Ab initio calculation of excitonic effects in the optical spectra of semiconductors

Stefan Albrecht and Lucia Reining

Laboratoire des Solides Irradiés, URA 1380 CNRS – CEA/CEREM, École Polytechnique, F-91128 Palaiseau, France

Rodolfo Del Sole and Giovanni Onida

Istituto Nazionale per la Fisica della Materia, Dipartimento di Fisica dell'Università di Roma "Tor Vergata", Via della Ricerca Scientifica, I–00133 Roma, Italy

An *ab initio* approach to the calculation of excitonic effects in the optical absorption spectra of semiconductors and insulators is formulated. It starts from a quasiparticle bandstructure calculation and is based on the relevant Bethe–Salpeter equation. An application to bulk silicon shows a substantial improvement with respect to previous calculations in the description of the experimental spectrum, for both peak positions and lineshape.

71.35.Cc, 71.45.Gm, 78.20.Bh, 78.40.Fy

Recent advances in *ab initio* calculations, mostly Density Functional Theory – Local Density Approximation (DFT-LDA) applications, allow to determine the ground state properties and the Kohn-Sham (KS) electronic structure [1] for even complicated systems. In order to treat excited states, realistic quasiparticle (QP) energies are then in general obtained by applying self-energy corrections to the KS energies, usually evaluated in the *GW* approximation [2]. Excellent agreement of the resulting bandstructure with experimental data has been found for a wide range of materials [3,4]. However, spectroscopic properties involving two-particle excitations are often only poorly described at this one-particle level. The main example is absorption spectroscopy, where a simultaneously created electron–hole pair interacts more or less strongly. As a consequence, in addition to bound exciton states which occur within the gap, the spectral lineshape above the continuous–absorption edge is distorted.

The reported qualitative agreement with experiment of many computed KS-LDA absorption spectra, obtained from one-electron transitions between KS states [5], is indeed due to a partial cancellation between two principal errors: namely, the compensation of the large KS-LDA underestimation of the valence–conduction bandgap, with an overestimation of the absorption onset induced by calculating the dielectric function entirely within the one-particle picture. The situation often worsens when only the first error is corrected by replacing the KS eigenvalues with the realistic QP energies [6,7]. On the other hand, going beyond the one-particle picture through inclusion of local field and/or exchange–correlation effects within DFT-LDA in the calculation of the absorption spectrum does generally not remove the observed discrepancy [8]. In fact, most of the residual error stems from the neglect of the electron–hole interaction.

Up to now, excitonic effects have been rarely calculated from first principles. Some information about energetic changes can be extracted from an LDA-based Δ-SCF approach [9]. Large excitonic effects on the spectral properties have been calculated *ab initio* in the case of a small sodium cluster [6]. This approach has consequently been generalized with the calculation of the absorption onset for an infinite system, the Li_2O crystal [10]. The calculation of the entire optical spectrum of a solid, finally, still remains a major challenge [11]. Quantitatively correct theoretical absorption spectra are indeed needed as a reference for the interpretation and prediction of experimental results.

A paradigmatic case is bulk silicon, which is representative for the group IV, III-V, and II-VI semiconductors. These materials show qualitatively similar optical spectra, with two major structures at 3–5 eV. The first peak (E_1) has been interpreted as a M_1 type critical point transition, and the second peak (E_2) as a M_2 type one [13]. Theoretical work based on the one-electron approximation, ranging from early empirical pseudopotential approaches [14] to *ab initio* DFT-LDA work [8], all yielded the same qualitative result, i.e. an underestimation of the E_1 peak by as much as 50%, reducing it to a weak shoulder of the generally overestimated E_2 peak. In order to go beyond, Louie *et al.* [15] included local field effects in the calculation of the dielectric matrix. The resulting spectrum is significantly improved at higher energies (above 15 eV), but not in the region of interest around 4 eV.

Several authors suggested that strong contributions to the E_1 peak could arise from saddle point excitons [16–18]. Excitonic effects allowed to explain the measured temperature and pressure dependence of the lineshape and the symmetry in wavelength modulation reflectance spectra [17]. Until now, the most sophisticated calculation of excitonic effects on the spectral lineshape of silicon was done by Hanke and Sham [18]. They performed a semi-empirical LCAO calculation, including local field effects and the screened electron–hole attraction. As in Ref. [15], local field effects alone were shown to transfer oscillator strength to higher energies and hence to increase the discrepancy with experiment at lower energies. On the contrary, the electron–hole interaction shifted the position of the E_1 peak to lower energies, and almost doubled its intensity, while the oscillator strength of the higher energy peaks was decreased. The overall agreement with experiment was hence improved, and clear evidence was given for the importance of excitonic effects. However, the final intensity ratio between the E_1 and E_2 peaks was reversed, in disagreement with the experimental spectrum. As pointed out by Wang *et al.* [13], the reliability of semi-empirical approaches is limited. For instance, there are important differences, already at the one-electron level, between the spectra of Refs. [15] and [18].

