Accurate Quasiparticle Spectra from Self-Consistent GW Calculations with Vertex Corrections

M. Shishkin, M. Marsman, and G. Kresse
Faculty of Physics, Universität Wien and Center for Computational Materials Science, Sensengasse 8/12, A-1090 Wien, Austria

Self-consistent GW calculations, maintaining only the quasiparticle part of the Green’s function $G$, are reported for a wide class of materials, including small gap semiconductors and large gap insulators. We show that the inclusion of the attractive electron-hole interaction via an effective nonlocal exchange correlation kernel is required to obtain accurate band gaps in the framework of self-consistent GW calculations. If these are accounted for via vertex corrections in $W$, the band gaps are found to be within a few percent of the experimental values.

DOI: 10.1103/PhysRevLett.99.246403
PACS numbers: 71.10.–w, 71.20.Nr

The accurate prediction of band gaps is a long-standing challenge to computational materials science. This concerns particularly extended systems, such as bulk semiconductors and insulators, for which the widely successful density functional theory (DFT) method generally yields much too small band gaps, and in specific cases, even the wrong band order [1]. The most common approach to deal with this problem is the $GW$ approximation [2–4], which is nowadays usually applied perturbatively on top of a computationally less demanding scheme relying typically on DFT wave functions and DFT one-electron energies. The corresponding approximation is referred to as $G_0W_0$ [3–5]. Although this approximation was originally vastly successful, recent calculations clearly demonstrate that part of the success was related to a fortuitous cancellation of errors between systematic shortcomings introduced by the $G_0W_0$ approximation itself and technical errors incurred by the pseudopotential approximation, on which many $G_0W_0$ calculations were based. This understanding originates mainly from accurate all-electron calculations relying on full potential methods that allow a consistent and accurate treatment of core and valence electrons on the same footing [6–10]. As shown recently the $G_0W_0$ approximation yields particularly large errors for systems with shallow $d$ states such as GaAs, ZnO, and CdS [10,11]. The problem can be partly remedied by iterating the one-electron energies in the Green’s function $G$ until self-consistency is reached, but even this approach underestimates band gaps in the aforementioned materials by up to 15% (e.g. Ref. [10]). This has prompted Faleev, van Schilfgaarde, and Kotani to suggest a self-consistent way of approximating $G$ in order to determine the quasiparticle peaks in the subsequent $GW$ calculations. This method one needs to determine all solutions $n$ pairs of six, which is responsible for the excitonic features in the adsorption spectra [14]. Unfortunately, including these effects, which is usually done via vertex corrections in the $GW$ approximation, is not an easy matter. Here a recent development by Reining and co-workers comes in, who suggested to recast this computationally demanding term into an effective nonlocal exchange correlation kernel $f_{sc}(r,r')$ [15–18]. We demonstrate that the inclusion of these many-body correlation effects allows for the prediction of band gaps with very good precision.

Here we present scGW calculations including a parameter-free $f_{sc}$ for small gap and large gap materials, comprising typical semiconductors (Ge, Si, SiC, GaAs, GaN, ZnO, ZnS, CdS, and AlP), insulators (C, BN, MgO, LiF), and noble gas solids (Ar, Ne). The calculations are performed using the projector-augmented-wave method and the VASP code. The first step in our calculations consists of a hybrid Hartree-Fock density functional calculation using the HSE03 functional [19]. This functional mixes 25% of the nonlocal Hartree-Fock exchange to an otherwise semilocal functional and predicts reasonable band gaps for semiconductors, although the gaps remain seriously underestimated for ZnO and rare gas solids [20]. The final results are independent of the choice of the initial functional, but a good starting point allows us to cut down on the number of iterations required to achieve self-consistency in the subsequent $GW$ calculations.

