Screened exchange density functional applied to solids

Stewart J. Clark[1] and John Robertson[2]
[1]*Department of Physics, Science Laboratories, University of Durham, South Road, Durham DH1 3LE, United Kingdom*
[2]*Department of Engineering, Cambridge University, Cambridge CB2 1PZ, United Kingdom*
(Received 11 February 2010; revised manuscript received 18 June 2010; published 27 August 2010)

The screened exchange (SX) hybrid functional, which mixes a Thomas-Fermi screened Hartree-Fock exchange into the local-density approximation (LDA), is applied to solids. SX can be used as a variational functional in total-energy minimizations and it represents an efficient way to improve the accuracy of band calculations. Here we summarize the computational implementation within the plane-wave, pseudopotential formalism and compare results to related methods. This representation of the exchange-correlation energy improves the incorrect treatment of the self-interaction in the LDA. SX also improves the calculated band gap of a wide range of semiconductors and insulators compared to the LDA with a mean relative error of 7.4% compared to near 30% for LDA or generalized gradient approximation.

DOI: 10.1103/PhysRevB.82.085208 PACS number(s): 71.15.Dx, 71.15.Mb, 71.20.−b

I. INTRODUCTION

The prime role of quantum-mechanical calculations in solids and molecules is to determine total-energy differences, atomic structure, and electronic charge density. For many years, the local-density approximation (LDA) of the density functional theory (DFT) provided an efficient method for such calculations in solids, giving both lattice constants and bulk moduli with reasonable accuracy. However, a major weakness of LDA is that it underestimates the band gaps for semiconductors, insulators, and strongly correlated systems. DFT expresses the many-body Schrödinger equation in terms of the electron density, which necessitates an approximation to the many-body exchange and correlation effects. The LDA approximates the exchange-correlation (XC) interaction by a functional of the local electron density. However, the eigenvalues of the resulting variational equation do not represent quasiparticle energies causing an underestimate of the band-gap energy in semiconductors and insulators. This arises from the discontinuity of the XC energy across the band gap as a function of electron occupancy.[1] This applies both in terms of the eigenvalue difference between valence and conduction states, and if the band gap is found by total-energy differences.

The LDA can also give the incorrect localization of electron and hole states due to the incorrect treatment of the electron self-interaction.[2] There have been improvements to the LDA such as the many generalized gradient approximations (GGAs) but these have not corrected the errors.

Several methods have been proposed to overcome the band-gap problem within DFT. The first was the self-interaction correction method of Perdew and Zunger.[3] However, the difficulty of its implementation means that it is rarely used even today.[4] A second method is the GW method, based on Green's function, which calculates the quasiparticle energies from an expansion of the electron self-energies and the dielectric function.[5–10] This widely used method gives generally reliable values for the band gaps of semiconductors and insulators. However, it has three disadvantages; it is very costly, it is a perturbative expansion which is often only used in the so-called G_0W_0 approximation whose result can depend on the starting configuration, and finally it cannot be used for self-consistent geometry optimizations because forces are not available.

A third method is LDA+U.[11,12] U is a Coulomb (Hubbard) energy parameter, which introduces a repulsion between the localized electrons on a given atom (often the d electrons) which can cause a symmetry breaking and thereby can open up a gap. It is a low cost method for open shell systems and it can be used as an energy functional for structural relaxations for open shell systems. It is not a generic method for correcting the error in band gaps and it works mainly for highly correlated systems. Therefore, the method does not solve the problem for closed shell systems, which include most semiconductors of technological interest. If it is used to fix closed shell systems, it requires unphysically large values of U to empirically fit the band gap, and as such has often been used incorrectly.

This indicates the need for a universally applicable method, which should be a functional whose density derivative exists, so that it can be used both for electronic and geometric structural relaxations, give accurate band gaps, and be fairly computationally efficient. These conditions, allowing for some increase in computational cost, are satisfied by the so-called hybrid functionals. Hybrid functionals mix a fraction of Hartree-Fock (HF) exchange with (semi-) local-density functionals and have given some excellent results for molecules and solids.

The local exchange and correlation functionals of LDA and GGA lead to a spurious self-interaction. The HF method uses a nonlocal exchange, so that it can be self-interaction free, but HF lacks electronic correlation, and its exchange is unrealistically long ranged due to an absence of screening.

Perdew *et al.*[13] gave some first-principles arguments based on the adiabatic linkage of the HF and LDA limits to argue that 25% is an appropriate amount of HF exchange to be mixed with local exchange-correlation functionals. Their PBEh (formerly PBE0) functional incorporates this fraction of HF exchange,[13–15] which also gives improved band gaps. The development of B3LYP functional[16] to correct the overbinding of LDA for small molecules led to a conclusion that 20% is the appropriate fraction of HF. Muscat *et al.*[17] found that B3LYP also gave good band gaps for many semi-

conductors. However, B3LYP gives poorer cohesive energies for compounds of heavier elements, for which B3LYP was not calibrated.[18] Furthermore, Becke[19] also argued for 50% HF exchange in his half-and-half functional. This has been mainly used on molecular systems where is overestimates the strength of hydrogen bonds.[20]

The Heyd-Scuseria-Erzenhof (HSE) functional modified the mixing of local and nonlocal functionals of PBEh by separating the local potential into long- and short-range parts while using a fraction (again, 25%) of HF exchange.[21–25] This is based on the notion that the exchange and correlation terms cancel at long range. The retention of only short-range HF exchange allows faster calculations. This was implemented for a local orbital basis and has been tested on various molecules and solids,[23,24] and later for a plane-wave basis with projector augmented waves.[26]

However, earlier, Bylander and Kleinman[27] proposed an analogous separation of long- and short-ranged parts of the screened exchange (SX) based on ideas of Phillips[28] and Hedin.[5] They represented the exchange interaction by a screened, nonlocal exchange potential, and mix with the LDA. Seidel *et al.*[29] noted the functional properties of SX. However, the method was not used extensively, except for band-structure calculations by the Freeman group,[30,31] because its implementation for a plane-wave basis was still quite costly. Important advances were then made by Gibson[32] who calculated the Hellman-Feynman stress terms for a plane-wave basis, which enabled its efficient implementation for full structural energy minimization. It has also since been used for a number of band structure and defect calculations.[33–36] This paper describes the implementation, testing, and some calculation results using the SX method which belongs to the general family of screened hybrid functionals.[37]

