Adequacy of approximations in *GW* theory

Mark van Schilfgaarde,[1] Takao Kotani,[1] and Sergey V. Faleev[2]
[1]*School of Materials, Arizona State University, Tempe, Arizona, 85284, USA*
[2]*Sandia National Laboratories, Livermore, California 94551, USA*

(Received 22 August 2005; revised manuscript received 25 August 2006; published 28 December 2006)

Following the usual procedure of the *GW* approximation (GWA) within the first-principles framework, we calculate the self-energy from eigenfunctions and eigenvalues generated by the local-density approximation. We analyze several possible sources of error in the theory and its implementation, using a recently developed all-electron approach based on the full-potential linear muffin-tin orbital (LMTO) method. First we present some analysis of convergence in some quasiparticle energies with respect to the number of bands, and also their dependence on different basis sets within the LMTO method. We next present a new analysis of core contributions. Then we apply the GWA to a variety of materials systems to test its range of validity. For simple *sp* semiconductors, GWA always underestimates band gaps. Better agreement with experiment is obtained when the renormalization (Z) factor is not included, and we propose a justification for it. We close with some analysis of difficulties in the usual GWA procedure.

DOI: 10.1103/PhysRevB.74.245125 PACS number(s): 71.15.Qe, 71.10.−w, 71.15.Ap

I. INTRODUCTION

Even though the *GW* approximation (GWA) of Hedin[1] is as old as the local-density approximation (LDA), it is still in its early stages because of serious difficulties in its implementation. In the usual *ab initio* procedure, G and W are constructed from the LDA potential, which generate the self-energy $\Sigma = iG \times W$. Additionally, the quasiparticle energies (QPEs) are usually approximated as a perturbation correction to the LDA from the matrix elements of the diagonal parts of $\Sigma - V_{xc}^{\mathrm{LDA}}$ [see Eqs. (18) and (19) below]. In principle, it is well defined as a procedure. However, there is a controversy regarding what the numerical result of this procedure is for semiconductors. Nearly always the GWA is implemented in conjunction with the pseudopotential (PP) approximation, which we will call PPGW. It was widely thought that PPGW predicts band gaps in semiconductors to rather high accuracy. However, recent all-electron *GW* calculations that survey band gaps in semiconductors using the full-potential linear muffin-tin orbitals (FP-LMTO) basis by Kotani and van Schilfgaarde[2] result in band gaps which are generally smaller than experimental values. The result is confirmed by other calculations using two independently developed full-potential linear augmented plane-wave (FP-LAPW) codes: one by Usuda, Hamada, Kotani, and van Schilfgaarde[3] and another by Friedrich, Schindlmayr, Blügel, and Kotani.[4] These methods all use essentially the same *GW* codes originally developed in conjunction with the LMTO method;[2] they differ only in the input eigenfunctions. Calculations from other, independently developed all-electron *GW* methods[5–7] are consistent with this conclusion.[8]

Tiago, Ismail-Beigi, and Louie[9] used the PPGW scheme that included Si 2s and 2p cores in the valence to analyze the dependence of some semiconductor band gaps on the number of unoccupied states N' used to construct the self-energy. They suggested that the discrepancy between all-electron *GW* and PPGW gaps could be attributed to incomplete convergence in the all-electron calculations. To address this point, the convergence in N' is taken up in Sec. III. We begin with an outline of our all-electron *GW* method (Sec. II); it includes a comparison of the energy bands in Si to those of an APW calculation taken from Friedrich *et al.*[4] and establishes the method's ability to reproduce near-exact LDA eigenvalues. In Sec. IV we show how selected QPEs change with increasingly larger LMTO basis sets for a variety of semiconductors. The results are weakly dependent on basis even for relatively small basis sets. We present some rationale for why this should be so, and note the implications for both precision and efficiency in implementations of the GWA for basis sets in general.

Because our results are well converged for either kind of test, we still think that PPGW is problematic, in contradistinction to the conclusions in a recent paper by Delaney, García-Gonzaléz, Rubio, Rinke, and Godby,[10] who showed that all-electron *GW* and PPGW give essentially the same result for the Be free atom. However, the Be atom is a special case in part because the all-electron and pseudo radial functions should closely correspond to each other (the 2p radial function has no nodes, and the only core that is orthogonalized or pseudized is the deep 1s core ($\epsilon^{\mathrm{LDA}} \approx -105$ eV); moreover, the PP is constructed with the atom itself as reference. Pseudopotentials are constructed to solve LDA reliably, but not to solve the GWA. There are now many detailed checks comparing PP-LDA results against the corresponding all-electron values, but there are few similar comparisons for *GW*. The discrepancies between all-electron and PPGW appear to be much smaller when PPGW includes the highest lying core states in the valence.

Section V analyzes different core contributions to the QPE in several semiconductors. This provides some insight as to what approximations may be made concerning the core; we also briefly consider some aspects of PPGW in this context.

In Sec. VI we show some new results for a variety of materials, as well as repeating some previously reported calculations[2,11–14] with rather tight tolerances. We confirm that the usual GWA procedure generally underestimates band gaps. We also show that a partial self-consistency can be accomplished by calculating QPEs without the renormaliza-

©2006 The American Physical Society

tion factor Z (i.e., $Z=1$). Semiconductor band gaps are systematically improved using $Z=1$, though they continue to be underestimated. An important reason for this is that the LDA overestimates the screening of W, resulting in an underestimate of $\Sigma=iGW$ and band gaps. We show that the adequacy of the GWA varies from system to system: only when the starting LDA is reasonably good does the GWA reasonably predict QP energies. Thus, some kind of self-consistency is necessary to obtain reasonable results for a wide range of materials.[12]

II. METHODOLOGY

A. All-electron LDA in FP-LMTO

Before turning to the analysis, we briefly describe the LDA method we use as input for the *GW* calculations. (Readers interested in the conclusions of this paper *not* related to basis-set issues can skip this section.) An early version of this method was presented in Ref. 15; we describe here how additional local orbitals are included to extend the linear method. Local orbitals are essential to the analysis because QPE in GWA are sensitive to a wider range of states than in LDA (e.g., the LDA depends only on occupied states). One consequence is that the linear approximation inherent in standard linear and pseudopotential methods is less reliable for the GWA than for the LDA. The basis functions used in the present technique are a generalization[15] of the standard[16] LMTO basis. Conventional LMTOs consist of atom-centered envelope functions augmented around atomic sites by a linear combination of radial wave functions φ and their energy derivatives $\dot{\varphi}$. $\varphi=\varphi_{Rl}(\varepsilon_{Rl},r)$ is the solution of the radial Schrödinger equation at site R at some linearization energy ε_{Rl}. A linear method matches the $\{\varphi,\dot{\varphi}\}$ pair to value and slope of the envelope function at each augmentation sphere boundary, which means that the LDA Schrödinger equation can be solved more or less exactly to first order in $\varepsilon-\varepsilon_{Rl}$ inside each augmentation sphere. Envelope functions in the standard LMTO method consist of Hankel functions. In the present basis[15] the envelope functions are smooth, nonsingular generalizations[17] of the Hankel functions: the $l=0$ smooth Hankel satisfies the equation

$$(\nabla^2+\varepsilon)H_0(\varepsilon,r^s;\mathbf{r})=-4\pi g_0(r^s,r),$$

$$g_0(r^s;r)=(\sqrt{\pi}r^s)^{-3}\exp[-(r/r^s)^2]\rightarrow\delta(\mathbf{r})\quad\text{as } r^s\rightarrow 0, \tag{1}$$

and reduces to a usual Hankel function in the limit $r^s\rightarrow 0$. H_L for higher $L=(l,m)$ are obtained by recursion.[17] The basis can be divided into three types of functions:

(i) A muffin-tin orbital (MTO) χ_{RjL}, which consists of a smoothed Hankel centered at nucleus $\mathbf{R}$ and augmented by linear combinations of φ_{Rl} and $\dot{\varphi}_{Rl}$ for each L channel inside every augmentation sphere

$$\chi_{RjL}(\mathbf{r})=H_L(\varepsilon_{Rjl},r^s_{Rjl};\mathbf{r}-\mathbf{R})+\sum_{R'k'L'}C^{RjL}_{R'k'L'}\{\tilde{P}_{R'k'L'}(\mathbf{r})-P_{R'k'L'}(\mathbf{r})\}. \tag{2}$$

$P_{R'k'L'}$ is a one-center expansion of $H_L(\varepsilon_{Rjl},r^s_{Rjl};\mathbf{r}-\mathbf{R})$, and $\tilde{P}_{R'k'L'}$ is a linear combination of $\varphi_{Rl}(\varepsilon_{Rjl},r)$ and $\dot{\varphi}_{Rl}(\varepsilon_{Rjl},r)$ that matches $P_{R'k'L'}$ at the augmentation sphere radius. Expansion coefficients $C^{RjL}_{R'k'L'}$ are chosen to make $\chi_{RjL}(\mathbf{r})$ smooth across each augmentation boundary. ε_{Rjl} is chosen to be at or near the center of the occupied part of that particular l channel. Products of such functions enters into the construction of the Hamiltonian and output density. The present method differs in significant ways from the usual LMTO and LAPW methods: the density is not generated from simple products of MTOs, Eq. (2), but bears some resemblance to the projector augmented wave prescription.[18] It greatly facilitates l convergence in the augmentation; see Ref. 15.

(ii) "Floating orbitals" consisting of the same kind of function as (i), but not centered at a nucleus. Thus, there is no augmentation sphere where the envelope function is centered. There is no fundamental distinction between this kind of function and the first type, except that the distinction is useful when analyzing convergence. Floating orbitals make little difference in LDA calculations, but a basis consisting of purely atom-centered envelope functions is not quite sufficient to precisely represent the interstitial over the wide energy window needed for *GW* calculations. Without their inclusion, errors in QPEs of order 0.1 eV cannot be avoided, as we will show.

(iii) A kind of "local orbital," which has a structure similar to (i). The fundamental distinction is that the "head" (site where the envelope is centered) consists of a new radial function ϕ^z_l evaluated at energy ε^z_l either far above or far below the linearization energy ε_l. For deep, corelike orbitals, ϕ^z_l is integrated at the core energy; for high-lying orbitals, ε^z_l is typically taken 1–2 Ry above the Fermi level E_F. In the former case a tail is attached, with its smoothing radius r^s chosen to make the kinetic energy continuous across the head augmentation sphere. It is thus atypical of conventional local orbitals, as it is nearly an eigenstate of the LDA Hamiltonian without requiring other basis functions.

As is well known, the reason for using augmented wave methods in general (and especially the LMTO method), is that the Hilbert space of eigenfunctions in the energy range of interest is spanned by much fewer basis functions than with other basis sets. In the present method two envelope functions $j=1,2$ are typically used for low l channels s, p, and d, and one for higher l channels (f and sometimes g). The augmentation + local orbital procedure ensures that the basis is reasonably complete inside each augmentation sphere within a certain energy window; the envelope functions + floating orbitals ensures completeness of the basis in the interstitial. The distinction between standard LMTO envelope functions and the smoothed ones used here is important because the generalized form significantly improves this convergence. Core states not treated as local orbitals are handled by integrating the radial Schrödinger equation inside an augmentation sphere and attaching a smoothed Hankel tail, allowing it to spill into the interstitial. Thus the Hartree and exchange-correlation potentials are properly included; only the matrix element coupling core and valence states are neglected.

