Assessment of correction methods for the band-gap problem and for finite-size effects in supercell defect calculations: Case studies for ZnO and GaAs

Stephan Lany and Alex Zunger
National Renewable Energy Laboratory, Golden, Colorado 80401, USA
(Received 8 March 2008; revised manuscript received 23 July 2008; published 4 December 2008)

Contemporary theories of defects and impurities in semiconductors rely to a large extent on supercell calculations within density-functional theory using the approximate local-density approximation (LDA) or generalized gradient approximation (GGA) functionals. Such calculations are, however, affected by considerable uncertainties associated with: (i) the "band-gap problem," which occurs not only in the Kohn-Sham single-particle energies but also in the quasiparticle gap (LDA or GGA) calculated from total-energy differences, and (ii) supercell finite-size effects. In the case of the oxygen vacancy in ZnO, uncertainties (i) and (ii) have led to a large spread in the theoretical predictions, with some calculations suggesting negligible vacancy concentrations, even under Zn-rich conditions, and others predicting high concentrations. Here, we critically assess (i) the different methodologies to correct the band-gap problem. We discuss approaches based on the extrapolation of perturbations which open the band gap, and the self-consistent band-gap correction employing the LDA+U method for d and s states simultaneously. From the comparison of the results of different gap-correction, including also recent results from other literature, we conclude that to date there is no universal scheme for band gap correction in general defect systems. Therefore, we turn instead to classification of different types of defect behavior to provide guidelines on how the physically correct situation in an LDA defect calculation can be recovered. (ii) Supercell finite-size effects: We performed test calculations in large supercells of up to 1728 atoms, resolving a long-standing debate pertaining to image charge corrections for charged defects. We show that once finite-size effects not related to electrostatic interactions are eliminated, the analytic form of the image charge correction as proposed by Makov and Payne leads to size-independent defect formation energies, thus allowing the calculation of well-converged energies in fairly small supercells. We find that the delocalized contribution to the defect charge (i.e., the defect-induced change of the charge distribution) is dominated by the dielectric screening response of the host, which leads to an unexpected effective $1/L$ scaling of the image charge energy, despite the nominal $1/L^3$ scaling of the third-order term. Based on this analysis, we suggest that a simple scaling of the first order term by a constant factor (approximately 2/3) yields a simple but accurate image-charge correction for common supercell geometries. Finally, we discuss the theoretical controversy pertaining to the formation energy of the O vacancy in ZnO in light of the assessment of different methodologies in the present work, and we review the present experimental situation on the topic.

DOI: 10.1103/PhysRevB.78.235104 PACS number(s): 71.15.Mb, 61.72.Bb

I. INTRODUCTION

In semiconductors, the incorporation of desired dopant impurities and formation of undesired defects, such as recombination centers or compensating defects, controls the electrical and optical properties of these technologically important materials.[1] While numerous experimental methods exist for the identification and characterization of defects,[2] experiment often probes only specific defect properties, as accessible by the respective spectroscopic method, thereby providing only isolated aspects of the complete picture of defect-related effects. Theoretical studies of defects, hence, play an important complementary role. There, the pivotal quantity is the defect formation energy ΔH,[1,3–5] from which one can calculate the defect concentrations[1,3,6] and the electrical[3] and optical[7] transition levels of electrically active defects. Combining these theory-derived data with thermodynamic modeling of the host+defect+carrier system, one can simulate the materials system with all its lattice imperfections under realistic thermochemical conditions (growth conditions), thus obtaining the concentrations of all desired and undesired impurities and defects at equilibrium, including the carrier densities and the Fermi level.[1,4,6,8]

Calculations of the defect formation energy are often performed within density-functional theory (DFT), employing the local-density or generalized gradient approximation (LDA or GGA) and modeling the defect systems by construction of supercells with periodic boundary conditions. Such LDA or GGA supercell calculations owe their popularity for defect systems to their capability to calculate fairly accurate total energies in large systems on the order of 100 atoms needed to simulate isolated defects in solids. There are two classes of corrections needed, however, in such calculations:

(i) *Band-edge corrections due to the approximate DFT functional* (see Sec. III). Both the LDA and the GGA generally exhibit a considerable underestimation of the semiconductors' band gap, which in general affects the calculated defect formation energy.[5,9] Thus, defect calculations based on LDA or GGA generally require *ex post facto* corrections which are applied to supercell total energies after the self-consistent calculation. Recent advances in electronic structure theory hold promise for band-gap-corrected *ab initio* methods, such as GW,[10–12] model GW,[13] screened exchange,[14–16] exact-exchange or optimized effective potentials (OEPs),[17–20] and hybrid DFT.[21–23] Also, the self-

