_{i\sigma,i\sigma}(\omega)$. Note that for $\mathbf{k}$ points that are not included in the original $\mathbf{k}$ mesh, Wannier interpolation [107] is performed on self-energies in the GTO basis, and interpolated Green's functions are computed via the Dyson equation before continuation.

E. Complexity analysis

We now discuss the computational scaling and memory requirements of our sc*GW* with density-decomposed interactions. All our calculations are performed in double precision. We define N_k as the number of $\mathbf{k}$ points sampled in the Brillouin zone, N_{orb} as the number of atomic orbitals in the unit cell, N_{aux} as the number of auxiliary basis functions used to decompose two-electron integrals in the density-fitting procedure, and N_τ as the number of sparse sampling points on the imaginary-time axis. N_τ equals to the number of sampling points in the Matsubara domain (N_ω).

The calculation of the static Hartree-Fock part of the self-energy requires the evaluation of Eqs. (20) and (22). The evaluation of the Coulomb term scales as $O(N_k N_{\text{orb}}^2 N_{\text{aux}})$. The computational bottleneck at this step is the evaluation of the exchange potential $\mathbf{K}^{\mathbf{k}}$, which scales as $O(N_k^2 N_{\text{orb}}^3 N_{\text{aux}})$.

The evaluation cost of the dynamical part of the self-energy $(\tilde{\mathbf{\Sigma}}^{GW})^{\mathbf{k}}(\tau)$ is dominated by the complex dense matrix products in Eqs. (28b) and (31) as well as the Dyson-type linear equation for $\tilde{\mathbf{P}}^{\mathbf{q}}(i\Omega_n)$ in Eq. (29). The most costly steps in Eqs. (28b) and (31) scale as $O(N_\tau N_k^2 N_{\text{orb}}^2 N_{\text{aux}}^2)$, while the linear equation can be solved in $O(N_\omega N_k N_{\text{aux}}^3)$.

The evaluation of the linear system in Eq. (29) has a low scaling with respect to N_k and N_ω. Inefficient sampling on the Matsubara axis, such as on a uniform Matsubara grid, will result in a large N_ω and a slow evaluation of Eq. (29). This bottleneck is avoided with sparse frequency sampling techniques [58] (Sec. IV C) for the bosonic functions $\tilde{\mathbf{P}}^{\mathbf{q}}_0(i\Omega_n)$ and $\tilde{\mathbf{P}}^{\mathbf{q}}(i\Omega_n)$. The resulting N_ω is reduced to $10 \sim 200$ sampling points, depending on the temperature and bandwidth of the system.

Due to the lack of frequency dependence, the time to solution of the Hartree-Fock self-energies is typically $2 \sim 3$ orders of magnitude smaller than the one of the dynamical self-energy part. The dynamical part is typically dominated by Eqs. (28b) and (31).

The memory requirements are as follows. The largest objects in memory are the Green's function $G^{\mathbf{k}}_{i\sigma,j\sigma'}(\tau)$ and the self-energies $\Sigma^{\mathbf{k}}_{i\sigma,j\sigma'}(\tau)$, which scale as $O(N_\tau N_k N_{\text{orb}}^2)$. In our implementation, both objects are stored once per node in shared memory. The GDF integrals are precomputed and stored on disk. In the evaluation of the *GW* self-energy, they are read in a batch of $\mathbf{k}$ points at a time, such that the memory requirement is $O(N_{\text{orb}}^2 N_{\text{aux}})$. The disk storage requirement scales as $O(N_k^2 N_{\text{orb}}^2 N_{\text{aux}})$, and for large simulations, the storage needs may exceed 1 TB.

Algorithm 1 Pseudocode for the evaluation of the dynamical *GW* self-energy $(\tilde{\mathbf{\Sigma}}^{GW})^{\mathbf{k}}(\tau)$.

```
for $\mathbf{q} \leftarrow 1$ to $N_k$ do
    for $\mathbf{k} \leftarrow 1$ to $N_k$ do
        Read GDF Coulomb tensors $(\mathbf{k}, \mathbf{k}+\mathbf{q})$
        for $\tau \leftarrow 1$ to $N_\tau$ do
            Eq. (28b) $(\mathbf{q}, \mathbf{k}, \tau)$
    Eq. (28a) $(\mathbf{q})$
    for $\Omega_n \leftarrow 1$ to $N_\omega$ do
        Eq. (29) $(\mathbf{q}, \Omega_n)$
    Eq. (30) $(\mathbf{q})$
    for $\mathbf{k} \leftarrow 1$ to $N_k$ do
        Read GDF Coulomb tensors $(\mathbf{k}, \mathbf{k}-\mathbf{q})$
        for $\tau \leftarrow 1$ to $N_\tau$ do
            Eq. (31) $(\mathbf{q}, \mathbf{k}, \tau)$
```

F. GPU acceleration

The computational bottleneck of sc*GW* equations is the evaluation of the dynamical *GW* self-energy (the blue box in Fig. 4). To facilitate its calculation, this part is accelerated using MPI and CUDA.

Algorithm 1 shows a pseudocode for the evaluation of the dynamical *GW* self-energy. Parentheses indicate function arguments. A naive parallelization strategy would be a scheme that distributes the outermost loop over $\mathbf{q}$ points to different MPI processes and performs the calculations independently. However, such a scheme will only be able to reach peak performance for systems with extremely large sizes. To fully maximize the throughput of GPUs for arbitrary system sizes, multiple layers of parallelization are needed.

We first distribute small batches of $\mathbf{q}$ points to different MPI processes as the first layer of parallelization. Each process is assigned one GPU at which the modified bare polarization function $\tilde{\boldsymbol{P}}_0^{\mathbf{q}}$ [Eq. (28b)], the modified dielectric function $\tilde{\boldsymbol{P}}^{\mathbf{q}}$ [Eq. (29)], and the corresponding contributions to $(\tilde{\boldsymbol{\Sigma}}^{GW})^{\mathbf{k}}(\tau)$ [Eq. (31)] for a given $\mathbf{q}$ batch are calculated. The size of a $\mathbf{q}$ batch will depend on the number of GPU cards available.

Different $\mathbf{q}$ points in the same local $\mathbf{q}$ batch are processed serially. At this stage, the number of active MPI processes per node equals the number of available GPU per node. Within the two intermediate `for` loops over $\mathbf{k}$ points, multiple asynchronous streams are created over the $\mathbf{k}$ axis as a second layer of parallelization. This asynchronous stream handling allows overlaps between complex dense matrix multiplication (ZGEMM) with different k indices for a given $\mathbf{q}$ point. The number of streams is determined automatically by the available GPU memory. Note that the parallelization at this layer may be inhibited by I/O operations, and the memory copying for the GDF Coulomb tensors $V_{a\,b}^{\mathbf{kk+q}}(Q)$ between GPUs and CPUs. Asynchronous streams are used to hide this latency. When CPU memory is large enough to store fV, reading the entire $V_{a\,b}^{\mathbf{kk+q}}(Q)$ tensor at the beginning of the calculation is advantageous, such that the only overhead is the memory copy between CPUs and GPUs. Lastly, for loops over N_τ and N_ω, we use batched versions of ZGEMM [108] and a batched Cholesky linear system solver [109] as a third layer of parallelization. The size of the τ batch is set as an external parameter to allow further fine tuning.

On top of the first layer of parallelization using Message Passing Interface (MPI), the second and the third layers allow to fully utilize the computing resources on each GPU independent of system size. Once the computation of the local *GW* self-energy is completed, data reductions for both $(\boldsymbol{\Sigma}_\infty^{GW})^{\mathbf{k}}$ and $(\tilde{\boldsymbol{\Sigma}}^{GW})^{\mathbf{k}}(\tau)$ are performed.

V. RESULTS

In this section, we list results from our sc*GW* by analyzing sc*GW* band gaps. In this work, band gaps are defined as peak-to-peak distances in $\mathbf{k}$-resolved spectral functions, as defined in Sec. IV D. With Nevanlinna techniques [75], such quantities can be evaluated accurately from imaginary-time data. Except Sec. V B, the inverse temperature β is always chosen to be 700 a.u. (corresponds to ∼450 K).

