Quasiparticle Self-Consistent *GW* Theory

M. van Schilfgaarde,[1] Takao Kotani,[1] and S. Faleev[2]

[1]*Arizona State University, Tempe, Arizona, 85287, USA*
[2]*Sandia National Laboratories, Livermore, California 94551, USA*

In past decades the scientific community has been looking for a reliable first-principles method to predict the electronic structure of solids with high accuracy. Here we present an approach which we call the quasiparticle self-consistent *GW* approximation. It is based on a kind of self-consistent perturbation theory, where the self-consistency is constructed to minimize the perturbation. We apply it to selections from different classes of materials, including alkali metals, semiconductors, wide band gap insulators, transition metals, transition metal oxides, magnetic insulators, and rare earth compounds. Apart from some mild exceptions, the properties are very well described, particularly in weakly correlated cases. Self-consistency dramatically improves agreement with experiment, and is sometimes essential. Discrepancies with experiment are systematic, and can be explained in terms of approximations made.

PACS numbers: 71.15.−m, 71.10.−w, 71.20.−b

The Schrödinger equation is the fundamental equation of condensed matter, and the importance of being able to solve it reliably can hardly be overestimated. The most widely used theory in solids is the celebrated local density approximation (LDA) [1]. In spite of its successes, it is well known that the LDA suffers from many deficiencies, even in weakly correlated materials (see Figs. 1 and 2). This has stimulated the development of flavors of extensions to the LDA to redress one or another of its failures, such as the LDA+*U* method. Each of these methods improves one failing or another in the LDA, but they often bear a semiempirical character, and none can be considered universal and parameter-free. Thus we are far from a precise and universally applicable theory for solids, with attendant limits to their ability to predict materials properties.

The random phase approximation (RPA) or *GW* approximation (*GW*A, *G* is the Green's function, *W* is the screened coulomb interaction) of Hedin [2] is almost as old as the LDA. A major advance was put forward by Hybertsen and Louie [3] when they employed LDA eigenfunctions to generate the *GW* self-energy $\Sigma = iGW$, and showed that fundamental gaps in sp^3 bonded materials were considerably improved over the LDA. Since that seminal work, many papers and some reviews [4–6] have been published on *GW* theory and extensions to it. One problem that has plagued the *GW* community has been that calculated results of the same quantities tend to vary between different groups, much as what occurred in the early days of the LDA. This is because *further* approximations are usually employed which significantly affect results. Almost ubiquitous is the 1-shot approximation where (following Hybertsen) $\Sigma \approx iG^{\mathrm{LDA}}W^{\mathrm{LDA}}$; i.e., Σ is computed from LDA eigenfunctions. However, there is an emerging consensus [7–12] that, when cores are treated adequately [11], $G^{\mathrm{LDA}}W^{\mathrm{LDA}}$ bandgaps are underestimated even in (weakly correlated) semiconductors. (see top panel of Fig. 1; note especially CuBr).

In general, one-shot *GW* approaches are rather unsatisfactory. The quality of the $G^{\mathrm{LDA}}W^{\mathrm{LDA}}$ approximation is closely tied to the quality of LDA starting point [11], and is adequate to construct *G* and *W* only under limited circumstances [11]. It can fail even qualitatively in transition-metal and rare-earth-metal compounds such as CoO and ErAs [11]. Some kind of self-consistency is essential: the quasiparticle (QP) levels should not be an artifact of the starting conditions. The full self-consistent *GW* method (full SC*GW*) determines *G* self-consistently from $\Sigma = iGW$, which in turn generates *G*. Here $W = v(1 - vP)^{-1}$ where $P = -iG \times G$ and v are, respectively, the irreducible polarization function and (bare) Coulomb interaction. In the few cases where it has been applied, some difficulties were found: in particular, the valence bandwidth of the homogeneous electron gas [13] is ~15% *wider* than the noninteracting case, whereas the $G^{\mathrm{LDA}}W^{\mathrm{LDA}}$ width is ~15% *narrower*, in agreement with experiment for Na (see Fig. 2). A recent (nearly) full SC*GW* study of Ge and Si also overestimates the valence bandwidth [7], though the fundamental gaps are well described.