In the *ab initio* framework, on the other hand, the precision achievable for the computation of electronic spec-

tra is in general still poor when compared with the quality of calculated ground state properties. This work is aimed to shrink this gap, showing how a significant improvement of the *ab initio* calculation of absorption spectra can be obtained.

The absorption spectrum is given by the imaginary part of the macroscopic dielectric function ϵ_M

$$\epsilon_M(\omega) = 1 - \lim_{\mathbf{q}\to 0} v(\mathbf{q})\hat{\chi}_{\mathbf{G}=0,\mathbf{G}'=0}(\mathbf{q};\omega), \qquad (1)$$

where $\hat{\chi}(\mathbf{r},\mathbf{r}';\omega) = -iS(\mathbf{r},\mathbf{r},\mathbf{r}',\mathbf{r}';\omega)$. $S(1,1';2,2')$ is the part of the two-particle Green's function which excludes the disconnected term $-G(1,1')G(2,2')$, and $G(1,1')$ is the one-particle Green's function [18]. The notation (1,2) stands for two pairs of space and time coordinates, $(\mathbf{r}_1,t_1;\mathbf{r}_2,t_2)$.

Following Ref. [18], we start from the Bethe–Salpeter equation for S,

$$S(1,1';2,2') = S_0(1,1';2,2') + S_0(1,1';3,3')\Xi(3,3';4,4')S(4,4';2,2').$$
$$(2)$$

Repeated arguments are integrated over. The term $S_0(1,1';2,2') = G(1',2')G(2,1)$ yields the polarization function of independent quasiparticles χ_0, from which the standard RPA ϵ_M is obtained. The kernel Ξ contains two contributions:

$$\Xi(1,1',2,2') = -i\delta(1,1')\delta(2,2')v(1,2) + i\delta(1,2)\delta(1',2')W(1,1').$$
$$(3)$$

Considering the first term in the calculation of S is equivalent to the inclusion of local field effects in the matrix inversion of a standard RPA calculation. In order to obtain the macroscopic dielectric constant, the bare Coulomb interaction v contained in this term must, however, be used without the long range term of vanishing wave vector [19]. When spin is not explicitly treated, v gets a factor of two for singlet excitons. In the second term, W is the screened Coulomb attraction between electron and hole. It is obtained as a functional derivative of the self-energy in the GW approximation, neglecting a term $G\frac{\delta W}{\delta G}$. This latter term contains information about the change in screening due to the excitation, and is expected to be small [20]. We limit ourselves to static screening, since dynamical effects in the electron–hole screening and in the one particle Green's function tend to cancel each other [21], which suggests to neglect both of them.

In order to solve Eq. (2), we have to invert a 4-point function. In Ref. [18] this has been possible due to the use of a very limited basis set. In an *ab initio* plane wave calculation, such a procedure is clearly prohibitive, when plane waves are chosen as straightforward basis functions. Instead, the physical picture of interacting electron–hole pairs suggests to use a basis of LDA Bloch functions, $\psi_n(\mathbf{r})$, expecting that only a limited number of electron–hole pairs will contribute to each excitation.

In this basis, $\chi_0^{(n_1,n_2),(n_3,n_4)} = \delta_{n_1,n_3}\delta_{n_2,n_4}(f_{n_2} - f_{n_1})/(E_{n_2} - E_{n_1} - \omega)$ and, after solving for S, in the case of static screening, equation (2) can be written as

$$S_{(n_1,n_2),(n_3,n_4)} = (H_{exc} - I\,\omega)^{-1}_{(n_1,n_2),(n_3,n_4)}(f_{n_4} - f_{n_3}),$$
(4)

with

$$H_{exc}^{(n_1,n_2),(n_3,n_4)} = (E_{n_2} - E_{n_1})\delta_{n_1,n_3}\delta_{n_2,n_4} - i(f_{n_2} - f_{n_1}) \times$$
$$\int d\mathbf{r}_1 \int d\mathbf{r}_1' \int d\mathbf{r}_2 \int d\mathbf{r}_2' \; \psi_{n_1}(\mathbf{r}_1)\,\psi_{n_2}^*(\mathbf{r}_1')\,\Xi(\mathbf{r}_1,\mathbf{r}_1',\mathbf{r}_2,\mathbf{r}_2')\,\psi_{n_3}^*(\mathbf{r}_2)\,\psi_{n_4}(\mathbf{r}_2'). \quad (5)$$