The scGW calculations are performed using a variant of the method originally suggested by van Schilfgaarde et al. [12,14]. In order to determine the quasiparticle peaks in the $GW$ method one needs to determine all solutions $n = 1,\ldots,N_{states}$ (filled and unoccupied bands) of a nonlinear one-electron–like Schrödinger equation (the Brillouin zone index has been dropped for brevity):

$$[T + V + \Sigma(E_n)]\psi_n = E_n\psi_n,$$

(1)

where $T$ is the kinetic energy operator, the local potential $V$
corresponds to the electrostatic potential of the nuclei and electrons, and the self-energy $\Sigma(r, r', E)$ is energy dependent and nonlinear. In principle, the solutions can be determined by iteration, and we proceed by a linearization around some reference energy $E_n$ for state $n$ and iteration $i$:

$$
\left( T + V + \Sigma(E_n^i) + \frac{\partial \Sigma(E_n^i)}{\partial E_n^i} (E - E_n^i) \right) |\psi_n^i\rangle = E |\psi_n^i\rangle.
$$

With the abbreviation $\xi(E_n^i) = \frac{\partial \Sigma(E_n^i)}{\partial E_n^i}$, the equation can be recast as a generalized and non-Hermitian (but linear) eigenvalue problem:

$$
\left( T + V + \Sigma(E_n^i) - \xi(E_n^i) E_n^i \right) |\psi_n\rangle = E \left[ 1 - \xi(E_n^i) \right] |\psi_n\rangle.
$$

The fact that the quasiparticle peaks are not one-normalized is reflected by the presence of the overlap operator $S$: as a consequence of $S$, the quasiparticle peaks usually contain only 75% of the norm, whereas the rest of the intensity is found in satellites (lacking after linearization). Unfortunately, no unique method to map this problem onto a corresponding Hermitian eigenvalue problem exists, but one possible route is to determine the Hamiltonian and overlap operator in a suitable basis set.

$$
\text{id} E_n^i - \xi(E_n^i) E_n^i |\psi_n\rangle = E \left[ 1 - \xi(E_n^i) \right] |\psi_n\rangle.
$$

The second important issue is the inclusion of the electron-hole interaction in the dielectric matrices. These matrices are required in order to determine the screened Coulomb kernel $W = e^{-1} V$, which enters in the previously used self-energy $\Sigma = iGW$. In time-dependent density functional theory, the full polarizability $\chi$ is given by the Dyson-like equation

$$
\chi = [1 - \chi_0 (v + f_{xc})]^{-1} \chi_0.
$$

The kernel $\chi_0$ is the independent particle polarizability and $v$ the Coulomb kernel $1/|r - r'|$. We note that $\chi_0$, $\chi$, and $\sigma$ are frequency dependent nonlocal matrices that are expanded in a plane wave basis set including all plane waves with kinetic energies of up to 200 eV in the present calculations (for details see Ref. [8]). For local density functionals, $f_{xc}$ is simply the second variation of the exchange correlation energy functional with respect to the density:

$$
f_{xc}(r, r') = \delta^2 E_{xc}[\rho(r)] / \delta \rho(r) \delta \rho(r').
$$

Because of the locality of the exchange correlation term, it lacks any long range interactions and inclusion or omission of $f_{xc}$ in Eq. (4) changes the static screening properties only little in the long wave length limit. In a proper quasiparticle description, as applied in the present framework, the inclusion of a consistent exchange correlation kernel, however, modifies the dielectric properties significantly, as we will show below. To determine a consistent many-body $f_{xc}(r, r', \omega)$ in the $GW$ framework, we follow the recent suggestions of Reinig and Del Sole [15–18]. Here $f_{xc}$ is calculated only at zero frequency, since tests for the adsorption spectrum show that this is an excellent approximation; e.g., the exact time-dependent adsorption spectra can be recovered using a frequency independent $f_{xc}(r, r')$ (see for instance Ref. [17]). In the present case we have included the direct Coulomb-like electron-hole interaction, as well as the exchange-like contributions coupling the resonant and antiresonant part of the response. The present $scGW$ calculations are performed including $f_{xc}$ right from the first iteration. Alongside the wave functions and eigenvalues [see Eq. (3)], $f_{xc}$ is updated until self-consistency is reached. The kernel $f_{xc}$ is calculated initially for the HSE03 functional and then using the static, screened Coulomb kernel $W(\omega \rightarrow 0)$, as commonly done in Bethe-Salpeter calculations [21]. In the tables below we will report calculations without $f_{xc}$ (``RPA''—random phase approximation) and including $f_{xc}$ (labeled ``e-h'' for electron-hole).