When used in conjunction with *GW* calculations we typically add local orbitals for states not spanned by $\{\phi,\dot{\phi}\}$ and

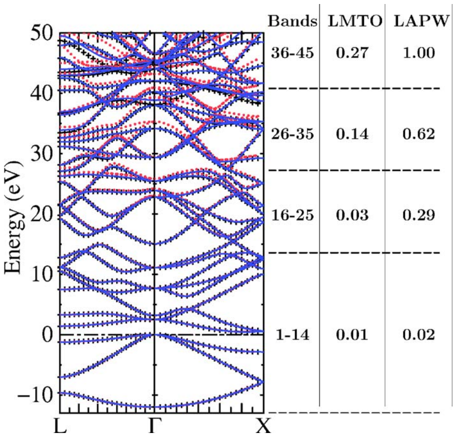


FIG. 1. (Color online) LDA energy bands in Si, computed by different methods. APW bands from Ref. 19 are denoted by "+" and can be regarded as near exact. Dotted lines denote bands calculated by the same authors using the LAPW method, without local orbitals. Solid lines denote bands computed by the present generalized LMTO method, including local and floating orbitals as described in the text (180 energy bands were included in the basis). On the right is a table of the RMS deviation relative to the APW bands for several energy windows, computed along the L-Γ line. The first column denotes the range of bands used to compute the RMS deviation; the horizontal dashed lines denote the approximate energy window for each range. The second and third columns denote how the present method and the LAPW method, respectively, deviate from the APW bands (in eV), as described in the text.

whose center of gravity falls within $\sim\pm 25$ eV of the Fermi level E_F. Both the low-lying and high-lying states can be important, and we shall return to it later. Figure 1 shows the effects of linearization in Si, where an APW calculation of the LDA energy bands is available.[19] Friedrich *et al.*[19] compared LDA bands generated by a full APW calculation to those generated by LAPW. They are reproduced in Fig. 1, together with the bands generated by the LMTO+local +floating orbitals described above. The LAPW and APW are nearly indistinguishable on the scale of figure for energies up to $\sim$25 eV. For energies above 25 eV, the LAPW begins to deviate from the other two, showing the effects of linearization. The APW and generalized LMTO bands are essentially indistinguishable on the scale of figure for energies up to $\sim$40 eV. Above that, slight differences begin to appear; the differences gradually increase for still higher energies. Figure 1 also tabulates the RMS deviation from the APW bands for several energy windows. The present method agrees with the APW bands to $\sim$0.01 eV for levels within $E_F\pm 1$ Ry, and to $\sim$0.25 eV for levels below E_F+4 Ry. Friedrich *et al.* report similar improvements to their LAPW bands when local orbitals are added.[4] They also compare bands generated by a PP, and show that the errors are comparable to the conventional LAPW method. Figure 1 establishes rather convincingly that the present method is nearly complete over a rather wide energy window in Si. When local orbitals are included, it is comparable to an LAPW method that includes local orbitals,[19] and it is superior both to PP and conventional LAPW methods.

B. All-electron *GW* with mixed basis for *W*

We briefly describe our all-electron implementation of the *GW* approximation. A more detailed account will be given elsewhere.[20] The self-energy Σ is

$$\Sigma(\mathbf{r},\mathbf{r}',\omega)=\frac{i}{2\pi}\int d\omega' G(\mathbf{r},\mathbf{r}',\omega-\omega')e^{i\delta\omega'}W(\mathbf{r},\mathbf{r}',\omega'). \tag{3}$$

In this paper G will be taken to be the one-body noninteracting Green function as computed by the LDA, and the screened Coulomb interaction W is computed in the random-phase approximation (RPA) from G. Both G and W are obtained from the LDA eigenvalues $\varepsilon_{\mathbf{k}n}$ and eigenfunctions $\Psi_{\mathbf{k}n}$. For a periodic Hamiltonian, we can restrict $\mathbf{r}$ and $\mathbf{r}'$ to a unit cell and write G as

$$G_{\mathbf{k}}(\mathbf{r},\mathbf{r}',\omega)=\sum_{n}^{\text{All}}\frac{\Psi_{\mathbf{k}n}(\mathbf{r})\Psi^{*}_{\mathbf{k}n}(\mathbf{r}')}{\omega-\varepsilon_{\mathbf{k}n}\pm i\delta}. \tag{4}$$

The infinitesimal $-i\delta$ is to be used for occupied states, and $+i\delta$ for unoccupied states. W is written as

$$W=\epsilon^{-1}v=(1-v\Pi)^{-1}v, \tag{5}$$

where $\Pi=-iG\times G$ is the bare polarization function shown below, $v=e^2/|\mathbf{r}-\mathbf{r}'|$ is the bare Coulomb interaction, and ϵ is the dielectric function. For simplicity, the spin degree of freedom is omitted.

Neglecting the off-diagonal part of Σ, we can evaluate QPE $E_{\mathbf{k}n}$ from

$$E_{\mathbf{k}n}=\varepsilon_{\mathbf{k}n}+Z_{\mathbf{k}n}[\langle\Psi_{\mathbf{k}n}|\Sigma(\mathbf{r},\mathbf{r}',\varepsilon_{\mathbf{k}n})|\Psi_{\mathbf{k}n}\rangle-\langle\Psi_{\mathbf{k}n}|V^{\text{LDA}}_{xc}(\mathbf{r})\times|\Psi_{\mathbf{k}n}\rangle]. \tag{6}$$

$Z_{\mathbf{k}n}$ is the quasiparticle (QP) renormalization factor

$$Z_{\mathbf{k}n}=\left[1-\langle\Psi_{\mathbf{k}n}|\frac{\partial}{\partial\omega}\Sigma(\mathbf{r},\mathbf{r}',\varepsilon_{\mathbf{k}n})|\Psi_{\mathbf{k}n}\rangle\right]^{-1}, \tag{7}$$

and accounts for the fact that Σ is evaluated at the LDA energy rather than at the QPE. Equation (6) is the customary way QPEs are evaluated in *GW* calculations. In Sec. VI, we present an argument that using $Z=1$ (or neglecting the Z factor) is a better choice than Eq. (7), and shows how QPEs are affected in actual calculations. However, the results presented here use the Z factor except where noted.

In FP-LMTO, eigenfunctions of the valence states are expanded in linear combinations of Bloch-summed MTOs, Eq. (2),

$$\Psi_{\mathbf{k}n}(\mathbf{r})=\sum_{RjL}a^{n}_{RjL}\chi^{\mathbf{k}}_{RjL}(\mathbf{r}). \tag{8}$$

Inside augmentation sphere $\mathbf{R}$, the Hilbert space of the valence eigenfunction $\Psi_{\mathbf{k}n}(\mathbf{r})$ consists of the pair (or triplet) of orbitals (φ_{Rl}, $\dot{\varphi}_{Rl}$ or φ_{Rl}, $\dot{\varphi}_{Rl}$, φ^{z}_{Rl}) at that site,[21] and can be represented in a compact notation $\{\varphi_{Ru}\}$. u is a compound index for both L and one of the $(\varphi_{Rl},\dot{\varphi}_{Rl},\varphi^{z}_{Rl})$ triplet. The interstitial is comprised of linear combinations of envelope functions consisting of smooth Hankel functions, which can be expanded in terms of plane waves.[17] Therefore the $\Psi_{\mathbf{k}n}(\mathbf{r})$

can be written as a sum of augmentation and interstitial parts

$$\Psi_{\mathbf{k}n}(\mathbf{r}) = \sum_{Ru} \alpha^{\mathbf{k}n}_{Ru}\varphi^{\mathbf{k}}_{Ru}(\mathbf{r}) + \sum_{\mathbf{G}} \beta^{\mathbf{k}n}_{\mathbf{G}} P^{\mathbf{k}}_{\mathbf{G}}(\mathbf{r}), \tag{9}$$

where the interstitial plane wave (IPW) is defined as

$$P^{\mathbf{k}}_{\mathbf{G}}(\mathbf{r}) = 0 \quad \text{if } \mathbf{r} \in \text{any MT}$$
$$= \exp[i(\mathbf{k}+\mathbf{G})\mathbf{r}] \quad \text{otherwise}, \tag{10}$$

and the $\varphi^{\mathbf{k}}_{Ru}$ are Bloch sums of φ_{Ru},

$$\varphi^{\mathbf{k}}_{Ru}(\mathbf{r}) \equiv \sum_{\mathbf{T}} \varphi_{Ru}(\mathbf{r}-\mathbf{R}-\mathbf{T})\exp(i\mathbf{k}\cdot\mathbf{T}). \tag{11}$$

T and **G** are lattice translation vectors in real and reciprocal space, respectively.

Throughout this paper, we will designate eigenfunctions constructed from MTOs as "VAL." Below them are the core eigenfunctions, which we designate as "CORE." There are two fundamental distinctions between VAL and CORE: First, the latter are constructed independently by the integration of the spherical part of the LDA potential, and they are not included in the secular matrix. Second, the cores are confined to MT spheres.[22] CORE eigenfunctions are also expanded using Eq. (9) in a trivial manner; $\beta^{\mathbf{k}n}_{\mathbf{G}}=0$ and only one of $\alpha^{\mathbf{k}n}_{Ru}\neq 0$. The discussion below applies to all eigenfunctions, VAL and CORE.

Through Eq. (9), products $\Psi_{\mathbf{k}_1 n}\times\Psi_{\mathbf{k}_2 n'}$ can be expanded by $P^{\mathbf{k}_1+\mathbf{k}_2}_{\mathbf{G}}(\mathbf{r})$ in the interstitial region because $P^{\mathbf{k}_1}_{\mathbf{G}_1}(\mathbf{r}) \times P^{\mathbf{k}_2}_{\mathbf{G}_2}(\mathbf{r}) = P^{\mathbf{k}_1+\mathbf{k}_2}_{\mathbf{G}_1+\mathbf{G}_2}(\mathbf{r})$. Within sphere R, wave-function products can be expanded by $B^{\mathbf{k}_1+\mathbf{k}_2}_{Rm}(\mathbf{r})$, which is the Bloch sum of the product basis $\{B_{Rm}(\mathbf{r})\}$, which in turn is constructed from the set of products adapting the procedure by Aryasetiawan.[23] Equation (9) is equally valid in a LMTO or LAPW framework, and eigenfunctions from both types of methods have been used in this *GW* scheme.[3,4] We restrict ourselves to LMTO-derived basis functions here.

We define the mixed basis $\{M^{\mathbf{k}}_I(\mathbf{r})\}\equiv\{P^{\mathbf{k}}_{\mathbf{G}}(\mathbf{r}), B^{\mathbf{k}}_{Rm}(\mathbf{r})\}$, where the index $I\equiv\{\mathbf{G},Rm\}$ classifies the members of the basis. By construction, $M^{\mathbf{k}}_I$ is a good basis set for the expansion of products of $\Psi_{\mathbf{k}n}$. Complete information to calculate Σ and $E_n(\mathbf{k})$ are matrix elements of the products $\langle\Psi_{\mathbf{q}n}|\Psi_{\mathbf{q}-\mathbf{k}n'}M^{\mathbf{k}}_I\rangle$, the LDA eigenvalues $\varepsilon_{\mathbf{k}n}$, the Coulomb matrix $v_{IJ}(\mathbf{k})\equiv\langle M^{\mathbf{k}}_I|v|M^{\mathbf{k}}_J\rangle$, and the overlap matrix $\langle M^{\mathbf{k}}_I|M^{\mathbf{k}}_J\rangle$. (The overlap matrix of IPW is necessary because $\langle P^{\mathbf{k}}_{\mathbf{G}}|P^{\mathbf{k}}_{\mathbf{G}'}\rangle\neq 0$ for $\mathbf{G}\neq\mathbf{G}'$.) The Coulomb interaction is expanded as

$$v(\mathbf{r},\mathbf{r}') = \sum_{\mathbf{k},I,J} |\tilde{M}^{\mathbf{k}}_I\rangle v_{IJ}(\mathbf{k})\langle\tilde{M}^{\mathbf{k}}_J|, \tag{12}$$

where we define

$$|\tilde{M}^{\mathbf{k}}_I\rangle \equiv \sum_{I'} |M^{\mathbf{k}}_{I'}\rangle(O^{\mathbf{k}})^{-1}_{I'I}, \tag{13}$$

$$O^{\mathbf{k}}_{I'I} = \langle M^{\mathbf{k}}_{I'}|M^{\mathbf{k}}_I\rangle. \tag{14}$$

W and the polarization function Π shown below are expanded in the same manner as Eq. (12).

