Good semiconductor band gaps with a modified local-density approximation

D. M. Bylander and Leonard Kleinman
Department of Physics, University of Texas, Austin, Texas 78712
(Received 22 November 1989)

The exchange operator is divided into two parts, a Thomas-Fermi screened exchange operator and the remainder. The remainder and correlation are treated in the local-density approximation, while the screened exchange matrix elements are exactly evaluated. Calculations for Si result in much improved band gaps as well as an improved exchange contribution to the binding energy.

It has long been known[1] that Hartree-Fock energy-band calculations result in semiconductor band gaps too wide by a factor of 3 or more because the dynamic screening of the Fock operator is neglected. On the other hand, Kohn-Sham[2] (KS) calculations are well known to result in gaps too small by 0.7 eV or more which results in Ge being a zero-band-gap semiconductor.[3] This is attributed to the zeroth Fourier component of the KS exchange potential changing discontinuously across the energy gap.[4,5] In the original[2] KS theory

$$V_{\rm xc}(\mathbf{r})=\delta E_{\rm xc}[\rho]/\delta\rho(\mathbf{r}) \tag{1}$$

is defined with the subsidiary condition that the number of electrons is fixed so that

$$V_{\rm xc}(\mathbf{G})=\delta E_{\rm xc}[\rho]/\delta\rho_{\mathbf{G}} \tag{2}$$

is undefined for $\mathbf{G}=0$. Here $\rho_{\mathbf{G}}$ is the Gth Fourier component of the charge density and $E_{\rm xc}[\rho]$ is the exchange-correlation part of the Hohenberg-Kohn[6] energy-density functional. Since the eigenvalues ε_i of the Schrödinger equation containing the exchange-correlation potential $V_{\rm xc}$ are assigned no meaning in this theory and its eigenfunctions are independent of $V_{\rm xc}(0)$, there is no need for $V_{\rm xc}(0)$ to be defined. Replacing the variations which conserve number n by arbitrary ones via a Lagrange multiplier, Perdew *et al.*[7] were able to show that $\varepsilon_N=-I$ for $N-1<n\le N$ and $\varepsilon_{N+1}=-A$ for $N<n\le N+1$ where I and A are the ionization potential and electron affinity in either an atom with atomic number N or a semiconductor with N valence-band states. Here ε_N and ε_{N+1} are the Nth and $(N+1)$st eigenvalues and the number of electrons n is assumed to be a continuous variable. However, ε_{N+1} calculated for the N electron case is not the same as ε_{N+1} when $N<n\le N+1$ due to the discontinuity in $V_{\rm xc}(\mathbf{G}=\mathbf{0})$ which occurs as n passes through integer values.[4,5] Greater insight into this discontinuity can be obtained by examining $E_T[\rho]$ in the two band model where $E_T[\rho]$ is the kinetic energy functional. In that case, unlike the exchange case, explicit analytical results are obtained[8] for $V_T(\mathbf{G})$ and $V_T(\mathbf{G}=\mathbf{0})$ for states on either side of the gap and the discontinuity in $V_T(\mathbf{G}=\mathbf{0})$ thus exactly calculated.

Another way of looking at this discontinuity is to note that for any operator O one may write $O\psi=V_\psi(\mathbf{r})\psi(\mathbf{r})$ where $V_\psi(\mathbf{r})=O\psi/\psi$ is a wave-function-dependent potential. When ψ changes discontinuously as it does (from bonding to antibonding) across the energy gap, V_ψ will also change discontinuously. The only surprising thing is that in density-functional theory (DFT), it is only the zeroth Fourier component of V_ψ which is discontinuous. This is a consequence of the fact that only $\sum_i|\psi_i|^2$ and not the individual ψ_i have any physical significance in DFT. Be that as it may, it occurred to us that if the exchange-correlation operator could be separated into two parts, one which could be treated exactly and another which resulted in an only weakly ψ-dependent V_ψ, then if only the latter were replaced by a density functional, one could expect to obtain good energy gaps. We write

$$H_{\rm xc}=H_{sx}+\delta E_x/\delta\rho-\delta E_{sx}/\delta\rho+\delta E_c/\delta\rho\ . \tag{3}$$

