Phys. Status Solidi B **248**, No. 1, 19–27 (2011) / **DOI** 10.1002/pssb.201046290

Advances in electronic structure methods for defects and impurities in solids

Feature Article

Chris G. Van de Walle* **and Anderson Janotti**

Materials Department, University of California, Santa Barbara, CA 93106-5050, USA

Received 2 June 2010, revised 7 September 2010, accepted 21 October 2010
Published online 18 November 2010

Keywords calculation methods, density functional theory, electronic structure, impurities, point defects

* Corresponding author: e-mail vandewalle@mrl.ucsb.edu, Phone: +1-805-893-7144

Defects and impurities are often decisive in determining the physical properties of most materials. The process of defect identification and characterization is typically difficult and indirect, usually requiring an ingenious combination of different experimental techniques. First-principles calculations have emerged as a powerful microscopic tool that complements experiments or sometimes even serves as the sole source of atomistic information due to experimental limitations. Still, first-principles calculations based on density functional theory in the local density or generalized gradient approximations suffer from serious limitations when describing defects in solids. Recent advances in electronic structure methods, rapid increases in computing power, and the development of efficient algorithms indicate a promising future for computational defect physics. We review recent advances in the theory of defects in solids from the perspective of first-principles calculations. We focus in particular on methods that improve the description of band gaps, leading to results that can be directly compared to experiments on a quantitative level. We discuss the use of LDA+U in wide-band-gap materials, screened hybrid functionals, the quasiparticle GW method, and the use of modified pseudopotentials. Advantages and limitations of these methods are illustrated with examples.

1 Introduction First-principles studies of point defects and impurities in semiconductors, insulators, and metals have become an integral part of materials research over the last few decades [1–3]. Point defects and impurities often have decisive effects on materials properties. A prime example is doping of semiconductors: the addition of minute amounts (often at the ppm level) of donor or acceptor impurities renders the material n type or p type, enabling the functionality of electronic or optoelectronic devices [4, 5]. Control of doping is therefore essential, and all too often eludes experimental efforts. Sometimes high doping levels required for low-resistivity transport are limited by compensation effects; such compensation can be due to point defects that form spontaneously at high doping. In other cases, unintentional doping occurs. For instance, many oxides exhibit unintentional n-type doping, which due to its prevalence has often been attributed to intrinsic causes, i.e., to native point defects. Recent evidence indicates, however, that the concentration of native point defects may be lower than has conventionally been assumed, and that, instead, unintentional incorporation of impurities may cause the observed conductivity [6]. Last but not least, many materials resist attempts at ambipolar doping, i.e., they can be easily doped one type but not the other. Again, the oxides (or more generally, wide-band-gap semiconductors) that exhibit unintentional n-type doping often cannot be doped p-type. The question then is whether this is due to an intrinsic limitation that cannot be avoided, or whether specific doping techniques might be successful.

Aside from the issue of doping, the study of point defects is important because they are involved in the diffusion processes and act to mediate mass transport, hence contributing to equilibration during growth, and to diffusion of dopants or other impurities during growth or annealing [7–9]. In addition, an understanding of point defects is essential for characterizing or suppressing radiation damage, and for analyzing device degradation.

Experimental characterization techniques are available, but they are often limited in their application [10–12]. Impurity concentrations can be determined using secondary

ion mass spectrometry (SIMS), but some impurities (such as hydrogen) are hard to detect in low concentrations. Point-defect concentrations are even harder to determine. Electron paramagnetic resonance is an excellent tool that can provide detailed information about concentrations, chemical identity, and lattice environment of a defect or impurity, but it is a technique that requires dedicated expertise and possibly for that reason has few practitioners [12]. Other tools, such as Hall measurements or photoluminescence, can provide information about the effect of point defects or impurities on electrical or optical properties, but cannot by themselves identify their nature or character. For all these reasons, the availability of first-principles calculations that can accurately address atomic and electronic structure of defects and impurities has had a great impact on the field.

Obviously, to make the information obtained from such calculations truly useful, the results should be as reliable and accurate as possible. Density functional theory (DFT) [13, 14] has proven its value as an immensely powerful technique for assessing the structural properties of defects [1]. (In the remainder of this article, we will use the term "defects" to generically cover both native point defects and impurities.) Minimization of the total energy as a function of atomic positions yields the stable structure, including all relaxations of the host atoms, and most functionals [including the still most widely used local density approximation (LDA)] all yield results within reasonable error bars [15]. Quite frequently, however, information about electronic structure is required, i.e., the position of defect levels that are introduced in the band gap of semiconductors or insulators. Since DFT in the LDA or generalized gradient approximation (GGA) severely underestimates the gap, the position of defect levels is subject to large error bars and cannot be directly compared with experiment [16–18]. In turn, this affects the calculated formation energy of the defect, which determines its concentration. This effect on the energy is still not generally appreciated, since it is often assumed that the formation energy is a ground-state property for which DFT should give reliable results. However, in the presence of gap levels that can be filled with varying numbers of electrons (corresponding to the charge state of the defect), the formation energy becomes subject to the same type of errors that would occur when trying to assess excitation energies based on total energy calculations with N or $N+1$ electrons. Recently, major progress has been made in overcoming these inaccuracies, and the approaches for doing so will be discussed in Section 2.