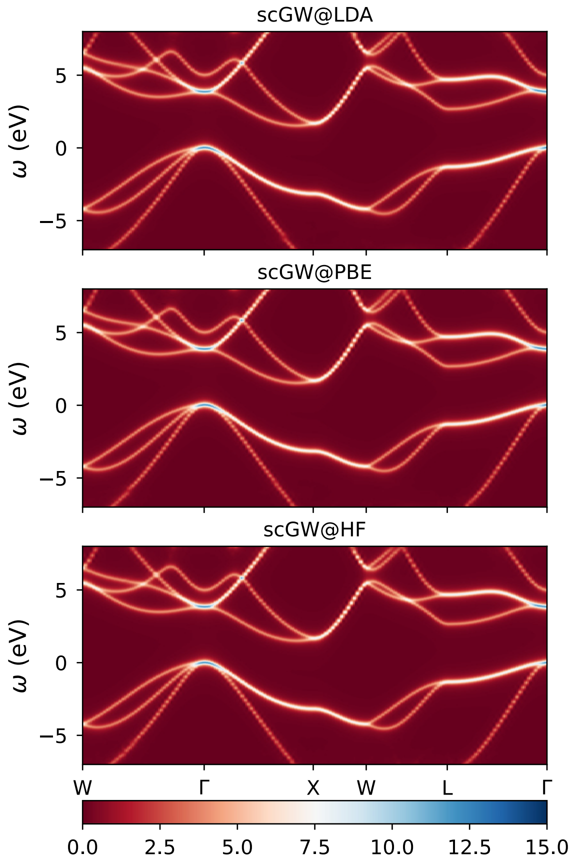

FIG. 5. sc*GW* $\mathbf{k}$-resolved spectral functions of Si calculated from the LDA (top), PBE (middle), and HF (bottom) solutions.

A. Validation of self-consistent solutions

One of the main advantages of sc*GW* is its independence of starting point on the final solution. Here, we verify this property by comparing the sc*GW* results calculated from different starting solutions.

Figure 5 shows the sc*GW* $\mathbf{k}$-resolved spectral functions of Si whose initial Green's functions are taken from LDA, PBE, and HF. Calculations are performed using a $6 \times 6 \times 6$ $\mathbf{k}$ mesh and the all-electron x2c-TZVPall basis set. The spectral functions are obtained via the prescription described in Sec. IV D. In spite of the different starting noninteracting Green's function, the sc*GW* spectral functions converge to the same results consistently along the high-symmetry $\mathbf{k}$ path.

B. Thermodynamic consistency

We continue our analysis of sc*GW* by verifying its thermodynamic consistency. Given self-consistent *GW* solutions and the corresponding Φ functionals, we evaluate the electronic contribution to the thermodynamic quantities, including total energy (E), free energy (F), entropy (S), and specific heat (C_V), as functions of temperature. In Fig. 6, we illustrate thermodynamic properties for BN calculated using a $4 \times 4 \times 4$ $\mathbf{k}$ mesh and the all-electron x2c-TZVPall basis set. Consistency

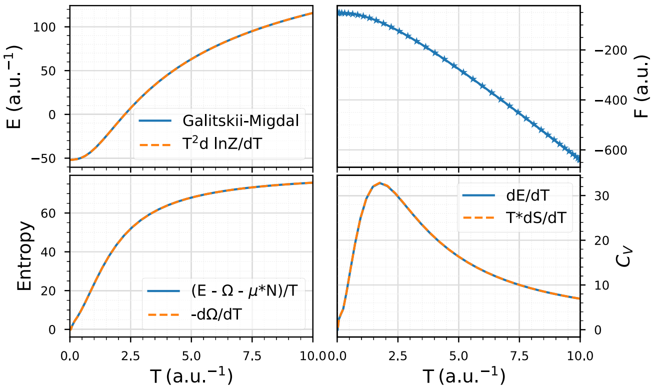

FIG. 6. Electronic thermodynamic quantities evaluated at fixed chemical potential $\mu = 0$, including total energy (E), free energy (F), entropy (S), and specific heat (C_V), as functions of temperature for BN at $4 \times 4 \times 4$ **k** mesh.

between different ways of evaluating total energy, entropy, and specific heat verifies that sc*GW* is thermodynamically consistent.

C. Finite-size corrections

In this section, we analyze sc*GW* by investigating the finite-size effects. By manually neglecting the singularity of the two-electron Coulomb integrals at $\mathbf{q} \rightarrow \mathbf{0}$ as discussed in Sec. III D, the integrable divergence is avoided. This will result in a slow convergence to the thermodynamic limit (TDL), scaling as $O(N_k^{-1/3})$. We investigate the effect of applying the head corrections described in Sec. III D to the *GW* screened exchange self-energy.

Figure 7 shows the convergence of band gaps with and without the head corrections to the dynamical *GW* self-energy $(\tilde{\mathbf{\Sigma}}^{GW})^{\mathbf{k}}(\tau)$ as a function of $N_k^{-1/3}$ for two systems BN and MgO. Note that the finite-size corrections to the HF exchange potential are always included.

In the absence of the head corrections, sc*GW* band gaps consistently exhibit a linear convergence with respect to $N_k^{-1/3}$. The band-gap values are far from converged even with the largest $7 \times 7 \times 7$ **k** mesh. While the convergence is slow, band gaps in the thermodynamic limit can be extrapolated as demonstrated in Fig. 7. We perform the finite-size extrapolation by fitting the sc*GW* band gap to $\Delta(N_k) = \Delta_{\rm TDL} + aN_k^{-1/3}$ for each system (orange lines) and extrapolate to the TDL value $\Delta_{\rm TDL}$ (blue dashed lines). The extrapolation yields the band gaps of 7.24 eV for BN and 9.31 eV for MgO.

When the head corrections are added to the dynamical sc*GW* self-energy, a much faster convergence is observed consistently for all systems tested. A $4 \times 4 \times 4$ **k** mesh already results in band gap that is very close to the TDL value. The band-gap values at the $7 \times 7 \times 7$ **k** mesh are 7.15 and 9.25 eV for BN and MgO which only differ with our extrapolated values by 0.1 eV. The same behavior is observed in all the test systems employed.

Although an extrapolation to the TDL values for band gaps is possible, such a strategy may become impractical when quantities other than band gaps are of interest. Different convergence patterns may be exhibited for these quantities.

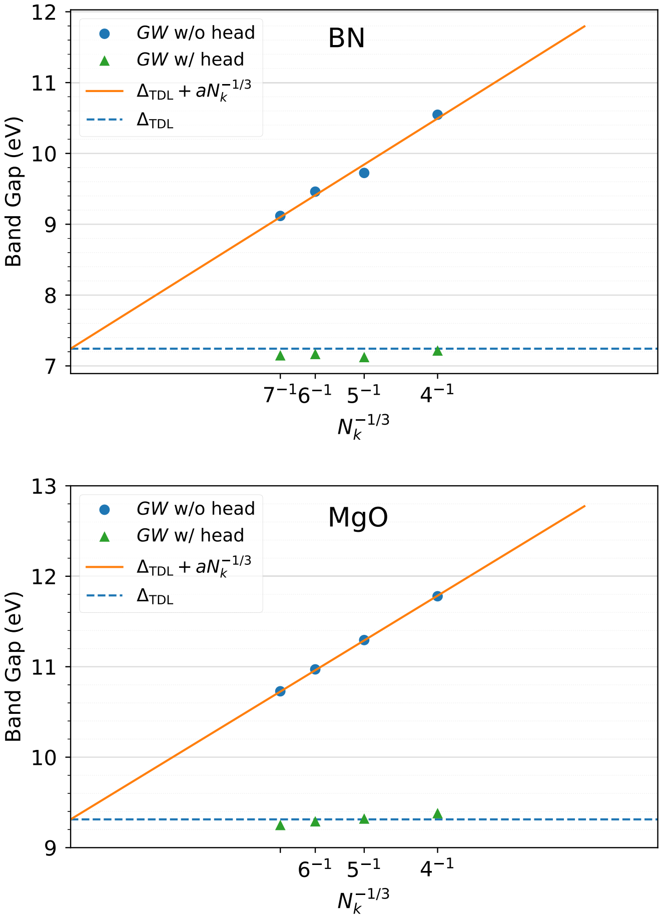

FIG. 7. sc*GW* band gap as a function of $N_k^{-1/3}$ with and without the head corrections to the dynamical part of the *GW* self-energy. A linear fit is performed for the uncorrected band gaps (orange lines) to extrapolate the TDL values (blue dashed lines).