The exchange part of Σ is written in the mixed basis as

$$\langle\Psi_{\mathbf{q}n}|\Sigma_x|\Psi_{\mathbf{q}n}\rangle = \sum_{\mathbf{k}}^{\mathrm{BZ}}\sum_{n'}^{\mathrm{occ}} \langle\Psi_{\mathbf{q}n}|\Psi_{\mathbf{q}-\mathbf{k}n'}\tilde{M}^{\mathbf{k}}_I\rangle v_{IJ}(\mathbf{k}) \times\langle\tilde{M}^{\mathbf{k}}_J\Psi_{\mathbf{q}-\mathbf{k}n'}|\Psi_{\mathbf{q}n}\rangle. \tag{15}$$

The screened Coulomb interaction $W_{IJ}(\mathbf{q},\omega)$ is calculated through Eq. (5), where the polarization function Π is written,

$$\Pi_{IJ}(\mathbf{q},\omega) = \sum_{\mathbf{k}}^{\mathrm{BZ}}\sum_{n}^{\mathrm{occ}}\sum_{n'}^{\mathrm{unocc}} \langle\tilde{M}^{\mathbf{q}}_I\Psi_{\mathbf{k}n}|\Psi_{\mathbf{q}+\mathbf{k}n'}\rangle\langle\Psi_{\mathbf{q}+\mathbf{k}n'}|\Psi_{\mathbf{k}n}\tilde{M}^{\mathbf{q}}_J\rangle \times\left(\frac{1}{\omega-\varepsilon_{\mathbf{q}+\mathbf{k}n'}+\varepsilon_{\mathbf{k}n}+i\delta} - \frac{1}{\omega+\varepsilon_{\mathbf{q}+\mathbf{k}n'}-\varepsilon_{\mathbf{k}n}-i\delta}\right). \tag{16}$$

Equation (16) assumes time-reversal symmetry. We use the tetrahedron method for the Brillouin zone (BZ) summation in Eq. (16) following Ref. 24. We first calculate the contribution to Π proportional to the imaginary part of the second line in Eq. (16), and determine the rest of Π by Hilbert transformation (Kramers-Krönig relation). Such an approach significantly reduces the computational time required to calculate Π.

The correlation part of Σ is

$$\langle\Psi_{\mathbf{q}n}|\Sigma_{\mathrm{c}}(\omega)|\Psi_{\mathbf{q}n}\rangle = \sum_{\mathbf{k}}^{\mathrm{BZ}}\sum_{n'}^{\mathrm{All}}\sum_{IJ} \langle\Psi_{\mathbf{q}n}|\Psi_{\mathbf{q}-\mathbf{k}n'}\tilde{M}^{\mathbf{k}}_I\rangle\langle\tilde{M}^{\mathbf{k}}_J\Psi_{\mathbf{q}-\mathbf{k}n'}|\Psi_{\mathbf{q}n}\rangle \times\int_{-\infty}^{\infty}\frac{i\,d\omega'}{2\pi}W^{\mathrm{c}}_{IJ}(\mathbf{k},\omega')\frac{1}{-\omega'+\omega-\varepsilon_{\mathbf{q}-\mathbf{k}n'}\pm i\delta}. \tag{17}$$

Here $-i\delta$ is for occupied states; $+i\delta$ is for unoccupied states. $W^{\mathrm{c}}\equiv W-v$.

GW calculations usually approximate Eq. (6) by

$$E_{\mathbf{k}n} = \varepsilon_{\mathbf{k}n} + Z_{\mathbf{k}n}[\langle\Psi_{\mathbf{k}n}|\Sigma^{\mathrm{VAL}}(\mathbf{r},\mathbf{r}',\varepsilon_{\mathbf{k}n})|\Psi_{\mathbf{k}n}\rangle - \langle\Psi_{\mathbf{k}n}|V^{\mathrm{LDA}}_{xc}([n^{\mathrm{VAL}}],\mathbf{r})|\Psi_{\mathbf{k}n}\rangle], \tag{18}$$

where Σ^{VAL} and $V^{\mathrm{LDA}}_{xc}([n^{\mathrm{VAL}}])$ are calculated only from eigenfunctions belonging to VAL. In the present method we calculate the $E_{\mathbf{k}n}$ including the core contributions from

$$E_{\mathbf{k}n} = \varepsilon_{\mathbf{k}n} + Z_{\mathbf{k}n}\times[\langle\Psi_{\mathbf{k}n}|\Sigma_x(\mathbf{r},\mathbf{r}')+\Sigma_c(\mathbf{r},\mathbf{r}',\varepsilon_{\mathbf{k}n})|\Psi_{\mathbf{k}n}\rangle - \langle\Psi_{\mathbf{k}n}|V^{\mathrm{LDA}}_{xc}([n^{\mathrm{total}}],\mathbf{r})|\Psi_{\mathbf{k}n}\rangle]. \tag{19}$$

Note that $V^{\mathrm{LDA}}_{xc}([n^{\mathrm{total}}],\mathbf{r})$ is for the entire density n^{total}. As n' in Eq. (15) can be divided between CORE and VAL, we have $\Sigma_x=\Sigma^{\mathrm{CORE}}_x+\Sigma^{\mathrm{VAL}}_x$. In this paper, we neglect the CORE contribution to Σ_c. In Sec. V we examine in some detail the contributions by shallow cores to correlation by including them in VAL using local orbitals, and will see that their contribution is small except for very shallow cores.

In short, no important approximation is made other than the *GW* approximation itself; and it is to the best of our knowledge the only implementation of GWA that makes no significant approximations. Results depend slightly on what kind of basis set is used to generate G and W, as we will show, and also on the tolerances in parameters used in the *GW*-specific part of the calculation.

LMTO-based calculations presented here employ a widely varying set of basis functions, ranging from ~20 to 90 orbitals per atom, as described in more detail below. They typically consist of a basis of *spdfg*+*spd* orbitals centered on each atom, some floating orbitals and sometimes local orbitals. In the Si calculations local p orbitals of either $2p$ character or of $4p$ character were used, as we describe below. For the *GW* part of the calculation, the Si results shown below use parameters representative of the various system studied: LMTO basis functions are reexpanded in plane waves to a cutoff of 3.3 a.u. in the interstitial region, i.e., $|\mathbf{k}+\mathbf{G}|<3.3\ \text{bohrs}^{-1}$ in the second term of Eq. (9). The IPW part of the mixed basis used to expand v, Π, and W used a cutoff $|\mathbf{k}+\mathbf{G}|<3.0\ \text{bohrs}^{-1}$; the product basis part consisted of 90–110 Bloch functions/atom (we use a different product basis for Σ_x^{CORE} and Σ_x^{VAL}). Augmentation sphere radii were chosen so that spheres approximately touched but did not overlap, and the product basis functions entering into the mixed basis $\widetilde{M}$ in Eqs. (16) and (17) were expanded to $l=5$. In the calculation of Eq. (17) the poles of G were Gaussian broadened by $\sigma=0.003$ or 0.01 Ry. These parameters correspond to rather conservative tolerances: tests at tighter tolerances in these parameters change the QPE by ~0.01 eV. (Systematic checks were performed for each material studied.) The tetrahedron method was used for Π with a $6\times6\times6$ k mesh (doubling the number of points in the energy denominator) except where noted. The same mesh is used to calculate Π and Σ. This k mesh is reasonably well converged, systematically overestimating conduction-band states by ~0.02 eV in Si and similar semiconductors relative to the fully k-converged result.[25]

III. CONVERGENCE IN QUASIPARTICLE ENERGIES: NUMBER OF UNOCCUPIED STATES

In Fig. 2, we show the $\Gamma_{25v}\rightarrow X_{1c}$ gap for Si computed by Eq. (19) as a function of the number of unoccupied states N' in Eqs. (16) and (17): i.e., summation n' over unoccupied states is restricted to $n'<N'$. Figure 2 depicts our main with pentagons (Si $2p$ treated as VAL). It tracks well the all-electron results of Ku and Eguiluz,[26] which used an LAPW method, except their data is ~0.05 eV less than ours. However, their results are limited to $N'<31$, which is not sufficient to analyze convergence for large N'. If one assumes the LAPW converges with N' at the same rate as the LMTO case, their best result with $N'=31$ should be ~0.1 eV less than the converged value. Indeed, a very recent calculation of the same system by Friedrich, Schindlmayr, Blügel, and Kotani,[4] based on LAPW with an LDA basis of ~300 basis functions, showed N'-convergence similar to LMTO.

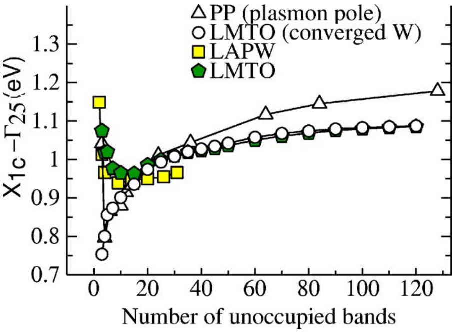


FIG. 2. (Color online) $\Gamma_{25v}\rightarrow X_{1c}$ gap in Si in GWA as a function of the number of unoccupied states N'. Filled (yellow) squares: LAPW GWA taken from Ku and Eguiluz (Ref. 26). The authors presented data only for $N'<31$. Also, their data was given for the minimum gap, so we shifted their results by +0.14 eV to estimate the $\Gamma_{25v}\rightarrow X_{1c}$ gap. Some checks show that the shift should be ~0.14 eV, approximately independent of N'. Filled (green) pentagons: LMTO results varying N' both G and W. LMTO results were shifted by −0.02 eV to correct for incomplete k convergence (Ref. 25). The Si $2p$ was included in the valence using a local orbital (the LAPW calculation of Ku and Eguiluz did not). The total dimension of the LMTO basis is 180. Results from the same basis are shown as filled (green) pentagons in Fig. 4, and also filled (green) pentagons No. (13) in Fig. 5. Open triangles: PPGW from Ref. 9, which included the Si $2p$ levels as part of the valence, and in which W was computed using the plasmon-pole approximation. Open circles: LMTO results varying N' in G but not W.

Tiago, Ismail-Beigi, and Louie[9] presented a PPGW calculation of some QPE in Si, Ge, and GaAs where they included the higher-lying core states into the valence so as to assess the effect of the core. They monitored the rate of convergence in QPE with N'; their data for Si are shown as open triangles in Fig. 2. There are some similarities, but also two discrepancies:

(i) For $N'\lesssim30$, the behavior is rather different.

(ii) In the asymptotic region $N'\gtrsim30$, the PPGW and LMTO results converge at somewhat different rates.

In order to examine point (i), we tried LMTO calculations where W is fixed [i.e., N' is truncated only in Eq. (4)]. This calculation (open circles in Fig. 2) tracks well the PPGW result for $N'\lesssim30$. This looks reasonable because the PPGW is combined with the plasmon-pole approximation, which satisfies the sum rule for Im ϵ^{-1} for any N'; thus W converges rather quickly with respect to N'. However, the two LMTO calculations show little difference in the asymptotic behavior, which means that it is controlled by N' in Eq. (17), as was already discussed by Tiago *et al.*

It can be seen that the N' dependence of either LMTO calculation is slightly different than the PPGW result for both intermediate and large N': the change in the $\Gamma_{25v}\rightarrow X_{1c}$ gap from $N'=35$ to $N'=60$ for PPGW is roughly twice the change obtained by the LMTO method. As we noted in Sec. II, LMTO-LDA eigenvalues are very close to the full APW results in this energy range (see Fig. 1). This indicates that the eigenfunctions are also precise. Moreover, Friedrich *et al.*[4] compare LDA-APW eigenvalues to an LAPW+local orbitals method; the three sets of eigenvalues (APW and LMTO+local orbitals, LAPW+local orbitals) are very close to one another. By contrast, the LDA energy bands computed by either conventional LAPW or PP methods correspond to

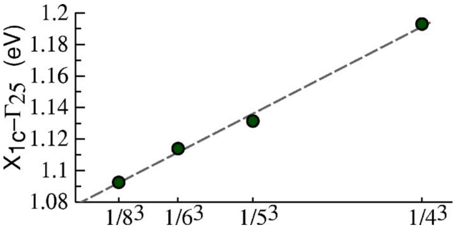


FIG. 3. (Color online) $\Gamma_{25v}\rightarrow X_{1c}$ gap in Si as a function $1/n_k^3$, where n_k^3 is the number of k points in the full Brillouin zone. The dependence on $1/n_k^3$ shown for $\Gamma_{25v}\rightarrow X_{1c}$ is essentially the same for all of the unoccupied QPEs we examined. Gaps for $n_k=6$, $n_k=5$, and $n_k=4$ exceed the $n_k=8$ case by 0.02, 0.04, and 0.10 eV, respectively. We can also estimate what the $n_k\rightarrow\infty$ gap would be by extrapolating the approximately linear dependence on $1/n_k^3$ to zero (dashed line). The $n_k=8$ case apparently overestimates the converged result by 0.01 eV.