Here H_{sx} is an exchange operator calculated with the screened Coulomb interaction $e^{-K_s r}/r$, $E_x[\rho]=-\frac{3}{2}(3/\pi)^{1/3}\rho^{4/3}(\mathbf{r})$ is the well-known local-density approximation (LDA) for the exchange energy-density functional[2] in Rydberg atomic units, $E_c[\rho]$ is one of the LDA correlation functionals (we use Wigner[9]), and $E_{sx}[\rho]$ is the LDA screened exchange density functional,[10]

$$E_{sx}[\rho]=-\frac{3}{2}\left[\frac{3}{\pi}\right]^{1/3}\times\left\{1-\frac{\gamma^2}{6}\left[1-\left[\frac{\gamma^2}{4}+3\right]\ln\left[1+\frac{4}{\gamma^2}\right]+\frac{8}{\gamma}\tan^{-1}\left[\frac{2}{\gamma}\right]\right]\right\}\rho^{4/3}(\mathbf{r}) \tag{4}$$

where $\gamma=K_s/k_F$. We call these three terms the modified (M)LDA. What is modified, of course, is not the LDA but that part of the Hamiltonian which is treated in the LDA. The matrix elements of H_{sx} are

$$\langle\mathbf{q}+\mathbf{G}|H_{sx}|\mathbf{q}+\mathbf{G}'\rangle=\frac{8\pi}{\mathcal{N}\Omega}\sum_{n,\mathbf{k},i}\frac{a_{n\mathbf{k}}(\mathbf{G}_i+\mathbf{G})a^*_{n\mathbf{k}}(\mathbf{G}_i+\mathbf{G}')}{(\mathbf{k}-\mathbf{q}+\mathbf{G}_i)^2+K_s^2} \tag{5}$$

where the sum is over occupied bands n at $\mathcal{N}$ different $\mathbf{k}$'s in the Brillouin zone, Ω is the unit-cell volume, and $a_{n\mathbf{k}}(\mathbf{G}_i)$ is the coefficient[11] of the $(\mathbf{k}+\mathbf{G}_i)$th plane wave in

$\psi_{n\mathbf{k}}$. The only remaining question is what the dependence of γ on $\rho(\mathbf{r})$ should be. One could argue that K_s^2 in Eq. (5) is independent of ρ so that $\gamma^2 = K_s^2/k_F^2 \sim \rho^{-2/3}$. Or one could argue that K_s is either the Thomas-Fermi wave vector, $K_{\mathrm{TF}}^2 = 4k_F/\pi$, or proportional to it, so that $\gamma^2 \sim \rho^{-1/3}$. We believe, however, that the correct argument is that E_{sx} must scale like an exchange energy density and that the dimensionless γ has no ρ dependence, i.e., $\gamma^2 \sim \rho_0^{-1/3}$ rather than $\rho(\mathbf{r})^{-1/3}$. Note, however, that all choices are equivalent in the free-electron-gas limit so all yield valid LDA's for the screened exchange operator.

We have performed self-consistent norm-conserving[12] pseudopotential calculations using both the MLDA with $K_s = K_{\mathrm{TF}}$ and the usual LDA. The ten-"special"-**k**-point sample[13] of an irreducible wedge of the Brillouin zone (BZ) was used which gives a sum over 256 **k**'s in the full BZ in Eq. (5). In order to facilitate comparison with the quasiparticle energies obtained by Hybertsen and Louie,[14] we expanded in the same set of plane waves with $k^2 < 17$ Ry as they. The results are given in Table I. We also tried the calculation with $\gamma^2 \sim [\rho(\mathbf{r})]^{-2/3}$. This gave a gap of only 0.773 eV and a bandwidth of 12.76 eV. We were able to increase the gap to 1.030 eV by taking $K_s^2 = \frac{1}{2}K_{\mathrm{TF}}^2$ but only at the cost of increasing the bandwidth to 13.44 eV. Thus our choice of γ independent of $\rho(\mathbf{r})$ as the most physically reasonable is verified by the numerical results. We also calculated the MLDA energies from the LDA eigenfunctions to see if the tedious repeated evaluation of Eq. (5) could be avoided.[15] The errors are small but not negligible. For example E_{gap} became 1.396 eV and the Γ'_{25v}-Γ'_{2c} gap became 3.77 eV.