Another type of error that may occur in defect calculations is related to the geometry in which the calculations are performed. Typically, one wishes to address the dilute limit in which the defect concentration is low and defect–defect interactions are negligible. Green's functions calculations would in principle be ideal, but in practice have proven quite cumbersome and difficult to implement. Another approach would be to use clusters, but surface effects are almost impossible to avoid, and quantum confinement effects may obscure electronic structure. Nowadays, point defect calculations are almost universally performed using the supercell geometry, in which the defect is embedded within a certain volume of material which is periodically repeated. This has the advantage of maintaining overall periodicity, which is particularly advantageous when using plane-wave basis sets which rely on Fast Fourier Transforms to efficiently move between reciprocal- and real-space representations. The supercells should be large enough to minimize interactions between defects in neighboring supercells. This is relatively straightforward to accomplish for *neutral* defects, but due to the long-range nature of the Coulomb interaction, interactions between charged defects are almost impossible to eliminate. This problem was recognized some time ago, and a correction was suggested based on a Madelung-type interaction energy [19]. It had been observed, however, that in many cases the correction was unreliable or "overcorrected," making the result less accurate than the bare values [20]. Recently, an approach based on a rigorous treatment of the electrostatic problem has been developed that outlines the conditions of validity of certain approximations and provides explicit expression for the quantities to be evaluated [21]. Issues relating to supercell-size convergence are addressed in detail in the article by Freysoldt et al. [22] in this volume.

We note that it is not the intent of the present paper to provide a comprehensive review of the entirety of this large and growing field. Rather, we attempt to introduce the main concepts of present-day defect calculations illustrated with a few select examples, and do not aspire to cover the countless important contributions to the field by many different research groups.

2 Formalism and computational approach

The key quantities that characterize a defect in a semiconductor are its concentration and the position of the transition levels (or ionization energies) with respect to the band edges of the host material. Defects that occur in low concentrations will have a negligible impact on the properties of the material. Only those defects whose concentration exceeds a certain threshold will have observable effects. The position of the defect transition levels with respect to the host band edges determines the effects on the electrical and optical properties of the host. Defect formation energies and transition levels can be determined entirely from first principles [1], without resorting to any experimental data for the system under consideration.

2.1 Defect formation energies and concentrations

In the dilute limit, the concentration of a defect is determined by the formation energy E^{f} through a Boltzmann expression:

$$c = N_{\mathrm{sites}} \exp\left(-E^{\mathrm{f}}/k_{\mathrm{B}}T\right). \tag{1}$$

N_{sites} is the number of sites (including the symmetry-equivalent local configurations) on which the defect can be incorporated, k_{B} is the Boltzmann constant, and T the

temperature. Note that this expression assumes thermodynamic equilibrium. While defects could also occur in nonequilibrium concentrations, in practice most of the existing bulk and epitaxial film growth techniques operate close to equilibrium conditions. Equilibration of defects is actually unavoidable if the diffusion barriers are low enough to allow easy diffusion at the temperatures of interest. In addition, even if kinetic barriers would be present, Eq. (1) is still relevant because obviously defects with a high formation energy are less likely to form.

Defect formation energies can be written as differences in total energies, and these can be obtained from first principles, i.e., without resorting to experimental parameters. The dependence on the chemical potentials (atomic reservoirs) and on the position of the Fermi level in the case of charged defects is explicitly taken into account [1, 5]. This is illustrated here with the specific example of an oxygen vacancy in a 2+ charge state in ZnO. The formation energy of V_O^{2+} is given by:

$$E^f\left(V_O^{2+}\right) = E_{tot}\left(V_O^{2+}\right) - E_{tot}(ZnO) + \mu_O + 2E_F, \qquad (2)$$

where $E_{tot}(V_O^q)$ is the total energy of the supercell containing the defect, and $E_{tot}(ZnO)$ is the total energy of the ZnO perfect crystal in the same supercell. E_F is the energy of the reservoir with which electrons are exchanged, i.e., the Fermi level. The O atom that is removed is placed in a reservoir, the energy of which is given by the oxygen chemical potential μ_O. Note that μ_O is a variable, corresponding to the notion that ZnO can in principle be grown or annealed under O-rich, O-poor, or any other condition in between. It is subject to an upper bound given by the energy of an O atom in an O_2 molecule. Similarly, the zinc chemical potential μ_{Zn} is subject to an upper bound given by the energy of a Zn atom in bulk Zn. The sum of μ_O and μ_{Zn} corresponds to the energy of ZnO, which is the stability condition of ZnO. An upper bound on μ_{Zn}, given by the energy of bulk Zn, therefore leads to a lower bound on μ_O, and *vice versa*. The chemical potentials thus vary over a range given by the formation enthalpy of the material being considered. Formation enthalpies are generally well described by first-principles calculations. For instance, the calculated formation enthalpy of -3.50 eV for ZnO [8] is in a good agreement with the experimental value of -3.60 eV [23].

Note that it is, in principle, the *free energy* that determines the defect concentration, and one should in principle take into account vibrational entropy contributions in Eq. (1). Such contributions are usually small, on the order of a few k_B, and there is often a significant cancellation between vibrational contributions in the solid and in the reservoir [1]. In rare instances, inclusion of vibration entropy has a distinct impact on which configuration is most stable for a given defect or impurity [24], but it hardly ever has a significant effect on the overall concentration. The reader is referred to Ref. [1] for a detailed discussion on the calculation of defect formation energies from first principles.

2.2 Transition levels or ionization energies Defects in semiconductors and insulators can occur in different charge states. For each position of the Fermi level, one particular charge state has the lowest energy for a given defect. The Fermi-level positions at which the lowest-energy charge state changes are called transition levels or ionization energies. The transition levels are thus determined by formation energy differences:

$$\varepsilon(q/q') = \frac{E^f(D^q; E_F = 0) - E^f\left(D^{q'}; E_F = 0\right)}{(q' - q)}, \qquad (3)$$

where $E^f(D^q; E_F = 0)$ is the formation energy of the defect D in the charge state q for the Fermi level at the valence-band maximum ($E_F = 0$). These are thermodynamic transition levels, i.e., atomic relaxations around the defect are fully included; for Fermi-level positions below $\varepsilon(q/q')$ the defect is stable in charge state q, while for Fermi-level positions above $\varepsilon(q/q')$, the defect is stable charge state q'. The thermodynamic transition levels are not to be confused with the single-particle Kohn–Sham states that result from band-structure calculations for a single charge state. They are also not to be confused with optical transition levels derived, for example, from luminescence or absorption experiments. In this case, the final state may not be completely relaxed, and the optical transition levels may significantly differ from the thermodynamic transition levels, as discussed in Ref. [1].