D. Basis-set convergence

In this section, we investigate the basis-set convergence of sc*GW* band gaps. Rather than employing nonrelativistic calculations, we focus on the scalar relativistic case with spin-free X2C1e (sfX2C1e) Hamiltonian since conventional nonrelativistic GTO basis sets can become inadequate especially for the description of core electrons. We employ a family of all-electron basis sets optimized with the X2C Hamiltonian [110,111]. The basis sets are systematically enlarged from a double-ζ [x2c-SV(P)all], to triple-ζ (x2c-TZVP), and finally to quadruple-ζ (x2c-QZVPall) basis set by adding additional high-lying atomic orbitals.

Table I shows the basis-set convergence of sc*GW* band gaps. The band gaps of Si converge very fast with the number of atomic basis functions. At x2c-TZVPall, the gaps are well converged, and only ~0.01 eV difference is observed from x2c-TZVPall to x2c-QZVPall. The slightly slower convergence in AlP is likely due to the missing diffuse functions in x2c-QZVPall that are removed from Al to avoid linear dependencies. On the other hand, the convergence behavior in the presence of transition metal elements is the slowest as demonstrated by ZnO and ZnS. An additional band-gap narrowing of about 0.1 eV is observed when going from

TABLE I. sc*GW* band gaps (eV) of Si, AlP, ZnO, and ZnS calculated using different basis sets. A 5 × 5 × 5 **k** mesh is used for Si and AlP, and a 4 × 4 × 4 **k** mesh is used for ZnO and ZnS. In x2c-QZVPall, the most diffuse *s* and *p* functions of Si, Al, and Zn are removed to avoid linear dependencies.

Basis sets	x2c-SV(P)all	x2c-TZVPall	x2c-QZVPall
Si	1.87	1.54	1.55
AlP	2.97	2.90	2.96
ZnO	5.19	4.59	4.50
ZnS	4.87	4.58	4.46

x2c-TZVPall to x2c-QZVPall. Note that the slower convergence is well known and is attributed to the *d* orbitals of the transition metal elements. Similar behavior has also been observed in compounds with 4*d* transition elements, such as silver halides [70,112].

Overall, a large improvement is observed when going from a double-ζ basis to a triple-ζ basis. The quantitative differences between x2c-TZVPall and x2c-QZVPall are typically minor as long as no transition metal element is present. In the presence of transition metal elements, an additional band-gap narrowing of ∼0.1 eV is expected. Lastly, we have also investigated the basis convergence of the valence band maximum (VBM) and the conduction band minimum (CBM) in Appendix B. We conclude that the convergence behavior observed in Table I is not due to fortunate error cancellation between VBM and CBM.

E. sc*GW* band gaps

We now analyze our sc*GW* by benchmarking the band gaps of systems for which experimental data exist. Table II shows the sc*GW* band gaps of selected semiconductors and insulators calculated using sc*GW* and DFT as well as theoretical [40,122] and experimental literature data [113,115,117,119,121]. Zero-point renormalization (ZPR) due to electron-phonon coupling from existing calculations [114,116,118,120] is taken into account with the raw experimental data to facilitate the comparison. For DFT calculations, the PBE density functional [123] is used. All calculations, including PBE, are based on all-electron sfX2C1e-Coulomb Hamiltonians with a 6 × 6 × 6 **k** mesh.

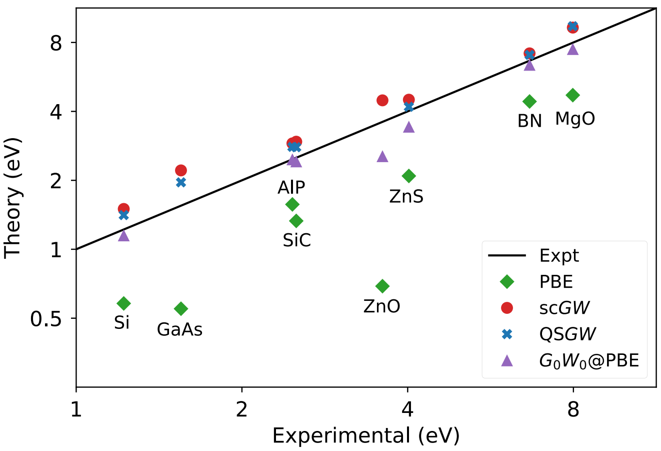

FIG. 8. Band gaps of selected semiconductors and insulators using all-electron sc*GW* with sfX2C1e-Coulomb Hamiltonian in comparison with PBE functional and experimental data with corrections from ZPR as shown in Table II. The QS*GW* band gaps taken from Ref. [44] and the G_0W_0 (based on PBE functional) band gaps taken from Ref. [68] are also shown.

The sfX2C1e Hamiltonian incorporates the exact scalar relativistic effects at the one-electron level while neglecting the spin-orbit interactions and all the relativistic corrections to the electron-electron interactions. All our results, denoted as “this work,” use all-electron triple-ζ bases optimized with respect to X2C Hamiltonians (x2c-TZVPall) [110]. For B and Mg atoms, the most diffuse *s* and *p* functions are removed to avoid linear dependencies.

As shown in Table II, the PBE functional significantly underestimates the experimental band gaps. This trend has been observed in other works (see, e.g., [40,68]). The many-body treatment from sc*GW* induces a gap widening. The largest difference between PBE and sc*GW* is observed for ZnO where the electron correlations from the transition metal *d* orbitals are strong. While an overall good agreement with experiment is observed, especially when corrections from ZPR are taken into account, sc*GW* systematically overestimates experimental band gaps as shown in Fig. 8. Since there is no starting-point dependence in this self-consistent approximation, we argue that this overestimation is due to the absence of high-order diagrams, such as “vertex corrections”

TABLE II. Band gaps (eV) of selected semiconductors and insulators calculated using all-electron sc*GW* with sfX2C1e-Coulomb Hamiltonian in comparison with the experimental data [113,115,117,119,121] with and without zero-point renormalization due to electron-phonon coupling [114,116,118,120].

System	PBE		sc*GW*		QS*GW*		G_0W_0@PBE		Expt.	Expt.+ZPR
		This work	Ref. [44]	Ref. [68]	Ref. [44]	Ref. [68]	Ref. [40]	Ref. [68]		
Si	0.58	1.50	1.55	2.18	1.41	1.49	1.08	1.15	1.17 [113]	1.22 [114]
SiC	1.33	2.95	2.89	3.29	2.79	2.88	2.42	2.42	2.40 [115]	2.51 [114]
GaAs	0.55	2.21	2.27		1.96				1.52 [113]	1.55 [116]
AlP	1.57	2.90	2.84	3.20	2.80	2.94	2.41	2.47	2.45 [117]	2.47 [118]
ZnO	0.69	4.47		4.92		4.29	2.91	2.55	3.44 [113]	3.60 [118]
ZnS	2.09	4.50	4.28	4.68	4.19	4.27	3.63	3.43	3.91 [113]	4.02 [118]
BN	4.42	7.17	7.06	7.67	7.06	7.50	6.41	6.39	6.40 [119]	6.66 [120]
MgO	4.71	9.29	9.31	9.53	9.42	9.58	7.43	7.49	7.83 [121]	7.98 [120]

in the *GW* self-energy and the polarizability that have been studied in the framework of bold diagrammatic expansions [42–44]. The effect of spin-orbit interaction is likely minor for the systems considered here. For instance, the strongest SOC effect is observed in GaAs, which exhibits a 0.1-eV band-gap narrowing in Ref. [124] and in our in-house two-component sc*GW* based on the the X2C1e-Coulomb Hamiltonian (not shown in this work). Additional uncertainties, such as those caused by finite-size effects and basis-set convergence, are only expected to result in small quantitative differences (see Secs. V C and V D).