 ©2008 The American Physical Society

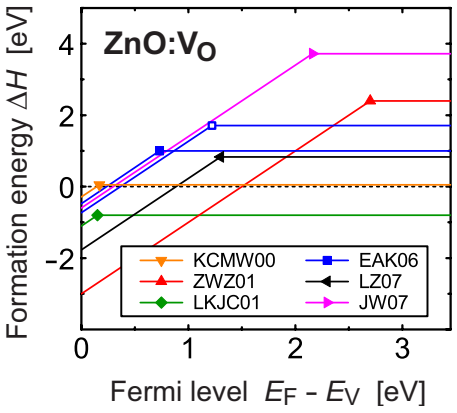


FIG. 1. (Color online) The formation energy ΔH of the O vacancy in ZnO under O-poor/Zn-rich conditions as calculated in recent theoretical works, using different schemes and procedures to account for LDA/GGA deficiencies and supercell finite-size effects. The references are KCMW00 (Ref. 33), ZWZ01 (Ref. 34), LKJC01 (Ref. 36), EAK06 (Ref. 39) (closed symbol: GGA; open symbol: GGA+U), LZ07 (Ref. 6), and JW07 (Ref. 40).

interaction-correction (SIC) (Ref. 24) method has been applied in various different formulations for band-gap correction.[25–28] While very accurate methods, such as the GW method, are yet not practically applicable to total-energy calculation of large-scale defect systems, approximate or model methods continue to be tested for their accuracy.[29] The advances and limitations of different orbital-dependent DFT approaches such as OEP, hybrid DFT, and SIC are discussed in a recent review.[30] At the present, such *post*-LDA methods have not matured to replace LDA based total-energy calculation of large, relaxed, and possibly charged defect systems, because of issues of both accuracy and computational cost.

(ii) *Corrections due to the supercell approximation* (see Sec. IV). Even large supercells at the limit of today's computational capabilities (~1000 atoms) for first-principles quantum-mechanical calculations correspond to very high concentrations of 10^{19}–10^{20} cm^{-3} for semiconductor standards. The calculation of the properties of isolated defects (e.g., 10^{14} cm^{-3}) requires, therefore, the correction of finite-size effects present in supercell calculations, especially in the case of charged defects[31] or when Moss-Burstein-type band-filling effects[32] occur, as in the case of shallow electron donors or acceptors.

Different schemes and procedures for correcting LDA errors and supercell-size effects have led in some cases to strongly varying predictions by different theory groups. Most notably, there is a recent controversy concerning oxygen vacancies in the wide-gap semiconductor ZnO,[6,7,33–41] which exhibits a particular severe band-gap problem. This controversy is illustrated in Fig. 1, showing recent theoretical results on the formation energy of V_O, which controls the O-deficient off-stoichiometry of ZnO. On one extreme end, Janotti and van de Walle[37,40] and Lee *et al.*[38] predicted very large formation energies for V_O in *n*-type ZnO, even under the most O-poor/Zn-rich conditions. Such high values of $\Delta H \approx 4$ eV lead to the prediction of very small concentrations of V_O below 10^{10} cm^{-3} under equilibrium conditions. In contrast, we[6,7] and Erhart *et al.*[39] found much lower formation energies of $\Delta H \approx 1$ eV, predicting considerable V_O concentrations of up to 10^{19} cm^{-3}.[6] Earlier, Kohan *et al.*[33] as well as Oba *et al.*[35] found still lower ΔH close to zero (under the O-poor condition), and Lee *et al.*[36] predicted even unphysically negative formation energies of V_O. Very recently, Pemmaraju *et al.*[42] and Oba *et al.*[43] confirmed our finding of a low V_O formation energy close to 1 eV, based on self-consistently band-gap corrected calculations employing SIC and hybrid-DFT, respectively.

While many approaches and schemes have been applied to address the band-gap problem and finite-size effects in the previous literature, the purpose of the present work is to assess the validity of such schemes in those cases where considerable uncertainties and controversies remain. We now provide a guide for the reader to the organization of this paper:

In Sec. II, we describe and review in some detail the general formalism of supercell total-energy calculation of defects in semiconductors and insulators. We also summarize the specific set of band-gap and supercell corrections we favored, as introduced before in the Appendix of Ref. 5.

In Sec. III, concerning the correction of the band-gap problem, we first illustrate the manifestation of this problem in terms of single-particle and total energies (Sec. III A), and the implication of the host-band-edge corrections for defect formation energies (Sec. III B). We then discuss different schemes for correcting the band gap in defect systems, such as the extrapolation of a band-gap-opening perturbation toward the experimental gap (Sec. III C) and the LDA+U method, which allows for gap correction in the self-consistent calculation when applied to s and d states simultaneously (Sec. III D; see also Ref. 44). Applying these methods to the specific example of V_O in ZnO (Sec. III E), we find that a universal method for accurate band-gap correction of defects remains elusive. Comparing the present results and recent literature data[42–44], we find that even self-consistently band gap corrected methods which do not require *a posteriori* corrections yield a rather wide spread of predicted transition levels for V_O. Therefore, we develop a general picture of different defect behaviors (Sec. III F), identified by the energies of the single-particle defect levels relative to the band edges, which require different types of corrections. Thereby, we provide guidelines that should generally aid in recovering the qualitatively correct physical situation from an LDA defect calculation.