For a defect to contribute to conductivity, it must be stable in a charge state that is consistent with the presence of free carriers. For instance, in order to contribute to n-type conductivity, the defect must be stable in a positive charge state and the transition level from the positive to the neutral charge state should occur close to or above the conduction-band minimum (CBM). A defect is a typical shallow donor when the transition level for a positive to the neutral charge state [e.g., the $\varepsilon(+/0)$ level], as defined based on formation energies, lies above the CBM. In this case, a neutral charge state in which the electron is localized in the immediate vicinity of the defect cannot be maintained if the corresponding electronic level is resonant with the conduction band; instead, the electron will be transferred to extended states, but may still be bound to the positive core of the defect in a hydrogenic effective-mass state. Similarly, shallow acceptors are defects in which the transition level from a negative to the neutral charge state [e.g., the $\varepsilon(-/0)$ level] is near or below the VBM. If the latter, the hole can be bound to the negative core of the defect in a hydrogenic effective-mass state [1, 25].

2.3 Practical aspects The total energies in Eq. (2) are often evaluated by performing DFT calculations within the LDA or its semi-local extension, the GGA [26, 27]. Defects are typically calculated by using a supercell geometry, in which the defect is placed in a cell that is a multiple of the primitive cell of the crystal. The supercell is then periodically repeated in three-dimensional space. The use

of supercells also has the advantage that the underlying band structure of the host remains properly described, and integrations over the Brillouin zone are replaced by summations over a discrete and relatively small set of special k-points. Supercell-size corrections for charged defects are addressed in Refs. [21] and [22]. Convergence with respect to the supercell size, number of plane waves in the basis set, and the number of special k-points should always be checked, to make sure that the quantities that are derived are representative of the isolated defect.

The number of atoms or electrons in the calculations is limited by the available computer power. For typical defect calculations, supercells containing 32, 64, 128, 216, and 256 atoms are used for materials with the zinc-blende structure, whereas supercells containg 32, 48, 72, and 96 atom cells are used for materials in the wurtzite structure. These fairly large cell sizes call for efficient computational approaches. Ultrasoft pseudopotential [28–30] and projector-augmented-wave [31] methods to separate the chemically active valence electrons from the inert core electrons have proven ideal for tackling such large systems. First-principles methods based on plane-wave basis sets have been implemented in many codes such as the Vienna *Ab initio* Simulation Program (VASP) [32–34], ABINIT [35, 36], and Quantum Expresso [37].

3 The DFT-LDA/GGA band-gap problem and possible approaches to overcome it

The LDA and the GGA in the DFT are plagued by the problem of large band-gap errors in semiconductors and insulators, resulting in values that are typically less than 50% of the experimental values [38–42]. It has often been assumed that the band-gap problem is not an issue when studying defects in semiconductors, since each individual calculation for a specific charge state of the defect could be considered to be a ground-state calculation. However, this notion is not correct, in the same way that the assumption that LDA calculations could yield reliable total-energy differences between N-electron versus $(N+1)$-electron systems is not correct [16]. Indeed, the change in the number of electrons elicits the issue of the lack of a discontinuity in the exchange-correlation potential, which is at the root of the band-gap problem [38–42]. Similarly, the formation energy expressed in Eq. (2) involves changes in the occupation of defect-induced states. In other words, if a specific charge state of a defect involves occupying a state in the band gap, and the band gap is incorrect in DFT–LDA/GGA, then the position of the defect state and hence the calculated total energy will suffer from the same problem [8, 16]. Careful practitioners have always been aware of this problem and refrained from drawing conclusions that might be affected by these uncertainties. The problem is exacerbated, of course, in the case of wide-band-gap semiconductors in which the band-gap errors can be particularly severe; for example, in ZnO the LDA band gap is only 0.8 eV, compared to an experimental value of 3.4 eV.

In the remainder of this section we address several approaches that have been, or are being, developed to overcome these problems.

3.1 LDA+U for materials with semicore states

Many of the wide-band-gap materials of interest have narrow bands, derived from semicore states, that play an important role in their electronic structure [43]. For example, in ZnO narrow bands derived from the Zn 3d states occur at $\sim$8 eV below the valence-band maximum (VBM) and strongly interact with the top of the valence band derived from O 2p states. Inclusion of the Zn d states as valence states (as opposed to treating them as core states) is therefore important for a proper description of the electronic structure of ZnO, as it affects structural parameters, band offsets, and deformation potentials [44, 45]. The DFT–LDA/GGA does not properly describe the energetic position of these narrow bands due to their higher degree of localization, as compared to the more delocalized s and p bands. One way to overcome this problem is to use an orbital-dependent potential that adds an extra Coulomb interaction U for these semicore states, as in the LDA+U (or GGA+U) approach [46, 47].

In the LDA+U the electrons are separated into localized electrons for which the Coulomb repulsion U is taken into account via a Hubbard-like term in a model Hamiltonian, and delocalized or itinerant electrons that are assumed to be well described by the usual orbital-independent one-electron potential in the LDA. Although this approach had been developed and applied for materials with partially filled d bands [46, 47], it has been recently demonstrated that it significantly improves the description of the electronic structure of materials with completely filled d bands such as GaN and InN, as well as ZnO and CdO [44, 45].

An important issue in the LDA+U approach is the choice of the parameter U. It has often been treated as a fitting parameter, with the goal of reproducing either the experimental band gap or the experimentally observed position of the d states in the band structure. Neither approach can be justified, because (a) LDA+U cannot be expected to correct for other shortcomings of DFT-LDA, specifically, the lack of a discontinuity in the exchange-correlation potential, and (b) experimental observations of semicore states may include additional (“final state”) effects inherent in experiments such as photoemission spectroscopy. An approximate but consistent and unbiased approach has been proposed in which the calculated U for the isolated atom is divided (screened) by the optical dielectric constant of the solid under consideration [44]. Tests on a number of systems have shown that applying LDA+U effectively lowers the energy of the narrow d bands, thus reducing their coupling with the p states at the VBM; simultaneously, it increases the energy of the s states that compose the CBM, due to the improved screening by the more strongly bound d states, leading to further opening of the band gap. Such improvements have been described in detail in the case of ZnO, CdO, GaN, and InN [44, 45].