We compare our results to some of the *GW* implementations [40,44,68,122] available in the literature, including sc*GW* [44,68], quasiparticle *GW* (QS*GW*) [68,122], and G_0W_0 [40,68] as shown in Table. II. Note that Ref. [44] is chosen since, to the best of our knowledge, it is the only fully self-consistent finite-temperature *GW* capable of calculating realistic solids in the LAPW basis. The basis set therefore marks the only major difference between the methodology of Ref. [44] and this work. The VASP implementation [68] is chosen because different variants of *GW* are reported in this work employing a projector augmented wave (PAW) basis. Calculations in Refs. [40,44,68,122] are performed using a $6 \times 6 \times 6$ **k** mesh (with the exception of $8 \times 8 \times 8$ **k** mesh for Si in Ref. [68]).

In general, good agreement is reached between our data and the sc*GW* data from Ref. [44]. This is somewhat expected since both the implementations are based on finite-temperature Green's function methods and executed on the imaginary axes exclusively. Both use no analytical continuation during the self-consistency loop, treat all electrons explicitly without the use of pseudopotentials, and define band gaps as the peak-to-peak distance of the spectral function. We attribute the main difference to Ref. [44] to the difference between the LAPW and GTO basis sets. Remaining small differences may therefore be attributed to finite-size effects, which include the treatment of the integral divergence, the basis-set error, and uncertainty in the analytical continuation procedure. For example, the larger deviations observed for ZnO and ZnS are consistent with the slower basis convergence for Zn atom as discussed in Sec. V D.

The comparison to VASP [68] is somewhat surprising. Overall, we found that the sc*GW* band gaps from VASP are generally larger than ours and those in Ref. [44]. Even for a simple system such as silicon, a 0.65-eV larger band gap is observed in Ref. [68]. Several aspects may be responsible for these differences. Numerically, VASP uses a PAW basis, which implies different basis-set convergence, and a different treatment of relativistic effects. In addition, the band gaps in VASP are defined as the quasiparticle gaps (evaluated in the last step of the algorithm). Given a self-energy calculated from sc*GW* without a quasiparticle approximation, the band gaps are determined by solving a quasiparticle equation in the HF canonical-orbital basis in the postprocessing step. In that case, all off-diagonal self-energies in the HF canonical-orbital basis are neglected, potentially resulting in overestimation of the gaps similar to the one observed in Ref. [85]. Another difference comes from the fact that the sc*GW* in VASP is the zero-temperature version unlike ours and the one in Ref. [44]. However, this difference should be negligible since the thermal excitations at the temperature ($\beta = 700$ a.u.) used here are not expected to affect the resulting spectral functions for systems studied in this work.

As for the two selected G_0W_0 results [40,68], a band-gap narrowing is observed in comparison to our sc*GW* results. Note that larger deviations between Refs. [40,68] appear in ZnO and ZnS. As Ref. [40] suggests, this is possibly due to different treatments for the core electrons. In calculations from Ref. [40], using all-electron GTO basis, the relativistic effects are completely ignored. Therefore, an additional band narrowing due to the scalar relativistic effect is expected. Note that, in spite of the numerical similarities due to the use of GTO basis sets, the core electrons are accurately treated with the scalar relativistic effects in the sfX2C1e Hamiltonian that is employed in our implementation.

Overall, as expected, sc*GW* results in larger band gaps when compared to G_0W_0. In G_0W_0, band gaps for semiconductors in the absence of transition metal elements are expected to be slightly underestimated, due to an error cancellation between the lack of self-consistency and the vertex corrections [42,44]. Such an error cancellation is missing both in sc*GW* and QS*GW*, and results in systematic overestimation of the band gap. With our data, we cannot confirm the observation of Ref. [68] that the overestimation is generally larger in sc*GW* than QS*GW*. Lastly, the good agreement between our results and those in Ref. [44] suggests that consistent and reproducible sc*GW* results independent of the basis set employed are now possible.

F. GPU performance

In this section, we demonstrate the performance of our implementation of sc*GW*. We will focus in particular on GPU kernels for the evaluation of the dynamical part of *GW* self-energy $(\tilde{\Sigma}^{GW})^{\mathbf{k}}_{ij}(\tau)$ which is the computational bottleneck in our implementation.

Figure 9 shows a profiling result of our GPU kernels on `summit` at the Oak Ridge Leadership Computing Facility. Each node consists of six Nvidia Volta V100s GPU cards. Antiferromagnetic (AFM) MnO is chosen as the test system with a $6 \times 6 \times 6$ **k** mesh centered at the Γ point, using the *gth-dzvp-molopt-sr* basis [125] and the *gth-pbe* pseudopotential [126]. Considering only inversion symmetry, the number of effective **q** points is 112. As shown in the first panel, the implementation exhibits almost ideal speedup consistently from a single GPU to 112 GPUs. The middle panel shows that around 70% of the theoretical peak performance is achieved consistently on up to 20 nodes. The third panel shows the total time of evaluating $(\tilde{\mathbf{\Sigma}}^{GW})^{\mathbf{k}}(\tau)$ per iteration. This value includes communication and I/O overhead. When the first layer of MPI parallelization over the **q** axis is fully exploited, one iteration of the *GW* self-energy evaluation takes ~3 min for this particular example.

VI. CONCLUSION

In this paper, we present implementation details and results for a fully self-consistent finite-temperature *GW* method in Gaussian Bloch orbitals for solids. The method employs finite-temperature Green's function on the imaginary

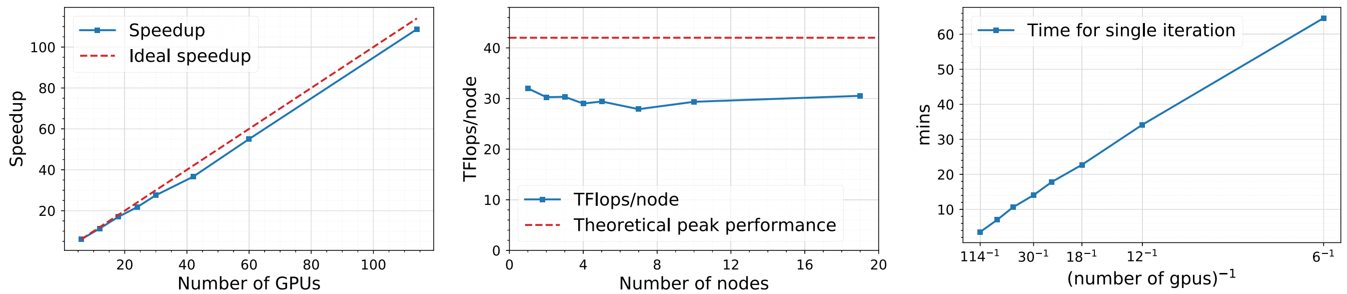

FIG. 9. Profiles of our GPU kernels for the evaluation of $(\tilde{\mathbf{\Sigma}}^{GW})^{\mathbf{k}}(\tau)$.

axis. The full self-consistency between Green's functions and self-energies guarantees that results are conserving and thermodynamically consistent. We do not employ the quasiparticle approximation, and all matrix elements of the *GW* self-energy at all Matsubara frequencies are evaluated explicitly. Instead of calculating a quasiparticle gap, single-particle excitation information is obtained directly from a spectral function, calculated using Green's function analytically continued from the imaginary to real frequency axis.

The finite-temperature self-consistent *GW* is computationally feasible due to various numerical developments employed in this work. In particular, Gaussian density fitting reduces the complexity of the sc*GW* algorithm to $O(N_\tau N_k^2 N_{\text{orb}}^2 N_{\text{aux}}^2)$. A compact representation of dynamical quantities using sparse sampling on imaginary axis with IR basis greatly reduces the memory requirement for dynamical quantities. More importantly, computational overheads of the Dyson-type equation for the bosonic function [the renormalized auxiliary function $\tilde{\mathbf{P}}^{\mathbf{q}}(i\Omega_n)$ in our case], as well as frequency integration along Matsubara axis, and Fourier transformation between two imaginary axes are negligible, compared to the evaluation of self-energy and polarization function (or the noninteracting auxiliary function $\tilde{\mathbf{P}}_0^{\mathbf{q}}$ in our case).