II. METHOD

The screened exchange functional is closely related to the Hartree-Fock method, which was derived before the local DFT functionals. These functionals are an implicit functional of the density and part of the XC operator is nonlocal. Hence the generalized Kohn-Sham orbitals are not eigenstates of a local Hamiltonian. Instead, the orbitals are solutions of a nonlocal Schrödinger equation,

$$-\frac{1}{2}\nabla^2\psi_i(r)+V_{loc}(r)\psi_i(r)+\int V_{nl}^{\mathrm{XC}}(r,r')\psi_i(r')dr'=\varepsilon_i\psi_i(r), \tag{1}$$

where $V_{nl}^{\mathrm{XC}}(r,r')$ is the nonlocal part of the XC potential and i labels the electronic states. The local potential, $V_{loc}(r)$, contains the Hartree potential, any local parts of the XC potential, and the external pseudopotential terms found in standard plane-wave calculations. The nonlocal XC potential is similar in form to the HF potential but it also incorporates the effects of correlation by screening the long-range interactions of exchange. This is achieved by introducing a factor which decays exponentially with electron separation. Labeling this nonlocal XC term with "sX," its contribution to the total energy in a periodic system of plane waves is

$$E_{nl}^{sX}=-\frac{1}{2}\sum_{ij,kq}\int\int \times\frac{\psi_{ik}^*(r)\psi_{ik}(r')\exp(-k_s|r-r'|)\psi_{jq}^*(r')\psi_{jq}(r)}{|r-r'|}drdr', \tag{2}$$

where i and j label electronic bands, k and q are the k points and k_s is a Thomas-Fermi screening length. However, it is advantageous to maintain the exact XC energy for the homogeneous electron gas. Therefore a local (*loc*) contribution is also required so that the total exchange-correlation energy in the screened exchange method is

$$E^{sX}=E_{nl}^{sX}+E_{loc}^{sX},$$

where E_{loc}^{sX} is this additional contribution which is parametrized using Perdew's expression for the LDA.[3] Thus, the local contribution to the exchange and correlation energy density is

$$\varepsilon_{loc}^{sX}(\rho)=\varepsilon_{loc}^{\mathrm{HEG}}(\rho)-\varepsilon_{nl}^{\mathrm{HEG}}(\rho), \tag{3}$$

where the local $\varepsilon_{loc}^{\mathrm{HEG}}(\rho)$ function is the same as the LDA [homogeneous electron gas (HEG)]. The second term is obtained by applying the nonlocal functional to the HEG, which is given by

$$\varepsilon_{nl}^{\mathrm{HEG}}(\rho)=V_X^{\mathrm{HEG}}(\rho)F(\rho), \tag{4}$$

where $F(\rho)$ is the function

$$F(\alpha)=1-\frac{4}{3}\tan^{-1}\left(\frac{2}{\alpha}\right)-\frac{\alpha^2}{6}\left[1-\left(\frac{\alpha^2}{4}+3\right)\ln\left(1+\frac{4}{\alpha^2}\right)\right],$$

$\alpha(\rho)=k_s/k_F(\rho)$ and $V_X^{\mathrm{HEG}}(\rho)$ is the pure exchange energy per electron. This ensures that both the LDA and HF limits are correct for the HEG at long and short screening lengths. This is analogous to the HSE term where the first term is represents the "$\alpha=0.25$" of HF, the second term is the long-ranged correlation, and the final term is similar to the short-ranged-screened local PBE correlation of HSE. An important factor is that SX reproduces the correct asymptotic limits of XC in both the free electron gas and the HF limit.

III. COMPUTATIONAL DETAILS

The SX method has been implemented within the plane-wave basis set and pseudopotential formalisms in the CASTEP code.[38,39] The valence electron wave functions are described by a plane-wave expansion and the ion-electron interactions are described using norm-conserving pseudopotentials of the Kleinman-Bylander form. The norm-conserving pseudopotentials used in our calculations are defined in Table I. Many of them are the default pseudopotentials of the CASTEP database. In other cases, more transferable pseudopotentials were constructed using the OPIUM code of Rappe *et al.*[40] This was found to be necessary for the Mg, Al, Ga, Ge, In, Ti, Cu, Zn, and O atoms.

The evaluation the screening length k_s is an important consideration. It is usually evaluated from the average elec-

TABLE I. Valence electron configurations and cutoff energies used for the norm-conserving pseudopotentials. The pseudopotentials are all generated within the LDA approximation. It should be noted that SX pseudopotentials are not well defined since the screening length varies from material to material (dependent on the material's electron density).

Element	Valence configuration	Cutoff energy (eV)
Al	$3s^2$, $3p^1$	380
As	$4s^2$, $4p^3$	380
B	$2s^2$, $2p^1$	380
Be	$2s^2$	850
Br	$4s^2$, $4p^5$	380
C	$2s^2$, $2p^2$	850
Ca	$3s^2$, $3p^6$, $4s^2$	850
Cd	$4d^{10}$, $5s^2$	850
Cl	$3s^2$, $3p^5$	380
Cu	$3d^{10}$, $4s^1$	850
F	$2s^2$, $2p^5$	900
Ga	$3d^{10}$, $4s^2$, $4p^1$	850
Ge	$4s^2$, $4p^2$	380
Hf	$5d^2$ $6s^2$	380
I	$5s^2$, $5p^5$	380
In	$5s^2$, $5p^1$	380
K	$3s^2$, $3p^6$, $4s^1$	850
Li	$2s^1$	500
Mg	$2s^2$, $2p^6$, $3s^2$	900
N	$2s^2$, $2p^3$	850
Na	$2s^2$, $2p^6$, $3s^1$	750
O	$2s^2$, $2p^4$	850
P	$3s^2$, $3p^3$	380
Pb	$5d^{10}$, $6s^2$, $6p^2$	680
Rb	$4s^2$, $4p^6$, $5s^1$	750
S	$3s^2$, $3p^4$	380
Sb	$5s^2$, $5p^3$	380
Se	$4s^2$, $4p^4$	380
Si	$3s^2$, $3p^2$	380
Sn	$5s^2$, $5p^2$	380
Sr	$4s^2$, $4p^6$, $5s^2$	680
Te	$5s^2$, $5p^4$	100
Ti	$3s^2$, $3p^6$, $3d^2$, $4s^2$	680
Zn	$3d^{10}$, $5s^2$	680
Zr	$4s^2$, $4p^6$, $5d^2$, $5s^2$	680

tron density but of which electrons? At the limits of evaluating the screening length with all electrons (core+valence) and no electrons, the SX method gives (almost) the LDA and HF approximations. The optimum band structure and band gap occurs between these two limits. Unfortunately we find that using the electrons which are considered valence with respect to the pseudopotential is not always optimum. When semicore d states are included in the evaluation of screening length, the band gap is underestimated by 20–50 %. Therefore, for elements such as Zn, Cd, or In with shallow filled d core states, the valence electron screening length is calculated from the outer s and p electrons only. The d cores are included as part of the valence electron calculation but their electron number does not contribute to the screening length.

As screening length varies with the valence density of the system, a SX pseudopotential is not well defined; there is no obvious choice of screening length to use when generating such a pseudopotential. Our pseudopotentials are generated within the LDA giving an inconsistency in the treatment of exchange and correlation between core and valence electrons. A recent investigation[41,42] into the variation in band gap with respect to core-valence partitioning within the LDA and G_0W_0 methods reported deviations of approximately 0.2 eV in the value of the band gap when comparing all-electron G_0W_0 results to that of using pseudopotentials were core-valence exchange corrections are ignored.

The nonlocal screened exchange energy is calculated efficiently by means of fast Fourier transform and the order of the transforms is arranged to minimize the storage requirements. The basis-set size is expressed in terms of a plane-wave kinetic-energy cutoff, which is given in Table I. The size of the plane-wave basis set is such that total-energy differences are converged to better than 1 meV/atom.

The Brillouin-zone integrations are performed on a Monkhorst-Pack k-point mesh. The size of the mesh is fixed such that total energies are converged to a similar energy accuracy as for the plane-wave expansion. The electronic energy minimizations are performed using a preconditioned conjugate-gradients scheme. The commonly used density mixing minimization scheme was found to be inappropriate for such calculations, causing numerical instabilities due to a significant part of the Hamiltonian not being a density functional. Geometries are relaxed under the influence of the self-consistent Hellmann-Feynman forces and stresses. Nonlocal stresses are evaluated by the scheme of Gibson *et al.*,[32] which now allows the efficient geometry optimization of the unit cell.