One can take advantage of the partial correction of the band gap by the LDA+U to study defects. Based on an extrapolation of LDA and LDA+U results, one can obtain transition levels and formation energies that can be directly compared with experiments. Such extrapolation schemes have been applied in other contexts as well; they are based on evaluation of defect properties for two different values of the band gap followed by a linear extrapolation to the experimental gap. A number of empirical extrapolation approaches were described by Zhang et al. [48], for instance based on use of different exchange and correlation potentials or different plane-wave cutoffs. Such extrapolation schemes are most likely to be successful if the calculations that produce different band gaps are physically motivated, ensuring that the shifts in defect states that give rise to changes in formation energies reflect the underlying physics of the system.

An extrapolation based on LDA and LDA+U calculations, as described in Refs. [8] and [17], has been shown to be particularly suitable for describing defect physics in materials with semicore d states. The LDA+U produces genuine improvements in the electronic structure related to the energetics of the semicore states; one of these effects is an increase in the band gap. The shifts in defect-induced states between LDA and LDA+U reflect their relative valence- and conduction-band character, and hence an extrapolation to the experimental gap is expected to produce reliable results. Such an approach has led to accurate predictions for point defects in ZnO, InN, and SnO_2 [8, 49, 50]. Figure 1(a) shows the result of this extrapolation scheme for the case of oxygen vacancy in ZnO. The success of this approach can be attributed to the fact that the defect states can in principle be described as a linear combination of host states, under the assumption that the latter form a complete basis. A defect state in the gap region will have contributions from both valence-band states and conduction-band states. The shift in transition levels with respect to the host band edges upon band-gap correction reflects the valence- versus conduction-band character of the defect-induced single-particle states. In the case of a shallow donor, the related transition level is expected to shift with the conduction band, i.e., the variation of the transition level is almost equal the band gap correction. For a shallow acceptor, the position of the transition level with respect to the valence band is expected to remain unchanged.

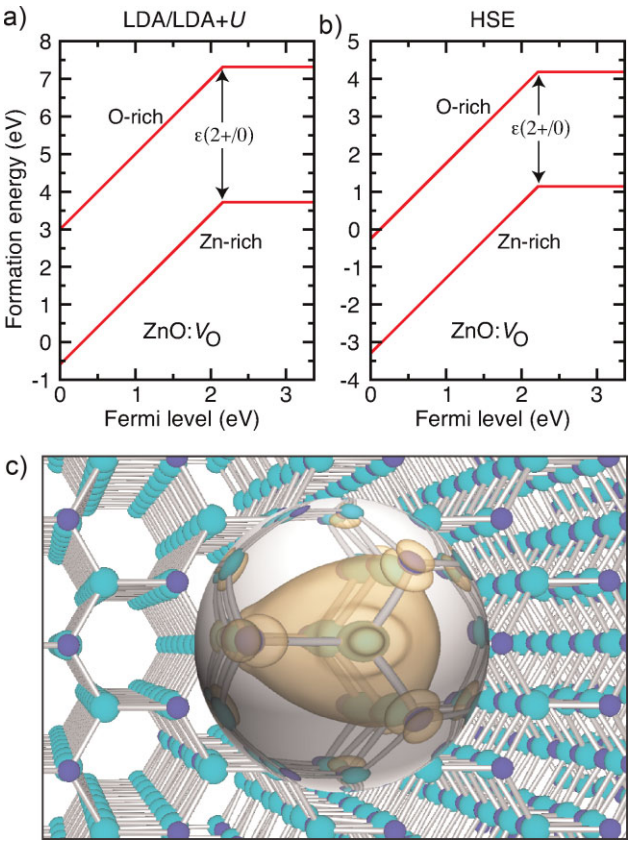

Figure 1 (online color at: www.pss-b.com) Formation energy as a function of Fermi level for an oxygen vacancy (V_O) in ZnO. (a) Energies according to the LDA/LDA+U scheme described in Section 3.1. (b) Energies according to the HSE approach [51]. The lower curve in each plot indicates Zn-rich conditions, and the upper curve O-rich conditions. The position of the transition level ε(2+/0) is also indicated. (c) Charge density of the V_O^0 gap state, which is occupied with two electrons. The isosurface corresponds to 10% of the maximum.

3.2 Hybrid functionals The use of hybrid functionals has been rapidly spreading in the study of defects in solids. In particular, hybrid functionals have proven reliable for describing the electronic and structural properties of defects in semiconductors. The method consists of mixing local (LDA) or semi-local (GGA) exchange potentials with the non-local Hartree–Fock exchange potential. The correlation potential is still described by the LDA or GGA. Hybrid functionals have been successful in describing structural properties and energetics of molecules in quantum chemistry, with Becke's three-parameter exchange functional (B3) with the Lee, Yang, and Parr (LYP) correlation (B3LYP) being the most popular choice [52]. However, the use of B3LYP for studying defects in solids has been limited due to its shortcomings in describing metals and narrow-gap semiconductors [53]. This issue is particularly important since formation enthalpies of metals usually enter the description of the chemical potential limits in the defect-formation-energy expressions (*cf.* Eq. (2)).