Moreover, we explore additional acceleration of the sc*GW* algorithm by migrating computationally intensive parts to a hybrid CPU/GPU platform. We demonstrate that this implementation scales to hundreds of GPUs, with good scalability on large systems.

Lastly, the Nevanlinna analytical continuation as a post-processing step makes the execution of *GW* exclusively on imaginary axis possible, by providing access to causal high-quality real-frequency data. We note that we did not yet explore additional optimizations, such as those based on the locality of the self-energy and basis functions [38,39,60,62], and optimum basis sets for Gaussian Bloch orbitals [127,128] as well as auxiliary bases in periodic systems. These are interesting options for future development that will facilitate large-unit-cell calculations for sc*GW*.

In our analysis of sc*GW*, we demonstrate its thermodynamic consistency in practice, investigate the finite-size effects with and without the head correction to the dynamical *GW* self-energy, and investigate the basis convergence of sc*GW* in Gaussian Bloch orbitals. Our benchmark employing band gaps of selected semiconductors and insulators shows consistent results when compared to the finite-temperature sc*GW* implementation reported in Ref. [44], where the numerical setup is substantially different. This agreement shows that sc*GW* is now routinely possible to reach high-quality results that are converged with respect to the basis-set and finite-size effects.

Without a quasiparticle approximation and with inclusion of the full self-consistency, our work provides a direct assessment of the fully self-consistent *GW* method when applied to realistic materials. Deviations from experimental data can be attributed to higher-order self-energy diagrams. These diagrams can be added either by employing vertex corrections [42,44,51] to the *GW* self-energy diagrams or by using embedding methods [5,8–16] on top of sc*GW*.

ACKNOWLEDGMENTS

This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. D.Z. and Ch.-N.Y. were supported by a grant from the Department of Energy under Award No. DE-SC0022198. This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research and Office of Basic Energy Sciences, Scientific Discovery through Advanced Computing (SciDAC) program under Award No. DE–SC0022198. E.G and S.I. were supported by the Simons Foundation via the Simons Collaboration on the Many-Electron problem.

APPENDIX A: INTEGRABLE DIVERGENCE TREATMENT

In this Appendix, we follow the procedure described in Ref. [94] and derive the finite-size corrections for both the HF exchange potential [Eq. (43)] and the dynamical *GW* self-energy [Eq. (42)] as shown in Sec. III D.

Considering a general numerical problem that involves an integral over the first Brillouin zone whose integrand contains a smooth function A and the bare Coulomb kernel expressed in the plane-wave basis ($\mathbf{G}$),

$$X = \frac{-1}{(2\pi)^3}\int_{\text{BZ}} d\mathbf{q}\sum_{\mathbf{G}}\frac{4\pi}{|\mathbf{q}+\mathbf{G}|^2}A(\mathbf{q},\mathbf{G}). \tag{A1}$$

Analytically, the integral is integrable although the integrand diverges as $1/\mathbf{q}^2$ at $\mathbf{G}=\mathbf{0}$ when $\mathbf{q}\to\mathbf{0}$. However, this singularity forbids a direct numerical evaluation using discretized $\mathbf{q}$ mesh. The numerical evaluation of Eq. (A1) directly resembles the evaluations of the HF exchange potential [Eq. (22)] and the dynamical *GW* self-energy [Eq. (23)]. We will show

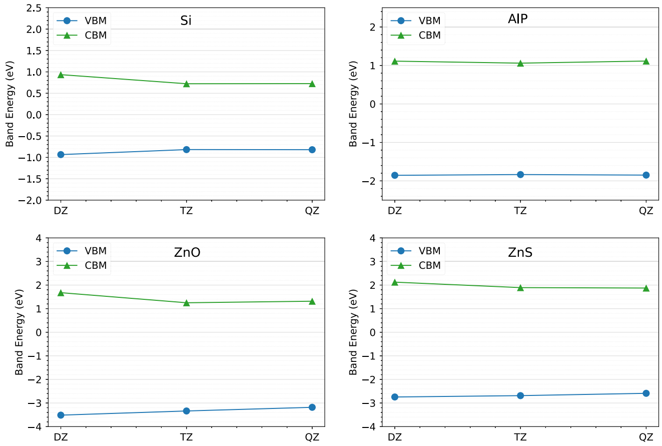

FIG. 10. The sc*GW* valence band maximum (VBM) and the conduction band minimum (CBM) of Si, AlP, ZnO, and ZnS calculated using different basis sets. A 5 × 5 × 5 **k** mesh is used for Si and AlP, and a 4 × 4 × 4 **k** mesh is used for ZnO and ZnS. In x2c-QZVPAll, the most diffuse *s* and *p* functions of Si, Al, and Zn are removed to avoid linear dependencies.

how to solve Eq. (A1) numerically and apply the same strategy to Eqs. (22) and (23).

We subtract and add the integrand of Eq. (A1) by an auxiliary function $F(\mathbf{q},\mathbf{G})$ that exhibits the same divergence $\sim 1/\mathbf{q}^2$ as $\mathbf{q}\rightarrow\mathbf{0}$ at $\mathbf{G}=\mathbf{0}$, i.e.,

$$X = \frac{-1}{2\pi^2}\int_{\mathrm{BZ}} d\mathbf{q}\sum_{\mathbf{G}}\left\{\frac{1}{|\mathbf{q}+\mathbf{G}|^2}A(\mathbf{q},\mathbf{G}) - F(\mathbf{q},\mathbf{G})A(\mathbf{0},\mathbf{0})\right\} + \frac{-1}{2\pi^2}\int_{\mathrm{BZ}} d\mathbf{q}\sum_{\mathbf{G}}F(\mathbf{q},\mathbf{G})A(\mathbf{0},\mathbf{0}). \tag{A2}$$

In the long-wavelength limit ($\mathbf{q}\rightarrow\mathbf{0}$), the singularity of bare Coulomb kernel is canceled by the one of the auxiliary function. The resulting smooth integrand in the curly brackets can therefore be evaluated accurately by a summation over a finite number of $\mathbf{q}$ points. On the other hand, the singularity has been transferred to the auxiliary function $F(\mathbf{q},\mathbf{G})$ in the last term in Eq. (A2) which can be evaluated analytically.

Approximating the integral by a discrete summation $\frac{1}{(2\pi)^3}\int_{\mathrm{BZ}} d\mathbf{q}\rightarrow\frac{1}{\Omega N_k}\sum_{\mathbf{q}}$ and rearranging different terms, Eq. (A2) can be expressed as

$$X \approx -\sum_{\mathbf{q}}\sum_{\mathbf{G}}\Phi(\mathbf{q},\mathbf{G})A(\mathbf{q},\mathbf{G}), \tag{A3}$$

where

$$\Phi(\mathbf{q},\mathbf{G}) = \begin{cases}\chi \text{ for } \mathbf{q}=\mathbf{G}=\mathbf{0},\\ \frac{1}{N_k\Omega}\frac{4\pi}{|\mathbf{q}+\mathbf{G}|^2} \text{ otherwise,}\end{cases} \tag{A4}$$

with

$$\chi = \frac{1}{2\pi^2}\int_{\mathrm{BZ}} d\mathbf{q}\sum_{\mathbf{G}}F(\mathbf{q},\mathbf{G}) - \frac{4\pi}{\Omega N_k}\sum_{\mathbf{q}}{\sum_{\mathbf{G}}}'F(\mathbf{q},\mathbf{G}) \tag{A5a}$$

$$= \frac{1}{2\pi^2}\int d\mathbf{Q}\,F(\mathbf{Q}) - \frac{4\pi}{\Omega N_k}\sum_{\mathbf{Q}\neq\mathbf{0}}F(\mathbf{Q}). \tag{A5b}$$

The summation with the prime symbol implies $\mathbf{G}=\mathbf{0}$ is not included when $\mathbf{q}=\mathbf{0}$. In the second line, we define $\mathbf{Q}=\mathbf{q}+\mathbf{G}$. The singularity at $\mathbf{G}=\mathbf{q}=\mathbf{0}$ is included in χ which is properly treated through analytical integration. The choice of $F(\mathbf{Q})$ will affect the smoothness of the integrand in the parentheses in Eq. (A2), and therefore affect the convergence with respect to the number of $\mathbf{q}$ points to approximate the integral as shown in Eq. (A3). In this work, the auxiliary function proposed in Ref. [94] is adopted which makes χ the supercell Madelung constant.