APW eigenvalues far less well; see Ref. 4. Finally, the dependence on N' as computed by the LAPW+local orbitals method is essentially similar to the present LMTO results. When these observations are considered as a whole, they suggest that what discrepancy does exist between LMTO +local orbitals (or LAPW+local orbitals) and PPGW may be an artifact of the pseudopotential construction in the PPGW method. We cannot rule out possible limitations to the present method, however. Differences with PPGW are small in absolute terms. Even though eigenvalues generated by LMTO and LAPW+local orbitals are very close to APW eigenvalues, eigenfunctions may be less well described. And even though the LMTO and LAPW Hmiltonians are very different, the QPEs are generated by a common *GW* code. If there is some limitation in the numerical procedure, it would be common to both LMTO and LAPW calculations.

It is also possible that the calculation by Tiago *et al.* suffers from incomplete k convergence. Their PPGW used a $4\times4\times4$ k mesh. k convergence is mainly limited by divergent behavior for $|\mathbf{k}|\rightarrow0$ in Eq. (15). To treat this divergence, we use the offset-Γ method, which was originally developed by ourselves[2] and is now used by other groups.[7,27] It is essentially equivalent to techniques that treat the divergent part analytically, as is typically done by PPGW practitioners. Figure 3 shows the dependence of $\Gamma_{25v}\rightarrow X_{1c}$ gap on n_k, but in general all of conduction bands shift nearly rigidly with changes in n_k (n_k=number of linear divisions of the k mesh in the Brillouin zone). Band gaps are approximately linear in the reciprocal of the total number of points, $1/n_k^3$. The figure shows that a $4\times4\times4$ k mesh overestimates the k-converged gap by ~0.1 eV.[25] This may explain most of the remaining discrepancy between the PPGW calculations of Tiago *et al.* and the present results.

In Sec. V we analyze the dependence of QPEs on the core treatment. Proper treatment of the core is somewhat subtle,[28] and we use the local orbitals for the analysis. Because they are already nearly exact solutions of the LDA for the states they constructed to represent, they minimally hybridize with other basis functions; consequently, any higher-lying CORE state can readily be converted into a valence state with minimal perturbation of the LDA basis. Use of local orbitals enables us to investigate how different kinds of core contributions affect QPEs in a well-controlled and systematic way. We show that differing treatments of the Si $2p$ core only slightly affect QPEs; similar results are found for other deep cores. A significantly larger dependence is apparently found using the PPGW method.

IV. CONVERGENCE IN QUASIPARTICLE ENERGIES: BASIS DEPENDENCE

Here we study the convergence in QPEs as the LMTO basis set changes, retaining all the eigenfunctions for a given basis in the calculation (N' encompasses all unoccupied states). A given LMTO basis defines a finite Hilbert space of eigenfunctions; the GWA is a well-defined procedure in that space, and we can study how the QPEs change as the Hilbert space is refined. This procedure corresponds more closely to analyses of basis-set convergence common in other kinds of calculations (e.g., LDA and Hartree-Fock). We can also anticipate that it will be smoother than the N' truncation of Sec. III; indeed, this will turn out to be the case (see especially Fig. 5): the band gaps are insensitive to the choice of basis once a certain level of completeness is reached. It is also obviously true that the Hilbert space depends on the choice of basis constructing it. Therefore, the results presented here are specific to the LMTO basis described in Sec. II, and in particular, *what kinds* of orbitals are included, e.g., orbitals of f or g character, or local orbitals that correct the linearization inherent in most of the standard methods (LMTO, LAPW, and the construction of a norm-conserving PP). By adding different kinds of orbitals we can identify how different parts of the Hilbert space (most notably corrections to linearization common to most methods), affect QPEs. Since the LMTO basis is tailored to the crystal potential, LDA eigenfunctions converge more rapidly with the basis dimension than do plane-wave-based basis sets.[31] Consequently, we might expect a more rapid convergence in the corresponding GWA QPEs. On the other hand, by transformation to, e.g., Wannier functions, it should be possible to design a generic scheme that exhibits similarly rapid convergence.

Initially, we compare in Fig. 4 the dependence of the $\Gamma_{25v}\rightarrow X_{1c}$ gap in Si on N' for two basis sets: one relatively small and another relatively large. The right panel of Fig. 4 compares the same data, but the horizontal scale corresponds the energy of state N' at Γ. Also in this panel, the small-basis data was artificially shifted down by 0.01 eV to make it easier to compare their energy dependence.[29] The difference, initially 0.01 eV, increases by an additional 0.01 eV as the energy E approaches $\sim E_f+50$ eV. For higher energies, the discrepancy between the two increases more rapidly. This is because the ability of the small LMTO basis to describe eigenvalues above $\sim E_f+40$ eV begins to degrade.

However, we can see that the gaps at the respective maximal N' (i.e., all unoccupied states included) are in good agreement. Including all unoccupied states in a limited basis is another kind of Hilbert-space truncation, but it is also well defined: it is Hilbert space of eigenfunctions basis consisting of LMTO eigenfunctions and their products (Sec. II), and apply the GWA within that Hilbert space. The LMTO basis can efficiently choose the important part of the Hilbert space

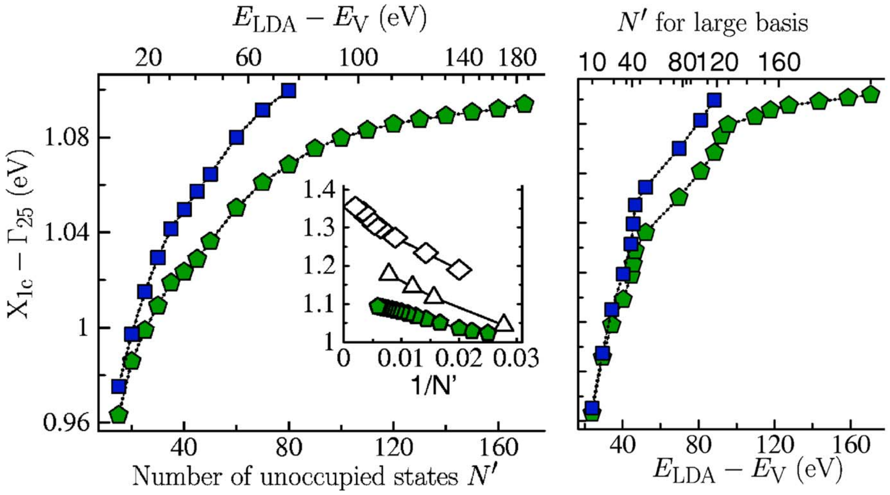


FIG. 4. (Color online) The left panel shows $\Gamma_{25v} \rightarrow X_{1c}$ gap in Si as a function of the number of unoccupied states N' for a smaller basis (filled squares) and a larger basis (pentagons). The latter are redrawn from Fig. 2. The top horizontal scale shows an approximate relation between energy and N' in the large basis (interpolated from levels at Γ). The right panel contains the same data but reverses the top and bottom horizontal scales. Had N' in the upper horizontal axis been drawn for the small basis (squares), the scale would be a little different: the last data point corresponds to $N'=82$ instead of 120. In the right panel the small-basis QPE's (filled squares) were shifted by −0.01 eV (Ref. 29) to clarify how large- and small-basis data diverge as the energy increases. The inset compares convergence in X_{1c} as a function of $1/N'$ to a PPGW calculation that includes $2p$ states in the valence (Ref. 9) (open triangles) and a PPGW calculation that does not (Ref. 30) (open diamonds). LMTO data were shifted by −0.02 eV to correct for incomplete k convergence (Ref. 25). Some of the differencess between PP and LMTO data (triangles and hexagons) may be related to incomplete k convergence; see Sec. III.

tailored to the crystal potential. Thus good agreement need not be some fortutitous artifact of this particular pair of the LMTO basis sets, even though the maximal N' is small in light of a traditional N' cutoff analysis.[32] Indeed, the N' cutoff of Sec. III may choose the Hilbert space less well, especially since that kind of truncation is not smooth. Below we present a detailed analysis of the dependence on the basis set to justify the good agreement in Fig. 4:[29] the band gaps are insensitive to the choice of basis once a certain level of completeness is reached.

Figure 5 shows the results of a systematic study of the convergence in the first unoccupied QPE at Γ, X, and L in Si with progressively larger basis sets. LDA eigenvalues are not shown because they are the same within ~0.01 eV for all cases (0.60 eV for X, 1.42 eV for L, 2.52 eV for Γ). These data comprise very diverse basis sets, particularly for the LMTO method, which traditionally uses a minimal basis. Some details concerning these sets help explain in what manner convergence is reached:

(i) Filled (yellow) diamonds (1) includes *spdfsp* atom-centered functions and is the only basis without floating orbitals. There are no local orbitals; Si $1s,2s,2p$ are CORE.

(ii) Filled (blue) circles (2 and 3) add floating orbitals of *sp* and of *spd* character, respectively. Their effect is to cause QPEs to *decrease* slightly relative to (1). Adding still more floating orbitals (even large numbers of them) shift QPEs by ~0.01 eV.

(iii) Other filled (blue) circles (5,6,10,12) include still more envelope functions comprised of a mixture of atom-centered functions and floating orbitals, but adding no local orbitals.

(iv) Filled (green) square and open square (4 and 7) correspond to (3 and 6), respectively, but adding a local orbital (green: Si $2p$, open: Si $4p$). When the $2p$ is included as VAL, CORE consists of Si $1s,2s$ only. A local orbital of *either* Si $2p$ or Si $4p$ shifts QPEs—in roughly equal but opposite directions.

(v) Filled (green) pentagon and open pentagon (8,9,11,13) include an additional Si $4d$ local orbital. (11) corresponds to (10)+(Si $2p$ or Si $4p$)+Si $4d$. (13) corresponds to $(12)+\cdots$ as well. (8) is (6)+Si $4p$+Si $4d$. The effect of Si $4d$ is small.

These points show in a compelling way that once the basis reaches a certain level of completeness, the change of

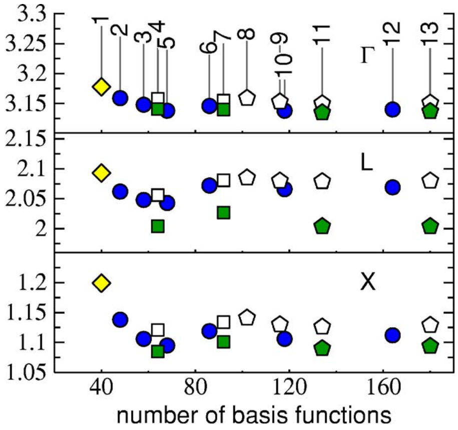


FIG. 5. (Color online) QPE in the GWA at Γ_{15c} (top), L_{1c} (middle), and X_{1c} (bottom) in Si relative to the valence-band maximum, using different basis sets in the present FP-LMTO. Abscissa is the total number of basis functions N. Yellow diamonds show a minimal basis (see text). All results depicted by blue (filled) circles contain no local orbitals. Those depicted by filled (empty) squares or pentagons include the Si $2p$ (Si $3p$) as local orbitals. See the text for further description.

QPE with further enlargement is very small. Set (1), which consists only of atom-centered functions, is somewhat incomplete except inside the augmentation spheres where the eigenfunctions are constructed out of linear combinations of $\{\phi,\dot{\phi}\}$. Considering the open structure of zinc blende, such a basis may be expected to be less complete in the interstitial. Comparing basis sets without local orbitals (circles) with set (1) shows that this particular purely atom-centered basis is slightly deficient for reliable calculation of QPEs, since the addition of floating orbitals induces a (k-dependent) reduction in the conduction band of ~0.02–0.10 eV. It is an open question whether a still more sophisticated atom-centered basis[33] would be adequate to describe the interstitial.