Except for ε_N, no KS eigenvalues represent excitation energies in the N-electron problem. By removing the discontinuity in $V_{\mathrm{xc}}(0)$ as $N \to N+1$ we have attempted to make ε_{N+1} an excitation energy as well,[16] but no other ε_i are expected to be meaningful. Therefore it is quite surprising that the MLDA eigenvalues at the bottom of the valence bands (i.e., Γ_1 relative to the top of the valence bands and the Γ_1-L_1 bandwidth) are in better agreement with the experimental results than the calculated quasiparticle energies. This we believe is peculiar to the semiconductors, for it is well known that the occupied valence bandwidth in the nearly free electron metals is reduced from its LDA value by many-body effects.[17-20] Since the free-electron Hartree-Fock energy at $k=0$ is twice that at k_F, exchange appreciably increases the bandwidth. The statically screened exchange that we use cuts that increase drastically but cannot turn it into a reduction of the bandwidth. There is currently some controversy in the literature concerning the magnitude of the reduction of the quasiparticle bandwidth. For example, for Na, Northrup, Hybertsen, and Louie[17] (NHL) obtain a reduction of 0.64 eV from the LDA bandwidth whereas Mahan and Sernelius[18] (MS) obtain 0.36 eV which they[19] claim is in agreement with the experimental value[20] of 0.51 eV because there is a surface contribution to the narrowing of the bandwidth measured by photoemission. MS pointed out that NHL included exchange corrections to their dielectric function but failed to include vertex corrections in the self-energy, which is inconsistent since the two corrections tend to cancel. We[21] noted the same thing over 20 years ago when we defined three different dielectric functions: ϵ_{tt} for a Coulomb interaction with no vertex corrections, which is appropriate to two test charges; $\epsilon_{\mathbf{k},t}$ for a Coulomb line with vertex corrections at one end, which is appropriate to an electron and a test charge or to an electron exchanging with itself if there are no intervening vertices; and $\epsilon_{\mathbf{k}',\mathbf{k}}$ for a Coulomb line with vertex corrections at both ends, appropriate in all other cases. Thus we believe MS are correct in their assertion that the LDA bandwidth of Na has an 11% reduction and Al only a $\frac{1}{2}$% reduction due to many-body effects. Hence it seems likely that the MLDA would yield an improved Fermi surface for Al along with an acceptable bandwidth.

TABLE I. Various energy gaps (in eV) in silicon calculated with the LDA, MLDA, quasiparticle (QP) theory (Ref. 13), and compared with experimental values in Ref. 13.

	LDA	MLDA	QP	Expt.
E_{gap}	0.439	1.323	1.29	1.17
$\Gamma_{1v} \to \Gamma'_{25v}$	12.00	12.54	12.04	12.5±0.6
$\Gamma'_{25v} \to \Gamma_{15c}$	2.52	3.34	3.35	3.4
$\Gamma'_{25v} \to \Gamma'_{2c}$	3.15	3.86	4.08	4.2
$X_{4v} \to \Gamma'_{25v}$	2.87	2.78	2.99	2.9, 3.3±0.2
$\Gamma'_{25v} \to X_{1c}$	0.57	1.48	1.44	1.3
$L'_{2v} \to \Gamma'_{25v}$	9.64	10.13	9.79	9.3±0.4
$L_{1v} \to \Gamma'_{25v}$	7.03	7.07	7.18	6.7±0.2
$L'_{3v} \to \Gamma'_{25v}$	1.20	1.16	1.27	1.2±0.2, 1.5
$\Gamma_{1v} \to L_{1v}$	4.97	5.46	4.86	5.8±0.8
$\Gamma'_{25v} \to L_{1c}$	1.41	2.12	2.27	2.1, 2.4±0.15
$\Gamma'_{25v} \to L_{3c}$	3.29	4.21	4.24	4.15±0.1
$L'_{3v} \to L_{1c}$	2.61	3.29	3.54	3.45
$L'_{3v} \to L_{3c}$	4.49	5.37	5.51	5.50