Details of the computational procedure are as follows: given a set of plane waves, we evaluate the expectation value of the nonlocal exchange-correlation operator for each state, which is given by

$$\varepsilon_{ik}^{NL} = -\frac{2\pi}{\Omega}\sum_{jq}\sum_{G}\frac{|C_{jqik}(G)|^2}{|q-k+G|^2+k_s^2}, \tag{5}$$

where i and j label bands, k and q label k points, Ω is the cell volume, G is the wave vector of the plane wave and

$$C_{jqik}(G) = \sqrt{\Omega} FT[\psi_{ik}^*(r)\psi_{jq}(r)], \tag{6}$$

where FT is the Fourier transform operator, transforming r space into G space. For each band (i index) and k point (k index), the computational procedure is (1) transform the plane-wave expansion of state ik, determined by expansion coefficients $c_{ik}(G)$, into real space. (2) For every band and k point [labeled j and q in Eq. (6)], evaluate the exchange integrand in real space and then Fourier transform to obtain reciprocal space the quantity $C_{jqik}(G)$. (3) Divide $|C_{jqik}(G)|^2$ by the reciprocal Coulomb factors plus screening, then per-

form the G-vector sum and increment the jq sum in Eq. (5). This completes the basic calculation.

To apply the nonlocal operator required for a self-consistent calculation, we have

$$\frac{\delta E_{NL}^{sX}}{\delta c_{ik}^{*}(G)} = \sqrt{\Omega} FT\left[-\sum_{jq} \psi_{jq}(r) f_{jqik}(r)\right] \tag{7}$$

with

$$f_{jqik}(r) = \frac{4\pi}{\sqrt{\Omega}} FT^{-1}\left[\frac{C_{jqik}^{*}(G)}{|q-k+G|^2 + k_s^2}\right]. \tag{8}$$

To perform this operation, we proceed similarly to that above for evaluating the eigenvalues. This gives the rate of change in the nonlocal energy term with respect to plane-wave coefficients. This is used to determine the steepest descents direction as part of a standard preconditioned conjugate-gradients minimization procedure.

However, in addition to the above formalism, note that Eq. (5) contains a singularity when $G=0$ and $k_s=0$ as $k \to q$ (and is sharply peaked at the same point when $k_s \neq 0$) and so the usual assumption that the potential is smoothly varying breaks down and integrations across the Brillouin zone require very fine sampling. However we introduce a divergence correction[41] in Eq. (5), which alleviates numerical instabilities in the evaluation of this term. We add a term to each expectation value of

$$\frac{2\pi}{\Omega}\left[\sum_{q} D(q-k) - \Omega \int_{B.Z.} D(q-k)dq\right] \tag{9}$$

for a suitable smearing function, D. This function D is Brillouin-zone periodic and has the same divergence rate as the exchange; it vanishes as $1/q^2$ as q vanishes. Other than this, the choice of D is arbitrary. Here we use

$$D(q) = \sum_{G} \frac{S(|q-G|)}{|q-G|^2},$$

where S is the sinusoidal envelope function,

$$S(x) = \begin{cases} \frac{1}{2}\left[\cos\left(\frac{\pi x}{\omega}\right) + 1\right], & |x| \leq \omega \\ 0, & |x| > \omega \end{cases}$$

and ω is the envelope width. The larger ω is the smoother D is away from the singularity. ω is set larger than the typical separation of k points.

Regarding symmetry, the use of symmetry has two main advantages in a (semi-) local functional calculation. First, it reduces the number of k points that need to be explicitly included, resulting in increased speed and lower memory requirements. Second, in certain situations, it causes the effective k-point set to be larger than the original Monkhorst-Pack grid, which may improve convergence, and ensures that symmetry-related degeneracies are exactly satisfied. This effective larger k-point set is implicit when a charge density symmetrization is performed. In a nonlocal functional calculation, use of symmetry has the advantage of reducing the k-point set on only one of the k-point sums in Eq. (5). Unlike

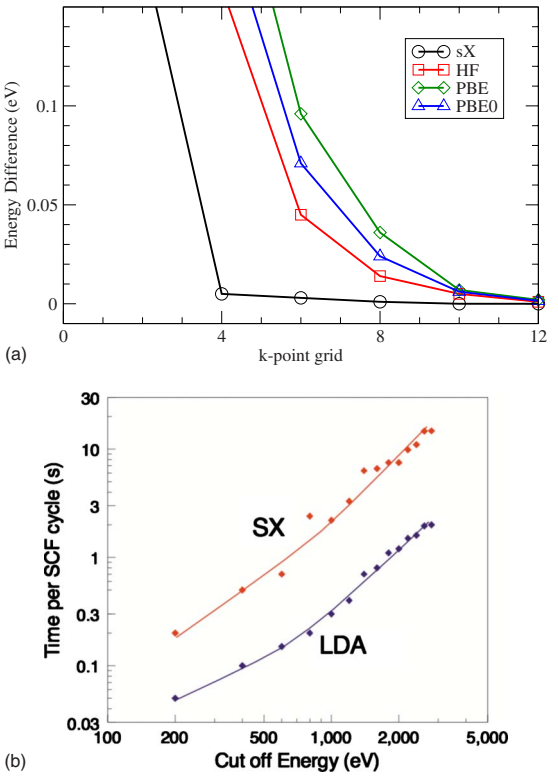

FIG. 1. (Color online) (a) A plot of convergence of the exchange and correlation energy's contribution to total energy with respect to k-point sampling grid size is shown for sX, PBE, PBE0, and HF functionals for silicon in the diamond structure is shown. The screening within the SX method makes the interactions short ranged and so has a faster convergence than the XC contribution to the total energy in HF and the PBE. (b) The computational scaling is shown with respect to basis size. We plot the time required for an SCF sX calculation with respect to plane-wave cutoff energy.

in a standard local functional calculation, the second k-point sum (labeled q, which we now call the q-point set) is not symmetry reduced. To impose full symmetry, which corresponds to charge-density symmetrization, the q-point set is symmetry expanded (this set being greater in size than the unreduced Monkhorst-Pack set) and the calculation then re-

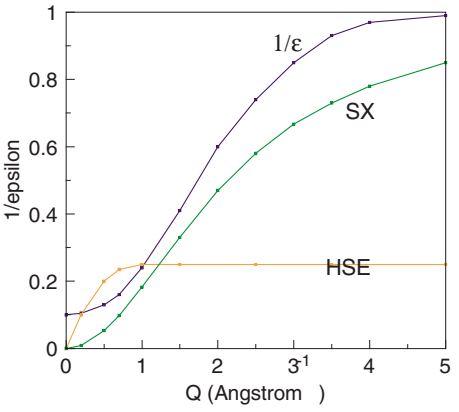

FIG. 2. (Color online) A comparison of the HF screening in the SX and in HSE functionals in reciprocal space is given by exampling the inverse dielectric function $(1/\varepsilon)$.

TABLE II. Comparison of the calculated PBE and SX band gaps to the experimental values are given. Also presented are the calculated PBE and SX lattice constants which are compared to the experimental values (Refs. 49–58).