Once the interstitial is reasonably complete [cf. sets (3) and higher], there is an almost negligible dependence on basis *provided* no orbitals are included that extend the linear method or alter how the core is treated. Basis sets marked by a common symbol (squares, circles, pentagons) share essentially the same Hilbert space in the *augmentation* regions; only the basis set corresponding to the *interstitial* region changes. The variation is ±0.01 eV for a wide range of basis sets.[29]

Figure 5 also gives us some insight into the limitations of the linear method. Basis sets (3) and (4), which differ only in how the Si p channel is treated inside the augmentation region, affect QPEs more strongly than radically enlarging the Hilbert space of the envelope functions—compare $(3)\rightarrow(4)$ and $(3)\rightarrow(12)$. Envelope functions affect only the interstitial; they negligibly affect the Hilbert space of the augmentation region. For the latter it is largely irrelevant how many envelope functions are used—and consequently, the size of N' entering into Eqs. (16) and (17). What *is* relevant is the completeness of $\{\phi,\dot{\phi}\}$, and results are independent of basis dimension provided that *the entire* $\{\phi,\dot{\phi}\}$ *Hilbert space is included*. Said another way, the LMTO method is by design reasonably complete over a certain energy window in the augmentation spheres, more or less independent of the envelope functions. A similar story may be told for the interstitial: sets (3,6,10,12) differ in the number of envelope functions by as much as a factor of three, but QPEs are unchanged within ±0.01 eV. QPEs do shift, in a consistent manner, when $\{\phi,\dot{\phi}\}\rightarrow\{\phi,\dot{\phi},\phi^{4p}\}$ or $\{\phi,\dot{\phi}\}\rightarrow\{\phi,\dot{\phi},\phi^{2p}\}$, essentially independent of the number of envelope functions ($3\rightarrow4$, $6\rightarrow7$, $10\rightarrow11$, $12\rightarrow13$).

Table I shows data for three other materials (CdO, Ge, and GaAs). We can see (1) rapid convergence in QPEs as the basis is enlarged for a *fixed* set of augmentation functions; and (2) extensions to a linear augmentation affect QPEs in a manner approximately independent of the total dimension of the Hamiltonian. (In GaAs, both 3d and 4d must be included as VAL. If not, significant errors result[2].) We have tested a number of other materials as well, and these trends appear to be rather general. As might be expected, the number of basis functions needed to make the Hilbert space reasonably complete depends somewhat on the elements involved. The heavier atoms have larger radii and consequently slower l convergence in the number of envelope functions needed; also, d orbitals often play an important role. More orbitals are required to make the basis complete when heavier atoms are involved.

TABLE I. QPEs of the first unoccupied state at Γ, L, and X, for different basis sets, in eV (relative to valence maximum). Columns n_a, n_f, and n_l denote the number of atom-centered functions, the number of floating orbitals, and the number of local orbitals, respectively. The Hamiltonian dimension is the sum of these numbers. Experimental data are adjusted for spin-orbit coupling by adding 1/3 of the splitting in the Γ_{15} valence bands. The first four CdO basis sets are identical to the last four except for the addition of local orbitals in the O 3p and Cd 5d channels. A 6×6×6 k mesh was used in these calculations.

Data type	n_a	n_f	n_l	Γ	L	X
			CdO			
Expt. +0				0.84		
LDA	59	18	12	−0.53	4.26	3.58
GW	59	18	12	0.14	5.18	4.97
	59	50	12	0.10	5.14	4.92
	82	66	12	0.10	5.16	4.89
	82	82	12	0.10	5.16	4.88
	59	18	3	−0.01	5.05	4.78
	59	50	3	−0.06	5.01	4.73
	82	66	3	−0.02	5.05	4.73
	82	82	3	−0.02	5.06	4.74
			Ge			
Expt. +0.10				1.00	0.88	1.20
LDA	50	18	10	−0.12	0.07	0.65
GW	50	18	10	0.80	0.65	0.94
	68	18	10	0.84	0.68	0.97
	82	50	10	0.83	0.67	0.96
	82	82	10	0.82	0.67	0.96
			GaAs			
Expt. +0.11				1.63	1.96	2.11
LDA	42	18	6	0.34	0.86	1.34
GW	42	18	6	1.44	1.68	1.79
	68	18	6	1.46	1.69	1.79
	82	50	6	1.44	1.66	1.77
	82	82	6	1.43	1.66	1.77
	82	82	11	1.43	1.68	1.81

As noted, the linear $\{\phi,\dot{\phi}\}$ Hilbert space is already reasonably complete in the case of Si. But this is not true in general: oxides and nitrides form a materials class where the effect is significantly larger. CdO is one such example (CdO forms in the NaCl structure; the valence-band maximum falls at L and the conduction-band minimum falls at Γ.) As happens for Si, there is a weak dependence on basis when the number of envelope functions is changed and the Hilbert space of the augmentation is held constant. But the QPEs change by ~0.15±0.05 eV when the O 3p and Cd 5d states

TABLE II. QPE of the first unoccupied state at Γ, L, and X relative to top of valence, for core treatments (i)–(iv) as described in the text, in eV. States and corresponding eigenvalues $\varepsilon_c^{\mathrm{LDA}}$ treated as *core* are: $2p$ in Si, $3d$ in Ga and Ge, and $2s$ in Mg. Si data corresponds to the basis set (13) in Fig. 5; GaAs data corresponds to the 68+18+6 orbital basis in Table I; Ge data corresponds to the 68+18+10 orbital basis in that table. Here G means $(G^{core}+G^{val})$, W means $W[\Pi]$ (see text). A $6\times6\times6$ k mesh was used in these calculations. For results with better k convergence and larger basis sets, see Table III.

		(i): $G^{val}, W[\Pi^{val}]$			(ii): $G, W[\Pi^{val}]$			(iii): G^{val}, W			(iv): G, W		
	$\varepsilon_c^{\mathrm{LDA}}$	Γ	L	X	Γ	L	X	Γ	L	X	Γ	L	X
Si	−89.6	3.17	2.09	1.14	3.17	2.09	1.14	3.17	2.06	1.15	3.16	2.02	1.11
Ge	−24.7	0.98	0.74	0.98	0.96	0.73	0.95	0.88	0.71	0.99	0.84	0.68	0.97
GaAs	−14.8	1.65	1.83	1.86	1.63	1.82	1.85	1.54	1.75	1.83	1.46	1.69	1.79
MgO	−71.4	7.31	10.55	11.62	7.36	10.56	11.62	7.30	10.55	11.60	7.36	10.55	11.60

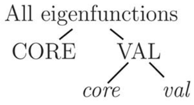
are added, as Table I shows. (In this particular case it is the O $3p$ contribution that is dominant; however, cases arise when the contributions from high-lying d or f orbitals can be of order 1–2 eV. NiO is such a case[11].)

The inset of Fig. 4 shows some PPGW results for reference. Based on the observation that cutoff in the Hilbert space should be important, PPGW data by Tiago *et al.* should be extrapolated to $1/N'\to0$ because they used a very large LDA basis.

V. CORE CONTRIBUTIONS TO Σ_c

In a series of papers, Shirley and co-workers analyzed the effects of the core on QPE in atoms (Shirley, Mitáš, and Martin,[34] and Shirley and Martin[35]) semiconductors (Shirley, Zhu, and S. G. Louie[36,37]) within the pseudopotential framework. Approximate core contributions to both Eqs. (16) and (17) were evaluated. They also compared pseudopotentials constructed from both LDA exchange and from Hartree-Fock (HF) exchange for atoms and molecules,[38] and incorporated pseudopotentials of both types in studying core effects.[36,37] They found sizable shifts in QPE in Si, and rather dramatic and k-dependent shifts in Ge and GaAs. These analyses highlight the importance of core effects. However, the decomposition of the various core contributions in Ref. 37 is somewhat involved, and it is rather closely tied to the pseudopotential construction that was a part of their implementation. This makes it a little difficult to disentangle the various contributions.

Here we examine contributions from the shallowest cores to Σ_c within the framework of our *GW*. As we noted, all the eigenfunctions are divided into two groups, CORE and VAL, as explained in Sec. II. Using local orbitals we can represent the shallowest cores in VAL. To distinguish true core effects from artifacts of implementation,[28] we include these cores in the valence with local orbitals and treat them in a special way, as described below. We denote such eigenfunctions as *core*, and the rest as *val*. Thus we distinguish three kinds of orbitals: CORE, *core*, and *val*;



(20)

In Si, for example, we use CORE=$\{1s,2s\}$, *core*=$\{2p\}$, and *val*=$\{3s,3p,3d,\ldots\}$. Because the *core* states are well separated from higher-lying states, G can be partitioned into $G=G^{\mathrm{CORE}}+G^{core}+G^{val}$. Σ_x is always calculated from the entire G, while Σ_c is calculated from *core* and *val* only (we do not consider any case where some portion of the self-energy is supplied by the LDA): $\Sigma_c=i(G^{core}+G^{val})W^c$, where $W^c=W[\Pi]-v$, and Π is calculated from $G^{core}+G^{val}$. Thus *core* states contribute to Σ_c directly through G in iGW^c, and also through W^c. We resolve these contributions; that is, we calculate Σ_c in one of four ways:

(i) Neglect the *core* contribution to Σ_c entirely: i.e., $\Sigma_c=iG^{val}(W[\Pi^{val}]-v)$, where $W[\Pi^{val}]$ means that Π is calculated from G^{val} only. We denote this as "exchange-only *core*."

(ii) Neglect the *core* contribution to screening: $\Sigma_c=i(G^{core}+G^{val})(W[\Pi^{val}]-v)$.

(iii) Neglect the *core* contribution to G: $\Sigma_c=iG^{val}(W[\Pi]-v)$.

(iv) $\Sigma_c=i(G^{core}+G^{val})(W[\Pi]-v)$: there is no distinction between *core* and *val* states.

Table II shows that the difference between exchange-only (i) and *GW* (iv) approximations to core treatment is small in Si (~0.03 eV for X_{1c}). As expected, the adequacy of an exchange-only core depends on how deep the core is. The exchange-only approximation for shallow cores, such as the Ga $3d$ and In $4d$, and the highest-lying p core in column I (Na, K, Rb) and column IIA alkali metals (Mg, Ca, and Sr), is rather crude. It is interesting that the core contributions to Π and to G are *not* always additive.

The difference between (iii) and (iv) is in general rather small; that is, the inclusion of *core* contributions to Π alone is sufficient to bring QPE results within 0.05 eV of the full results in Table II except for the very shallow Ga $3d$ channel. For moderately deep cores, exchange-only treatment (i) is generally adequate, as Aryasetiawan suggested. A rough rule of thumb seems to be: for cores whose total charge Q_{spill} outside the augmentation radius is less than 0.01 electrons, exchange-only treatment of them results in errors ~0.1 eV or less for the lowest excited states. (This radius may be taken as approximately half the nearest-neighbor bond length.)

Inclusion of core contributions to Π can significantly increase the computational cost (in the Si case, leaving out the $2p$ contribution to Π reduces the computational cost by ~40%). The relative smallness of corrections to exchange-only treatment, and the observation that core contributions to Π alone are adequate for all but the most shallow cores, suggests that a simple approximate inclusion of core contributions to Π [Eq. (16)], should be adequate for all but the most shallow cores such as the Ga $3d$. (Fleszar and Hanke proposed a construction for pseudopotentials when core states are not pseudized.[39]) Supposing the core was confined to the augmentation sphere at site R, we can eliminate all contributions to the matrix element $\langle \tilde{M}_I^{\mathbf{q}} \Psi_{\mathbf{k}n} | \Psi_{\mathbf{q}+\mathbf{k}n'} \rangle$ except from the product-basis contribution at R. Since also the augmented part of Ψ depends rather weakly on $\dot{\varphi}_{Rl}$, we can neglect the $\dot{\varphi}_{Rl}$ contribution to the eigenfunctions and assume that $\langle \tilde{M}_I^{\mathbf{q}} \Psi_{\mathbf{k}n} | \Psi_{\mathbf{q}+\mathbf{k}n'} \rangle$ only depends on n, n', $\mathbf{k}$, or $\mathbf{k}+\mathbf{q}$ through the coefficients, $(\alpha_{Ru}^{\mathbf{k}n})^* \alpha_{Ru}^{\mathbf{k}+\mathbf{q}n'}$. Moreover, the core-level energy is large and negative, and nearly independent of $\mathbf{k}$ or n. Since the dominant contributions to Π will come from coupling to low-lying states, we can approximate $\varepsilon_{kn} - \varepsilon_{k'n'}$ by a constant, e.g., $E_F - \varepsilon_{\text{core}}$. These approximations are all modest but can vastly simplify the computation of Π^{CORE}.