It is well known that the LDA results in too little exchange energy and too much correlation energy for systems with energy gaps such as atoms or semiconductors.[22] The errors have opposite signs and as a rule the exchange error is larger in magnitude but smaller in percentage. For example, for the neon atom[22] $E_{\mathrm{corr}} = -10.0$ eV, $E_{\mathrm{corr}}^{\mathrm{LDA}} = -19.9$ eV and $E_x = -329.5$ eV, $E_x^{\mathrm{LDA}} = -298.4$ eV. The exchange energy results from the exchange hole which has several wave-function-independent properties, such as removing exactly one electron and subtracting out all parallel spin charge density at its origin. Thus an approximation based on free electrons works reasonably well even in atoms. Correlation, on the other hand, results from the admixture of excited configurations into the one-electron ground-state configuration. In the limit of an infinite energy gap the correlation energy must vanish. Therefore we propose a modified correlation functional

$$\hat{E}_c[\rho] = W E_c[\rho]/(W+\Delta) \tag{6}$$

where $E_c[\rho]$ is any standard correlation density functional, W is the bandwidth, and Δ is some average energy gap. Since W and Δ are functionals of the density, $\hat{E}_c[\rho]$ is also a valid density functional and becomes equal to

$E_c[\rho]$ in the free-electron-gas limit. It is not a local-density functional, however, because W and Δ are not local functionals of ρ.

Because the total ground-state energy may be obtained from the single-particle Green's function, the binding energy of a solid is, in principle, obtainable from the quasiparticle calculations; however, as far as we know this has never been done. We here calculate the binding energy in the MLDA and compare with the LDA. [Since Eq. (6) is somewhat *ad hoc*, we used the usual Wigner correlation here.] The first row of Table II lists the one-electron eigenvalues summed over bands and averaged over the BZ. The next four rows subtract various electron-electron contributions to the one-electron eigenvalues. The V's are the input potentials of the final iteration while the ρ's are the output charge densities. In particular,

$$\langle\psi|H_{sx}|\psi\rangle=\mathcal{N}^{-1}\sum_{m,\mathbf{q},\mathbf{G},\mathbf{G}'}a^{\text{out}*}_{m\mathbf{q}}(\mathbf{G})a^{\text{out}}_{m\mathbf{q}}(\mathbf{G}')\langle\mathbf{q}+\mathbf{G}|H^{\text{in}}_{sx}|\mathbf{q}+\mathbf{G}'\rangle$$
$$=(8\pi/\mathcal{N}^2\Omega)\sum_{n,\mathbf{k},m,\mathbf{q},i}\left|\sum_{\mathbf{G}}a^{\text{in}*}_{n\mathbf{k}}(\mathbf{G}+\mathbf{G}_i)a^{\text{out}}_{m\mathbf{q}}(\mathbf{G})\right|^2[(\mathbf{k}-\mathbf{q}+\mathbf{G}_i)^2+K_s^2]^{-1}\,. \tag{7}$$

The sum of the first five rows, listed in the sixth, represents the sum of the kinetic and pseudopotential energies of the electrons. The difference between the LDA and MLDA is much smaller than the differences between the individual contributions to this term and is a measure of the differences between the LDA and MLDA eigenfunctions. The next three rows list the electron-electron Coulomb and correlation and exchange energies evaluated in the LDA. H_{sx} is the contribution of the screened exchange operator to the total energy and $-E_{sx}$ subtracts off the screened exchange energy in the LDA. H_{sx} is given by Eq. (7) with a factor of $\frac{1}{2}$ to avoid double counting and with both a's being output a's. The twelfth row is the Ewald-Madelung energy of point ions in a constant background of compensating charge density. The zero point vibrational energy row in the 13th row is taken from Ref. 23. The sum of the 6th through 13th rows is the total energy, whose negative is the binding energy E_B listed in the last row. Although the individual contributions to E_B are not well converged, as can be seen by comparing $2E_{\text{Coul}}$ with $\int V_{\text{Coul}}\rho$, E_B because it is calculated variationally,[24] is converged to at least five decimal places.

TABLE II. Various contributions to the total energy of Si in Ry/atom.