The introduction of a screening length in the exchange potential by Heyd, Scuseria, and Ernzerhof (HSE) [54, 55] and its implementation in a plane-wave code [56] have been instrumental in enabling the use of hybrid functionals in the study of defects in semiconductors. In the HSE the exchange potential is divided in short- and long-range parts. In the short-range part, the GGA exchange of Perdew, Burke, and Ernzerhof (PBE) [27] potential is mixed with non-local Hartree–Fock exchange potential in a ratio of 75/25. The long-range exchange potential as well as the correlation is described by the PBE functional. The range-separation is

implemented through an Error function with a characteristic screening length set to ∼10 Å [55], the variation of which can also affect band gaps [57]. The screening is essential for describing metals and insulators on the same footing. The HSE functional has been shown to accurately describe band gaps for many materials [56, 58]. We should note, however, that since the Hartree–Fock potential involves four-center integrals its implementation in plane-wave codes results in a high computational cost, and currently hybrid functional calculations take at least an order of magnitude more processing time than standard LDA calculations for systems with the same number of electrons.

As an example of hybrid functional calculations for defects in semiconductors, we show in Fig. 1(b) the formation energy as a function of Fermi level for the oxygen vacancy (V_O) in ZnO using the HSE functional [51]. These calculations were performed by setting the mixing parameter to 37.5% so to reproduce the experimental value of the band gap of ZnO. We note that the position of the transition level $\varepsilon(2+/0)$ with respect to the band edges is in remarkably good agreement with the value obtained using the LDA/LDA+U approach in Fig. 1(a). On the other hand, the absolute values of the formation energies are quite different, with the HSE results being more than 2 eV lower than the LDA/LDA+U results. This difference can be attributed to the effects of the HSE on the absolute position of the VBM in ZnO. In the LDA/LDA+U approach, U is applied only to the d states and the gap is corrected due to the effects of the coupling between the O 2p Zn d states, and the improved screening of the Zn 4s by the d states. Within this approach, it was assumed that the LDA+U would result in a correct position of the VBM. The HSE results show, however, that the position of the VBM on an absolute energy scale is affected by the inclusion of Hartree–Fock exchange [59]. That is HSE also corrects (at least in part) the self-interaction error in the LDA or GGA, which is still present in the LDA+U results, and this correction is significant for the O 2p bands that make up the VBM in ZnO. In Ref. [59] it was found that the VBM in ZnO is shifted down by 1.7 eV in HSE calculations, compared to PBE.

Other examples of the use of HSE include calculations for Si and Ge impurities in ZnO, which revealed that these impurities are shallow double donors when substituting on the Zn sites in ZnO, with relatively low formation energies [59]. Si can occur as a background impurity in ZnO, and these results indicate that it may give rise to unintentional n-type conductivity. Another example relates to p-type doping in ZnO. It has been long believed that incorporating N on the O site would lead to p-type ZnO. However, the effectiveness of N as a shallow acceptor dopant has never been firmly established. Despite many reports on p-type ZnO using N acceptors, the results have been difficult to reproduce, raising questions about the stability of the p-type doping and the position of the N ionization energy. Recent calculations for N in ZnO have shown that N is actually a very deep acceptor with a transition level at 1.3 eV above the VBM [60]. Therefore, it has been concluded that N cannot lead to p-type ZnO. For comparison and as a benchmark, HSE calculations correctly predicted that N in ZnSe is a shallow acceptor when substituting on Se sites, in agreement with experimental findings.

Hybrid functional calculations have also been performed for oxygen vacancies in TiO_2. Despite the fact that oxygen vacancies have frequently been invoked in the literature on TiO_2, their identification in bulk TiO_2 has remained elusive. First-principles calculations based on LDA or GGA suffer from band-gap problems and are unable to describe the neutral or the positively charged vacancy (V_O^+) in TiO_2 [61, 62]. In LDA or GGA, the Kohn–Sham single-particle states related to V_O are above the CBM, causing the electron(s) from V_O^0 or V_O^+ to occupy the CBM. Calculations based on the HSE, on the other hand, show that locally stable structures of V_O^0 and V_O^+ exist, in which the occupied single-article states lie within the band gap and the defect wave functions are localized within the vacancy. However, the formation energies of V_O^0 and V_O^+ are always higher in energy that of V_O^{2+} [62] as shown in Fig. 2(a); The atoms around V_O^{2+} relax outward as indicated in Fig. 2(b). Thus, oxygen vacancies are predicted to be shallow donors in TiO_2. This is in contrast to GGA+U calculations which indicate that V_O is a deep donor with transition levels in the gap [63]. The problem with GGA+U calculations for TiO_2 is that the conduction band in TiO_2 is derived from the Ti d states. The LDA/GGA+U approach was designed to be applied to narrow bands with localized electrons; hence its success when applied to semicore d states. The d states that constitute the conduction band of TiO_2, in contrast, are fairly delocalized, as evidenced by the high conductivity of this material. Applying LDA/GGA+U will always lead to an energy lowering of the occupied states, since that was what the approach was designed to do. Therefore, when the LDA/GGA+U approach is applied to a case in which electrons occupy the conduction band of TiO_2, localization will result. However, it is hard to distinguish whether this is a real physical effect or an artefact due to the nature of the LDA/GGA+U approach. We therefore feel that LDA/GGA+U

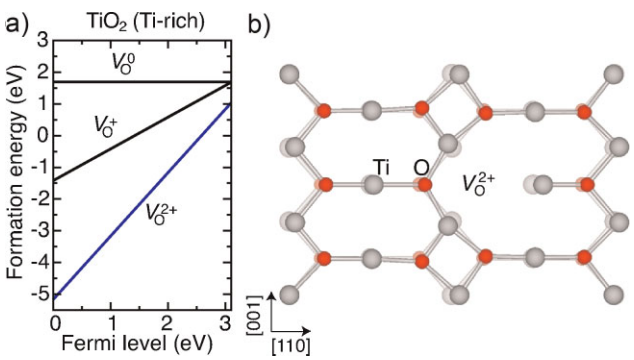

Figure 2 (online color at: www.pss-b.com) (a) Formation energy as a function of Fermi level for an oxygen vacancy (V_O) in TiO_2 in the Ti-rich limit, according to Ref. [62]. (b) Local lattice relaxations around V_O^{2+}. The positions of the atoms in the perfect crystal are also indicated (faded).

should not be applied in cases where the states are intrinsically extended states, such as the d states that make up the conduction band of TiO_2.