The fact that the core spills out slightly from the augmentation region needs to be taken into account.[22] This can readily be accomplished by integrating the core and corresponding valence φ_l to a larger radius, and orthogonalizing φ_l to the core. Checks show that the adjustment to ϕ_l is small unless the core is very shallow, in which case the core should be treated as a valence state.

A. Comparison to PPGW

When the highest cores are put explicitly into the valence as Tiago *et al.* did, there is reasonable agreement between PPGW and our results for *sp* semiconductors. Comparison with the paper of Tiago *et al.* to Table II above shows that there is agreement at the ~0.1 eV level in Si (Ref. 40) and a similar agreement is found for GaAs and Ge, with the PPGW results systematically higher than our results by ~0–0.1 eV. Similarly, Fleszar and Hanke[39] calculated QPEs in the GWA for a variety of II-VI semiconductors, including the highest s and p cores in the valence. Their values are also in reasonable agreement with the results presented in Table III, though the PPGW data are systematically higher by ~0.0–0.2 eV. (Part of the discrepancy can be traced to contributions from high-lying d states, which are included in the present calculation using local orbitals.) Even when the high-lying s and p core states are included explicitly in the valence, it still seems to be the case that PPGW band gaps are systematically slightly larger (by ~0.1 eV) in semiconductors than our *GW* predict.

Materials involving transition metals are rather more complicated. In a recent PPGW calculation, Marini, Onida, and del Sole analyzed the QP valence bands of Cu,[41] comparing in some detail the occupied d bands to photoemission experiments. The LDA places the position of these levels approximately 0.5 eV closer to E_F than the experiments show. The authors find that the d bands narrow and shift downward by approximately 0.5 eV, bringing the PPGW d bands into excellent agreement with photoemission experiments. They report that the PPGW results depend rather dramatically on the treatment of the Cu core $3s$ and $3p$ levels: that it is necessary to include both states explicitly in the valence to obtain reasonable results. They found that the correlation contribution Σ_c^{core} from these states shifts the d bands downward ~0.5 eV.

We conducted a similar calculation by using the present all-electron *GW*, and find a very different result. In our case, the *GW* correction to the LDA d bands is small—between 0 and 0.1 eV. Moreover, QPEs are essentially independent of how the Cu $3p$ state is treated: the $3d$ levels change by less than 0.05 eV when the Cu $3p$ state is explicitly included in the valence (using a $3p$ local orbital), as compared to being treated as core at the exchange-only level. The Cu $3p$ state is rather deep, and the weak dependence on correlation contributions from it is consistent with the rule of thumb indicated above: $Q_{\text{spill}} \approx 0.005$ electrons; $\epsilon_{3p}^{\text{LDA}} \approx -70$ eV. In the Cu case, it appears likely that the main discrepancy between PPGW and our *GW* (whether d bands shift by 0.5 eV or not) originates in the discrepancies in Σ_c^{core}.

VI. ADEQUACY OF GWA APPLIED TO A RANGE OF MATERIALS

In Ref. 10, Delaney *et al.* argued that GWA based on the LDA eigenfunctions and eigenvalues, is an adequate (or better) approximation than self-consistent *GW*. It is apparently the case that self-consistency worsens agreement with experiment for the Be atom. Moreover, Holm and von Barth[42] found that the valence bandwidth of the homogeneous electron gas is considerably worsened by self-consistency; similarly a self-consistent *GW* calculation by Ku and Eguiluz resulted in an overestimate of the valence bandwidth in Ge.[6] Thus, self-consistency of this type has shortcomings. On the other hand, even in simple materials such as *sp* semiconductors, GWA band gaps based on LDA eigenvalues and eigenfunctions are *always* underestimated when properly calculated.[2,39] The GWA based on LDA is evaluated as a perturbation relative to LDA; thus the band gap can be poor if the LDA itself is poor. Thus, some kind of self-consistency is necessary to reduce the dependence on the starting point.

In this section, we consider three points about the GWA based on LDA, Eq. (6):

(A) Use of the Z factor. We show that using $Z=1$ in Eq. (6) is a way to include partial self-consistency, and it should be a better approximation than including the Z factor.

(B) Off-diagonal Σ. Equation (6) is a perturbation treatment that involves only the diagonal matrix element of Σ. We consider the effect of the full Σ in a variety of systems analyzing how the adequacy of GWA is dependent on the adequacy of LDA. Even GWA with $Z=1$ fails for cases when the starting LDA is poor.

TABLE III. Fundamental gap, in eV. (For Gd, QPE correspond to the position of the majority and minority f levels relative to E_F; for Cu QPE corresponds to the d level.) Low-temperature experimental data were used when available. QPEs in the "*GW*" column are calculated with usual GWA Eqs. (6) and (7). In the "$Z=1$" column the Z factor is taken to be unity. In the "$\Sigma_{nn'}$" column the off-diagonal parts of Σ are included in addition to taking $Z=1$. k meshes of $8\times8\times8$ k and $6\times6\times6$ were used for cubic and hexagonal structures, respectively (symbol w indicates the wurtzite structure). *GW* calculations leave out spin-orbit coupling and zero-point motion effects. The former is determined from $\Delta/3$, where Δ is the spin splitting of the Γ_{15v} level (in the zinc-blende structure); it is shown in the "$\Delta/3$" column. Contributions to zero-point motion are estimated from Table 2 in Ref. 45 and are shown in the "*ZP*" column. The "adjusted" gap adds these columns to the true gap, and is the appropriate quantity to compare to *GW*.

			GW	*GW*				
	LDA	*GW*	$Z=1$	$\Sigma_{nn'}$	Expt.	$\Delta/3$	ZP	Adj
C	4.09	5.48	5.74	5.77	5.49	0	0.37	5.86
Si	0.46	0.95[a]	1.10	1.09	1.17	0.01	0.06	1.24
Ge	−0.13	0.66	0.83	0.83	0.78	0.10	0.05	0.93
GaAs	0.34	1.40	1.70	1.66	1.52	0.11	0.05	1.68
wAlN	4.20	5.83	6.24		6.28	0[b]	0.24	6.52
wGaN	1.88	3.15	3.47	3.45	3.49	0[b]	0.17	3.66
wInN	−0.24	0.20[a]	0.33		0.69	0[b]	0.1	0.79
wZnO	0.71	2.51	3.07	2.94	3.44	0[b]	0.16	3.60
ZnS	1.86	3.21	3.57	3.51	3.78	0.03	0.09	3.90
ZnSe	1.05	2.25	2.53	2.55	2.82	0.13	0.05	3.00
ZnTe	1.03	2.23	2.55		2.39	0.30	0.05	2.74
CuBr	0.29	1.56	1.98	1.96	3.1	0.04	0.01	3.15
CdO	−0.56	0.10	0.22	0.15	0.84	0.01[b]		
CaO	3.49	6.02	6.62	6.50	~7	0[b]		
*w*CdS	0.93	1.98	2.24		2.50	0.03	0.07	2.60
$SrTiO_3$	1.76	3.83	4.54	3.59	~3.3			
ScN	−0.26	0.95	1.24	0.96	~0.9	0.01[b]		
NiO	0.45	1.1	1.6		4.3			
Cu[c]	−2.33	−2.35	−2.23	−2.18	−2.78			
Cu[d]	−2.33	−2.85	−2.73	−2.18	−2.78			
Gd$^{\uparrow}$	−4.6	−5.6	−6.2	−4.1	−7.9			
Gd$^{\downarrow}$	0.3	0.2	1.8	1.5	4.3			

[a] See Ref. 40.

[b] LDA calculation.

[c] Position of Γ_{12} d level, with E_F set to charge-neutral point.

[d] Position of Γ_{12} d level, with E_F set to LDA value.

(C) Band-disentanglement problem. Even when LDA eigenfunctions are reasonable, if eigenvalues are wrongly ordered the perturbation treatment can have important adverse consequences.

A. *Z* factor

Let us consider a partial kind of self-consistency where only eigenvalues are updated: both eigenfunctions and W are unchanged from the LDA. This is a little different from the usual eigenvalue-only self-consistency, where eigenfunctions are frozen but W is updated. Updating eigenvalues widens semiconductor band gaps. This reduces the screening, which causes W to increase, which in turn causes gaps to increase still more. Thus we expect that results from such kind of partial self-consistency we are considering here should fall somewhere between the usual one-shot *GW* and the usual eigenvalue-only self-consistency. Partial self-consistency, while incomplete, should result in better QPEs than the standard one-shot *GW*, since eigenvalues shift in the right direction. The Appendix evaluates how this kind of self-consistency modifies Eq. (6) for a model two-level system. The result is that this kind of self-consistency can be approximately realized by putting $Z_{\mathbf{k}n}=1$ in Eq. (6). A different justification for omitting the Z factor emerged from a paper of Niquet and Gonze,[43] who calculated the interacting band-gap energy (within RPA) to obtain a correction to the Kohn-Sham gap. They found that the difference is essentially Eq. (6) with $Z=1$. Finally, a further justification for using $Z=1$ is discussed in Chapter 7 of Ref. 44. $Z=1$ corresponds to the Rayleigh-Schrödinger perturbation, Z from Eq. (19) to the Brillouin-Wigner perturbation. It shows the $Z=1$ scheme should be better for the Fröhlich Hamiltonian Mahan analyzed.[44]

The calculations in Table III support the argument that using $Z=1$ is a better approximation than including Z: semiconductor band gaps are in significantly better agreement with experiment. They continue to be smaller than experimental values, which can be qualitatively understood as follows. Using $Z=1$ corresponds to updating G, but leaving W determined from the LDA eigenfunctions and eigenvalues. Because the gap is underestimated in the construction of Π and W, Π is overestimated so that W is screened too strongly; thus Σ is too small. It is interesting, however, that QPEs evaluated with $Z=1$ can be rather good at times because of a fortuitous cancellation of errors. We can refine self-consistency by updating W in a manner similar to the updating of G; that is, using eigenvalues from Eq. (6) in the calculation of Π. However, ϵ computed in the RPA ($\epsilon=1-v\Pi$), omits excitonic effects. Inclusion of electron-hole correlations to Π (via e.g., ladder diagrams) increases Im $\epsilon(\omega)$ for ω in the vicinity of the gap for semiconductors. There is a concomitant increase in Re $\epsilon(\omega)$ for $\omega\rightarrow 0$, as is evident by the Kramers-Kronig relations; see e.g., Ref. 46. Errors resulting from the neglect of excitonic contributions to ϵ partially cancel errors resulting from LDA eigenvalues, as shown by Arnaud and Alouani.[46] Thus, W calculated from LDA eigenvalues is not so bad in many cases because of this cancellation. Often ϵ_∞ calculated from the LDA eigenvalues is better than ϵ_∞ calculated from LDA eigenvalues shifted by a scissors operator to match the experimental band gap (see Table III in Ref. 46). This cancellation means that $GW(Z=1)$ can often be rather good, since W itself is also better than what would be obtained from (eigenvalue) self-consistency. Table III shows that the fundamental gap for $GW(Z=1)$ is quite good for mostly covalent semiconductors such as Si or GaAs, but that the agreement deteriorates as the ionicity increases.