Compound	PBE band gap (eV)	SX band gap (eV)	Expt. band gap (eV)	PBE lattice constant (Å)	SX lattice constant (Å)	Expt. lattice constant (Å)
Diamond	4.27	5.38	5.5	3.537	3.501	3.567
Si	0.69	1.07	1.12	5.401	5.397	5.431
Ge	0.59	0.69	0.7	5.478	5.414	5.657
c-SiC	1.47	2.25	2.42	4.320	4.262	4.348
AlP	1.66	2.21	2.45	5.438	5.386	5.451
AlAs	1.58	2.30	2.24	5.640	5.581	5.62
AlSb	1.57	1.83	1.70	6.063	6.021	6.13
GaP	1.70	1.85	1.9	5.502	5.374	5.45
GaAs	0.87	1.47	1.45	5.707	5.570	5.66
GaSb	1.00	1.13	0.82	6.066	5.905	6.09
InP	1.21	1.44	1.42	5.723	5.769	5.86
InAs	0.60	1.00	0.37	5.918	5.908	6.05
InSb	1.27	1.67	0.238	6.314	6.320	6.47
ZnO(zb)	0.89	3.43	3.44	4.583	4.586	4.51
ZnO (wz)	0.8	3.41	3.44	3.268/5.299	3.27/5.25	3.25/5.21
ZnS	2.15	3.74	3.80	5.606	5.421	5.41
ZnSe	1.68	2.71	2.82	5.875	5.569	5.67
ZnTe	1.81	2.34	2.39	6.280	6.025	6.089
CdS	1.59	2.38	2.42	5.983	5.865	5.818
CdSe	1.33	1.88	1.84	6.245	6.113	6.05
CdTe	1.67	1.71	1.60	6.652	6.486	6.48
MgS (rs)	2.77	3.70	3.7	5.210	5.167	5.20
MgS (zb)	3.37	4.84	4.8	5.659	5.599	5.66
MgSe (zb)	2.95	3.91	4.0	5.949	5.893	5.91
BN	4.76	6.12	6.36	3.601	3.544	3.615
AlN (zb)	3.36	6.09	6.2	4.328	4.260	4.35
GaN (zb)	2.00	3.27	3.39	4.576	4.516	4.54
InN	0.26	0.66	0.7	5.130	4.677	4.98
CdO	–0.60	0.98	0.9	4.708	4.670	4.69
MgO	3.60	7.72	7.8	4.223	4.126	4.21
LiF	9.24	13.27	13.7	4.093	4.032	4.017
SiO_2	6.05	8.74	9	4.909/5.402	4.855/5.371	5.01/5.47
α-Al_2O_3	6.25	8.64	8.8	4.76/13.00	4.70/12.97	4.76/12.99
SnO_2	0.93	3.66	3.6	4.738/3.149	4.692/3.136	4.737/3.186
In_2O_3	0.90	3.03	2.9	10.118	10.016	10.12
Cu_2O	1.04	2.11	2.12	4.359	4.315	4.27
TiO_2	1.86	3.1	3.2	4.691/2.994	4.608/2.920	4.59/2.96
c-HfO_2	3.74	5.60	5.8	5.161	5.037	5.11
c-ZrO_2	3.43	5.76	5.7	5.131	5.022	5.07
La_2O_3	3.74	6.19	5.6	3.951/6.195	3.897/6.084	3.94/6.13
Y_2O_3	4.39	6.01	5.8	10.633	10.536	10.6
$SrTiO_3$	1.93	3.28	3.2	3.971	3.874	3.905
$PbTiO_3$	1.71	3.43	3.4	3.983	3.904	3.96
$LaAlO_3$	3.28	5.14	5.7	3.802	3.746	3.78

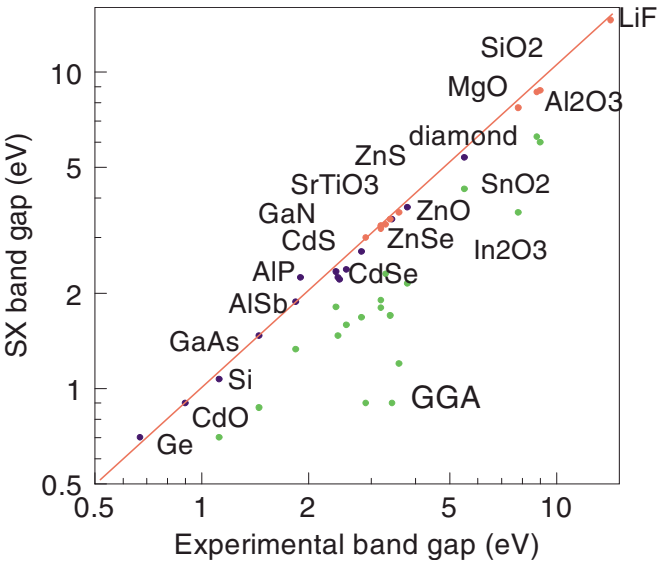

FIG. 3. (Color online) Comparison of the calculated SX and experimental minimum band gaps for various compounds are given. The GGA (PBE) values are also shown.

tains the full symmetry of the system but the q-point sum is increased in expense.

Note that in our implementation, the density of the k-point and q-point sets are the same as the full self-consistent-field (SCF) calculation, whereas in other implementations a lower density is used for these sets, for computational speed.[26,43]

All hybrid functionals (PBEh, HSE, B3LYP, etc.) contain a fraction of exact exchange. Implemented in plane waves, they must all evaluate the same types of integrals as in Eq. (5), which dominate the calculational cost. Once a code has this implemented, it can easily be expanded to evaluate this entire family of functionals.

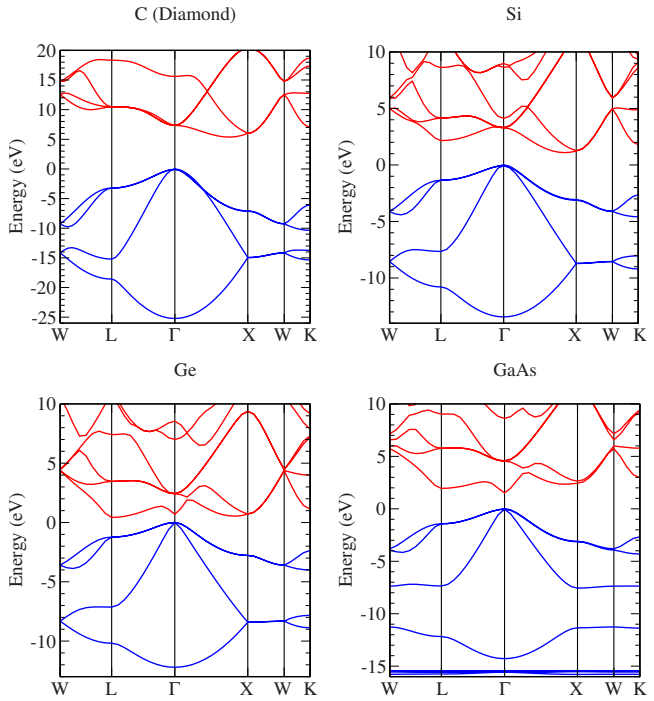

FIG. 4. (Color online) SX band structures of diamond, Si, Ge, and GaAs are shown. The top of valence band is set at 0 eV.