	MLDA	LDA
$\sum E_{n\mathbf{k}}$	−0.051 99	0.315 08
$-\int V_{\text{Coul}}\rho$	−1.226 26	−1.075 74
$-\int V_{xc}\rho$	3.118 69	3.080 42
$\int V_{sx}\rho$	−0.521 36	
$-\langle\psi\|H_{sx}\|\psi\rangle$	0.959 72	
$\sum E_{n\mathbf{k}}-\int\psi H_{ee}\psi$	2.278 80	2.319 76
E_{Coul}	0.614 50	0.537 87
E_c	−0.364 92	−0.364 20
E_x	−2.047 92	−2.019 64
$-E_{sx}$	0.394 60	
H_{sx}	−0.480 52	
E_{Mad}	−8.399 65	−8.399 65
E_{vib}	0.005	0.005
E_B	8.000	7.921

The experimental $E_B=7.917$ Ry is much closer to the LDA result than to the MLDA due to a fortuitous cancellation of exchange and correlation errors. Note though that when we took the LDA valence exchange-correlation energy to be of the form $E_{xc}(\rho_{\text{total}})-E_{xc}(\rho_{\text{core}})$ we[3] obtained $E_B=7.990$ Ry, which is much closer to our current MLDA result than to our current LDA result. Using the MLDA eigenfunctions, we have calculated[25] the unscreened Fock exchange energy to be −2.128 72 Ry. Our MLDA exchange is $E_x+H_{sx}-E_{sx}=-2.133\,84$ Ry which is in much better agreement with the Fock exchange than our LDA $E_x=-2.019\,64$ Ry. In order to obtain the experimental E_B with the MLDA exchange, E_c would have to be reduced to −0.282 Ry; to obtain it with the Fock exchange, it would have to be reduced to −0.287 Ry. Using the experimental bandwidth, $W=12.5$ eV, and the average dielectric[26] gap, $\Delta=4.8$ eV, in Eq. (6) yields $\hat{E}_c=-0.264$ Ry. If we take $\Delta=3.9$ eV, the weighted average of the experimental gaps at points Γ, X, and L, we obtain $\hat{E}_c=-0.278$ Ry. These estimates omit any changes in E_B arising from changes in the eigenfunctions due to replacing $E_c[\rho]$ with $\hat{E}_c[\rho]$. Another effect of using $\hat{E}_c[\rho]$ would be to reduce E_{gap} toward the experimental value. (See Table I.)

In conclusion we would like to point out that although there are some points of similarity, this work is basically different than that of Gygi and Baldereschi.[27] They separated the self-energy into two pieces, one of which could be evaluated in the LDA, in order to obtain an approximation to the GW approximation for quasiparticle energies; we separated the Fock exchange operator into two pieces, one of which could be treated in the LDA, in order to obtain one-electron eigenvalues which approximate excitation energies around the energy gap. For Si we found that this yields not only an improved energy gap (or the exact energy gap with a small adjustment of K_s) but it also yields KS eigenvalues throughout the bands which are in improved agreement with experimental excitation energies. Our approximation also makes feasible that which appears not to be feasible in the GW approximation[14] or approximations to it,[27] i.e., total-energy calculations. The total exchange energy we obtained for Si is larger in magnitude than the LDA exchange energy and in much better agreement with the exact exchange energy when that is defined to be the Fock

exchange energy of KS eigenfunctions. When an overestimate of the correlation energy inherent in the LDA was corrected, the total binding energy was in excellent agreement with experiment. Finally, we note that the *GW* approximation is based upon a dielectric matrix which is obtained from a LDA calculation. There are cases when the LDA is so bad that a semiconductor becomes a metal and the dielectric matrix becomes useless. This would have no effect on a MLDA calculation which then could, if so desired, be used as the starting point for a GW calculation.

This work was supported by the University of Texas Center for High Performance Computing, by the Robert A. Welch Foundation (Houston, Texas), and by the National Science Foundation under Grant No. DMR-87-18048.