Figure 1 reports the band gaps for a conventional semilocal DFT functional (here and in the following the acronym DFT is used for calculations using the Perdew Burke Ernzerhof [22] functional) and for the scGW with and without $f_{xc}$ in $W$. The vertex corrections are included via Eq. (4) only, but neglected in the self-energy (see below). All materials were considered in their zero temperature ground state structure, except for ZnO and BN, which were investigated in the zinc-blend structure. The Brillouin zone was sampled using $8 \times 8 \times 8$ points, although results are usually already converged using $6 \times 6 \times 6$ points (for test cases $10 \times 10 \times 10$ points were used). The total number of
valence and conduction band states was set to 150 for all materials. We estimate the technical accuracy of the calculations to be about 2% [10]. Finally, it is noted that the experimental temperature dependence is often significant, which test calculations indicate to be not related to thermal expansion, but more likely electron-phonon couplings not accounted for in the present framework. One thus has to compare with gaps measured at low temperature wherever possible (underlined experimental values in Table I).

As emphasized before, the scGW calculations overestimate the band gaps, with fractional errors being largest for GaAs and Ge, as already noticed by Chantis et al. [13]. To a large extent the inclusion of vertex corrections remedies this problem, with the mean absolute relative error becoming smaller than 4%. On average the band gaps are still slightly too large, even including the electron-hole interaction, with the notable exceptions of ZnS and ZnO, where the gaps are underestimated by about 6%. We furthermore observe that the d band position is roughly 1 eV too shallow for all materials with semicore 3d states (this includes GaAs and GaN; see exemplary results for ZnO and GaAs in Table I) and that self-consistency in the wave functions has only a small effect on the d band position. This might be the result of neglecting vertex corrections in the calculation of the self-energy (see Ref. [23]), and we will return to this point later. We furthermore believe that the incorrect 3d position is the main reason for too small band gaps in ZnS and ZnO.

The ion-clamped macroscopic dielectric constants $\varepsilon_m$,

$$\varepsilon_m^{-1} = \lim_{\omega \to 0, \mathbf{q} \to 0} \varepsilon^{-1}(\mathbf{q}, \mathbf{q}, \omega),$$

are reported in Tables I and II. Agreement with experiment is again very good at the present level of theory. The present results allow for an assessment of why the commonly used random phase approximation for the calculation of $W$ was so successful in combination with DFT wave functions ($W_{0}^{\text{DFT}}$). In DFT, the band gaps are way too small, but if $f_{xc}$ is neglected in Eq. (4), the resultant dielectric constants agree well with the experimental values and the present high level theory (see Table II). If the wave functions and eigenvalues are updated in $G$ (but not $W$) until self-consistency is reached, almost the same band gaps as in the scGW calculations are obtained. This indicates that $W_{0}^{\text{DFT}}$ is very similar to the self-consistently determined $W$ with vertex corrections. Only for ZnO and GaAs the final band gaps are smaller, which is clearly related to the strong overestimation of the static dielectric constant in DFT-RPA (see Table II). The success of the $W_{0}^{\text{DFT}}$ approximation has to result from a cancellation of errors—too small DFT band gaps and neglect of electron-hole interactions—but this cancellation seems to be rather universal for gradient corrected DFT functionals. On the other hand, the neglect of the attractive electron-hole term ($f_{xc}$) in the scGW