In Sec. IV, addressing supercell finite-size effects, we test and validate our previously developed set of size-effect corrections (see the Appendix of Ref. 5) by calculating specific examples in very large supercells (up to 1728 atoms). Such an assessment is needed, in particular, as there is no consensus in literature to date[5,9,39,40,44–57] about whether one should apply corrections for the screened electrostatic interactions of image charges, as proposed by Makov and Payne[31] (Sec. IV A). We show that after elimination of finite-size effects that are not related to electrostatic interactions (in particular, after correction for the undefined potential reference in momentum-space calculations), the image charge correction of Ref. 31 provides essentially the same accuracy as finite-size scaling methods[50,54,56] but at a much reduced computa-

tional effort. In Sec. IV B, we test and illustrate the importance of Moss-Burstein-type band-filling effects[5] that occur when electrons (or holes) occupy strongly dispersive host-derived band states. The slow convergence with supercell size necessitates the correction of these band-filling effects if one is interested in defect formation energies in the dilute limit. We further discuss the cell-size dependence of the defect-state–host-band hybridization and the implications for the correct determination of the single-particle energies of the genuine defect states, which have to be distinguished from the host-derived bands that are perturbed by the presence of the defect.

Finally, we review in Sec. V the experimental situation of O deficiency in ZnO in the light of the theoretical controversy, finding that experimental evidence strongly suggests the thermodynamic formation of O vacancies in ZnO under O-poor/Zn-rich conditions at concentrations on the order of 10^{17} (Ref. 58) or 10^{18} cm^{-3}.[59,60] Thus, ZnO shows a similar tendency toward O deficiency as the related oxides In_2O_3,[61] SnO_2,[62] and MgO.[63] The ubiquitous existence of O vacancies in main-group oxides is thus a crucial benchmark of the validity of different methodologies to correct for band-gap and finite-size effects in supercell defect calculations.

II. GENERAL FORMALISM OF SUPERCELL DEFECT CALCULATIONS

A. Defect formation energies

Within the supercell formalism for the representation of defects in a host lattice, the defect formation energy of a defect D in charge state q is defined as

$$\Delta H_{D,q}(E_F,\mu) = [E_{D,q} - E_H] + q(E_V + \Delta E_F) + \Sigma n_\alpha(\mu_\alpha^0 + \Delta\mu_\alpha), \quad (1)$$

where E_D and E_H are the total energies of the host+defect and host-only supercells, respectively.

1. Energy of the valence-band maximum

$E_V = E_H(0) - E_H(+1)$ is defined as the energy difference between the pure host ($q=0$) and the host with one hole ($q=+1$) in the valence band in the dilute hole gas limit.[5] Thus, Eq. (1) describes the enthalpy of the defect formation reaction conserving the charge. E.g., for a singly charged donor and $\Delta E_F = 0$, this reaction is

$$D_R + H_H + h^+ \rightarrow D_H^+ + H_R,$$

where D_R denotes the donor atom in its chemical reservoir (before defect formation), H_H denotes a host atom at its native lattice site, and h^+ denotes a hole at the valence-band maximum (VBM).

2. Fermi energy

E_F is conventionally defined with respect to VBM, $E_F = E_V + \Delta E_F$, and is usually bounded between the VBM and the conduction-band minimum (CBM), i.e., $E_V < E_F < E_C$, except in the case of degenerate doping, in which additional energy contributions due to electron-concentration-dependent band-filling effects have to be considered.[6,64]

3. Chemical potentials

The growth conditions are reflected in the chemical potentials $\mu_\alpha = \mu_\alpha^0 + \Delta\mu_\alpha$ of the atoms removed ($n_\alpha = +1$) or added ($n_\alpha = -1$) to the host crystal when the defect is formed. For example, O-poor/Zn-rich conditions in ZnO are present when the Zn chemical potential equals that of elemental Zn metal, $\mu_{Zn} = \mu_{Zn}^0$ ($\Delta\mu_{Zn} = 0$). For such metal-rich conditions, which facilitate the formation of anion vacancies, Fig. 1 shows $\Delta H(E_F)$ for V_O in ZnO and compares recent literature results that are based on different assumptions about the corrections of band gap and supercell finite-size errors.