Both the HF exchange potential and the dynamical *GW* self-energy can be written in a similar format as in Eq. (A1). The HF exchange potential in the plane-wave basis reads as

$$K^{\mathbf{k}}_{i\sigma,j\sigma'} = \frac{-1}{(2\pi)^3}\int_{\mathrm{BZ}} d\mathbf{q}\sum_{\mathbf{G}}\sum_{ab}\frac{4\pi}{|\mathbf{q}+\mathbf{G}|^2} \times \rho^{\mathbf{k-qk}*}_{a\;\;i}(\mathbf{G})\gamma^{\mathbf{k-q}}_{a\sigma,b\sigma'}\rho^{\mathbf{k-qk}}_{b\;\;j}(\mathbf{G}) \tag{A6}$$

with

$$A(\mathbf{q},\mathbf{G};\mathbf{k},i\sigma,j\sigma') = \sum_{ab}\rho^{\mathbf{k-qk}*}_{a\;\;i}(\mathbf{G})\gamma^{\mathbf{k-q}}_{a\sigma,b\sigma'}\rho^{\mathbf{k-qk}}_{b\;\;j}(\mathbf{G}). \tag{A7}$$

The corresponding finite-size correction reads as

$$(\Delta^{\mathrm{HF}})^{\mathbf{k}}_{i\sigma,j\sigma'} = -\chi\sum_{ab}S^{\mathbf{k}}_{ia}\gamma^{\mathbf{k}}_{a\sigma,b\sigma'}S^{\mathbf{k}}_{bj}. \tag{A8}$$

Note that

$$\rho^{\mathbf{k-qk}}_{i\;\;j}(\mathbf{G})\Big|_{\mathbf{q}=\mathbf{G}=\mathbf{0}} = \int_{\Omega} d\mathbf{r}\,\rho^{\mathbf{k-qk}}_{i\;\;j}(\mathbf{r})e^{-i(\mathbf{q}+\mathbf{G})}\Big|_{\mathbf{q}=\mathbf{G}=\mathbf{0}} = \int_{\Omega} d\mathbf{r}\,\rho^{\mathbf{kk}}_{ij} = S^{\mathbf{k}}_{ij}. \tag{A9}$$

Similarly, we express the dynamical *GW* self-energy in the plane-wave basis,

$$(\tilde{\Sigma}^{GW})^{\mathbf{k}}_{i\sigma,j\sigma'}(\tau) = \frac{-1}{(2\pi)^3}\int_{\mathrm{BZ}} d\mathbf{q}\sum_{\mathbf{GG'}}\sum_{ab}G^{\mathbf{k-q}}_{a\sigma,b\sigma'}(\tau), \tag{A10}$$

$$\rho^{\mathbf{k-qk}*}_{a\;\;i}(\mathbf{G})\frac{\sqrt{4\pi}}{|\mathbf{q}+\mathbf{G}|}\left(\epsilon^{\mathbf{q},-1}_{\mathbf{GG'}}(\tau)-\delta_{\mathbf{GG'}}\right)\frac{\sqrt{4\pi}}{|\mathbf{q}+\mathbf{G'}|}\rho^{\mathbf{k-qk}}_{b\;\;j}(\mathbf{G'}).$$

Since we are only interested in the correction to the head of $(\tilde{\Sigma}^{GW})^{\mathbf{k}}$ which corresponds to $\mathbf{G}=\mathbf{G'}=\mathbf{0}$, we consider only the diagonal terms in the plane-wave basis, i.e., $\mathbf{G}=\mathbf{G'}$:

$$(\tilde{\Sigma}^{GW}_{\mathrm{diag}})^{\mathbf{k}}_{i\sigma,j\sigma'}(\tau) = \frac{-1}{(2\pi)^3}\int_{\mathrm{BZ}} d\mathbf{q}\sum_{\mathbf{G}}\sum_{ab}\frac{4\pi}{|\mathbf{q}+\mathbf{G}|^2} \times [\epsilon^{\mathbf{q},-1}_{\mathbf{GG}}(\tau)-1]\rho^{\mathbf{k-qk}*}_{a\;\;i}(\mathbf{G})G^{\mathbf{k-q}}_{a\sigma,b\sigma'}(\tau)\rho^{\mathbf{k-qk}}_{b\;\;j}(\mathbf{G}), \tag{A11}$$

with

$$A(\mathbf{q},\mathbf{G};\mathbf{k},i\sigma,j\sigma',\tau) = \sum_{ab}[\epsilon^{\mathbf{q},-1}_{\mathbf{GG}}(\tau)-1] \times \rho^{\mathbf{k-qk}*}_{a\;\;i}(\mathbf{G})G^{\mathbf{k-q}}_{a\sigma,b\sigma'}(\tau)\rho^{\mathbf{k-qk}}_{b\;\;j}(\mathbf{G}). \tag{A12}$$

The head correction then reads as

$$(\Delta^{GW})^{\mathbf{k}}_{i\sigma,j\sigma'}(\tau) = -\chi[\epsilon^{\mathbf{0},-1}_{\mathbf{00}}(\tau)-1]S^{\mathbf{k}}_{ia}G^{\mathbf{k}}_{a\sigma,b\sigma'}(\tau)S^{\mathbf{k}}_{bj}. \tag{A13}$$

APPENDIX B: BASIS CONVERGENCE OF BAND ENERGIES

Figure 10 shows the basis convergence of the valence band maximum (VBM) and the conduction band minimum (CBM) calculated using sc*GW*. Similar to Sec. V D, the basis sets are systematically enlarged from x2c-SV(P)all (DZ) to x2c-TZVPall (TZ), and finally to x2c-QZVPall (QZ) basis set. Both VBM and CBM show similar convergence behavior compared to Table I. This suggests that the basis-set convergence of band gaps observed in Sec. V D is not due to fortunate error cancellation.

[1] L. Hedin, Phys. Rev. **139**, A796 (1965).

[2] W. Kohn and L. J. Sham, Phys. Rev. **140**, A1133 (1965).

[3] F. Aryasetiawan and O. Gunnarsson, Rep. Prog. Phys. **61**, 237 (1998).

[4] G. Onida, L. Reining, and A. Rubio, Rev. Mod. Phys. **74**, 601 (2002).

[5] A. Georges, G. Kotliar, W. Krauth, and M. J. Rozenberg, Rev. Mod. Phys. **68**, 13 (1996).

[6] P. Sun and G. Kotliar, Phys. Rev. B **66**, 085120 (2002).

[7] S. Biermann, F. Aryasetiawan, and A. Georges, Phys. Rev. Lett. **90**, 086402 (2003).

[8] G. Kotliar, S. Y. Savrasov, K. Haule, V. S. Oudovenko, O. Parcollet, and C. A. Marianetti, Rev. Mod. Phys. **78**, 865 (2006).

[9] D. Zgid and E. Gull, New J. Phys. **19**, 023047 (2017).

[10] A. A. Kananenka, E. Gull, and D. Zgid, Phys. Rev. B **91**, 121111(R) (2015).

[11] S. Iskakov, C.-N. Yeh, E. Gull, and D. Zgid, Phys. Rev. B **102**, 085105 (2020).

[12] C.-N. Yeh, S. Iskakov, D. Zgid, and E. Gull, Phys. Rev. B **103**, 195149 (2021).