I is the identity operator. The energies E_n are the QP levels. Together with the above form of χ_0 this is consistent with the use of LDA wavefunctions and updated energy denominators in the Green's function used to construct the self-energy in the GW calculation. The f_n are Fermi-Dirac occupation numbers. We avoid to invert the matrix $(H_{exc} - I\,\omega)$ for each absorption frequency ω by applying the identity

$$(H_{exc} - I\,\omega)^{-1} = \sum_{\lambda,\lambda'} \frac{|\lambda > M_{\lambda,\lambda'}^{-1} < \lambda'|}{(E_\lambda - \omega)},$$
(6)

which holds for a system of eigenvectors and eigenvalues of a general, non-hermitian matrix defined by

$$H_{exc}|\lambda> = E_\lambda|\lambda>.$$
(7)

$M_{\lambda,\lambda'}$ is the overlap matrix of the (in general non-orthogonal) eigenstates of H_{exc}.

Equation (7) is the effective two-particle Schrödinger equation which we solve by diagonalization. The explicit knowledge of the coupling of the various two-particle channels, given by the coefficients $A_\lambda^{(n_1,n_2)}$ of the state $|\lambda>$ in our LDA basis, allows to identify the character of each transition. (This analysis would be much more cumbersome if a matrix inversion instead of the spectral representation was chosen, as in Ref. [18].)

The macroscopic dielectric function in Eq. (1) is obtained as

$$\epsilon_M(\omega) = 1 - \lim_{\mathbf{q}\to 0} v(\mathbf{q}) \sum_{\lambda,\lambda'} M_{\lambda,\lambda'}^{-1} \sum_{n_1,n_2} < n_1|e^{-i\mathbf{q}\cdot\mathbf{r}}|n_2 > A_\lambda^{(n_1,n_2)} \times$$
$$\sum_{n_3,n_4} < n_4|e^{+i\mathbf{q}\cdot\mathbf{r}}|n_3 > A_{\lambda'}^{*(n_3,n_4)} \frac{(f_{n_4} - f_{n_3})}{(E_\lambda - \omega)}. \quad (8)$$

In practice, the KS eigenvalues and eigenfunctions from a DFT-LDA calculation serve as input to the evaluation of the RPA screened Coulomb interaction W and the GW self-energy Σ. The KS eigenfunctions, together with the QP energies and W, are then used in the exciton calculation. Here each pair of indices (n_1, n_2) stands for a pair of bands and one Bloch vector $\mathbf{k}$ in the Brillouin zone (BZ), since we are interested in direct transitions only.

In principle, all combinations of bands should be considered. It can, however, be shown exactly that only pairs containing one filled and one empty LDA state contribute to (8). Still, the portion of the matrix H_{exc} to be considered is in general non-hermitian, being of the form [22]

$$\mathbf{H} = \begin{pmatrix} H^{(v_1,c_1),(v_2,c_2)} & H^{(v_1,c_1),(c_2,v_2)} \\ -H^{*(v_1,c_1),(c_2,v_2)} & -H^{*(v_1,c_1),(v_2,c_2)} \end{pmatrix}.$$

The off-diagonal coupling matrices do not contain the QP transition energies, but only the interaction elements, which are much smaller in the case of silicon. Hence, we neglect the latter and separate the Hamiltonian into two block-diagonal parts: the resonant contributions, which are active for positive frequencies, and the antiresonant ones, only contributing to negative frequencies. The matrix of the resonant part by its own is hermitian, and we therefore obtain the simpler formula

$$\epsilon_M(\omega) = 1 + \lim_{\mathbf{q}\to 0} v(\mathbf{q}) \sum_\lambda \frac{|\sum_{v,c;\mathbf{k}} <v,\mathbf{k}|e^{-i\mathbf{q}\mathbf{r}}|c,\mathbf{k}> A_\lambda^{(v,c;\mathbf{k})}|^2}{(E_\lambda - \omega)}.$$

(9)

(7) and (9) constitute a set of equations which has been frequently used in the non-*ab initio* framework [20,24]. Here, it appears as a particular approximation to the more general formula (8), with well-defined *ab initio* ingredients which are consistent with the *GW* approach.