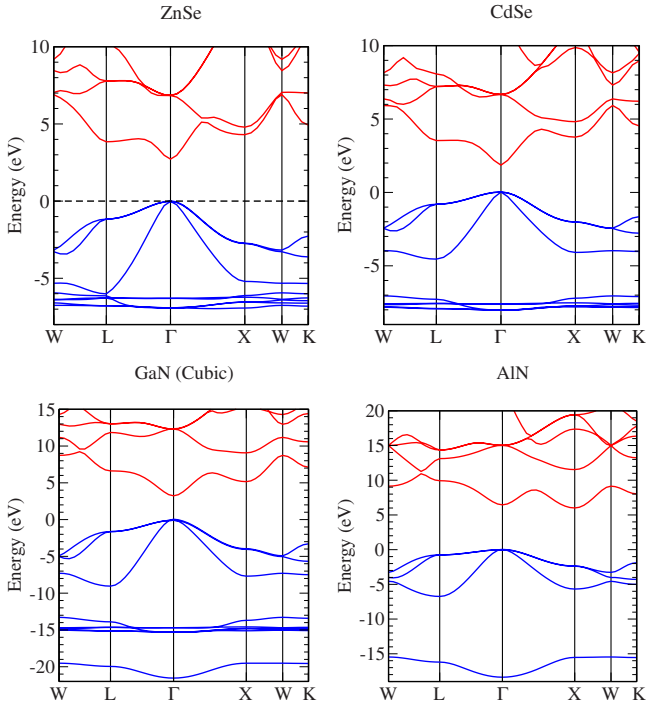

FIG. 5. (Color online) SX band structures of the zinc-blende phases of ZnSe, CdSe, GaN, and AlN are shown.

IV. SCALING

We start with a few points on the convergence of calculations and computational scaling. A major issue with nonlocal functionals is their higher computational cost compared to the semilocal DFT functionals commonly used in the plane-wave formalism. It is worth noting the convergence rate of the nonlocal exchange correlation energy with respect to k-point sampling. The convergence of the energy contribution due to the SX exchange-correlation functional with respect to k points is compared to that of LDA (PBE), PBEh, and HF for Si in Fig. 1(a). Note, this is not the convergence of the total energy but just the nonlocal energy contribution. It shows the relative nonlocal energy of per atom versus the Monkhorst-Pack k-point grid. SX is seen to converge much faster than PBEh due to the short-range nature of the screened exchange term. This result is similar to that in HSE.

The computational performance is investigated as a function of the plane-wave basis set size, N_p. We performed a series of total-energy calculations, using SX, on a two-atom primitive cell of silicon, using a single k point at [0.5,0.5,0.5] to sample the Brillouin zone. (The calculation is not converged with respect to k-point sampling but this is not important when simply evaluating the scaling properties). The speed of the calculation is determined in terms of the average time for one conjugate-gradients line search during the total-energy minimization. The results are shown in Fig. 1(b). We would expect this calculation to scale as $N_p \log(N_p)$ as confirmed by the line fit. The scaling of CPU time and storage

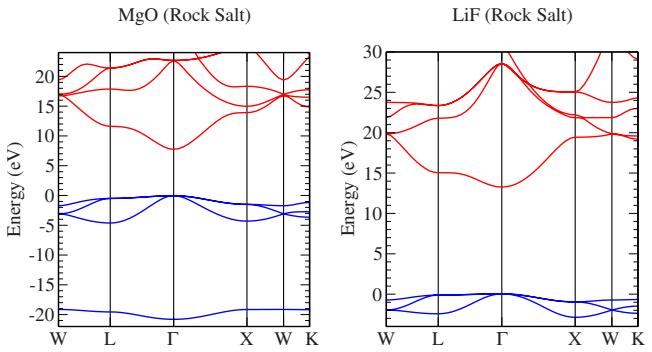

FIG. 6. (Color online) SX band structures of the insulators MgO and LiF are shown.

requirements of plane waves also follow a $O[N_p \log(N_p)]$ dependence. Figure 1(b) shows the LDA results for comparison. For calculations of this size, an LDA calculation also scales roughly as $N_p \log(N_p)$, but the prefactor is an order of magnitude smaller.

The scaling with respect to number of k points N_k for hybrid functionals including SX varies as N_k^2 from Eq. (5), compared to the linear scaling of local functionals. In local functionals, the linear scaling with N_k favors using the smallest possible unit cell with a large number of k points. In contrast, with hybrid functionals, there is poorer scaling with N_k than with N_p. Thus, even for small unit cells, where large k-points sets are necessary to converge the calculation, the calculation can be costly. As discussed above, the use of symmetry complicates the scaling of a calculation with respect to Monkhorst-Pack grid size; the k-point set is symmetry reduced while the q-point set is symmetry expanded.

Considering the number of states (bands), N_b, Eq. (5) shows a quadratic dependence here too. Overall, nonlocal functionals scale as $N_b^2 N_k^2 N_p \log(N_p)$ which leads to a cost typically an order of magnitude or more greater than a semilocal functional using a plane-wave basis set. The cost factor ratio between the hybrid functionals and local functionals is less for a localized basis due to the small number of basis functions in a local basis set and the manner in which the HF integral is evaluated but the basic cost of LDA in a local basis can be larger than for plane waves.

HSE is becoming one of the more widely employed functionals since its implementation in plane-wave codes, for example, as described in Ref. 26. It is interesting to compare the screening range in HSE and SX as in Fig. 2. SX has the correct asymptotic limits of V_{XC} at the free electron gas and atomic limits. We see that the range of the Thomas-Fermi screening in SX is closer to the screening length of the dielectric function $\varepsilon^{-1}(Q)$ of the semiconductor, e.g., GaAs.[44,45] This is expected, as the TF screening is a zeroth-order approximation of the semiconductor's screening but it is interesting that the screening length in HSE is longer ranged in real space. The screening length in HSE reflects the cancellation of the long-range correlation and exchange energies but it is also chosen in terms of computational speed.

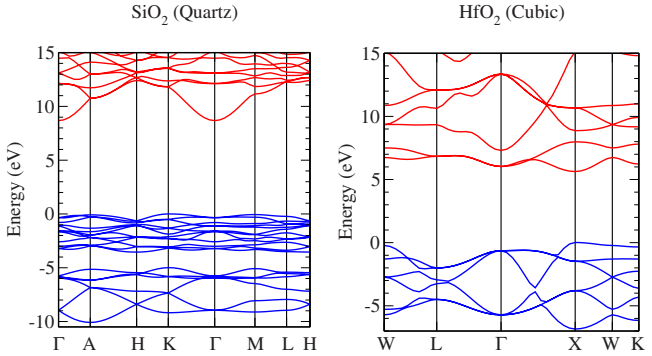

FIG. 7. (Color online) SX band structures of the insulating oxides SiO_2 and HfO_2 are shown.

