We present a new all-electron, augmented-wave implementation of the GW approximation using eigenfunctions generated by a recent variant of the full-potential LMTO method. The dynamically screened Coulomb interaction \( \mathcal{W} \) is expanded in a mixed basis set which consists of two contributions, local atom-centered functions confined to muffin-tin spheres, and plane waves with the overlap to the local functions projected out. The former can include any of the core states; thus the core and valence states can be treated on an equal footing. Systematic studies of semiconductors and insulators show that the GW fundamental bandgaps consistently fall low in comparison to experiment, and also the quasiparticle levels differ significantly from other, approximate methods, in particular those that approximate the core with a pseudopotential.

71.15.-m, 71.15.Qe ,71.15.Mb

The GW approximation (GWA) of Hedin has been applied to many kinds of materials. In customary \textit{ab initio} implementations, the self-energy \( \Sigma \) is generated from eigenvalues and eigenfunctions calculated within the self-consistent local-density approximation (LDA). It has been shown that quasiparticle energies computed in this way are in significantly better agreement with experiment than the LDA eigenvalues.

The various implementations of GWA by may classified by what kinds of basis sets are used in the expansion of the LDA eigenfunctions, and the expansion of the bare and screened Coulomb interactions \( v \) and \( W \). The most common implementations make an additional pseudopotential approximation for the core, which makes it possible to expand all these quantities in plane waves. However, a plane-wave basis is poorly suited to localized orbitals such as \( d \)- or \( f \)-states. Moreover, as we show here, it appears that the pseudopotential approximation is somewhat inadequate when used in conjunction with the GW approximation. Two other GW implementations that do not use pseudopotentials have also been published. In both of these methods, \( v \) and \( W \) are expanded in plane waves; owing to the difficulty in a plane-wave expansion of localized orbitals they did not take into account core contributions. Aryasetiawan and collaborators implemented a method that expands \( v \) and \( W \) in a linear combination of augmented wave function products (product basis) and has applied it to several kind of materials, including NiO, with reasonable results. However this implementation requires the atomic spheres approximation (ASA) for the LDA, which approximates space by a superposition of atom-centered ("muffin-tin") spheres, neglecting the interstitial. Thus, its reliability is uncertain.

We present a new GW implementation which uses a mixed-basis expansion for \( v \) and \( W \). \( v \) and \( W \) are expanded in Aryasetiawan’s product-basis in the muffin-tin (MT) spheres, and the interstitial plane waves (IPW) in the interstitial region. An IPW is a plane wave with the MT contributions projected out. This basis can be an efficient one applicable for the localized electrons and cores. Together with the rather accurate eigenfunctions given by the all-electron full-potential LMTO method, our GW method can be both efficient and accurate. It is our knowledge the first implementation of GW that makes no significant approximations beyond the lack of self-consistency (pseudopotential approximation for the core, shape approximation in the potential, plasmon-pole approximation for the dynamical screening).

The GW self-energy is

\[
\Sigma(r, r', \omega) = \frac{i}{2\pi} \int d\omega' G(r, r', \omega + \omega') e^{i\omega'} W(r, r', \omega'),
\]

(1)

where \( W \) is the RPA screened Coulomb interaction and \( G \) is the Green’s function. \( W \) and \( G \) are calculated directly from the self-consistent LDA eigenvalues and eigenfunctions, with exchange-correlation potential \( V_{xc}(r) \). Then the quasi-particle energy \( E_n(k) \) \((n\) is the band index and \( k \) is the wave vector) is given by

\[
E_n(k) - \epsilon_n(k) = Z_{nk} \frac{1}{\langle \Psi_{kn} | \Sigma(r, r', \epsilon_n(k)) | \Psi_{kn} \rangle - \langle \Psi_{kn} | V_{xc}^{\text{LDA}}(r) | \Psi_{kn} \rangle}
\]