B. Defect transition energies

The thermodynamic transition energy $\varepsilon(q/q')$ between two charge states q and q' describes the Fermi level E_F at which $\Delta H(E_F,q) = \Delta H(E_F,q')$, i.e.,

$$\varepsilon(q/q') - E_V = [\Delta H(E_V,q') - \Delta H(E_V,q)]/(q - q'). \quad (2)$$

The thermal ionization energy of simple donors and acceptors equals the distance of $\varepsilon(+/0)$ from the CBM and that of $\varepsilon(0/-)$ from the VBM, respectively. Optical transitions between defect states in the gap and the band-edge energies can also be calculated from $\Delta H_{D,q}$ [Eq. (1)] when the atomic positions of the initial state are kept during the optical (vertical) excitation or recombination process, according to the Franck-Condon principle. Thus, the optical absorption energy $\varepsilon_O(q/q+n;ne)$ due to the excitation ($n=+1$) of an electron from the defect level into the CBM or the respective optical emission (luminescence; $n=-1$) energy due to the recombination of an electron from the CBM into the defect level is calculated as[7]

$$\varepsilon_O(q/q+n;ne) = \Delta H(E_C,q+n) - \Delta H(E_C,q). \quad (3)$$

Analogously, the absorption ($n=+1$) energy from the VBM into the defect state and the emission ($n=-1$) energy from the defect level into the VBM, i.e., the recombination with a free hole, are given by

$$\varepsilon_O(q/q-n;nh) = \Delta H(E_V,q-n) - \Delta H(E_V,q). \quad (4)$$

It should be emphasized that the optical (vertical) transition energies ε_O calculated from total-energy differences [Eqs. (3) and (4)] are in general different from the respective single-particle energy level e_D of the defect, i.e., the respective eigenvalue obtained from the solution of the Kohn-Sham equation $\mathbf{H}^{KS}\psi_i = e_i\psi_i$. Consider, for example, the case of a finite system (atom, molecule, or cluster), where the (optical) excitation energy ε_O for the excitation of an electron from state i into the vacuum level is related to the initial-state single-particle energy $e_i(q)$, according to Refs. 65 and 66, by

$$\varepsilon_O(q/q+1;e) = -e_i(q) + \Pi_i + \Sigma_i. \quad (5)$$

The difference, i.e., the term $\Pi_i + \Sigma_i$, is the result of two electronic relaxation effects upon electron removal from the state i. First, the self-interaction term $\Pi_i = e_i(q) - \int_q^{q+1} e_i^*(q')dq'$ reflects the energy lowering of the eigen-

value e_i^* due to the elimination of self-interaction and the reduction in screening upon the electron removal, during which the initial-state wave functions are kept fixed (denoted by the asterisk). Second, the relaxation contribution Σ_i is the energy gain during relaxation of the initial-state wave functions. Comparing with the Hartree-Fock (HF) theory, where the Koopmans theorem[67] holds, i.e., $\Pi_i=0$, the difference between single-particle and total-energy transition levels is generally larger in common DFT functionals where $\Pi_i \neq 0$. Considering LDA or GGA calculations in semiconductor defect systems, the optical excitation energy, e.g., from a defect level in the gap into the CBM, differs from the respective single-particle energy usually by few tenths of an eV (Ref. 68) and becomes larger with increasing localization of the defect state, e.g., exceeding 1 eV in the case of transition-metal impurities in ionic oxides such as MgO.

C. Electronic structure methods employed in the present work

1. LDA and GGA supercell methods and pseudopotentials

We calculate the total energy via the pseudopotential-momentum-space formalism,[69] using projector-augmented-wave (PAW) potentials,[70] as implemented in the VASP code.[71] Exchange and correlation effects within DFT are treated in the LDA or the GGA, using the parameterizations of Refs. 24 and 72, respectively. Since the GGA is generally considered more accurate for molecules and surfaces,[73] we used in Ref. 6 the GGA for ZnO, where the elemental reference state of oxygen is the O_2 molecule. We find, however, that GGA and LDA give rather similar results for intrinsic defects in ZnO (Refs. 6 and 7) under the growth conditions that support formation of these defects (see also Sec. III E). Further, in order to calculate large supercells of up to 576 atoms in ZnO and up to 1728 atoms in GaAs, we employ pseudopotentials (PPs) which are particularly suitable for such large systems: For oxygen in ZnO, we use a particularly soft PP, which requires an energy cutoff of only 283 eV. In calculations using the typical supercell size of 72 atoms, we use the standard PP (400 eV) for O. Testing the soft PP against the standard PP for such small supercells, we found good agreement in formation energies, atomic relaxations, and single-particle-energies for V_O in ZnO.[74] For calculations in GaAs, where the effect of the Ga $3d$ electrons is very small,[75] we used a PP where the $3d$ shell is omitted from the valence.

2. LDA+U calculations

A computationally expedient post-LDA method is LDA $+U$,[76–78] which was originally developed to improve the LDA description of Mott insulators[76] by introducing Hubbard-type interactions into LDA via an adjustable Coulomb parameter U. In conventional (band-)semiconductor systems, we first applied LDA$+U$ for partial band-gap correction in the photovoltaic chalcopyrite $CuGaSe_2$,[79] where the valence band has strong Cu d character and lies too high in energy in LDA.[80] By applying LDA$+U$ on the Cu d shell, the self-interaction within the d shell is approximately corrected, thereby lowering the d-band energy and opening the band gap. When used only for the metal d states, a full band-gap correction is generally not achieved for physically meaningful values of U.[81] A full band-gap correction can be empirically achieved, however, when LDA$+U$ is used for the cation d states and, simultaneously, for anion s (Refs. 82 and 83) or cation s (Ref. 84) states. The band-gap correction of ZnO by simultaneous applications of LDA$+U$ on Zn d and Zn s orbitals was recently independently employed by Paudel and Lambrecht[44] within a linear-muffin-tin-orbital (LMTO) method.