In Ref. [14] we proposed an ansatz for a different kind of SC*GW*, and demonstrated that it radically improves the QP levels in the oxides MnO and NiO. In this Letter, we ground the idea on an underlying principle—namely optimization of the effective one-body Hamiltonian H^0 by minimizing the perturbation to it—and propose it as a universal approach to the reliable prediction of the electronic structure. We show that this approach, which we call the quasiparticle self-consistent *GW* (QS*GW*) method, results in accurate predictions of excited-state properties for a large number of weakly and moderately correlated materials. QP levels are uniformly good for all materials studied: not just fundamental gaps in semiconductors but for nearly all levels where reliable experimental data are available. Even in strongly correlated *d* and *f* electron systems we studied, errors are somewhat larger but still systematic.
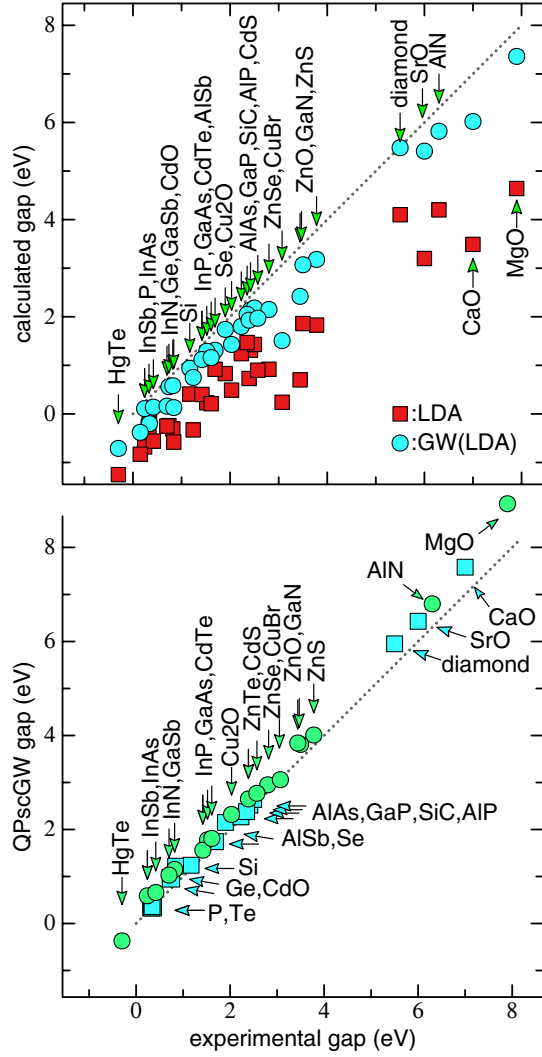
FIG. 2 (color online). Comparison of LDA (dashes), $G^{\mathrm{LDA}}W^{\mathrm{LDA}}$ (dots) and QSGW (solid lines) energy bands in GaAs (left) and Na (right). Circles are experimental data, with spin-orbit coupling subtracted by hand. The QSGW fundamental gap and conduction-band effective mass ($E_g = 1.77$ eV and $m_c^* = 0.077m_0$) are slightly overestimated, and the optical dielectric constant underestimated ($\epsilon_\infty = 8.4$): $E_g^{\mathrm{expt}}(0 \text{ K}) = 1.52$ eV, $m_c^{*,\mathrm{expt}} = 0.065m_0$, and $\epsilon_\infty^{\mathrm{expt}} = 10.8$. For comparison, $E_g^{G^{\mathrm{LDA}}W^{\mathrm{LDA}}} = 1.29$ eV and $m_c^{*,G^{\mathrm{LDA}}W^{\mathrm{LDA}}} \approx 0.059m_0$, while $E_g^{\mathrm{LDA}} = 0.21$ eV and $m_c^{*,\mathrm{LDA}} = 0.020m_0$. The correspondence between QSGW and experiment at other known levels at Γ, L, and X, the Ga 3d level near -18 eV is representative of nearly all available data for sp systems. For Na, the QSGW occupied bandwidth is 15% smaller than the LDA. Circles are taken from photoemission data [20]; square from momentum electron spectroscopy [21].