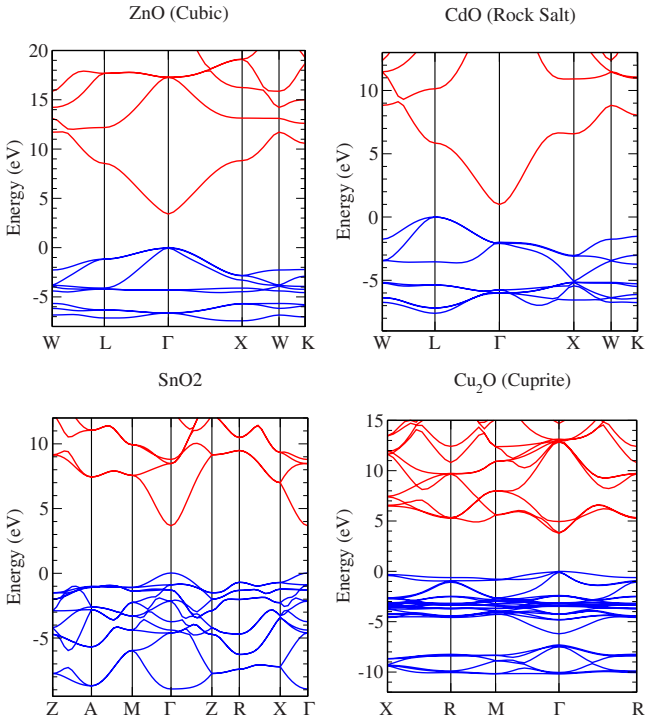

FIG. 8. (Color online) SX band structures for the conducting oxides ZnO, CdO, SnO_2, and Cu_2O are shown.

V. TEST SETS

Table II compares the calculated lattice constants of various semiconductors in GGA-PBE and SX to their experi-

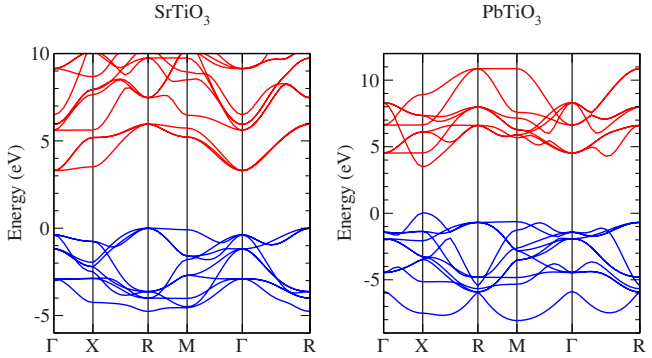

FIG. 9. (Color online) SX band structures of the perovskites (cubic phase) $SrTiO_3$ and $PbTiO_3$ are shown.

TABLE III. Comparison of calculated minimum band gaps (in eV) by different methods are shown. PBE and SX are calculated for this work while B3LYP, HSE03, GW, and experimental results are taken from various Refs. 17 and 48–58.

	PBE	SX	HSE03 (Ref. 24)	HSE03 (Ref. 48)	B3LYP (Ref. 17)	GW (PBE) (Refs. 46–48)	Experiment (Refs. 48–58)
Diamond	4.12	5.38	5.49	5.84	5.8	5.50	5.48
Si	0.62	1.07	1.28	1.32	1.57	1.17	1.17
Ge		0.69	0.59	0.79		0.54	0.74
c-SiC	1.35	2.25	2.39	2.60		2.64	2.40
GaAs	0.49	1.47	1.21	1.66	1.5	1.52	1.52
AlP	1.66	2.21	2.52	2.69		2.77	2.45
GaN	1.62	3.27	3.03	3.29		3.32	3.20
BN	4.45	6.12	5.98	6.54		6.73	6.37
ZnS	2.07	3.74	3.42	3.69	3.5	3.86	3.91
CdS	1.14	2.38	2.14	2.55		2.55	2.42
MgO	4.76	7.72	6.50	7.94	7.3	8.47	7.8
ZnO	0.67	3.41		2.86	3.2	3.2	3.44
LiF	9.2	13.27		14.1		15.1	14.2
SiO_2	6.05	8.74				10.1	9.0
ZrO_2	3.43	5.76				5.3	5.7
HfO_2	3.74	5.60				5.5	5.8
Al_2O_3	6.25	8.64			8.5		8.8

mental values. The mean relative error (MRE) is found to be 1.1% for PBE, compared to 2.3% for SX. This is to be compared with 0.2% for HSE for a different test set used by Paier[26] and Marsman *et al.*[43]

Table II also shows the calculated minimum band-gap energies in PBE and in SX, compared to the experimental values.[46–58] The various band gaps are plotted on a logarithmic scale in Fig. 3. Overall, the MRE for PBE is 38% and only 7.4% for SX while the mean absolute error is 1.4 eV for PBE and 0.13 eV for SX. A selection of band structures calculated self-consistently using the SX functional is plotted in Figs. 4–9.

Figure 4 shows the band structures of diamond, Si, Ge, and GaAs. There is a remarkable improvement in band gaps for these semiconductors. For Si, the gap improves from 0.69 to 1.07 eV, which is now very close to the experimental value of 1.12 eV. This is the typical 30% change. A similar change is found for diamond, from 4.27 eV in PBE to 5.37 eV in SX, compared to 5.54 eV experimentally.[49] For GaAs, a Ga pseudopotential including Ga 3*d* states is used. The band gap increases from 0.87 eV in PBE to 1.47 eV in SX.

Figure 5 shows the band structures of ZnSe, CdSe, GaN, and AlN in the zinc-blende (zb) structure. The ZnSe and CdSe valence bands retain the metal *d* core states, so the band gaps include the *p*-*d* repulsion between the core states and the valence-band top.

The development of the GW method for elements with shallow *d* core states such as CdS or CdSe required considerable efforts. Rohlfing *et al.*[59,60] found that it was necessary to include all the 4*s*, 4*p*, and 4*d* shells as valence states to obtain a good band structure. In the case of SX, we found that it is possible to obtain a reasonable band gap including just the Cd 4*d* semicore states, without including 4*s* and 4*p* states. This is seen in the band structures of CdO, CdS, and CdSe that are all correctly reproduced.

The group III nitrides will be discussed in detail elsewhere. Here, SX also improves the band gap greatly. For zinc-blende AlN, the gap increases from 3.36 eV in PBE to 6.0 eV in SX, compared to 6.2 eV experimentally. For GaN, using a Ga pseudopotential including Ga 3*d* states (but not included in the screening length), the band gap increases from 2.0 eV in PBE to 3.27 eV in SX, close to the experimental value of 3.37 eV.[50]

For the insulators, in SiO_2 the gap improves from 6.0 eV in PBE to 8.74 eV in sX, very close to the 9.0 eV experimental value. For face-centered cubic HfO_2 (c-HfO_2) (fluorite structure) the band gap changes from 3.5 eV in PBE to 5.8 eV in sX, which is the experimental value of 5.8 eV.[55]

The case of MgO was interesting. The Mg pseudopotential is usually constructed to fit the properties of metallic Mg. A typical Mg pseudopotential including only 3*s* states does not reproduce all the properties of MgO. We fitted a new norm-conserving pseudopotential including Mg 2*s*, Mg 2*p*, and Mg 3*s* as valence states. This produces greater transferability, and now gives an accurate 7.8 eV band gap for MgO, and analogous to the materials with *d* electrons, the 2*s* and 2*p* states are not included in the screening length. The SX method is also successful for insulators with very wide band gaps such as LiF. Its gap is 13.27 eV in SX, 9.24 eV in PBE (and 8.7 eV in LDA), compared to 13.7 eV experimentally.

The most startling improvements in the band structures are found for the transparent conducting oxides, such as ZnO, CdO, In_2O_3, and SnO_2, as shown in Fig. 6. These band structures are characterized by a single, broad conduction-

TABLE IV. A comparison of valence bandwidths (in eV) calculated using the PBE, SX (calculated here), and GW approximations (Refs. 46–48 and 59–63) are given. Experimental values (Refs. 49–58) are also given for comparison.