Since the LDA$+U$ method requires adjustable parameters for the Coulomb and exchange energies U and J, physically meaningful values for these parameters have to be found. (Note that the decisive parameter is actually the difference $U-J$.[85]) Possible strategies to determine suitable parameters are the adjustment of U_d so to reproduce experimental photoemission spectra,[79,86] constrained LDA calculations,[76,87] thermochemical considerations,[81,88] or a self-consistency requirement between U and the orbital (partial) occupancies.[89,90] The Hubbard U energy of a free atom (which can be directly calculated in LDA) is reduced in a semiconductor due to electronic screening.[76,91] Since, however, the d electrons are considerably localized within the typical screening length (on the order of 1 Å), the screening is incomplete. Therefore, dividing the free-atom U value by the dielectric constant of the solid[40] most likely underestimates the appropriate value of U. Indeed, finite values for U are generally used even in metals, e.g., $U=2-6$ eV in metallic Fe,[89,92] where the division by the dielectric constant yields $U=0$.[40] For Zn d orbitals in II-VI semiconductors, we found $U=7$ eV (for $J=0$) (Refs. 7 and 79) by comparison with photoemission experiments (a similar value of $U=7.5$ eV was used by Erhart *et al.* in Ref. 39). Concerning this strategy, it should be noted that generally in DFT calculations the single-particle energies e_i do not represent excitation energies as measured by photoemission [cf. Eq. (5), Sec. II B]. However, the approximate correction of self-interaction effects within the d shell in LDA$+U$ re-establishes a physical meaning for the single-particle energies of these states in the sense of the Koopmans theorem of HF theory, where the ionization energy (IE) is approximately equal to the respective eigenvalue, $\mathrm{IE}_i \approx e_i$ (note close relation between LDA$+U$ and HF theory[78,89]). Therefore, the adjustment of U to photoemission data is a justified approach.

We further note that there exists an additional complication with the LDA$+U$ method for defect formation energy calculations because of the need to calculate elemental reference energies [cf. Eq. (1), Sec. II A], e.g., of the Zn metal and of the O_2 molecule in the case of ZnO: On one hand the appropriate U value for the Zn d states should be smaller in the metallic element than in the semiconductor ZnO, due to stronger screening. On the other hand, total energies should in practice be compared only for the same U.[90,93] Thus, the LDA$+U$-calculated heat of formation may not be accurate, whether same or different values of U are taken for the semiconductor and the metal. In the specific case of a binary compound where LDA$+U$ is used only for one constituent, one can, of course, use the experimental heat of formation to determine defect formation energies under both the metal-rich and the anion-rich conditions, which is done here for

those results in Sec. III E that rely on LDA+U. Since we used in Refs. 6 and 7 LDA+U only to determine the band-edge shifts (see Sec. III B) but not to calculate supercell energies, these results did not suffer from the problem of an undefined heat of formation in LDA+U.

D. "Postprocessor" corrections to supercell energies

In Secs. III and IV below, we assess corrections for defect supercell energies that are related to the band-gap error (BGE) and to supercell-size effects (SSEs). While various approaches for such corrections have been suggested and used in the previous literature,[5,9,31,34,39,40,45,48–50,52,54,94] we give here a summary for the specific formulation of the set of corrections used by us, as introduced before (except Sec. II D 5 below) in the Appendix of Ref. 5.

1. Shifting the individual band-edge energies of the host (BGE)

Due to the band-gap problem of the approximate LDA and GGA functionals (see Sec. III A), one needs to determine the corrections ΔE_V for the VBM and ΔE_C for the CBM such that the experimental band gap is recovered, $E_g(\text{expt})=(E_C+\Delta E_C)-(E_V+\Delta E_V)$. In the case of charged defects, these corrections increase the range of possible formation energies, as ΔH depends linearly on the Fermi level E_F inside the gap [see Eq. (1)]. The determination of the required shifts of the band-edge energies is discussed in Sec. III B.