(2)

where \( \epsilon_n(k) \) and \( \Psi_{kn} \) denotes the LDA eigenvalues and eigenfunctions, and \( Z_{nk} \) is the quasi-particle (QP) renormalization factor

\[
Z_{nk} = \left[ 1 - \langle \Psi_{kn} | \frac{\partial}{\partial \omega} \Sigma(r, r', \epsilon_n(k)) | \Psi_{kn} \rangle \right]^{-1}.
\]

(3)

Our method to calculate the quasi-particle energy, is based on the FP-LMTO method which expands the eigenfunction \( \Psi_{kn} \) (we omit the spin index to simplify expressions) as

\[
\Psi_{kn} = \sum \xi_s^{kn} \chi_s^k(r).
\]

(4)
Here $\chi^k(r)$ denotes the Bloch sum of the MT orbitals $\chi_a(r)$. $s \equiv (a, n, L)$ is a composite index labelling the atom in the unit cell $a$, the angular momentum $L$, and another index $n$ to specify which MT orbitals (multiple envelope functions per $L$ are permitted) in the interstitial, $\chi^k(r)$ may be expanded in plane waves, while at the augmentation spheres it is matched smoothly and onto of a linear combination $A_{au}(r) \equiv \{\phi_{aL}(r)Y_L(\hat{r}), \phi_{\bar{a}L}(r)Y_L(\hat{r})\}$ of solutions of the radial Schrödinger equation, $\phi_{\bar{a}L}(r)$, and its energy derivative, $\phi_{aL}(r)$. Finally, $u \equiv (L, I\Gamma)$ is a composite of $L$ and $I\Gamma$, where $I\Gamma$ assumes value 0 for $\phi$ and 1 for $\phi$. Therefore, $\Psi_{kn}$ can be written as

$$
\Psi_{kn}(r) = \sum_{au} \alpha^{kn}_{au} A^k_{au}(r) + \sum_{G} \beta^{kn}_{G} P^k_{G}(r),
$$

where the IPW $P^k_{G}(r)$ is defined as

$$
P^k_{G}(r) = 0 \quad \text{if} \quad r \in \text{Any MT} = \exp(i(k + G)r) \quad \text{Otherwise.}
$$

In the interstitial, it is evident that products $\Psi_{k1n} \times \Psi_{k2n'}$ can similarly be expanded in IPW's because $P^k_{G1}(r) \times P^k_{G2}(r) = P^k_{G1+G2}(r)$. Within the augmentation sphere products of $\Psi$ are expanded in $P^k_{G1} + P^k_{G2}(r)$, which is the Bloch sum of the product basis $B_{am}(r)$. They are constructed from the products of $A_{au}(r) \times A_{a',w'}(r)$ following the procedure by Aryasetiawan [6].

Thus a basis suitable for expansion of wave function products, and therefore also $v$ and $W$, is the mixed basis $\{M^k_f(r)\} \equiv \{P^k_{G}(r), B^k_{G1,\ldots,G2}(r)\}$, where $P^k_{G}$ is an orthonormal basis rendered from $P^k_{G}$. $I \equiv \{G, am\}$ is a composite index labelling the basis. The complete information to calculate the self-energy and $E_n(k)$ are the Coulomb matrix $\alpha_{IJ}(k) = \langle M^l_I | v | M^k_J \rangle$, the matrix elements of the products $\langle \Psi_{qn} | \Psi_{q-kn'} | M^k_f \rangle$ and the eigenvalues $\epsilon_{kn}$. The exchange part of the self-energy is written as

$$
\langle \Psi_{qn} | \Sigma^{\text{ex}}(\omega) | \Psi_{qn} \rangle = \sum_{k} \sum_{n'} \langle \Psi_{qn} | \Psi_{q-kn'} | M^k_f \rangle v_{IJ}(k) \times \langle M^l_I \Psi_{q-kn'} | \Psi_{qn} \rangle.
$$

The screened Coulomb interaction $W_{IJ}(q, \omega)$ is calculated through $W = (1 - vD)^{-1}v$, where the polarization function $D$ is

$$
D_{IJ}(q, \omega) = \sum_{k} \sum_{n} \sum_{n'} \langle M^l_I \Psi_{qn} | \Psi_{q-kn'} \rangle \langle \Psi_{qn} | \Psi_{q-kn'} | M^k_J \rangle \times \left( \frac{1}{\omega - \epsilon_{kn} + \epsilon_{q-kn'} + i\delta} - \frac{1}{\omega + \epsilon_{kn} - \epsilon_{q-kn'} - i\delta} \right).
$$