[1]J. C. Phillips and L. Kleinman, Phys. Rev. **128**, 2098 (1962).
[2]W. Kohn and L. J. Sham, Phys. Rev. **140**, A1133 (1965).
[3]Y. T. Shen, D. M. Bylander, and L. Kleinman, Phys. Rev. B **36**, 3465 (1987).
[4]J. P. Perdew and M. Levy, Phys. Rev. Lett. **51**, 1884 (1983).
[5]J. J. Sham and M. Schluter, Phys. Rev. Lett. **51**, 1888 (1983).
[6]P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964).
[7]J. P. Perdew, R. G. Parr, M. Levy, and J. L. Balduz, Jr., Phys. Rev. Lett. **49**, 1691 (1982).
[8]L. Kleinman, Phys. Rev. B **33**, 7299 (1986).
[9]E. Wigner, Phys. Rev. **46**, 1002 (1934).
[10]A. R. E. Mohammed and V. Sahni, Phys. Rev. B **29**, 3687 (1984).
[11]When iterating to self-consistency the input and output potentials for the nth iteration are averaged to obtain the input potential of the $(n+1)$st iteration. The product $a_{n\mathbf{k}}(\mathbf{G}_l+\mathbf{G})a^*_{n\mathbf{k}}(\mathbf{G}_l+\mathbf{G}')$ is averaged in exactly the same way.
[12]A. Baldereschi, Phys. Rev. B **7**, 5212 (1973); D. J. Chadi and M. L. Cohen, *ibid.* **8**, 5747 (1973).
[13]L. Kleinman and D. M. Bylander, Phys. Rev. Lett. **48**, 1425 (1982).
[14]M. S. Hybertsen and S. G. Louie, Phys. Rev. B **34**, 5390 (1986).
[15]Note that Eq. (5) must be evaluated for all $\mathbf{G}$ and $\mathbf{G}'$. On the other hand, if the eigenfunctions are assumed known, one needs to evaluate

$$\sum_{\mathbf{G},\mathbf{G}'} a^*_{m\mathbf{q}}(\mathbf{G})a_{m\mathbf{q}}(\mathbf{G}')\langle \mathbf{q}+\mathbf{G}|H_{sx}|q+\mathbf{G}'\rangle$$
$$=(8\pi/\mathcal{N}\Omega)\sum_{n,\mathbf{k},l}\left|\sum_{\mathbf{G}} a^*_{n\mathbf{k}}(\mathbf{G}_l+\mathbf{G})a_{m\mathbf{q}}(\mathbf{G})\right|^2$$
$$\times[(\mathbf{k}-\mathbf{q}+\mathbf{G}_l)^2+K_s^2]^{-1}$$

which does not involve $\mathbf{G}'$.
[16]Although there is no *a priori* way of doing this, with a somewhat larger screening constant we would have obtained the exact value for the energy gap. That would mean that the discontinuity in $V_{xc}(0)$ cancelled the error in the LDA in this calculation, but, in principle, if one knew the exact density functional for exchange, correlation, and screened exchange, one could exactly remove the discontinuity in $V_{xc}(0)$ from the density functional.
[17]J. E. Northrup, M. S. Hybertsen, and S. G. Louie, Phys. Rev. B **39**, 8198 (1988).
[18]G. D. Mahan and B. E. Sernelius, Phys. Rev. Lett. **62**, 2718 (1989).
[19]K. W.-K. Shung, B. E. Sernelius, and G. D. Mahan, Phys. Rev. B **36**, 4499 (1987).
[20]I. W. Lyo and E. W. Plummer, Phys. Rev. Lett. **60**, 1558 (1988).
[21]L. Kleinman, Phys. Rev. **172**, 383 (1968).
[22]A. Zunger, J. P. Perdew, and G. L. Oliver, Solid State Commun. **34**, 933 (1980).
[23]M. T. Yin and M. L. Cohen, Phys. Rev. B **26**, 5668 (1982).
[24]J. R. Chelikowsky and S. G. Louie, Phys. Rev. B **29**, 3470 (1984).
[25]When the $(\mathbf{k}-\mathbf{q}+\mathbf{G}_l)^2$ denominator vanished, we replaced it with the average value of k^{-2} in a sphere whose volume is $\frac{1}{256}$ that of the BZ.
[26]J. A. van Vechten, Phys. Rev. **182**, 891 (1969).
[27]F. Gygi and A. Baldereschi, Phys. Rev. Lett. **62**, 2160 (1989).