2. Shifting shallow levels with the respective host bands (BGE)

Once the band-edge states are corrected, the question arises as to how defect levels would be affected by the band-gap correction. While it is common practice to refer donor states to the CBM and acceptor states to the VBM, it is important to realize in which situations this procedure is justified and in which it is not. A lattice defect in a semiconductor generally creates a *primary*, *defect-localized* state (DLS).[7] If this DLS occurs in the gap, the defect is deep. In contrast, the hallmark of shallow defects is that their DLS occurs as a resonance inside the continuum of host bands; e.g., the DLS of a shallow donor lies inside the conduction band. In this case the introduced electron relaxes from the DLS to the band edge, occupying a *secondary*, *delocalized* perturbed-host state (PHS),[7] which is essentially the electronic state of the CBM of the host, perturbed only by the screened Coulomb potential of the charged dopant ion. Thus, in the case of shallow defects, the occupied donor (acceptor) states can be expected to shift along with the CBM (VBM) during the band-gap correction, leading to an energy correction of $z_e\Delta E_C$ ($-z_h\Delta E_V$) for ΔH when the donor (acceptor) state is occupied by z_e electrons (z_h holes). This correction is applied, e.g., to the shallow Te_{As} donor in GaAs (see Sec. IV B). In the case of deep defects, the primary defect state, i.e., the DLS, occurs as a state inside the gap. The gap correction for this class of defects cannot be directly linked to the behavior of the host-band edges. In Sec. III, we discuss methods of determining corrections for such deep defects and propose a general classification scheme for distinguishing the cases that need different treatment for LDA correction.

3. Band-filling correction (SSE)

Due to the high defect concentrations implied by typical supercell calculations, Moss-Burstein-type band-filling effects[32] are present in the case of shallow defects where the carriers occupy the strongly dispersive PHS. In order to recover the dilute limit for ΔH_D, we eliminate these band-filling effects by a correction; e.g., for shallow donors,

$$\Delta E_{\text{bf}} = -\sum_{n,\mathbf{k}} \Theta(e_{n,\mathbf{k}} - \tilde{e}_C)(w_{\mathbf{k}}\eta_{n,\mathbf{k}}e_{n,\mathbf{k}} - \tilde{e}_C). \quad (6)$$

Here $e_{n,\mathbf{k}}$ are the band energies in the defect calculation, $\tilde{e}_C$ is the CBM energy of the pure host after potential alignment with the defect calculation (see below), Θ is the Heaviside step function, $w_{\mathbf{k}}$ is the $\mathbf{k}$-point weight, and $\eta_{n,\mathbf{k}}$ is the band occupation. In Sec. IV B, we study the convergence of band-filling effects with supercell size for Te_{As} in GaAs.

4. Potential-alignment correction for supercells with a net charge (SSE)

While the total energy of a periodic, *charge-neutral* system is well defined, the total energy of a periodic system with a net charge in a unit cell diverges. Even though the total-energy expression[69] of the pseudopotential momentum-space formalism has been derived under the explicit assumption of charge neutrality [see Eq. (21) of Ref. 69], it is usually applied without change also for charged defects. In this formalism, the average Hartree and ionic potentials V_H and V_i are individually set to zero, i.e., $V_H(\mathbf{G}=0)=V_i(\mathbf{G}=0)=0$. In the charge-neutral case, this procedure is justified by the exact cancellation of the respective electron-electron and ion-ion contributions.[69] In the charged calculation, the omission of the $\mathbf{G}=0$ terms can be viewed as an effective compensation for the net charge by a homogeneous (jellium) background charge. (Note, however, that the compensation charge is generally not explicitly introduced in calculations of charged supercells.) The total energy is no longer a well-defined quantity in a charged system:[5] When evaluated with the total-energy expression of Ref. 69, charged supercell energies exhibit the same arbitrary shifts as the Kohn-Sham eigenvalues, which are defined only up to a constant. Indeed, when we calculate the energy of a hole in the semiconductor valence band, $E_V=E_H(0)-E_H(+1)$, we find that E_V converges in the limit of a dilute hole gas toward the Kohn-Sham single-particle energy e_V of the VBM.[5] Thus, we use a potential-alignment technique[5,95,96] to ensure that the "charged energies" $E_{D,q}$ and $E_H(+)$ entering Eq. (1) are treated consistently, and we use $E_V=e_V$ in Eq. (1). The potential-alignment correction energy for a defect D with in the charge state q is

$$\Delta E_{\text{PA}}(D,q) = q(V^r_{D,q} - V^r_H), \quad (7)$$

where the reference potentials V^r in the charged-defect (D,q) and pure-host (H) calculations are determined from the (local) atomic-sphere-averaged electrostatic potentials[97] at atomic sites farther away from the defect.

5. Charged supercells without a reference point for potential alignment (SSE)

As we demonstrate in Sec. IV A, the potential-alignment correction described above is an essential part of our robust scheme for correcting supercell finite-size effects for ΔH of charged defects. However, this method is not applicable in a situation where there is no hostlike reference point far from the perturbation by the defect. Consider, for example, the case of an alloy where one adds electronic charge which is supposed to be compensated by a jellium background. Due to the compensating background, the system as a whole is charge neutral, so the total energy should be well defined. Our finding that the total energy of charged systems shows the same arbitrary offsets as the single-particle energies[5] implies that the energy evaluated with the usual expressions in Ref. 69 *does not* represent the energy of the (overall neutral) charge+jellium system. The energy contribution due to this interaction between the additional electronic charge and the jellium background can evidently occur only in the $\mathbf{G}=0$ terms. On the other hand, in an overall charge-neutral system one can set the average electrostatic potential to zero (see above), in which case the only $\mathbf{G}=0$ contribution arises due to the "$Z\alpha$" term.[69] Thus, this term has to be modified to account for the electron-jellium interaction.