An important issue regarding the use of hybrid functionals is the amount of Hartree–Fock exchange potential that is mixed with the GGA exchange [64]. Although a value of 25% was initially proposed, there is no *a priori* justification for this amount and this single value is not capable of correctly describing all semiconductors and insulators. For instance, in ZnO the experimental value of the band gap is obtained with HSE only when a mixing parameter of 37% is used. In GaN, a mixing parameter of 31% is necessary, and for MgO 32%. Since the position of transition levels in the band gap depends on the band-gap value, quantitative predictions require that the functional accurately describes band gaps, and an adjustment of the mixing parameter is the most straightforward way to achieve this.

3.3 Many-body perturbation theory in the *GW* approximation Quasiparticle calculations in the *GW* approximation produce band structures that are in close agreement with experiments [65]. However, at present the calculation of total energies within the *GW* formalism [66] is still a subject of active research and currently not available for studying defects in solids. We note that the *GW* quasiparticle energies are defined as removal and addition energies. In the case of defects, the *GW* quasiparticle energies that appear in the band gap correspond to the transition levels, provided that the geometry of the defects remains unchanged. For instance, the highest occupied quasiparticle state in a calculation for a defect in charge state q represents the $\varepsilon(q+1/q)$ level, and the lowest unoccupied state represents the $\varepsilon(q/q-1)$ level for a fixed geometry of the defect. It is possible to combine these transition levels determined from *GW* calculations with relaxation energies from LDA or GGA calculations to extract thermodynamic transition levels for defects in semiconductors and insulators. Recent *GW* calculations for the self-interstitial in Si have demonstrated the effectiveness of this approach [67].

The LDA or GGA underestimates the formation energy of the self-interstitial in Si by more than 1 eV compared to values extracted from self-diffusion experiments. Calculations based on Quantum Monte Carlo can yield more accurate formation energies but are very expensive computationally. Calculating removal and addition energies for Si self-interstitials in *GW* and combining with relaxation energies from LDA calculations lead to formation energies that are in good agreement with Quantum Monte Carlo results [69]. The only assumption was that LDA gives correct formation energies for charge state configurations with no occupied states above the VBM, such as the 2+ charge state of the Si self-interstitial in the tetrahedral configuration. A similar approach has been used to study oxygen-related defects in SiO_2 [68].

As a drawback in the *GW* approach, it has been recently argued that for systems with semicore d states such as ZnO a very large number of unoccupied bands is necessary for a proper description of the band structure [70]. This result, if confirmed, indicates that *GW* calculations for defects in these systems may be prohibitivly expensive in practice. This unusually large number of unoccupied states required is likely related to the underbinding of the semicore d states which, as discussed in Section 3.1, can make a significant contribution to the band-gap error.

3.4 Modified pseudopotentials In the pseudopotential formalism, once a separation between valence electrons and the inert core electrons is adopted, there is still some flexibility in constructing the ionic cores. Indeed, within this approach, there is no unique scheme for generating pseudopotentials, and a number of different generation schemes have been proposed over the years, often aimed at creating computationally efficient, ''softer'' potentials which can be described with a smaller plane-wave basis set. This flexibility can in principle be exploited to generate potentials that produce a more accurate band structure. However, past attempts did not succeed in producing such improvements while still maintaining a proper description of atomic structure and energetics [71].

A new approach was recently proven to be remarkably successful in describing nitride semiconductors [72, 73]. It was based on a proposal by Christensen, first implemented within the linearized muffin-tin orbital method [74], to add a highly localized (delta-function-like) repulsive potential centered on the atomic nucleus of each atom. Such a potential only affects *s* states, and since the CBM in compound semiconductors has largely cation *s* character one expects an upward shift of the corresponding eigenstates. At the same time, the highly localized character of the added potential leads one to expect only minimal changes in other aspects of the pseudopotential. These expectations were indeed borne out in the case of GaN and InN, where the modified pseudopotentials produced atomic structures and energetics that are as reliable as those obtained with standard potentials, but simultaneously producing band structures in very good agreement with experiment [73]. Even though the fitting procedure only aimed to produce the experimental value of the direct gap, the modified potentials actually produced improvements for other aspects of the band structure as well, including the position of higher-lying indirect conduction-band minima as well as the position of semicore d states [73]. This leads us to believe that the seemingly *ad hoc* modifications introduced by the repulsive potential are capturing some essential physics, justifying the expectation that similarly good results can be obtained for other materials. An application of the modified pseudopotentials to the calculation of the electronic structure of nitride surfaces produced results in very good agreement with experiment [72, 75].

4 Summary We have discussed recent progress in first-principles approaches to study defects in semiconductors and insulators. Emphasis was given to methods that overcome the band-gap problem in traditional DFT in the LDA; such

approaches include LDA+U, hybrid functionals, *GW*, and modified pseudopotentials. While the LDA+U approach is very efficient computationally, it should be limited to systems with semicore states for which LDA provides a poor description. Furthermore, the LDA+U only partially corrects the band gap, and futher extrapolation is needed. The HSE hybrid functional on the other hand is general and has been demonstrated to be a reliable method that result in accurate band gaps and seems to be describing the properties of defects correctly. The HSE functional contains two parameters, the Hartree–Fock mixing ratio and the screening length, which offer some flexibility in obtaining correct band gaps; however, the consequences of changes in these parameters on the physics of the system has not been fully explored yet. The *GW* method offers a formal approach for describing excited-state properties and defect physics, but its applicability is limited by the lack of an efficient way to extract total energies. Combining *GW* excitation energies with LDA/GGA relaxation energies offers a promising way to address thermodynamic transition levels. Finally, modified pseudpotentials is an *ad hoc* but remarkably reliable approach, which has been demonstrated very effective at describing the properties of nitride semiconductors.