Finally, the correlation part of self-energy is calculated by

$$
\langle \Psi_{qn} | \Sigma^{\text{cor}}(\omega) | \Psi_{qn} \rangle = \sum_{k} \sum_{n'} \sum_{IJ} \langle \Psi_{qn} | \Psi_{q-kn'} | M^k_f \rangle \times \langle M^l_I \Psi_{q-kn'} | \Psi_{qn} \rangle \times \int_{-\infty}^{\infty} \frac{i\omega'}{2\pi} W_{IJ}(k, \omega') \frac{1}{\omega' + \epsilon_{q-kn'} + i\delta}.\tag{9}
$$

Here the denominator $-i\delta$ is for occupied states, and $+i\delta$ for unoccupied states. We use the $\omega'$-integral method given by Aryasetiawan [6]. Our GW code is developed starting from his code.

Details of the method will be described elsewhere. Applying the method to a wide range of semiconductors, we find a number of systematic tendencies. The most important ones are: (i) there is a systematic underestimate of the fundamental bandgap (ii) The bandgap error tends to increase with bond polarity, and for compounds with relatively shallow cation $d$ states. (iii) the GW rather badly underestimates the deepening (relative to LDA) of occupied cation $d$ states found in III-V and II-V compounds. (iv) For zincblende lattices, the bandgap error tends to be slightly larger at the X point than at $\Gamma$. (v) the bandgap is sensitive to what approximations are used for shallow core states.

GaAs illustrates many of these general findings. It has a direct gap at 0K of 1.52 eV; however to compare against the present non-spin-polarized calculations, we assign the valence band maximum $\Gamma_15 \approx \frac{3}{2} \Gamma_7 \approx \frac{1}{2} \Gamma_7 \approx \frac{1}{2} \Gamma_7$. (Spin-orbit coupling splits $\Gamma_15$ by shifting the $\Gamma_8$ states $+\frac{1}{2} \Delta_0$ and the $\Gamma_7$ state $-\frac{1}{3} \Delta_0$. This was confirmed by an LDA-ASA calculation including spin-orbit coupling, which also yielded $\Delta_0=0.36$ eV, in good agreement with the observed $\Delta_0=0.34$ eV.) Table II shows gaps at $\Gamma$, X and L; the spin-orbit-corrected fundamental gap is 1.63 eV. An accurate LDA calculation must include both 3d states (which push upwards on $\Gamma_15$, narrowing the gap) and the 4d states (which push downwards, widening the gap). At present the GW implementation can include only one of these states for the input wave functions (the GW itself includes both), but to compute the bands within our present GW method we must select either 3d or 4d to input to the GW. Therefore, all three LDA cases are presented (the 3+4d case uses local orbitals [1]), and are seen to fall within $\sim 0.15$ eV of each other.

Similarly the GW gaps, whether the LDA 3d basis or 4d basis is input, lie within $\sim 0.15$ eV of each other (albeit with a slightly different $k$-dependence). By comparing LDA bands to the 3+4d case, the 4d is a better choice for GaAs, especially in the conduction bands where the Ga 4d begin to play an important role. In any case, the direct gap ($\Gamma_15 \rightarrow \Gamma_1c$) is $\sim 0.3$ eV smaller than the experimental value; while the X and L point ($\Gamma_8c$ and $\Gamma_1c$) are underestimated by a somewhat larger amount, $\sim 0.4$ eV. We have found this tendency to be rather systematically followed in the III-V semiconductors, as shown below.