The deviation of the local part of the atomic pseudopotentials $V_i^{\mathrm{PS}}(r)$ from the bare ionic Coulomb potential Z_i/r causes an offset of the average potential in the cell with volume Ω,

$$\alpha_{\mathrm{PS}} = \frac{1}{\Omega}\sum_i \int_\Omega [V_i^{\mathrm{PS}}(r) + Z_i/r]d^3r, \tag{8}$$

by which the electronic single-particle energies are downshifted when the average electrostatic potential is set arbitrarily to zero, as conventionally done in plane-wave calculations. As discussed in Ref. 69, this shift needs to be compensated by an energy contribution

$$E_{Z\alpha} = \alpha_{\mathrm{PS}}\sum_i Z_i. \tag{9}$$

This expression of the original work[69] was, however, derived by considering an uncharged system, where the sum of the ionic charges equals the number of electrons, $\Sigma_i Z_i = N$. Since this energy contribution represents an electron-ion interaction, the respective term should actually read $E_{Z\alpha}=N\alpha_{\mathrm{PS}}$ for a charged system. In order to compensate for the common implementation [Eq. (9)], we apply a correction

$$\Delta E_{q\alpha} = -q\alpha_{\mathrm{PS}} \tag{10}$$

to the supercell energies. Note that this contribution depends on the supercell volume [see Eq. (8)] and therefore changes in general the equilibrium lattice constant. Comparing with the "potential-alignment correction," we emphasize that only *either* ΔE_{PA} *or* $\Delta E_{q\alpha}$ can be applied. The potential alignment is more appropriate for defect systems where one is interested in the dilute limit and where one generally keeps the equilibrium volume of the host crystal constant, whereas the $\Delta E_{q\alpha}$ correction is more appropriate for charged systems where the nonhostlike perturbation is dispersed, such as, e.g., in the case of charged alloy systems.

6. Image charge correction for charged defects (SSE)

Makov and Payne[31] suggested an image charge correction to $O(L^{-5})$ in the linear supercell dimension $L=V_{\mathrm{SC}}^{-1/3}$ (supercell volume V_{SC}),

$$\Delta E_{\mathrm{MP}} = \frac{q^2\alpha_M}{2\varepsilon L} + \frac{2\pi q Q_r}{3\varepsilon L^3}, \tag{11}$$

where α_M is the (supercell) lattice-dependent Madelung constant, ε is the static dielectric constant of the host, and Q_r is the second radial moment of the electron-density difference $\tilde{\rho}_{D,q}(\mathbf{r})=\rho_{D,q}(\mathbf{r})-\rho_H(\mathbf{r})$ between the defect+host and pure-host systems,

$$Q_r = \int_{V_{\mathrm{SC}}} d^3r\tilde{\rho}_{D,q}(\mathbf{r})r^2. \tag{12}$$

In Sec. IV A, we test this correction method by studying the scaling behavior of several charged defects in GaAs supercells of up to 1728 atoms.

III. CORRECTION OF BAND-GAP ERRORS

A. Manifestation of the band-gap problem in the quasiparticle gap

The arguably greatest shortcoming of LDA or GGA for defect calculations in semiconductors is the underestimation of the band gap, typically by about 50%. ZnO is a particular severe case, where the experimental gap is $E_g=3.45$ eV at low temperature but the calculated Kohn-Sham single-particle gaps $e_{\mathrm{KS}}=e_C-e_V$ are only 0.80 eV in LDA and 0.73 eV in GGA at the respective equilibrium lattice constant. Of course, the Kohn-Sham single-particle energies are not physical quantities[98,99] and the "true" (quasiparticle) band gap E_{QP} is defined as the difference between the ionization potential I and the electron affinity A, both being ground-state quantities, $E_{\mathrm{QP}}=I-A$. In the case of the "exact" DFT functional, the difference Δ between the quasiparticle and single-particle gaps, $\Delta=E_{\mathrm{QP}}-e_{\mathrm{KS}}$, equals the derivative discontinuity exhibited by the exact exchange-correlation functional.[98,99] Note that the actual magnitude of this discontinuity and, hence, of Δ for semiconductor systems in the exact functional is subject of a unresolved debate.[100,101]

The fact that even the exact DFT functional shows a difference between the single-particle and quasiparticle band gaps is frequently viewed as an argument that the band-gap problem does not reflect the approximation of the DFT functional but is inherent to the DFT formalism.[30,100,102] This view, however, may easily lead to the misinterpretation that the band-gap problem is only an apparent problem caused by the nonphysical meaning of the Kohn-Sham eigenvalues. Since the exact functional would certainly give the correct quasiparticle gap $E_{\mathrm{QP}}=I-A$, we illustrate the LDA and GGA band-gap problem by calculating the quasiparticle gap $E_{\mathrm{QP}}^{\mathrm{GGA}}=E_H(-1)+E_H(+1)-2E_H(0)$ in GGA for the example of ZnO, shown in Fig. 2(a) as a function of the number N of