Acknowledgements We acknowledge fruitful collaborations and discussions with C. Freysoldt, G. Kresse, J. Lyons, J. Neugebauer, P. Rinke, M. Scheffler, A. Singh, N. Umezawa, and J. Varley. This work was supported by the NSF MRSEC Program under Award No. DMR05-20415, by the UCSB Solid State Lighting and Energy Center, and by the MURI program of the Army Research Office under Grant No. W911-NF-09-1-0398. It made use of the CNSI Computing Facility under NSF grant No. CHE-0321368 and Teragrid.

References

[1] C. G. Van de Walle and J. Neugebauer, J. Appl. Phys. **95**, 3851 (2004).
[2] D. A. Drabold and S. K. Estreicher (eds.), Theory of Defects in Semiconductors (Springer-Verlag, Berlin, 2007).
[3] M. Asato, T. Mizuno, T. Hoshino, K. Masuda-Jindo, and K. Kawakami, Mater. Sci. Eng. A **312**, 72 (2001).
[4] H. J. Queisser and E. E. Haller, Science **281**, 945 (1998).
[5] C. G. Van de Walle, D. B. Laks, G. F. Neumark, and S. T. Pantelides, Phys. Rev. B **47**, 9425 (1993).
[6] A. Janotti and C. G. Van de Walle, Rep. Prog. Phys. **72**, 126501 (2009).
[7] S. Limpijumnong and C. G. Van de Walle, Phys. Rev. B **69**, 035207 (2004).
[8] A. Janotti and C. G. Van de Walle, Phys. Rev. B **76**, 165202 (2007).
[9] A. Janotti, M. Krčmar, C. L. Fu, and R. C. Reed, Phys. Rev. Lett. **92**, 085901 (2004).
[10] M. Lannoo and J. Bourgoin, Point Defects in Semiconductors I: Theoretical Aspects (Springer-Verlag, Berlin, 1981); Point Defects in Semiconductors II: Experimental Aspects (Springer-Verlag, Berlin, 1983).
[11] S. T. Pandelides (ed.), Deep Centers in Semiconductors: A State-of-the-Art Approach, second edition (Gordon and Breach Science, Yverdon, 1992).
[12] M. Stavola (ed.), Identification of Defects in Semiconductors, Semiconductors and Semimetals, Vol. 51A, 51B (Academic, San Diego, 1999).
[13] P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964); W. Kohn, L. J. Sham, Phys. Rev. **140**, A1133 (1965).
[14] W. Kohn, Rev. Mod. Phys. **71**, 1253 (1999).
[15] M. Payne, M. P. Teter, D. C. Allan, T. A. Arias, and J. D. Joannopoulos, Rev. Mod. Phys. **64**, 1045 (1992).
[16] C. Stampfl, C. G. Van de Walle, D. Vogel, P. Krüger, and J. Pollmann, Phys. Rev. B **61**, R7846 (2000).
[17] A. Janotti and C. G. Van de Walle, Appl. Phys. Lett. **87**, 122102 (2005).
[18] A. Janotti and C. G. Van de Walle, J. Cryst. Growth **287**, 58 (2006).
[19] G. Makov and M. C. Payne, Phys. Rev. B **51**, 4014 (1995).
[20] J. Shim, E.-K. Lee, Y. J. Lee, and R. M. Nieminen, Phys. Rev. B **71**, 035206 (2005).
[21] C. Freysoldt, J. Neugebauer, and C. G. Van de Walle, Phys. Rev. Lett. **102**, 016402 (2009).
[22] C. Freysoldt, J. Neugebauer, and C. G. Van de Walle, Phys. Status Solidi B, DOI: 10.1002/pssb.201046289 (2010).
[23] J. A. Dean (ed.), Lange's Handbook of Chemistry, 14th edition (McGraw-Hill, Inc., New York, 1992).
[24] S. Limpijumnong, C. G. Van de Walle, and J. E. Northrup, Phys. Rev. Lett. **87**, 205505 (2001).
[25] J. Neugebauer and C. G. Van de Walle, J. Appl. Phys. **85**, 3003 (1999).
[26] J. P. Perdew and Y. Wang, Phys. Rev. Lett. **66**, 508 (1991).
[27] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).
[28] D. Vanderbilt, Phys. Rev. B **41**, 7892 (1990).
[29] K. Laasonen, A. Pasquarello, R. Car, C. Lee, and D. Vanderbilt, Phys. Rev. B **47**, 10142 (1993).
[30] G. Kresse and J. Hafner, J. Phys.: Condens. Matter **6**, 8245 (1994).
[31] P. E. Blöchl, Phys. Rev. B **50**, 17953 (1994).
[32] G. Kresse and J. Hafner, Phys. Rev. B **47**, 558 (1993).
[33] G. Kresse, J. Furthmüller, and Phys. Rev. B. **54**, 11169 1996.
[34] G. Kresse and J. Furthmüller, Comput. Mat. Sci. **6**, 15 (1996).
[35] X. Gonze, G.-M. Rignanese, M. Verstraete, J.-M. Beuken, Y. Pouillon, R. Caracas, F. Jollet, M. Torrent, G. Zerah, M. Mikami, Ph. Ghosez, M. Veithen, J.-Y. Raty, V. Olevano, F. Bruneval, L. Reining, R. Godby, G. Onida, D. R. Hamann, and D. C. Allan, Zeit. Kristallogr. **220**, 558 (2005).
[36] X. Gonze, B. Amadon, P.-M. Anglade, J.-M. Beuken, F. Bottin, P. Boulanger, F. Bruneval, D. Caliste, R. Caracas, M. Cote, T. Deutsch, L. Genovese, Ph. Ghosez, M. Giantomassi, S. Goedecker, D. R. Hamann, P. Hermet, F. Jollet, G. Jomard, S. Leroux, M. Mancini, S. Mazevet, M. J. T. Oliveira, G. Onida, Y. Pouillon, T. Rangel, G.-M. Rignanese, D. Sangalli, R. Shaltaf, M. Torrent, M. J. Verstraete, G. Zerah, and J. W. Zwanziger, Comput. Phys. Commun. **180**, 2582 (2009).
[37] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, A. Dal Corso, S. Gironcoli, S. Fabris, G. Fratesi, R. Gebauer, U. Gerstmann, C. Gougoussis, A. Kokalj, M. Lazzeri, L. Martin-Samos, N. Marzari, F. Mauri, R. Mazzarello, S. Paolini, A. Pasquarello, L. Paulatto, C. Sbraccia, S. Scandolo, G. Sclauzero, A. P. Seitsonen, A. Smogunov, P. Umari, and R. M. Wentzcovitch, J. Phys.: Condens. Matter **21**, 395502 (2009).