The effect of the core is particularly important. Line
core0 in Table I shows what gaps result when valence states only are included in the calculation of $D$ and $\Sigma$. For this case the LDA potential we subtract corresponds to the valence-only density, $V_{xc}^{\text{LDA}}(n_{\text{val}}; \mathbf{r})$. This approximates what is typically done in pseudopotential GW (GW:PP) calculations. It is seen that this approximation leads to much too strong a $k$-dependence on the gap shift, something also seen when compared to GW:PP calculations. Table I shows some data of Shirley et. al. [11], (marked GW:PP). They rather closely track the core0 results except for an approximately $k$-independent shift of 0.32 eV. Shirley also included an approximate core polarization term (marked GW:PP+CP). It is seen that the addition of core-polarization terms to the PP have roughly similar effects (increasing the shift at $\Gamma$, decreasing it at X), but the shifts are larger than the all-electron results, and depend more strongly on $k$. Similar overestimates of gaps by the GW:PP method are seen in Si and AlAs. For example Shirley computed the $\Gamma_{15v} \rightarrow X_{1c}$ transition to be 1.31 eV, while our all-electron result is 1.04 eV, close to the GW:PAW result (1.10 eV) of Arnaud [8], and an early GW:LAPW result (1.14 eV) by Hamada [4]. Moreover, we find for GaN (and indeed generally for all semiconductors studied), a weak energy-dependence of the GW-LDA shift of the conduction bands up to 10 eV above the conduction-band minimum, in accord with a careful analysis of both UV reflectivity and near-edge x-ray absorption spectra [12]. However, using the GW:PP method, Rubio et. al. [14] found a significant energy-dependence of this shift. Inspecting the bottom of Table I, it is evident that the Ga 3$d$ states contribute in an important way to the dielectric response. The row marked core2 shows that the levels change rather significantly when the Ga 3$d$ is omitted from the calculation of $D$.

Fig. 1 illustrates the systematics of the errors in the fundamental gap for a range of semiconductors and insulators. With the sole exception of diamond [14], the GW gaps are systematically smaller than experiment. As noted for GaAs above, there is some $k$-dependence to the errors: the lowest conduction band levels at L and more especially at X are generally in worse agreement than those at $\Gamma$. Two sources of error are readily identifiable, and as will be shown elsewhere [13]; they account for nearly all of the gap errors in these $sp$ bonded semiconductors. As noted by Maksimov [16], an important contribution to the gap error arises from the nonlocal screened exchange missing from LDA. In Hartree-Fock theory, the bare exchange is used, therefore it rather severely overestimates the bandgap and bandwidths in general. The GW does screen the exchange, and qualitatively speaking GW replaces in addition to $V_{xc}^{\text{LDA}}$, a term which corresponds to the nonlocal screened exchange, with the nonlocal screened exchange written as

$$\Sigma \approx V_{xc}^{\text{LDA}} + \frac{1}{\epsilon_{\infty}}(\Sigma_x - V_{xc}^{\text{LDA}}).$$

In the present non self-consistent GW implementation, $\epsilon$ is computed from the LDA, with $\epsilon_{\infty} \sim$10 to 20% overestimated on account of the LDA gaps being too small. As we will show elsewhere [14], self-consistency reduces $\epsilon$ slightly. This enhances the nonlocal part of the self-energy operator, and further increases the GW-LDA gap correction, as can be qualitatively understood by Eq. (10). This correction is 0.2 eV in Si [17] but increases in the more ionic materials where $\epsilon$ is smaller. That the correction is larger in the latter case might be expected because the initial LDA potential (from which GW is constructed) is a poorer approximation, as again can be qualitatively seen from Eq. (10).

Table I shows the underbinding by GW of the occupied cation $d$ levels. As is shown, the LDA rather badly underestimates this binding. The LDA error is well known and also reasonably well understood [15], namely that the LDA eigenvalues are inappropriately interpreted as excitation energies. It is significant that, while the GW shifts these levels in the direction of experiment, the shift is much too small. Comparison with the exchange-only (Hartree-Fock) column shows clearly that the GW is strongly overscreening the bare exchange for these cation $d$ levels. (Similar difficulties are found for the so-called charge transfer antiferromagnets such as NiO and MnO, as will be discussed elsewhere.) This underbinding also contributes an important term to the gap error; shows importance increases for shallow $d$ states. It is not clear whether self-consistency in the GW will remedy this error. It is interesting that, even within the LDA, these levels can be reasonably computed using Slater transition-state theory.