[13] C.-N. Yeh, A. Shee, S. Iskakov, and D. Zgid, Phys. Rev. B **103**, 155158 (2021).

[14] F. Nilsson, L. Boehnke, P. Werner, and F. Aryasetiawan, Phys. Rev. Mater. **1**, 043803 (2017).

[15] F. Petocchi, F. Nilsson, F. Aryasetiawan, and P. Werner, Phys. Rev. Res. **2**, 013191 (2020).

[16] L. Boehnke, F. Nilsson, F. Aryasetiawan, and P. Werner, Phys. Rev. B **94**, 201106(R) (2016).

[17] T. Zhu and G. K.-L. Chan, Phys. Rev. X **11**, 021006 (2021).

[18] G. Baym and L. P. Kadanoff, Phys. Rev. **124**, 287 (1961).

[19] G. Baym, Phys. Rev. **127**, 1391 (1962).

[20] G. Kresse and J. Furthmüllerb, Comput. Mater. Sci. **6**, 15 (1996).

[21] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, A. D. Corso, S. de Gironcoli, S. Fabris, G. Fratesi, R. Gebauer, U. Gerstmann, C. Gougoussis, A. Kokalj, M. Lazzeri, L. Martin-Samos *et al.*, J. Phys.: Condens. Matter **21**, 395502 (2009).

[22] R. Dovesi, A. Erba, R. Orlando, C. M. Zicovich-Wilson, B. Civalleri, L. Maschio, M. Rérat, S. Casassa, J. Baima, S. Salustro, and B. Kirtman, WIREs Computat. Mol. Sci. **8**, e1360 (2018).

[23] Q. Sun, X. Zhang, S. Banerjee, P. Bao, M. Barbry, N. S. Blunt, N. A. Bogdanov, G. H. Booth, J. Chen, Z.-H. Cui, J. J. Eriksen, Y. Gao, S. Guo, J. Hermann, M. R. Hermes, K. Koh, P. Koval, S. Lehtola, Z. Li, J. Liu *et al.*, J. Chem. Phys. **153**, 024109 (2020).

[24] T. D. Kühne, M. Iannuzzi, M. D. Ben, V. V. Rybkin, P. Seewald, F. Stein, T. Laino, R. Z. Khaliullin, O. Schütt, F. Schiffmann, D. Golze, J. Wilhelm, S. Chulkov, M. H. Bani-Hashemian, V. Weber, U. Borštnik, M. Taillefumier, A. S. Jakobovits, A. Lazzaro, H. Pabst *et al.*, J. Chem. Phys. **152**, 194103 (2020).

[25] P. Blaha, K. Schwarz, F. Tran, R. Laskowski, G. K. H. Madsen, and L. D. Marks, J. Chem. Phys. **152**, 074101 (2020).

[26] T. Rangel, M. D. 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, Comput. Phys. Commun. **255**, 107242 (2020).

[27] M. J. van Setten, F. Caruso, S. Sharifzadeh, X. Ren, M. Scheffler, F. Liu, J. Lischner, L. Lin, J. R. Deslippe, S. G. Louie, C. Yang, F. Weigend, J. B. Neaton, F. Evers, and P. Rinke, J. Chem. Theory Comput. **11**, 5665 (2015).

[28] E. Maggio, P. Liu, M. J. van Setten, and G. Kresse, J. Chem. Theory Comput. **13**, 635 (2017).

[29] F. Bruneval, N. Dattani, and M. J. van Setten, Front. Chem. **9**, 749779 (2021).

[30] M. S. Hybertsen and S. G. Louie, Phys. Rev. Lett. **55**, 1418 (1985).

[31] M. S. Hybertsen and S. G. Louie, Phys. Rev. B **34**, 5390 (1986).

[32] R. W. Godby, M. Schlüter, and L. J. Sham, Phys. Rev. B **37**, 10159 (1988).

[33] C. Friedrich, S. Blügel, and A. Schindlmayr, Phys. Rev. B **81**, 125102 (2010).

[34] J. Klimeš, M. Kaltak, and G. Kresse, Phys. Rev. B **90**, 075125 (2014).

[35] M. Govoni and G. Galli, J. Chem. Theory Comput. **11**, 2680 (2015).

[36] J. Wilhelm, M. Del Ben, and J. Hutter, J. Chem. Theory Comput. **12**, 3623 (2016).

[37] J. Wilhelm and J. Hutter, Phys. Rev. B **95**, 235123 (2017).

[38] J. Wilhelm, D. Golze, L. Talirz, J. Hutter, and C. A. Pignedoli, J. Phys. Chem. Lett. **9**, 306 (2018).

[39] J. Wilhelm, P. Seewald, and D. Golze, J. Chem. Theory Comput. **17**, 1662 (2021).

[40] T. Zhu and G. K.-L. Chan, J. Chem. Theory Comput. **17**, 727 (2021).

[41] T. Kotani, M. van Schilfgaarde, and S. V. Faleev, Phys. Rev. B **76**, 165106 (2007).

[42] A. Grüneis, G. Kresse, Y. Hinuma, and F. Oba, Phys. Rev. Lett. **112**, 096401 (2014).

[43] A. L. Kutepov, Phys. Rev. B **94**, 155101 (2016).

[44] A. L. Kutepov, Phys. Rev. B **95**, 195120 (2017).

[45] X. Zhu and S. G. Louie, Phys. Rev. B **43**, 14142 (1991).

[46] O. Zakharov, A. Rubio, X. Blase, M. L. Cohen, and S. G. Louie, Phys. Rev. B **50**, 10780 (1994).

[47] M. Shishkin and G. Kresse, Phys. Rev. B **75**, 235102 (2007).

[48] S. V. Faleev, M. van Schilfgaarde, and T. Kotani, Phys. Rev. Lett. **93**, 126406 (2004).

[49] M. van Schilfgaarde, T. Kotani, and S. Faleev, Phys. Rev. Lett. **96**, 226402 (2006).

[50] F. Bruneval, N. Vast, and L. Reining, Phys. Rev. B **74**, 045102 (2006).

[51] M. Shishkin, M. Marsman, and G. Kresse, Phys. Rev. Lett. **99**, 246403 (2007).

[52] L. Boehnke, H. Hafermann, M. Ferrero, F. Lechermann, and O. Parcollet, Phys. Rev. B **84**, 075145 (2011).

[53] M. Kaltak, J. Klimeš, and G. Kresse, J. Chem. Theory Comput. **10**, 2498 (2014).

[54] H. Shinaoka, J. Otsuki, M. Ohzeki, and K. Yoshimi, Phys. Rev. B **96**, 035147 (2017).

[55] E. Gull, S. Iskakov, I. Krivenko, A. A. Rusakov, and D. Zgid, Phys. Rev. B **98**, 075127 (2018).

[56] X. Dong, D. Zgid, E. Gull, and H. U. R. Strand, J. Chem. Phys. **152**, 134107 (2020).

[57] M. Kaltak and G. Kresse, Phys. Rev. B **101**, 205145 (2020).

[58] J. Li, M. Wallerberger, N. Chikano, C.-N. Yeh, E. Gull, and H. Shinaoka, Phys. Rev. B **101**, 035144 (2020).

[59] J. Kaye, K. Chen, and O. Parcollet, Discrete lehmann representation of imaginary time green's functions, Phys. Rev. B **105**, 235115 (2022).

[60] M. Kaltak, J. Klimeš, and G. Kresse, Phys. Rev. B **90**, 054115 (2014).

[61] P. Liu, M. Kaltak, J. Klimeš, and G. Kresse, Phys. Rev. B **94**, 165109 (2016).

[62] A. Kutepov, Comput. Phys. Commun. **257**, 107502 (2020).

[63] A. Stan, N. E. Dahlen, and R. van Leeuwen, J. Chem. Phys. **130**, 114105 (2009).

[64] F. Caruso, P. Rinke, X. Ren, M. Scheffler, and A. Rubio, Phys. Rev. B **86**, 081102(R) (2012).

[65] P. Koval, D. Foerster, and D. Sánchez-Portal, Phys. Rev. B **89**, 155417 (2014).