FIG. 1 (color online). Fundamental gaps of sp compounds from LDA (squares) and $G^{\mathrm{LDA}}W^{\mathrm{LDA}}$ (circles) in top panel, and from QSGW, Eq. (2), in bottom panel. The spin-orbit coupling was subtracted by hand from the calculations. The $G^{\mathrm{LDA}}W^{\mathrm{LDA}}$ gaps improve on the LDA, but are still systematically underestimated. For QSGW data, zinc-blende compounds with direct Γ-Γ transitions are shown as circles; All other gaps are shown as squares. Errors are small and highly systematic, and would be smaller than the figure shows if the electron-phonon renormalization were included.

The GWA is usually formulated as a perturbation theory starting from a noninteracting Green's function G^0 for given one-body Hamiltonian $H^0 = \frac{-\nabla^2}{2m} + V^{\mathrm{eff}}$. H^0 is noninteracting, so V^{eff} is static and Hermitian but it can be nonlocal. Because the GWA is an approximation to the exact theory, the one-body effective Hamiltonian $H(\omega) = \frac{-\nabla^2}{2m} + V^{\mathrm{ext}} + V^{\mathrm{H}} + \Sigma(\omega)$ depends on V^{eff} and is a functional of it: the Hartree potential V^{H} is generated through $G^0 = 1/(\omega - H^0 \pm i\epsilon)$, and the GWA generates $\Sigma(\omega)$. $H(\omega)$ determines the time evolution of the one-body amplitude for the many-body system.
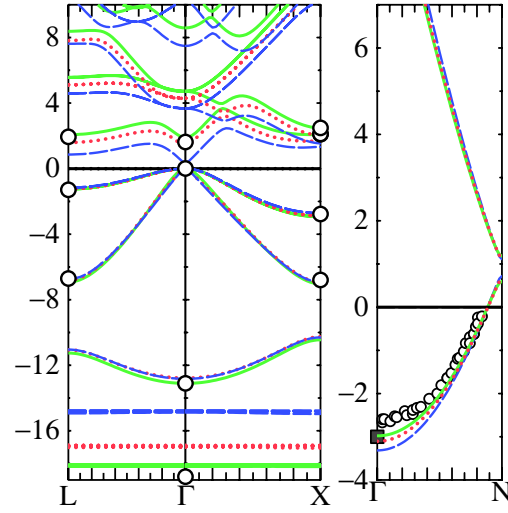
QSGW is a prescription to determine the optimum H^0: we choose V^{eff} based on a self-consistent perturbation theory so that the time evolution determined by H^0 is as close as possible to that determined by $H(\omega)$, within the RPA. This idea means that we have to introduce a norm M to measure the difference $\Delta V(\omega) = H(\omega) - H^0$; the optimum V^{eff} is then that potential which minimizes M. A physically sensible choice of norm is

$$M[V^{\mathrm{eff}}] = \mathrm{Tr}[\Delta V \delta(\omega - H^0)\{\Delta V\}^\dagger]$$
$$+ \mathrm{Tr}[\{\Delta V\}^\dagger \delta(\omega - H^0)\Delta V] \quad (1)$$

where the trace is taken over $\mathbf{r}$ and ω. Exact minimization M is apparently not tractable, but an approximate solution can be found. Note that M is positive definite. If we neglect the second term and ignore the restriction that V^{eff} is Hermitian, we have the trivial minimum $M[V^{\mathrm{eff}}] = 0$ at $V^{\mathrm{eff}} = V^{\mathrm{ext}} + V^{\mathrm{H}} + V^{\mathrm{xc}}$ where $V^{\mathrm{xc}} = \sum_{ij}|\psi_i\rangle\Sigma(\varepsilon_j)_{ij}\langle\psi_j|$. Here $\Sigma(\varepsilon_i)_{ij} = \langle\psi_i|\Sigma(\varepsilon_i)|\psi_j\rangle$, and $\{\psi_i, \epsilon_i\}$ are eigenfunctions and eigenvalues of H^0. The second term is similarly a minimum with $\Sigma(\varepsilon_i) \to \Sigma(\varepsilon_j)$. An average of the

Hermitian parts of these solutions results in

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

Re signifies the Hermitian part. This result is the same as Eq. (2) in Ref. [14].