B. Off-diagonal contributions of Σ

The usual GWA in Eq. (6) does not include the off-diagonal contribution of $\Sigma-V_{xc}^{\mathrm{LDA}}$. A simple way to take into account the contribution of off-diagonal parts is to replace the energy-dependent matrix Σ with some static Hermitian matrix V^{xc} as in the following, and to solve the eigenvalue problem, replacing V_{xc}^{LDA} in the LDA Hamiltonian with this potential. We take

$$V^{xc}=\frac{1}{2}\sum_{ij}|\psi_i\rangle\{\mathrm{Re}[\Sigma(\varepsilon_i)]_{ij}+\mathrm{Re}[\Sigma(\varepsilon_j)]_{ij}\}\langle\psi_j|, \qquad (21)$$

for Σ. Here Re signifies the Hermitian part; the eigenvalues ε_i and the eigenfunctions ψ_i are in LDA. This V^{xc} is used in our QP self-consistent GW method.[11–13] This V^{xc} retains the diagonal part contribution as in Eq. (6) (we now consider the $Z=1$ case). From the perspective of the QP self-consistent GW method, including the off-diagonal Σ corresponds to the first iteration, and the LDA corresponds to the zeroth iteration. Table III shows how the fundamental gap is affected by the off-diagonal parts of V^{xc} for selected semiconductors. Because the semiconductor eigenfunctions and density are already rather good, the off-diagonal contributions are small. Contributions from the off-diagonal part of V^{xc} significantly increase when eigenfunctions have significant d character (see $\mathrm{SrTiO_3}$ and ScN in Table III). For correlated systems the effects can be rather dramatic; see Ref. 12 for how the QPEs are affected by the off-diagonal parts of Σ in $\mathrm{CeO_2}$.

In general, GWA errors are rather closely tied to the quality of LDA starting point. In the covalent sp semiconductors C, Si, and Ge, GW gaps are rather good for $Z=1$. In the series Zn(Te, Se, S, O), the deviation between the LDA and experimental gap steadily worsens, and so does the GW gap. For ZnO, and even more so in CuBr, the GW gap falls far below experiment. For these simple sp materials, errors are related to their ionicity, which can be seen qualitatively as follows. As ionicity increases, the dielectric response becomes smaller; consequently the nonlocality missing from the LDA exchange-correlation potential[47] becomes progressively more important. Roughly speaking, a reasonable picture of electronic structure in sp systems resembles an interpolation between the LDA, which has no nonlocality in the exchange and underestimates gaps, and Hartree-Fock, which has nonlocality but wildly overestimates gaps because the nonlocal exchange is not screened. As ionicity increases the gap widens and the dielectric function decreases. As the screening is reduced the LDA becomes a progressively worse approximation. Thus, the LDA is not an adequate starting point for GW in the latter cases.

Discrepancies between GW and experiments become drastic when electronic correlations are strong. The GWA band gap for the antiferromagnetic-II NiO is far from experiment, and moreover the conduction-band minimum falls at the wrong place (between Γ and X [11]). As Table III shows, the LDA puts f levels in Gd too close to E_F. GWA results are only moderately better: shifts in the Gd f level relative to the LDA are severely underestimated (see Table III).

GW based on the LDA fails even qualitatively in CoO: it predicts a metal with E_F passing through an itinerant band of d character. In this case, the GWA gives essentially meaningless results. To get reasonable results it is essential to apply the GWA with a starting point that already has a gap. To get a band gap for CoO in the single-band picture, a nonlocal potential, which breaks time-reversal symmetry is required, something which is not built into the local potential of the LDA. Similar problems occur with ErAs: the LDA predicts a very narrow minority f band straddling E_F, whereas in reality the minority f manifold is exchange-split into several distinct levels well removed from E_F.[48] GWA shifts the minority f levels only slightly relative to the LDA: the entire f manifold remains clustered in a narrow band at the Fermi level, appearing once again qualitatively similar to the LDA.

Generally speaking, the GWA [even GWA($Z=1$)] is reasonable only under limited circumstances—when the LDA itself is already reasonable.

C. Band disentanglement problem

Even for the simple sp semiconductors, there can be a “band disentanglement problem” as a consequence of the

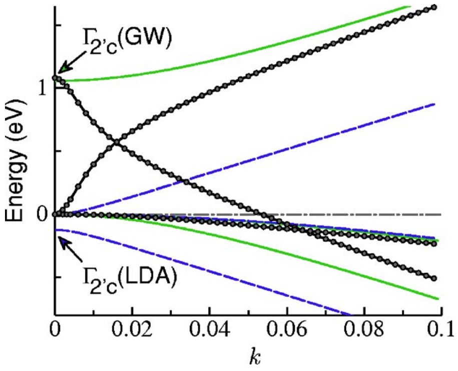


FIG. 6. (Color online) Energy bands in Ge for $\mathbf{k}=2\pi/a[00k]$ for small k within the GWA, using $Z=1$. Spin-orbit coupling was omitted. Three approximations are compared: LDA (dashed blue line), *GW* in the diagonal-Σ-only approximation, Eq. (6) (black line with circles), and *GW* with Σ computed according to Eq. (21) (solid green line). In all three cases the three states of p character ($\Gamma_{25'}$ symmetry) form the valence-band maximum; this was taken to be the energy zero. The LDA predicts the conduction band, the $\Gamma_{2'}$ state of s character, to be slightly negative, causing the energy bands to be wrongly ordered at Γ. For $k>0$, the $\Gamma_{25'}$ state of p_z symmetry couples to the $\Gamma_{2'}$ state, and the two repel each other. Both kinds of *GW* put $\Gamma_{2'c}$ at approximately the correct position, 1 eV. However, the diagonal-only *GW* must follow the topology of the LDA: the eigenvectors are unchanged from the LDA. Therefore the band starting out at $\Gamma_{2'c}$ sweeps downward, while the p_z band starting out at $\Gamma_{25'v}$ sweeps upward, and the two bands cross near $k=0.02$. When the off-diagonal parts of Σ are included, these two bands repel each other as they should.

diagonal-only approximation. At times the LDA orders energy levels wrongly: in hcp Co, for example, it inverts the order of the minority Γ_5 and Γ_3 levels, which correspond to states of L_3 and $L_{2'}$ symmetry in the fcc structure. Wrong ordering of levels is a particularly serious difficulty for narrow-gap semiconductors such as Ge, InAs, InSb, and InN. Because the LDA underestimates band gaps, the energy-band structure around Γ has an inverted structure: the s-like conduction band of Γ_1 symmetry (labeled as $\Gamma_{2'}$ in the homopolar case) incorrectly falls below the p-like states of $\Gamma_{25'}$ symmetry.

When the GWA is evaluated from Eq. (6), the energy bands retain the same connectivity as in the LDA, as Fig. 6 shows. Consequently, the conduction band has a nonsensical negative mass near Γ, and it crosses with one of the valence bands. The diagonal-only approximation cannot make Ge an insulator in principle, even though the levels are properly ordered at Γ. This problem is avoided if the off-diagonal parts of Σ are included, as Fig. 6 shows. The conduction-band effective mass in the latter case is computed to be $m^*=0.042m_0$, in good agreement with a value of $m^*=0.038m_0$ estimated from magnetopiezoreflectance spectra.[49] This shows that the off-diagonal contributions of Σ are reasonably well described by Eq. (21).

VII. CONCLUSIONS

To conclude, we have analyzed various possible sources of error in implementations of the GWA, using calculations based on an all-electron method with generalized linear muffin-tin orbitals as a basis. We analyzed convergence in QPEs with the number of unoccupied states N': the rate of convergence for intermediate N' (where the LMTO energy bands were shown to precisely replicate APW bands), was qualitatively similar to, but roughly twice that of a PP analysis by Tiago, Ismail-Beigi, and Louie.[9] On the other hand, it closely tracked the convergence calculated by an LAPW +local orbitals method, which had a very similar LDA band structure. More generally, those GWA that properly subtract V_{xc}^{LDA} calculated from the full density are in reasonable agreement with each other;[2,4,6,11,50] those that subtract valence density only[5,7] are also in reasonable agreement for cases such as Si and SiC where the cores are sufficiently deep. Our own experience suggests that the LDA treatment of core levels, where QPE are computed from Eq. (18), will be problematic for *GW* (Ref. 2) unless the cores are very deep. Since a PP construction is an approximation whose justification is grounded in an all-electron theory, we should expect *GW* calculations based on a LDA PP should be similarly problematic. There is apparently a significant dependence on how cores are treated in PP implementations,[37,39,41,51] even in Si and Cu with their deep $2p$ and $3p$ cores.

We then presented a new analysis of convergence that is of particular importance for minimal-basis implementations, and argued that measuring convergence in the traditional cutoff procedure—by the number of unoccupied states N' as given in Figs. 2 and 4—is not particularly meaningful for a minimal basis. We presented an alternative truncation of the full Hilbert space of eigenfunctions, namely, to use the entire Hilbert space of a relatively small basis. We showed that a suitably constructed minimal basis is sufficient to precisely describe the GWA QPE within 1 Ry or so of the Fermi level, and that this kind of cutoff procedure seems to be more efficient than the traditional N' cutoff of a large basis. We also showed that traditional linearization of basis functions, either explicit in an all-electron method or implicit through the construction of a pseudopotential, result in errors approximately independent of the size of basis. The addition of local orbitals to extend the linear approximation results in modest shifts in sp nitride and oxide compounds, and shifts of the order 1–2 eV in transition-metal oxides.

We analyzed core contributions to the self-energy, and showed that an exchange-only treatment of the core is adequate in most cases. For all but the most shallow cores (such as Na $2p$ and Ga $3d$), we showed that it is sufficient to include the core contribution to the polarization only; an approximate and rather painless implementation was suggested. These results can provide a framework for improved treatment of the core within a pseudopotential approximation.

Finally, we considered the adequacy of GWA based on the LDA, for different kinds of materials, and also Eq. (6) as an approximation to the GWA. We presented logical and numerical justifications that using $Z=1$, and showed that it generally gives better band gaps in insulators. In general, inclusion of the off-diagonal part of Σ and some kind of self-consistency is essential to make the GWA a universally applicable and predictive tool. Taking into account both theoretical and practical aspects, the quasiparticle self-consistent *GW* scheme we have proposed[11–13] has the potential to be an excellent candidate for such a tool: it obviates some of the difficulties seen in the standard self-consistency, it no longer depends on the LDA, and it appears to predict QPEs in a consistently reliable way for broad classes of materials.

ACKNOWLEDGMENTS

This work was supported by ONR Contract No. N00014-02-1-1025. S.F. was supported by DOE, BES Contract No. DE-AC04-94AL85000. We thank the Ira A. Fulton High Performance Computing Initiative for the use of their computer facilities. Our GW code was developed out of an LMTO-ASA GW code originally written by Ferdi Aryasetiawan, and whose contribution to this work we gratefully acknowledge.

APPENDIX: JUSTIFICATION FOR $Z=1$

Let us consider a limited self-consistency within GWA as follows. We restrict self-consistency as follows:

(1) We make only the QPE self-consistent. Eigenfunctions are constrained to be the LDA eigenfunctions.

(2) W is assumed to be fixed. Thus only the eigenvalues entering into G are made self-consistent.

Under these assumptions, we can show that QPEs are rather well approximated by Eq. (6) with $Z=1$. To illustrate it, consider a two-states model whose LDA eigenvalues and eigenfunctions are given by ψ_1,ε_1, and ψ_2,ε_2, and the Fermi energy falls between these states: $\varepsilon_2 > E_{\mathrm{F}} > \varepsilon_1$. Then the LDA Green's function is

$$G^{\mathrm{LDA}}(\omega) = \frac{|\psi_1\rangle\langle\psi_1|}{\omega-\varepsilon_1-i\delta} + \frac{|\psi_2\rangle\langle\psi_2|}{\omega-\varepsilon_2+i\delta}. \tag{A1}$$

After the limited self-consistency is attained, we will have eigenvalues

$$G(\omega) = \frac{|\psi_1\rangle\langle\psi_1|}{\omega-E_1-i\delta} + \frac{|\psi_2\rangle\langle\psi_2|}{\omega-E_2+i\delta}, \tag{A2}$$

where E_1 is given by

$$E_1 = \varepsilon_1 + \mathrm{Re}\langle\psi_1|\Sigma(E_1,[G]) - V_{xc}^{\mathrm{LDA}}|\psi_1\rangle. \tag{A3}$$

There is a similar equation for E_2. Note that $\Sigma(E_1,[G])$ is calculated in GWA from G of Eq. (A2) at E_1.