The underbinding of the $d$ levels makes an important contribution to the gap error where the levels are relatively shallow (e.g. CdTe, and GaAs). There is a coupling between the $d$ state and and the valence band maximum, which pushes the latter upwards and reduces the bandgap. When the energy separation between these states are too small, the coupling them is overestimated. This will be shown in some detail in a future work [15]. In addition, we will publish the GW results on the wurtzite ZnO where we prepare the input eigenfunctions by the full-potential LAPW method [23].

In summary, an all-electron implementation of GW presented here leads to systematic errors in semiconductor and insulator bandgaps, whose origin were identified.
TABLE I. Selected energy eigenvalues, in eV, at Γ, L and X for GaAs. Experimental data are taken from Ref. [19]; the line “Expt-SO” subtracts the spin-orbit coupling from the raw experimental data to compare with the calculated levels, as discussed in the text. The GW:PP+CP and GW:PP data are GW results [11] using an LDA pseudopotential, with and without a core polarization term added. The GW results are quoted with the \( Z \) (underlined), and with \( Z=1 \) immediately below. Also shown are changes in the GW levels owing to varying degrees of approximation for the Ga and As 3\( d \) states.

|        | Γ_{1c} | Γ_{15c} | L_{1c} | X_{1c} | X_{3c} |
|--------|--------|--------|--------|--------|--------|
| Expt   | 1.52   | 4.72   | 1.84   | 2.01   | 2.41   |
| Expt-SO| 1.63   | 4.83   | 1.95   | 2.12   | 2.52   |
| LDA    | 0.35   | 3.68   | 0.86   | 1.34   | 1.54   |
| LDA(4d)| 0.41   | 3.72   | 0.91   | 1.34   | 1.58   |
| LDA(3d)| 0.24   | 3.57   | 0.76   | 1.33   | 1.57   |
| GW,4d(Z)| 1.35  | 4.20   | 1.54   | 1.56   | 1.97   |
| GW,4d(Z=1)| 1.50 | 4.28   | 1.63   | 1.59   | 2.02   |
| GW,3d(Z)| 1.34   | 4.38   | 1.59   | 1.71   | 2.04   |
| GW,core0\({}^a\) | 0.94   | 4.18   | 1.37   | 1.72   | 2.01   |
| GW:PP  | 1.29   | 1.69   | 2.05   | 2.34   |
| GW:PP+CP| 1.69  | 1.88   | 1.93   | 2.29   |
| HF,4d  | 4.84   | 8.50   | 5.10   | 5.26   | 5.63   |
| \( \delta \)GW,core0\({}^a\) | -0.41  | -0.02  | -0.17  | 0.16   | 0.04   |

\( ^{a} \)core3 treats the As 3\( d \) at the Hartree-Fock level (bare exchange) and omits this state in the computation of \( D \)

\( ^{b} \)core2 additionally omits the Ga 3\( d \) in the calculation of \( D \)

\( ^{c} \)core1 additionally approximates the Ga 3\( d \) at the Hartree-Fock level

\( ^{d} \)core0 computes both \( D \) and \( W \) with valence states only, completely neglecting the core beyond the LDA treatment.

TABLE II. Occupied cation \( d \) band energy levels in selected II-VI and III-V semiconductors, relative to the valence band maximum. Column ‘HF’ is the exchange-only (Hartree-Fock) result.

|        | Expt | LDA | GW | HF |
|--------|------|-----|----|----|
| GaN    | -17.1\({}^a\) | -13.6 | -16.4 | -23.3 |
| GaAs   | -18.8\({}^c\) | -15.0 | -18.1 | -27.2 |
| InP    | -16.8\({}^c\) | -14.2 | -15.7 | -21.2 |
| InAs   | -17.1\({}^c\) | -14.4 | -16.1 | -22.6 |
| ZnS    | -8.7\({}^b\) | -6.2  | -7.1  | -14.1 |
| ZnSe   | -9.0\({}^c\) | -6.7  | -7.7  | -15.2 |
| CdS(ZB) | -9.2\({}^d\) | -7.5  | -8.2  | -12.3 |
| CdTe   | -10.5\({}^c\) | -8.1  | -9.3  | -14.4 |
| HgTe   | -8.6\({}^c\) | -7.1  | -7.6  | -12.8 |

\( ^{a} \) Ref. [18]

\( ^{b} \) Ref. [20]

\( ^{c} \) Ref. [21]

\( ^{d} \) Ref. [22], average of X and Γ, “turning points” method.