We identify solutions to H^0 as "bare QP," which interact via the (bare) v. The dressed QP consists of the central bare QP plus induced polarized clouds of the other bare QPs— this is nothing but the physical picture in RPA to calculate poles of G from G^0. In the charged Fermi liquid theory [15] of Landau and Silin, the QP interact via v in addition to the short-range Landau interaction [$f_{pp'}$ in Ref. [15]; see Eq. (3.41)]. We can virtually construct the Landau-Silin QP from G^0 by the calculation of the non-RPA contributions to Σ. Or we can identify our bare QP as the Landau-Silin QP if we assume they are minimally affected by such contributions.

Our QSGW is conceptually very different from the full SCGW. In the latter case the electron-hole mediated state making $P = -iG \times G$ is suppressed by the square of the renormalization factor $Z \times Z$, and includes physically unclear contributions such as QP × (incoherent parts) [14,16]. P loses its physical meaning as the density response function, $P = \delta n/\delta V$, but is merely an intermediate construction in the self-consistency cycle. Such a construction does not give reasonable W even in the electron gas [13,17], resulting in a poor G.

We now turn to QSGW results, focusing on the QP energies given by H^0. Figure 2 shows that the QSGW valence band in Na properly *narrows* relative to the LDA by 15%. Indeed, for nearly all the sp semiconductors studied, calculated QP levels generally agree very closely with available experimental data. The best known are the fundamental gaps, shown in Fig. 1. QSGW data is divided into circles for materials whose gap is a Γ-Γ transition and squares for all other kinds. Roughly, Γ-Γ transitions are overestimated by ~0.25 eV. Most of the remaining gaps are Γ-X transitions, which are overestimated by ~0.1 eV. Errors appear to be larger for wide-gap, light-mass compounds (bearing elements C, N, and especially O); however, the calculations omit reduction in the gaps by the nuclear zero-point motion. This effect has been studied through varying isotopic mass in some tetrahedral semiconductors [18]. It is largest for light compounds: the $T = 0$ gap is reduced by ~0.3 eV in diamond and ~0.2 eV in AlN, but $\lesssim 0.1$ eV for heavier compounds. Because the renormalization been measured only for a few cases, we do not include it here.

Apart from some mild exceptions, QSGW generates a consistently precise description of the electronic structure in sp systems, including other known excitations. This is illustrated in Fig. 2, where GaAs was chosen because of the abundance of available experimental data. It is notable that the errors are not only small, but unlike the

$G^{LDA}W^{LDA}$ or LDA, they are highly systematic: compare, for example, the fundamental gaps (Fig. 1). We may expect that the bandgaps should be overestimated, because the RPA dielectric function omits electron-hole correlation effects. Thus ϵ^{RPA} should be too small and underscreen W. Indeed, the optical dielectric constant ϵ_∞ is systematically underestimated slightly (Fig. 2). Similar consistencies are found in the effective masses. The conduction-band mass at Γ, $m_{c\Gamma}^{*,QSGW}$ consistently falls within a few percent of experimental data for wide-gap materials, but as the gap becomes smaller (induced by, e.g., scaling Σ), $m_{c\Gamma}^{*,QSGW}/m_{c\Gamma}^{*,expt}$ scales essentially as the ratio of the QSGW gap to the experimental one, as expected when the gap becomes small. Taking data for GaAs from Fig. 2 for example, we obtain $E_g^{QSGW}/E_g^{expt} = 1.16$, and $m_{c\Gamma}^{*,QSGW}/m_{c\Gamma}^{*,expt} = 1.18$.