[66] W.-D. Schöne and A. G. Eguiluz, Phys. Rev. Lett. **81**, 1662 (1998).

[67] A. Kutepov, K. Haule, S. Y. Savrasov, and G. Kotliar, Phys. Rev. B **85**, 155129 (2012).

[68] M. Grumet, P. Liu, M. Kaltak, J. Klimeš, and G. Kresse, Phys. Rev. B **98**, 155143 (2018).

[69] H. Jiang and P. Blaha, Phys. Rev. B **93**, 115203 (2016).

[70] M.-Y. Zhang and H. Jiang, Phys. Rev. B **100**, 205123 (2019).

[71] R. Gómez-Abal, X. Li, M. Scheffler, and C. Ambrosch-Draxl, Phys. Rev. Lett. **101**, 106404 (2008).

[72] O. Vahtras, J. Almlöf, and M. Feyereisen, Chem. Phys. Lett. **213**, 514 (1993).

[73] H. F. Schurkus and C. Ochsenfeld, J. Chem. Phys. **144**, 031101 (2016).

[74] I. Duchemin, J. Li, and X. Blase, J. Chem. Theory Comput. **13**, 1199 (2017).

[75] J. Fei, C.-N. Yeh, and E. Gull, Phys. Rev. Lett. **126**, 056402 (2021).

[76] C.-N. Yeh, A. Shee, Q. Sun, E. Gull, and D. Zgid, Phys. Rev. B **106**, 085121 (2022).

[77] J. McClain, Q. Sun, G. K.-L. Chan, and T. C. Berkelbach, J. Chem. Theory Comput. **13**, 1209 (2017).

[78] S. F. Boys and A. C. Egerton, Proc. R. Soc. London A **200**, 542 (1950).

[79] L. Boman, H. Koch, and A. S. de Merás, J. Chem. Phys. **129**, 134107 (2008).

[80] H.-J. Werner, F. R. Manby, and P. J. Knowles, J. Chem. Phys. **118**, 8149 (2003).

[81] X. Ren, P. Rinke, V. Blum, J. Wieferink, A. Tkatchenko, A. Sanfilippo, K. Reuter, and M. Scheffler, New J. Phys. **14**, 053020 (2012).

[82] H.-Z. Ye and T. C. Berkelbach, J. Chem. Phys. **154**, 131104 (2021).

[83] G. L. Stoychev, A. A. Auer, and F. Neese, J. Chem. Theory Comput. **13**, 554 (2017).

[84] J. G. Hill, Int. J. Quantum Chem. **113**, 21 (2013).

[85] J. Fei, C.-N. Yeh, D. Zgid, and E. Gull, Phys. Rev. B **104**, 165111 (2021).

[86] J. M. Luttinger and J. C. Ward, Phys. Rev. **118**, 1417 (1960).

[87] B. Holm and F. Aryasetiawan, Phys. Rev. B **62**, 4858 (2000).

[88] A. Fetter and J. Walecka, *Quantum Theory of Many-Particle Systems*, Dover Books on Physics (Dover, Mineola, NY, 2003).

[89] N. E. Dahlen, R. van Leeuwen, and U. von Barth, Phys. Rev. A **73**, 012511 (2006).

[90] S. Iskakov, A. A. Rusakov, D. Zgid, and E. Gull, Phys. Rev. B **100**, 085112 (2019).

[91] A. R. Welden, A. A. Rusakov, and D. Zgid, J. Chem. Phys. **145**, 204106 (2016).

[92] F. Gygi and A. Baldereschi, Phys. Rev. B **34**, 4405 (1986).

[93] J. Paier, R. Hirschl, M. Marsman, and G. Kresse, J. Chem. Phys. **122**, 234102 (2005).

[94] P. Broqvist, A. Alkauskas, and A. Pasquarello, Phys. Rev. B **80**, 085114 (2009).

[95] M. Shishkin and G. Kresse, Phys. Rev. B **74**, 035101 (2006).

[96] F. Hüser, T. Olsen, and K. S. Thygesen, Phys. Rev. B **87**, 235132 (2013).

[97] P. Pokhilko and D. Zgid, J. Chem. Phys. **155**, 024101 (2021).

[98] D. G. Anderson, J. ACM **12**, 547 (1965).

[99] P. Pulay, Chem. Phys. Lett. **73**, 393 (1980).

[100] P. Pulay, J. Comput. Chem. **3**, 556 (1982).

[101] H. F. Walker and P. Ni, SIAM J. Numer. Anal. **49**, 1715 (2011).

[102] M. Véril, P. Romaniello, J. A. Berger, and P.-F. Loos, J. Chem. Theory Comput. **14**, 5220 (2018).

[103] A. Förster and L. Visscher, Front. Chem. **9**, 736591 (2021).

[104] P. Pokhilko, C.-N. Yeh, and D. Zgid, J. Chem. Phys. **156**, 094101 (2022).

[105] N. Chikano, K. Yoshimi, J. Otsuki, and H. Shinaoka, Comput. Phys. Commun. **240**, 181 (2019).

[106] P.-O. Löwdin, *Advances in Quantum Chemistry, Volume 5* (Elsevier, Amsterdam, 1970), pp. 185–199.

[107] N. Marzari, A. A. Mostofi, J. R. Yates, I. Souza, and D. Vanderbilt, Rev. Mod. Phys. **84**, 1419 (2012).

[108] A. Abdelfattah, M. Baboulin, V. Dobrev, J. Dongarra, C. Earl, J. Falcou, A. Haidar, I. Karlin, T. Kolev, I. Masliah, and S. Tomov, High-Performance Tensor Contractions for GPUs, Technical Report No. UT-EECS-16-738, 2016 (unpublished).

[109] A. Haidar, T. Dong, S. Tomov, P. Luszczek, and J. Dongarra, *ISC High Performance* (Springer, Frankfurt, 2015).

[110] P. Pollak and F. Weigend, J. Chem. Theory Comput. **13**, 3696 (2017).

[111] Y. J. Franzke, L. Spiske, P. Pollak, and F. Weigend, J. Chem. Theory Comput. **16**, 5658 (2020).

[112] W. Gao, W. Xia, Y. Wu, W. Ren, X. Gao, and P. Zhang, Phys. Rev. B **98**, 045108 (2018).

[113] C. Kittel, *Introduction to Solid State Physics*, 8th ed. (Wiley, Nashville, TN, 2004).

[114] B. Monserrat and R. J. Needs, Phys. Rev. B **89**, 214304 (2014).

[115] P. Y. Yu and M. Cardona, *Fundamentals of Semiconductors*, 4th ed. (Springer, Berlin, 2010).

[116] G. Antonius, S. Poncé, P. Boulanger, M. Côté, and X. Gonze, Phys. Rev. Lett. **112**, 215501 (2014).

[117] S. M. Sze, *Physics of Semiconductor Devices* (Wiley, New York, 1981).

[118] M. Cardona and M. L. W. Thewalt, Rev. Mod. Phys. **77**, 1173 (2005).

[119] O. Madelung, *Semiconductors: Data Handbook* (Springer, Berlin, 2004).

[120] G. Antonius, S. Poncé, E. Lantagne-Hurtubise, G. Auclair, X. Gonze, and M. Côté, Phys. Rev. B **92**, 085137 (2015).

[121] R. Whited, C. J. Flaten, and W. Walker, Solid State Commun. **13**, 1903 (1973).

[122] A. Kutepov, V. Oudovenko, and G. Kotliar, Comput. Phys. Commun. **219**, 407 (2017).

[123] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).

[124] W. Chen and A. Pasquarello, Phys. Rev. B **92**, 041115(R) (2015).

[125] J. VandeVondele and J. Hutter, J. Chem. Phys. **127**, 114105 (2007).

[126] S. Goedecker, M. Teter, and J. Hutter, Phys. Rev. B **54**, 1703 (1996).

[127] Y. Zhou, E. Gull, and D. Zgid, J. Chem. Theory Comput. **17**, 5611 (2021).

[128] H.-Z. Ye and T. C. Berkelbach, J. Chem. Theory Comput. **18**, 1595 (2022).