	PBE	SX	GW	Experiment
Diamond	22	25.2	23	21
Si	12.5	13.45	12.3	12.5
Ge	12.6	12.4	12.8	12.6
GaAs	12.4	14.2		12.6
MgO	4.5	4.53	4.8	
SiO_2	9.1	9.0	11	9
Al_2O_3	6.85	7.26		7
ZnO	4.2	3.9		4
$SrTiO_3$	4.52	4.68		4.3
ZrO_2	5.95	5.8	6.5	6.5

band minimum at Γ derived from the metal *s* states. The minimum gap of SnO_2 is 0.9 eV in PBE, and this becomes 3.66 eV in sX, compared to 3.6 eV found experimentally. Similarly, for ZnO, the minimum gap is about 0.9 eV in PBE, and this improves to 3.43 eV in sX,[36] compared to 3.44 eV experimentally. The reason for this is that the direct gap at Γ is unrepresentative of the zone-averaged gap, due to the deep conduction-band minimum at Γ. The average gap opens up by the typical 20% but this translates into a very large fractional change at Γ. SnO_2 in the rutile structure has long been known to have a direct, forbidden gap.[56]

Recently, In_2O_3 has also been found to have an indirect forbidden gap so that its gap of about 2.9 eV is less than that previously quoted value of 3.5 eV.[57] Cu_2O is an interesting *p*-type conducting oxide with a direct, forbidden band gap, whose gap SX also reproduces well.

It is interesting that the complex band structure of rocksalt (rs) CdO is reproduced. This conducting oxide has an indirect gap with a conduction-band minimum at Γ and a valence-band maximum at L. In PBE, CdO has a negative band gap of −0.6 eV. In SX, the minimum indirect gap is calculated to be 0.98 eV compared to 0.9 eV experimentally.

Cu_2O is a *p*-type semiconducting oxide. Its band gap is direct forbidden with a value of 2.12 eV. The calculated band gap in SX is now 2.11 eV, very close to the experimental value, whereas the PBE value was only 1.04 eV.

We also compare the band gaps from screened exchange to those calculated by hybrid functional methods (HSE03) and by GW (Refs. 46–48 and 59–63) in Table III. There are two sets of band gaps for HSE03, those of Heyd *et al.*,[24] and those found by Kresse and Bechstedt *et al.*[46–48] We see that the band gaps calculated by SX are generally closer to experiment than those found by HSE03, for the tests of Kresse and Bechstedt.[46–48] The GW results based on PBE is also included for comparison.

B3LYP has also been found to reproduce band gaps better than GGA (Refs. 17 and 64) and the localization of wave functions in critical cases. Table III compares our SX band gaps to those of B3LYP for a different set of solids. Generally, SX comes out closer to experiment.

TABLE V. Bulk properties of ZnO in the wurtzite structure are given for both PBE and SX theory and compared to experimental results (Refs. 52 and 65).

	PBE	SX	Experiment
a (Å)	3.286	3.267	3.2495
c (Å)	5.299	5.245	5.2069
Free energy (eV)	−2.82	−3.31	−3.63
Direct gap (eV)	0.9	3.41	3.44
Zn 3*d* (eV)	−4.8	−7.0	−7.3

Table IV compares the calculated valence bandwidths for some crystalline solids, compared to PBE, to GW and to experiment. SX is seen to give a slightly wider valence band than experiment for the simple semiconductors such as Si or diamond. In this respect, this is somewhat like GW.

A further factor of improvement in SX is the placement of the shallow core states, such as the Zn 3*d* states in ZnO. These are always placed too high in energy in GGA. SX along with other methods HSE and GW places them deeper below the valence-band maximum, in closer agreement with experiment, as in ZnO,[65] shown in Table V.

SX has also been used on some correlated oxides. We previously calculated that the band gap of the multiferroic $BiFeO_3$ with its correlated Fe 3*d* bands was 2.7 eV,[35] compared to the experimental value of 2.6–2.8 eV. Results for the correlated oxides such as NiO, etc., will be reported later.

VI. CONCLUSIONS

We have described the computational implementation and application of the screened exchange method within the density functional, plane-wave pseudopotential formalism. We have use the method to evaluate the lattice parameters and electronic band structures of a wide range of different semiconducting an insulating materials. In general, the lattice parameters of the materials are described to a similar accuracy as in the local functional methods. However, in all cases considered here, the screened exchange method has corrected the band gap underestimation found in the local-density methods, and gives band gaps very close to those found experimentally. This is particularly true for the transparent conduction oxides such as ZnO and SnO_2 where the error was of order 70%. The screened exchange method is also compared to other hybrid density functional methods such as HSE and B3LYP in terms of implementation and its results. Screened exchange (and hybrid functionals) are more efficient than full GW methods and are very suitable for supercell calculations of point defects, surface, and interfaces including self-consistent geometry relaxation.