Table I shows that the 3d bandwidth, the relative position of s band, and exchange splittings ΔE_x, are systematically improved relative to the LDA in elemental 3d metals, and Gd. QSGW magnetic moments are systematically overestimated slightly. ΔE_x is overestimated in Ni, presumably owing to the neglect of spin fluctuations [19]. QSGW also predicts with reasonable accuracy the QP levels of all magnetic 3d compounds studied, in particular, correlated oxides such as MnO and NiO where the LDA fails dramatically. As might be expected, the accuracy deteriorates somewhat relative to sp systems. For example, the QSGW optical gap in NiO (4.8 eV) was found to be larger than experiment (~4.3 eV). Table I compares the magnetic moments, and Ref. [14] shows in detail that the QP levels are consistently well described. However, for

TABLE I. Valence d bandwidths W_d (calculated at Γ for Ti, Cr, and Co, and at N for Fe, and at X for Ni), relative position of s and d band bottoms ϵ_{sd}, splittings ΔE_x between majority and minority d (or f) states, and magnetic moments in 3d compounds and Gd.

	W_d (eV)			ϵ_{sd} (eV)		
	LDA	QSGW	Expt	LDA	QSGW	Expt
Ti	6.0	5.7		3.5	4.3	
Cr	6.6	6.2		3.5	4.3	
Fe	5.2	4.6	4.6	3.6	4.4	4.6
Co	4.1	3.8	3.7	4.6	5.3	4.9 ± 1
Ni	4.4	4.0	4.0	4.4	5.0	5.5

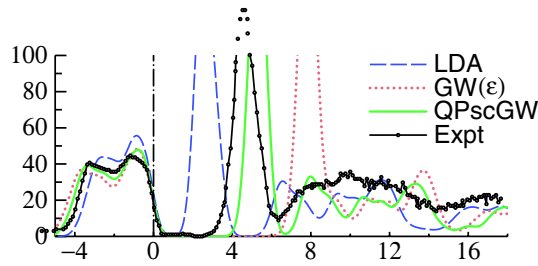|	moment (μ_B)			ΔE_x (eV)		
	LDA	QSGW	Expt	LDA	QSGW	Expt
Fe	2.2	2.2	2.2	1.95	1.67	1.75
Co	1.6	1.7	1.6	1.70	1.21	1.08
Ni	0.6	0.7	0.6	0.6	0.5	0.3
MnO	4.5	4.8	4.6			
NiO	1.3	1.7	1.9			
MnAs	3.0	3.5	3.4			
Gd	7.7	7.8	7.6	4.9	16.1	~12.5

FIG. 3 (color online). DOS in CeO_2. Line with dots shows PES + BIS data [22]. Calculated DOS were broadened with Gaussian of width 0.35 eV. $G^{LDA}W^{LDA}$ + eigenvalue-only self-consistency (dotted line) severely overestimates the position of the Ce f level, while it is slightly overestimated by QSGW (solid line). Broadening of the valence bands relative to LDA is found in all oxides studied, e.g., MgO and TiO_2, and has important consequences, e.g., in determining valence-band offsets.

Gd, (and for GdP and GdAs) QSGW overestimates the position of the (empty) minority Gd f shell by ~4 eV, and hence the exchange splitting ΔE_x.

Nonmagnetic oxides $SrTiO_3$, TiO_2, and CeO_2, with conduction bands of d or f character, overestimate fundamental gaps slightly more than their sp counterparts. The QSGW gaps were found to be 4.19 eV and 3.78 eV in $SrTiO_3$ and TiO_2, ~0.8 eV larger than the experimental gaps (~3.3 eV and ~3.1 eV). Figure 3 compares the QSGW density of states (DOS) of CeO_2 with spectroscopic data: the Ce f band is similarly overestimated by QSGW. This is reasonable, because electron-hole correlation effects are stronger in the narrow d (f) conduction bands. Figure 3 also shows DOS computed by $G^{LDA}W^{LDA}$, but with eigenvalue-only self-consistency, where only the diagonal part in Eq. (2) is kept. This constrains the eigenfunctions to the starting (LDA) eigenfunctions; thus the charge and spin densities do not change. While the off-diagonal parts of Σ add a small effect in, e.g., GaAs, their contribution is essential in CeO_2, even though the occupied states contain only a small amount of Ce f character.