We evaluate expression (9) for bulk silicon. The DFT-LDA calculation is performed using norm-conserving pseudopotentials [25], an energy cutoff of 15 Ry, and 256 special $\mathbf{k}$ points in the BZ [26]. Next, *GW* corrections to the KS band structure are obtained following the approach of [27]. The quite smooth *GW* corrections are interpolated for the denser $\mathbf{k}$ point mesh needed for the absorption spectrum. We evaluate equations (7) and (9) using different sets containing up to 2048 $\mathbf{k}$ points in the BZ. In order to handle such large matrices, the symmetry properties of the crystal are exploited. One has to be very careful in doing so, since the spectrum turns out to be extremely sensitive to any inconsistency in the phases which may appear when wavefunctions are rotated, notably for degenerate bands. A safe way to proceed is to make only partial use of symmetry, considering only those operations which form an abelian subgroup of the point group, and which altogether allow to reconstruct the whole zone from a corresponding reduced part. In the case of silicon, we found it convenient to use the 180^0 rotations around the x and the y axis, respectively. These two operations T allow us to break the equation $H^{\mathbf{k}\mathbf{k}'} A^{\mathbf{k}'} = E A^{\mathbf{k}}$ (band indices have been suppressed, and repeated indices are summed over) into four equations to be used for points $\mathbf{k}_i$ in the reduced zone only. These equations are of the form $h_{\mathbf{k}_i \mathbf{k}'_i} a^{\mathbf{k}'_i} = E a^{\mathbf{k}_i}$, where $h_{\mathbf{k}_i \mathbf{k}'_i} = H^{\mathbf{k}_i \mathbf{k}'_i} \pm H^{\mathbf{k}_i T \mathbf{k}'_i}$. The $A^{\mathbf{k}}$ are then reconstructed from the reduced eigenvectors $a^{\mathbf{k}_i}$. Moreover, we apply time reversal and hermiticity in order to accelerate the calculation of the matrix elements.

A set with 864 **k** points in the full BZ is used to check the various ingredients of our calculation, in particular the number of bands and the importance of the off-diagonal elements of the inverse dielectric matrix in the evaluation of W. In the inset of Fig. 1, the continuous line shows the results of a calculation with 4 valence and 4 conduction bands, and the full ϵ^{-1}. In the region of interest (below 4.5 eV), a 6 bands calculation (4 valence + 2 conduction, dotted line) appears to be sufficient. Neglecting the off-diagonal elements of $\epsilon^{-1}_{\mathbf{GG'}}(\mathbf{q})$ yields an indistinguishable curve. We then use 6 bands and the diagonal ϵ^{-1} to compute the spectrum with 2048 **k** points in the full BZ. In the main part of Fig. 1 the experimental spectrum (dotted line) [28] is compared to: i) an RPA calculation [29] taking only into account the QP shifts, but not the excitonic or local field effects (short–dashed curve): the result is, as generally observed, in great discrepancy with experiment; ii) a calculation including local field effects (i.e. using equation (3) with W set to 0, long–dashed curve): the agreement is worsened, since the oscillator strength is slightly shifted to higher energies and both the E_1 and E_2 peaks are lowered, thus confirming previous findings in the literature [15,18]; iii) finally, the full calculation including the electron–hole attraction (continuous curve): absolute intensities now agree well with experiment. The remaining slight overestimate is of the order of what has been predicted by Ref. [21] to be the contribution of dynamical effects. More important, the peak positions and the relative intensity of the main structures are both in good agreement with experiment. Also the structure at 3.8 eV, even though slightly overestimated due to a finite **k** point sampling, has been repeatedly confirmed in theoretical and experimental work [30].

In conclusion, we have shown how excitonic effects can be included in an *ab initio* calculation of optical absorption spectra of semiconductors. At the example of bulk silicon, we have demonstrated that good agreement with experiment can be obtained for a case where the inclusion of self-energy and local field effects alone still gives rise to a rather poor theoretical spectrum. In this context, bulk silicon is not particularly easy to handle, since the bottleneck of the calculation is the number of **k** points (high in silicon, due to large dispersion) and not the energy cutoff. Even though, the computational effort, mostly steming from diagonalization, is reasonable and demands only a few hours on a Cray C98. The present work opens hence the way to first-principles calculations of optical absorption spectra with a precision comparable to that typically achieved in ground state calculations.

[1] P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964); W. Kohn and L. Sham, Phys. Rev. **140**, A1133 (1965).

[2] L. Hedin, Phys. Rev. A **139**, 796 (1965).

[3] M.S. Hybertsen and S.G. Louie, Phys. Rev. Lett. **55**, 1418 (1985); Phys. Rev. B **34**, 5390 (1986).