[1] L. J. Sham and M. Schluter, Phys. Rev. Lett. **51**, 1888 (1983); J. P. Perdew and M. Levy, *ibid.* **51**, 1884 (1983).
[2] G. Pacchioni, F. Frigoli, D. Ricci, and J. A. Weil, Phys. Rev. B **63**, 054102 (2000).
[3] J. P. Perdew and A. Zunger, Phys. Rev. B **23**, 5048 (1981).
[4] A. Filippetti and N. A. Spaldin, Phys. Rev. B **67**, 125109 (2003).
[5] L Hedin, Phys. Rev. **139**, A796 (1965).
[6] M. S. Hybertsen and S. G. Louie, Phys. Rev. Lett. **55**, 1418 (1985).
[7] M. S. Hybertsen and S. G. Louie, Phys. Rev. B **34**, 5390 (1986).
[8] F. Aryasetiawan and O. Gunnarsson, Rep. Prog. Phys. **61**, 237 (1998).
[9] W. G. Aulbur, L. Jonsson, and J. Wilkins, in *Solid State Physics*, edited by H. Ehrenreich and F. Spaepen (Academic Press, NY, 2000), Vol. 54, p. 1.
[10] M. van Schilfgaarde, T. Kotani, and S. Faleev, Phys. Rev. Lett. **96**, 226402 (2006).
[11] V. I. Anisimov, J. Zaanen, and O. K. Andersen, Phys. Rev. B **44**, 943 (1991).
[12] V. I. Anisimov, F. Aryasetiawan, and A. I. Lichtenstein, J. Phys.: Condens. Matter **9**, 767 (1997).
[13] J. E. Perdew, M. Erzenhof, and A. D. Becke, J. Chem. Phys. **105**, 9982 (1996).
[14] M. Ernzerhof and G. E. Scuseria, J. Chem. Phys. **110**, 5029 (1999).
[15] C. Adamo and V. Barone, J. Chem. Phys. **110**, 6158 (1999).
[16] A. D. Becke, J. Chem. Phys. **98**, 5648 (1993).
[17] J. Muscat, A. Wander, and N. M. Harrison, Chem. Phys. Lett. **342**, 397 (2001).
[18] J. Paier, M. Marsman, and G. Kresse, J. Chem. Phys. **127**, 024103 (2007).
[19] A. D. Becke, J. Chem. Phys. **98**, 1372 (1993).
[20] K. Gkionis, J. G. Hill, S. P. Oldfield, and J. A. Platts, J. Mol. Model. **15**, 1051 (2009).
[21] J. Heyd, G. E. Scuseria, and M. Ernzerhof, J. Chem. Phys. **118**, 8207 (2003).
[22] J. Heyd and G. E. Scuseria, J. Chem. Phys. **120**, 7274 (2004).
[23] J. Heyd and G. E. Scuseria, J. Chem. Phys. **121**, 1187 (2004).
[24] J. Heyd, J. E. Peralta, E. Scuseria, and R. L. Martin, J. Chem. Phys. **123**, 174101 (2005).
[25] J. E. Peralta, J. Heyd, G. E. Scuseria, and R. L. Martin, Phys. Rev. B **74**, 073101 (2006).
[26] J. Paier, M. Marsman, K. Hummer, G. Kresse, I. C. Gerber, and J. G. Angyan, J. Chem. Phys. **124**, 154709 (2006).
[27] D. M. Bylander and L. Kleinman, Phys. Rev. B **41**, 7868 (1990).
[28] J. C. Phillips and L. Kleinman, Phys. Rev. **128**, 2098 (1962).
[29] A. Seidl, A. Gorling, P. Vogl, J. A. Majewski, and M. Levy, Phys. Rev. B **53**, 3764 (1996).
[30] C. B. Geller, W. Wolf, S. Picozzi, A. Continenza, R. Asahi, W. Mannstadt, A. J. Freeman, and E. Wimmer, Appl. Phys. Lett. **79**, 368 (2001); R. Asahi, A. Wang, J. R. Babcock, N. L. Edelman, A. W. Metz, M. A. Lane, V. P. Dravid, C. R. Kannewurf, A. J. Freeman, and T. J. Marks, Thin Solid Films **411**, 101 (2002).
[31] A. J. Freeman, J. Comput. Appl. Math. **149**, 27 (2002).
[32] M. C. Gibson, S. Brand, and S. J. Clark, Phys. Rev. B **73**, 125120 (2006).
[33] K. Xiong, J. Robertson, M. C. Gibson, and S. J. Clark, Appl. Phys. Lett. **87**, 183505 (2005).
[34] J. Robertson, K. Xiong, and S. J. Clark, Phys. Status Solidi B **243**, 2054 (2006).
[35] S. J. Clark and J. Robertson, Appl. Phys. Lett. **90**, 132903 (2007).
[36] S. J. Clark, J. Robertson, S. Lany, and A. Zunger, Phys. Rev. B **81**, 115311 (2010).
[37] S. Kümmel and L. Kronik, Rev. Mod. Phys. **80**, 3 (2008).
[38] M. D. Segall, P. J. D. Lindan, M. J. Probert, C. J. Pickard, P. J. Hasnip, S. J. Clark, and M. C. Payne, J. Phys.: Condens. Matter **14**, 2717 (2002).
[39] S. J. Clark, M. D. Segall, C. J. Pickard, P. J. Hasnip, M. J. Probert, K. Refson, and M. C. Payne, Z. Kristallogr. **220**, 567 (2005).
[40] A. M. Rappe, K. M. Rabe, E. Kaxiras, and J. D. Joannopoulos, Phys. Rev. B **41**, 1227 (1990).
[41] R. Gómez-Abal, X. Li, M. Scheffler, and C. Ambrosch-Draxl, Phys. Rev. Lett. **101**, 106404 (2008).
[42] F. Gygi and A. Baldereschi, Phys. Rev. B **34**, 4405 (1986).
[43] M. Marsman, J. Paier, A. Stropa, and G. Kresse, J. Phys.: Condens. Matter **20**, 064201 (2008).
[44] J. P. Walter and M. L. Cohen, Phys. Rev. B **2**, 1821 (1970).
[45] Z. H. Levine and S. G. Louie, Phys. Rev. B **25**, 6310 (1982).
[46] M. Shishkin and G. Kresse, Phys. Rev. B **75**, 235102 (2007).
[47] F. Fuchs, J. Furthmuller, F. Bechstedt, M. Shishkin, and G. Kresse, Phys. Rev. B **76**, 115109 (2007).
[48] F. Bechstedt, F. Fuchs, and G. Kresse, Phys. Status Solidi B **246**, 1877 (2009).
[49] F. J. Himpsel, J. F. van der Veen, and D. E. Eastman, Phys. Rev. B **22**, 1967 (1980).
[50] S. Strite and H. Morkoc, J. Vac. Sci. Technol. B **10**, 1237 (1992).
[51] V. Y. Davydov, A. A. Klochikhin, R. P. Seisyan, V. V. Emtsev, S. V. Ivanov, F. Bechstedt, J. Furthmüller, H. Harima, A. V. Mudryi, J. Aderhold, O. Semchinova, and J. Graul, Phys. Status Solidi B **229**, R1 (2002); J. Wu, W. Walukiewicz, K. M. Yu, J. W. Ager, E. E. Haller, H. Lu, W. J. Schaff, Y. Saito, and Y. Nanishi, Appl. Phys. Lett. **80**, 3967 (2002).
[52] U. Özgür, Y. I. Alivov, C. Liu, A. Teke, M. A. Reshchikov, S. Doğan, V. Avrutin, S. J. Cho, and H. Morkoç, J. Appl. Phys. **98**, 041301 (2005).
[53] D. A. Evans, A. G. McGlynn, B. M. Towlson, M. Gunn, D. Jones, T. E. Jenkins, R. Winter, and N. R. J. Poolton, J. Phys.: Condens. Matter **20**, 075233 (2008).
[54] R. H. French, J. Am. Ceram. Soc. **73**, 477 (1990).
[55] N. V. Nguyen, A. V. Davydov, D. Chandler-Horowitz, and M. M. Frank, Appl. Phys. Lett. **87**, 192903 (2005).
[56] J. Robertson, J. Phys. C **12**, 4767 (1979).
[57] A. Walsh, J. L. F. DaSilva, S. H. Wei, C. Korber, A. Klein, L. F. J. Piper, A. DeMasi, K. E. Smith, G. Panaccione, P. Torelli, D. J. Payne, A. Bourlange, and R. G. Egdell, Phys. Rev. Lett. **100**, 167402 (2008).
[58] N. B. Brookes, D. S. L. Law, T. S. Padmore, D. R. Warburton, and G. Thornton, Solid State Commun. **57**, 473 (1986).
[59] M. Rohlfing, P. Kruger, and J. Pollmann, Phys. Rev. Lett. **75**, 3489 (1995).
[60] M. Rohlfing, P. Kruger, and J. Pollmann, Phys. Rev. B **57**, 6485 (1998).
[61] B. Králik, E. K. Chang, and S. G. Louie, Phys. Rev. B **57**, 7027 (1998).
[62] J. Dabrowski, V. Zavodinsky, and A. Fleszar, Microelectron.

Reliab. **41**, 1093 (2001).
[63] E. K. Chang, M. Rohlfing, and S. G. Louie, Phys. Rev. Lett. **85**, 2613 (2000).
[64] S. Tomic, B. Montanari, and N. M. Harrison, Physica E **40**, 2125 (2008).
[65] R. T. Girard, O. Tjernberg, G. Chiaia, S. Soderholm, U. O. Jarlsson, C. Wigren, H. Nylen, and I. Lindau, Surf. Sci. **373**, 409 (1997).