FIG. 1. Top panel: experimental values of the fundamental gap in a variety of semiconductors and insulators, approximately ordered by ionic character. Bottom panel: deviations from experiment in the fundamental gap. Red points are LDA errors; green points are GW errors. The light and green denotes the smallest direct gap; the dark green denotes the fundamental gap when it differs from the direct gap. A term \( 1/3\Delta_0 \) was subtracted from the calculated GW and LDA data to account for spin-orbit coupling, as discussed in the text.

[1] L. Hedin, Phys. Rev. 139, A796 (1965); L. Hedin and S. Lundqvist, in Solid State Physics, edited by H. Ehren-
reich, F. Seitz, and D. Turnbull (Academic, New York, 1969), Vol. 23, p. 1.

[2] F. Arayasetiawan and O. Gunnarsson, Rep. Prog. Phys. 61, 237-312 (1998).

[3] W. G. Aulbur, L. Jönsson, and J. W. Wilkins, Solid State Physics, edited by H. Ehrenreich, F. Saepen (Academic, New York, 2000), Vol. 54, p. 1.

[4] N. Hamada, M. Hwang and A.J. Freeman, Phys. Rev. B 41, 3620 (1990).

[5] B. Arnaud and M. Alouani, Phys. Rev. B 62, 4464 (2000).

[6] F. Arayasetiawan and O Gunnarsson, Phys. Rev. B 49, 16214 (1994). Rep. Prog. Phys. B 61, 16214 (1998)

[7] F. Arayasetiawan and O Gunnarsson, Phys. Rev. Lett, 74, 3221, (1995); Phys. Rev. B54, 5353, (1996)

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

[9] See Refs. in [2,3] and in F. Arayasetiawan, in Advances in Condensed Matter Science, edited by I.V. Anisimov (Gordon and Breach, 2000), Vol. 1, p.33.

[10] M. van Schilfgaarde, unpublished. Local orbitals were originally designed for the LAPW method; see D. Singh, Phys. Rev. B 43, 6388 (1991).

[11] E.L. Shirley, Z. Zhu, S.G. Louie, Phys. Rev. B 56, 6648 (1997).

[12] W. R. L. Lambrecht, et. al, Phys. Rev. B 55, 2612 (1997).

[13] A. Rubio, et. al, Phys. Rev. B 48, 11810 (1993).

[14] Diamond appears to be the one material where the PP:GW and all-electron results agree. This is not surprising, since diamond has only one (very deep) core state.

[15] M. van Schilfgaarde, W. R. L. Lambrecht, and T. Kotani, in preparation.

[16] E. G. Maksimov et. al, J. Phys. Condens. Matter 1, 2493 (1989).

[17] Wei Ku found a similar gap for Si, and correction for self-consistency, computed by GW implemented within the LAPW method. Wei Ku, private communication.

[18] W. R. L. Lambrecht et. al, Phys. Rev. B 50, 14155 (1994).

[19] Crystal and Solid State Physics, Vol. 22a of Landolt-Börnstein, Numerical Data and Functional Relationships in Science and Technology, edited by O. Madelung (Springer, Berlin, 1988).

[20] L. Zhou et. al, Phys. Rev B 55, 5051 (1997).

[21] L. Ley et. al, Phys. Rev. B 9, 600 (1974).

[22] A. P. J. Stampf et. al, Phys. Rev. B 55, 9679 (1997).

[23] M. Usuda, N. Hamada, T. Kotani, and M. van Schilfgaarde, submitted to PRB.