[4] R.W. Godby, M. Schlüter, and L.J. Sham, Phys. Rev. Lett. **56**, 2415 (1986); Phys Rev. B **37**, 10 159 (1988).

[5] See, for example, G.E. Engel and B. Farid, Phys. Rev. B **46**, 1812 (1992).

[6] G. Onida, L. Reining, R.W. Godby, R. Del Sole, and W. Andreoni, Phys. Rev. Lett. **75**, 818 (1995).

[7] R. Del Sole and R. Girlanda, Phys. Rev. B **48**, 11 789 (1993).

[8] V.I. Gavrilenko and F. Bechstedt, Phys. Rev. B **54**, 13 416 (1996).

[9] F. Mauri and R. Car, Phys. Rev. Lett. **75**, 3166 (1995).

[10] S. Albrecht, G. Onida, and L. Reining, Phys. Rev. B **55**, 10 278 (1997).

[11] The difficulties and some preliminary results were illustrated in Ref. [12].

[12] S. Albrecht, G. Onida, L. Reining, and R. Del Sole, Comp. Mat. Sci. **572** (1997) (in press).

[13] C.S. Wang and B.M. Klein, Phys. Rev. B. **24**, 3417 (1981).

[14] D.J. Stukel, R.N. Euwema, T.C. Collins, F. Herman, and R.L. Kortum, Phys. Rev. **179**, 740 (1969); J.P. Walter, M.L. Cohen, Y. Petroff, and M. Balkanski, Phys. Rev. B **1**, 2661 (1970).

[15] S.G. Louie, J.R. Chelikowsky, and M.L. Cohen, Phys. Rev. Lett. **34**, 155 (1975).

[16] J.C. Phillips, Phys. Rev. A **136**, 1705 (1964); B. Velicky and J. Sak, Phys. Status Solidi **16**, 147 (1966).

[17] R.R.I. Zucca and Y.R. Shen, Phys. Rev. B **1**, 2668 (1970); J.E. Rowe, F.H. Pollack, and M. Cardona, Phys. Rev. Lett. **22**, 933 (1969).

[18] W. Hanke and L.J. Sham, Phys. Rev. Lett. **43**, 387 (1979); Phys. Rev. B **21**, 4656 (1980).

[19] W. Hanke and L. J. Sham, Phys. Rev. B **12**, 4501 (1975); R. Del Sole and E. Fiorino, Phys. Rev. B **29**, 4631 (1984).

[20] G. Strinati, Phys. Rev. Lett. **49**, 1519 (1982); Phys. Rev. B **29**, 5718 (1984).

[21] F. Bechstedt, K. Tenelsen, B. Adolph, and R. Del Sole, Phys. Rev. Lett. **78**, 1528 (1997).

[22] This form for H was already used by Ekhardt *et al.* [23], who, however, did not give further details.

[23] W. Ekardt and J.M. Pacheco, Phys. Rev. B **52**, 16 864 (1995).

[24] F. Bassani and G. Pastori Parravicini, *Electronic States and Optical Transitions in Solids* (Pergamon, Oxford, 1975), Chap. 6.

[25] G.B. Bachelet, D.R. Hamann, and M. Schlüter, Phys. Rev. B **26**, 4199 (1982).

[26] H.J. Monkhorst and J.D. Pack, Phys. Rev. B **13**, 5188 (1976).

[27] R.W. Godby and R.J. Needs, Phys. Rev. Lett. **62**, 1169 (1989).

[28] D.E. Aspnes and A.A. Studna, Phys. Rev. B **27**, 985 (1983).

[29] For comparison, all curves in the main part of Fig. 1 correspond to calculations with 2048 **k** points in the full BZ. The calculation without excitonic effects can also be (and usually is) performed with larger **k** points sets. Doing this, only the high-energy part of the spectra is modified, namely the anomalous peak at about 5 eV, which reduces to a small shoulder at full convergence.

[30] J.R. Chelikowsky and M.L. Cohen, Phys. Rev. B **14**, 556 (1976), and references therein.

FIG. 1. Absorption spectra of Si. Inset: calculation according to equation (9) with 864 **k** points in the BZ, using 8 bands (continuous curve) or only 6 bands (dotted curve). Main part: Calculation according to equation (9) with 2048 **k** points in the BZ, 6 bands and the diagonal approximation to ϵ^{-1}: with both electron–hole attraction and local field effects in the Hamiltonian (continuous curve), inclusion of local field effects alone (long–dashed curve) and RPA with QP shifts only (short–dashed curve). Experimental curve (dots) [28].