As we can expect that W is dominated by diagonal terms $W_1(\omega)=\langle\psi_1\psi_1|W(\omega)|\psi_1\psi_1\rangle$ and $W_2(\omega)=\langle\psi_2\psi_2|W(\omega)|\psi_2\psi_2\rangle$, we neglect other matrix elements of $W(\omega)$. Then Σ becomes

$$\begin{aligned}
&\mathrm{Re}\langle\psi_1|[\Sigma(E_1,[G])]|\psi_1\rangle \\
&\quad = \mathrm{Re}\int \langle\psi_1|iG(E_1+\omega')W(\omega')|\psi_1\rangle d\omega' \\
&\quad \approx \mathrm{Re}\int \frac{iW_1(\omega')d\omega'}{E_1+\omega'-E_1-i\delta} \\
&\quad = \mathrm{Re}\int \frac{iW_1(\omega')d\omega'}{\varepsilon_1+\omega'-\varepsilon_1-i\delta} \\
&\quad = \mathrm{Re}\langle\psi_1|\Sigma(\varepsilon_1,[G^{\mathrm{LDA}}])|\psi_1\rangle.
\end{aligned} \tag{A4}$$

A similar equation applies for E_2. The energy shift $E_1 \to \varepsilon_1$ entering into the evaluation Σ is exactly compensated by the energy shift in $G \to G^{\mathrm{LDA}}$, or equivalently using $Z=1$ is an approximate way to obtain self-consistency. Equation (A4) corresponds to Eq. (6) with $Z=1$.

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

[2] T. Kotani and M. van Schilfgaarde, Solid State Commun. **121**, 461 (2002).

[3] M. Usuda, N. Hamada, T. Kotani, and M. van Schilfgaarde, Phys. Rev. B **66**, 125101 (2002).

[4] C. Friedrich, A. Schindlmayr, S. Blügel, and T. Kotani (unpublished).

[5] N. Hamada, M. Hwang, and A. J. Freeman, Phys. Rev. B **41**, 3620 (1990); URL http://link.aps.org/abstract/PRB/v41/p3620

[6] W. Ku and A. G. Eguiluz, Phys. Rev. Lett. **89**, 126401 (2002).

[7] S. Lebègue, B. Arnaud, M. Alouani, and P. E. Bloechl, Phys. Rev. B **67**, 155208 (2003).

[8] Of the several all-electron implementations of the *GW* method that have been developed to date, some of them treat the core at the LDA level, and subtract the valence-only part of the LDA exchange-correlation potential (Ref. 7). Thus, care must be taken when comparing different all-electron calculations. In the Si and SiC cases the cores are very deep, and it matters little whether only the valence electrons or all the electrons are included in the *GW* potential.

[9] M. L. Tiago, S. Ismail-Beigi, and S. G. Louie, Phys. Rev. B **69**, 125212 (2004).

[10] K. Delaney, P. García-Gonzaléz, A. Rubio, P. Rinke, and R. W. Godby, Phys. Rev. Lett. **93**, 249701 (2004).

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

[12] M. van Schilfgaarde, T. Kotani, and S. Faleev, Phys. Rev. Lett. **96**, 226402 (2006); URL http://link.aps.org/abstract/PRL/v96/e226402

[13] A. N. Chantis, M. van Schilfgaarde, and T. Kotani, Phys. Rev. Lett. **96**, 086405 (2006); URL http://link.aps.org/abstract/PRL/v96/e086405

[14] The *GW* fundamental gap for Si was incorrectly quoted as 0.84 eV in Ref. 11. The correct value is 0.91 eV for the basis used there.

[15] M. Methfessel, M. van Schilfgaarde, and R. A. Casali, *Lecture Notes in Physics*, Vol. 535, edited by H. Dreysse (Springer-Verlag, Berlin, 2000).

[16] O. K. Andersen, Phys. Rev. B **12**, 3060 (1975).

[17] E. Bott, M. Methfessel, W. Krabs, and P. C. Schmidt, J. Math. Phys. **39**, 3393 (1998).

[18] P. E. Blochl, Phys. Rev. B **50**, 17953 (1994).

[19] A recent LAPW-*GW* calculation by C. Friedrich, A. Schindlmayr, S. Blügel, and T. Kotani, Phys. Rev. B **74**, 045104 (2006). in-

cluding eigenfunctions with about 300 valence bands, obtained a band gap slightly larger (by ~0.04 eV) than the results presented here, when similar input conditions for the *GW* calculation are used.

[20] T. Kotani, M. van Schilfgaarde, and S. Faleev, cond-mat/061102 (unpublished).

[21] In the LDA a small amount of the envelope function remains inside the augmentation spheres to handle incomplete *l* convergence of the augmentation functions. In the present *GW* implementation this feature is not retained because of the nonlocal character of the potential. As a consequence, wave-function products must be augmented to higher *l* than in the LDA case.

[22] In the *GW* scheme as we have currently implemented it, the LDA core states are approximated by truncating them outside the MT spheres and scaling the part inside the MT sphere to conserve charge. An alternative would be to integrate the core wave function subject to the boundary condition that the value and slope vanish at the MT boundary. In either case, the core is not strictly orthogonal to the valence; this appears to be the primary source of error in the present treatment of the core. Experience with the more precise local orbitals treatment shows that this approximation is not serious at the "exchange-only" level. However, their contribution to correlation is reliable only for relatively deep cores such as the Si $2p$.

[23] F. Aryasetiawan and O. Gunnarsson, Phys. Rev. B **49**, 16214 (1994).

[24] J. Rath and A. Freeman, Phys. Rev. B **11**, 2109 (1975).

[25] Using the linear tetrahedron method, a mesh of $6\times6\times6$ divisions (doubling the number of points in the energy denominator) resulted in an approximately rigid shift of +0.02 eV relative to a calculation with an $8\times8\times8$ mesh, while a $4\times4\times4$ mesh was found to overestimate the conduction bands by 0.10 eV. These shifts were approximately unchanged when varying, e.g., the basis set or the number of unoccupied states.

[26] W. Ku and A. G. Eguiluz, Phys. Rev. Lett. **93**, 249702 (2004).

[27] A. Yamasaki and T. Fujiwara, J. Phys. Soc. Jpn. **72**, 607 (2003).

[28] We can treat the core state explicitly as valence states through the use of local orbitals, or solve for the core independently (see Sec. II A). For a deep core such as the Si $2p$ it matters little whether the core eigenfunctions are computed separately or generated through a local orbital, even with the approximate core treatment (Ref. 22). This is expected since the core level is very deep. Since the LDA eigenfunctions are represented completely different in the two cases, the stability is a confirmation of the robustness of the method. For shallow cores with their larger radial extent, approximations in our present treatment of core states (Ref. 22) do not enable us to distinguish true core contributions from artifacts of the implementation, so we use the local orbital approach here.

[29] Comparing the two basis sets of Fig. 4, the LDA-occupied eigenvalues $\varepsilon_{\mathbf{k}n}$ and total energy differ by $\sim10^{-3}$ and $\sim10^{-2}$ eV, respectively. Thus, even though the LDA eigenvalues used to generate Σ are nearly identical for small N', the gap changes by ~0.01 eV. In principle, both curves in Fig. 4 should converge to a common value as N' becomes small. That they deviate slightly is not a measure of convergence in the basis, but is connected with limitations to numerical precision in the method. It is apparently unavoidable because the eigenfunctions are represented in very different ways in the two basis sets: slight differences in them are responsible for the gap difference. Thus, the gap is ~0.01 eV higher in the small-basis case *both* for small N' and at their maximum N': it is an indication that incompleteness of basis is not the dominant source of the discrepancy between the two basis sets. Apparently random changes in eigenvalues of order ±0.01 eV are also seen in Fig. 5 (for a fixed choice of $\{\varphi_{Ru}\}$) for a wide variety of basis sets. Similar discrepancies are also seen when comparing to the gap generated by an LAPW basis: with similar input conditions for the *GW* part—similar tolerances and approximately similar augmentation functions $\{\varphi_{Ru}\}$—the LAPW gap falls ~0.04 eV higher than those calculated by the present method (Ref. 19). Moreover, there is a slight dependence of the gap on augmentation radius: shrinking augmentation sphere volume by 30% has the effect of increasing the gap by ~0.01 eV. Thus numerical accuracy is limited to order ~0.1 eV or a little less for Si.

[30] L. Steinbeck, A. Rubio, L. Reining, M. Torrent, I. White, and R. Godby, Comput. Phys. Commun. **125**, 105 (2000).

[31] We reiterate the present LMTO method has a key advantage in that it can be made small and also highly precise when needed (see Fig. 1).

[32] Those familiar with the usual N' truncation analysis might ask why if applying it to a small-basis set calculation results in a different convergence rate, whether the small-basis calculation is adequately converged. This can be understood as follows. High-lying eigenvalues in the small basis tends to some kind of average of the true higher states; unlike a basis that spans the Hilbert space of those states, individual states do not have physical meaning. Only the collective contribution from the sum over higher-lying states to Eqs. (16) and (17) has physical meaning in a finite basis, namely, $\Pi_{IJ}(\omega)$ and $\Sigma_c(\omega)$ within the Hilbert space of eigenfunctions. All that is necessary for completeness is that the basis have enough freedom in it to describe $\Pi_{IJ}(\omega)$ and $\Sigma_c(\omega)$ on the energy scales of relevance. The completeness of a particular basis must be checked numerically by making it evermore complete and monitoring the convergence, as is done in traditional LDA calculations.

[33] O. K. Andersen, T. Saha-Dasgupta, R. W. Tank, and G. K. C. A. O. Jepsen, *Lecture Notes in Physics*, Vol. 535, edited by H. Dreysse (Springer-Verlag, Berlin, 2000).

[34] E. L. Shirley, L. Mitás, and R. M. Martin, Phys. Rev. B **44**, 3395 (1991).

[35] E. L. Shirley and R. M. Martin, Phys. Rev. B **47**, 15404 (1993).

[36] E. L. Shirley, X. Zhu, and S. G. Louie, Phys. Rev. Lett. **69**, 2955 (1992).

[37] E. L. Shirley, X. Zhu, and S. G. Louie, Phys. Rev. B **56**, 6648 (1997).

[38] E. L. Shirley, R. M. Martin, G. B. Bachelet, and D. M. Ceperley, Phys. Rev. B **42**, 5057 (1990).

[39] A. Fleszar and W. Hanke, Phys. Rev. B **71**, 045207 (2005).

[40] QP levels calculated by ourselves in Ref. 2 did not have local orbitals included. Mostly for that reason, the QP levels are a little different from those cited here. Also, the gap in wurtzite InN cited here is 0.17 eV larger than the 0.02 eV reported by Usuda (Ref. 50), who used LAPW eigenfunctions that did not have local orbitals. (Leaving out the N $3p$ and In $5d$ local orbitals, we obtained a gap of 0.01 eV with an LMTO basis in Ref. 2.) The basis sets used in the present calculations are substantially larger, and the tolerances are set tighter than in what Ref. 2 used. These changes make additional corrections of ~0.05 eV. Finally, the fundamental gap for Si was incorrectly tabulated as

0.84 eV in Ref. 11: the actual value for the basis used in that paper was 0.91 eV.

[41] A. Marini, G. Onida, and R. De Sole, Phys. Rev. Lett. **88**, 016403 (2001).

[42] B. Holm and U. von Barth, Phys. Rev. B **57**, 2108 (1998).

[43] Y. M. Niquet and X. Gonze, Phys. Rev. B **70**, 245115 (2004).

[44] G. D. Mahan, *Many-Particle Physics* (Plenum Press, New York, 1990).

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

[46] B. Arnaud and M. Alouani, Phys. Rev. B **63**, 085208 (2001).

[47] E. G. Maksimov, I. I. Mazin, S. Y. Savrasov, and Y. A. Uspenski, J. Phys.: Condens. Matter **1**, 2493 (1989).

[48] T. Komesu, H.-K. Jeong, J. Choi, C. N. Borca, P. A. Dowben, A. G. Petukhov, B. D. Schultz, and C. J. Palmstrom, Phys. Rev. B **67**, 035104 (2003).

[49] R. L. Aggarwal, Phys. Rev. B **2**, 446 (1970).

[50] M. Usuda, H. Hamada, K. Siraishi, and A. Oshiyama, Jpn. J. Appl. Phys., Part 2 **43**, L407 (2004).

[51] M. Rohlfing, P. Kruger, and J. Pollmann, Phys. Rev. Lett. **75**, 3489 (1995).