[38] J. P. Perdew and M. Levy, Phys. Rev. Lett. **51**, 1884 (1983).
[39] L. J. Sham and M. Schlüter, Phys. Rev. Lett. **51**, 1888 (1983).
[40] J. P. Perdew, Int. J. Quant. Chem. **28**, 497 (1985).
[41] R. W. Godby, M. Schlüter, and L. J. Sham, Phys. Rev. Lett. **56**, 2415 (1986).
[42] P. Mori-Sanchez, A. J. Cohen, and W. Yang, Phys. Rev. Lett. **100**, 146401 (2008).
[43] S. H. Wei and A. Zunger, Phys. Rev. B **37**, 8958 (1988).
[44] A. Janotti, D. Segev, and C. G. Van de Walle, Phys. Rev. B **74**, 045202 (2006).
[45] A. Janotti and C. G. Van de Walle, Phys. Rev. B **75**, 121201 (2007).
[46] V. I. Anisimov, J. Zaanen, and O. K. Andersen, Phys. Rev. B **44**, 943 (1991).
[47] V. I. Anisimov, F. Aryasetiawan, and A. I. Liechtenstein, J. Phys.: Condens. Matter **9**, 767 (1997).
[48] S. B. Zhang, S. H. Wei, and A. Zunger, Phys. Rev. B **63**, 075205 (2001).
[49] A. Janotti and C. G. Van de Walle, Appl. Phys. Lett. **92**, 032104 (2008).
[50] A. K. Singh, A. Janotti, M. Scheffler, and C. G. Van de Walle, Phys. Rev. Lett. **101**, 055502 (2008).
[51] F. Oba, A. Togo, I. Tanaka, J. Paier, and G. Kresse, Phys. Rev. B **77**, 245202 (2008).
[52] A. D. Becke, J. Chem. Phys. **98**, 1372 (1993).
[53] J. Paier, M. Marsman, and G. Kresse, J. Chem. Phys. **127**, 024103 (2007).
[54] J. Heyd, G. E. Scuseria, and M. Ernzerhof, J. Chem. Phys. **118**, 8207 (2003).
[55] J. Heyd, G. E. Scuseria, and M. Ernzerhof, J. Chem. Phys. **124**, 219906 (2006).
[56] J. Paier, M. Marsman, K. Hummer, G. Kresse, I. C. Gerber, and J. G. Àngyàn, J. Chem. Phys. **124**, 154709 (2006).
[57] H.-P. Komsa, P. Broqvist, and A. Pasquarello, Phys. Rev. B **81**, 205118 (2010).
[58] M. Marsman, J. Paier, A. Stroppa, and G. Kresse, J. Phys.: Condens. Matter **20**, 064201 (2008).
[59] J. L. Lyons, A. Janotti, and C. G. Van de Walle, Phys. Rev. B **80**, 205113 (2009).
[60] J. L. Lyons, A. Janotti, and C. G. Van de Walle, Appl. Phys. Lett. **95**, 252105 (2009).
[61] J. M. Sullivan and E. C. Erwin, Phys. Rev. B **67**, 144415 (2003).
[62] A. Janotti, J. B. Varley, P. Rinke, N. Umezawa, G. Kresse, and C. G. Van de Walle, Phys. Rev. B **81**, 085212 (2010).
[63] J. Osorio-Guillen, S. Lany, and A. Zunger, Phys. Rev. Lett. **100**, 036601 (2008).
[64] A. Alkauskas, P. Broqvist, and A. Pasquarello, Phys. Status Solidi B, DOI: 10.1002/pssb.201046195 (2010).
[65] M. S. Hybertsen and S. G. Louie, Phys. Rev. B **34**, 5390 (1986).
[66] P. Sánchez-Friera and R. W. Godby, Phys. Rev. Lett. **85**, 5611 (2000).
[67] P. Rinke, A. Janotti, M. Scheffler, and C. G. Van de Walle, Phys. Rev. Lett. **102**, 026402 (2009).
[68] L. Martin-Samos, G. Roma, P. Rinke, and Y. Limoge, Phys. Rev. Lett. **104**, 075502 (2010).
[69] E. R. Batista, J. Heyd, R. G. Hennig, B. P. Uberuaga, R. L. Martin, G. E. Scuseria, C. J. Umrigar, and J. W. Wilkins, Phys. Rev. B **74**, 121102 (2006).
[70] P. Zhang and B. Shih, Abstract Q23, American Physical Society Meeting, March 2010, http://meetings.aps.org/Meeting/MAR10/Event/120825.
[71] L. W. Wang, Appl. Phys. Lett. **78**, 1565 (2001).
[72] D. Segev and C. G. Van de Walle, Europhys. Lett. **76**, 305 (2006).
[73] D. Segev, A. Janotti, and C. G. Van de Walle, Phys. Rev. B **75**, 035201 (2007).
[74] N. E. Christensen, Phys. Rev. B **30**, 5753 (1984).
[75] C. G. Van de Walle and D. Segev, J. Appl. Phys. **101**, 081704 (2007).