<table>
<thead>
<tr>
<th>Material</th>
<th>$\Delta$ (eV)</th>
<th>$\varepsilon_m$ (eV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ge</td>
<td>0.95</td>
<td>0.75</td>
</tr>
<tr>
<td>Si</td>
<td>1.41</td>
<td>1.28</td>
</tr>
<tr>
<td>GaAs</td>
<td>1.85</td>
<td>1.55</td>
</tr>
<tr>
<td>SiC</td>
<td>2.88</td>
<td>2.62</td>
</tr>
<tr>
<td>CdS</td>
<td>2.87</td>
<td>2.37</td>
</tr>
<tr>
<td>AlP</td>
<td>2.90</td>
<td>2.27</td>
</tr>
<tr>
<td>GaN</td>
<td>3.82</td>
<td>3.30</td>
</tr>
<tr>
<td>ZnO</td>
<td>3.80</td>
<td>3.20</td>
</tr>
<tr>
<td>ZnS</td>
<td>4.15</td>
<td>3.59</td>
</tr>
<tr>
<td>C</td>
<td>6.18</td>
<td>5.88</td>
</tr>
<tr>
<td>BN</td>
<td>7.14</td>
<td>6.66</td>
</tr>
<tr>
<td>MgO</td>
<td>9.16</td>
<td>8.25</td>
</tr>
<tr>
<td>LiF</td>
<td>15.9</td>
<td>14.8</td>
</tr>
<tr>
<td>Ar</td>
<td>14.9</td>
<td>14.0</td>
</tr>
<tr>
<td>Ne</td>
<td>22.1</td>
<td>21.1</td>
</tr>
</tbody>
</table>

TABLE I. Band gaps $\Delta$ and averaged $d$ band positions $\varepsilon_d$ for scGW calculations without (RPA) and with ($e$-h) attractive electron-hole interaction (vertex corrections in $W$), and for $W_{0}^{\text{DFT}}$ calculations, where $W_{0}^{\text{DFT}}$ is calculated from DFT wave functions and eigenvalues, and wave functions and eigenvalues in $G$ are updated until self-consistency is reached (values in parentheses are results for an update of the eigenvalues only in $G$ from Ref. [10]). References for experimental values are collected in Ref. [10]; underlined values correspond to measurements at low temperature. Theoretical values are corrected for spin-orbit coupling (0.10 eV for GaAs and Ge). Also reported are the theoretical (scGW) and experimental ion-clamped (high frequency) dielectric constants $\varepsilon_m$. |
calculations yields unreliable dielectric constants (scGW—RPA in Table II). The underestimation is typically 20%, which is the origin for the overestimated band gaps (5–10%) in the scGW approximation [14].

From a theoretical point of view, the present calculations leave one open question. We have included the vertex corrections via the electron-hole interaction $f_{xc}$ in W only, but we have neglected them in the construction of the self-energy ($\Sigma = iGW$), since the latter approach turned out to be numerically rather unstable and tended to bring the band gaps back to those obtained without vertex corrections (see also Ref. [18]). The approximation applied here is often termed $GW^{TC-TC}$, as opposed to the more concise $GW^{T}$ approximation. Reining has recently shown that the present form for the nonlocal exchange correlation kernel $f_{xc}$ is not optimal for approximating the vertices in the self-energy, but better kernels are not yet available [25]. Based on the good agreement of the present values with experiment, we believe the effect of appropriate vertex corrections in the self-energy should be rather small, and these corrections should mostly affect localized d states.

In summary, we have demonstrated that the inclusion of approximate many-electron vertex corrections in W yields excellent band gaps in the framework of the self-consistent GW method. This allows for a consistent and entirely parameter-free calculation of band gaps using the GW method with a precision approaching experimental methods. The approach opens a wide variety of applications, for instance band gap engineering in hypothetical compounds and prescreening of dopants, and it might be applicable to molecular systems as well. Since fully self-consistent GW calculations with vertex corrections are exceedingly time-consuming (in average a single calculation takes 1 day on four Pentium Duo processors), the other result of the present study is equally important. The static screening properties calculated from gradient corrected functionals in the random phase approximation agree very well with the self-consistently determined screening properties in the scGW method. This offers a convenient shortcut and legitimizes the often applied $GW_o$ approximation.

This work was supported by the Austrian Fonds zur Förderung der wissenschaftlichen Forschung.