To summarize, the QSGW theory (apart from some mild exceptions) appears to be an excellent predictor of QP levels for a variety materials selected from the entire periodic table. Self-consistency is an essential part of the theory. From the results obtained so far, this approach shows promise to be a universally applicable scheme, sufficient in its own right for QP levels in many materials. In contrast to the LDA or any other popular theory of electronic structure of solids in the literature today, the method is truly *ab initio* with errors that are generally small and highly systematic across many different materials classes. The errors can be attributed to missing electron-hole correlation contributions to ϵ. When better calculations are necessary (usually where the physics lies completely outside the domain of a one-particle picture, such as the description of excitons, multiplets, or Mott transitions),

QSGW can be taken as an optimum starting point where the relevant many-body contributions to the Hamiltonian are (nearly) as small as possible. The systematic character of the error suggests that the dominant terms left out can be described by a few diagrams, in particular, the ladder diagrams coupling electrons and holes; the smallness of the error suggests that the additional terms can be added as a perturbation around the QSGW H^0, without the need for further self-consistency.

[1] W. Kohn and L. J. Sham, Phys. Rev. **140**, A1133 (1965).
[2] L. Hedin, Phys. Rev. **139**, A796 (1965).
[3] M. S. Hybertsen and S. G. Louie, Phys. Rev. B **34**, 5390 (1986).
[4] F. Aryasetiawan and O. Gunnarsson, Rep. Prog. Phys. **61**, 237 (1998).
[5] W. G. Aulbur, L. Jönsson, and J. Wilkins, in *Solid State Physics*, edited by H. Ehrenreich and F. Saepen (Academic Press, New York, 2000), Vol. 54, p. 1.
[6] G. Onida, L. Reining, and A. Rubio, Rev. Mod. Phys. **74**, 601 (2002).
[7] W. Ku and A. G. Eguiluz, Phys. Rev. Lett. **89**, 126401 (2002).
[8] S. Lebègue, B. Arnaud, M. Alouani, and P. E. Bloechl, Phys. Rev. B **67**, 155208 (2003).
[9] M. L. Tiago, S. Ismail-Beigi, and S. G. Louie, Phys. Rev. B **69**, 125212 (2004).
[10] A. Fleszar and W. Hanke, Phys. Rev. B **71**, 045207 (2005).
[11] M. van Schilfgaarde, T. Kotani, and S. Faleev, cond-mat/0508295 (to be published).
[12] T. Kotani and M. van Schilfgaarde, Solid State Commun. **121**, 461 (2002).
[13] B. Holm and U. von Barth, Phys. Rev. B **57**, 2108 (1998).
[14] S. V. Faleev, M. van Schilfgaarde, and T. Kotani, Phys. Rev. Lett. **93**, 126406 (2004).
[15] D. Pines and P. Nozieres, *The Theory of Quantum Liquid* (W. A. Benjamin Inc., New York, 1966), Vol. I.
[16] F. Bechstedt, K. Tenelsen, B. Adolph, and R. Del Sole, Phys. Rev. Lett. **78**, 1528 (1997).
[17] D. Tamme, R. Schepe, and K. Henneberger, Phys. Rev. Lett. **83**, 241 (1999).
[18] F. J. Manjón, M. A. Hernández-Fenollosa, B. Marí, S. F. Li, C. D. Poweleit, A. Bell, J. Menéndez, and M. Cardona, Eur. Phys. J. B **40**, 453 (2004).
[19] F. Aryasetiawan, Phys. Rev. B **46**, 13 051 (1992).
[20] I.-W. Lyo and E. W. Plummer, Phys. Rev. Lett. **60**, 1558 (1988).
[21] M. Vos, A. S. Kheifets, V. A. Sashin, E. Weigold, M. Usuda, and F. Aryasetiawan, Phys. Rev. B **66**, 155414 (2002).
[22] E. Wuilloud, B. Delley, W. D. Schneider, and Y. Baer, Phys. Rev. Lett. **53**, 202 (1984).