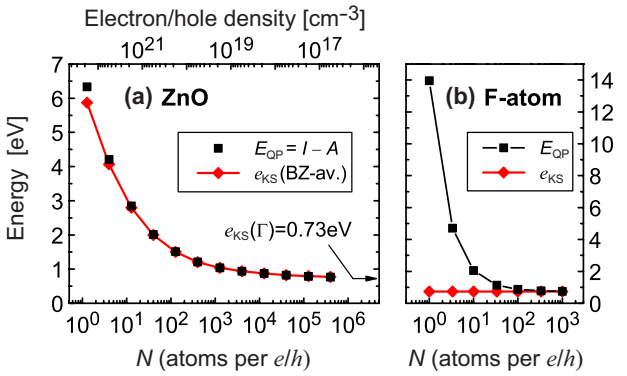


FIG. 2. (Color online) (a) The quasiparticle gap $E_{QP}=I-A=E_H(-1)+E_H(+1)-2E_H(0)$ and the Kohn-Sham single-particle gap e_{KS} (Brillouin-zone average) of ZnO in GGA as a function of the number N of atoms per additional electron/hole. (b) The quasiparticle and single-particle gaps for a separated open system of F atoms.

atoms (cell size) over which the additional electron or hole[103] is distributed, i.e., as a function of the carrier density. We see in Fig. 2(a) that E_{QP} *converges to the single-particle gap at the Brillouin-zone center*, $e_{KS}(\Gamma)=e_C(\Gamma)-e_V(\Gamma)$, in the limit of a dilute gas of free electrons and holes. Thus, for approximate functionals such as LDA or GGA, the quasiparticle gap E_{QP}^{LDA} shows the same band-gap error as the single-particle gap e_{KS}^{LDA}. For finite carrier densities, the apparent band gap is larger than the direct ZnO gap at the Γ point, due to band-filling effects [see Sec. II D 3, Eq. (6)]. When comparing the quasiparticle gap with the appropriate Brillouin-zone average (BZ av) of the single-particle energies,

$$e_g(\mathrm{BZ\ \ av}) = \sum_{\mathbf{k}} w_{\mathbf{k}}(\eta_{C,\mathbf{k}}e^*_{C,\mathbf{k}} - \eta_{V,\mathbf{k}}e^*_{V,\mathbf{k}}), \qquad (13)$$

we see in Fig. 2(a) that the quasiparticle and single-particle gaps are still practically identical in GGA except for extremely large electron and hole densities. [As in Sec. II B, the asterisks in Eq. (13) denote that the eigenvalues e^*_C and e^*_V are determined with the wave functions of the initial neutral state.]

Considering that $I=-e_V(q)+\Pi_V+\Sigma_V$ and $A=-e_C(q)+\Pi_C+\Sigma_C$ [see Eq. (5) in Sec. II B] and the fact that Π and Σ approach zero in the limit of very delocalized wave functions,[66] the equality $E_{QP}=I-A=e_{KS}$ can actually be expected for band semiconductors in general when the band-edge states e_V and e_C are extended over all N atoms. It is interesting to note that in cases where carriers are more localized, such as, for example, the self-trapped hole polarons in halides,[104] the quasiparticle gap would be even smaller than the Kohn-Sham gap: Since, in this case the localized polaronic hole state must be lower in energy than the delocalized bandlike state, i.e., $E_H^{loc}(+1)<E_H^{deloc}(+1)$, it follows that $I^{loc}<I^{deloc}=e_V$ and, hence, $E_{QP}<e_{KS}$. Note, however, that due to residual self-interaction in LDA or GGA, these functionals generally tend to find delocalized hole states lower in energy than localized states, which in some cases leads to a qualitatively wrong description, such as, e.g., in the case of the Al_{Si} impurity in SiO_2.[105]

The finding that the LDA or GGA description of the quasiparticle and the single-particle energy gaps is equally poor in extended periodic systems such as ZnO may be surprising in view of the fact that in isolated small systems, such as free atoms, the total-energy differences describe rather accurately the atomic ionization energies.[106] Indeed, when we calculate the QP band gap, for example, for an isolated F atom, we obtain $I-A=E(F^+)+E(F^-)-2E(F^0)=13.95$ eV in GGA,[107] in very good agreement with the experimental value of 14.02 eV and much larger than the single-particle gap, which is calculated as $e_{KS}=0.74$ eV in the symmetry-broken solution of the F atom. However, when we consider a *periodic array* of F atoms, the calculated quasiparticle gap $E_{QP}=I-A$ is strongly reduced with the number N of F atoms per additional electron and hole, as shown in Fig. 2(b) (due to the separation of the F atoms,[107] there is no band dispersion and, hence, no N dependence of the single-particle gap e_{KS}, as in the case of ZnO). Since the additional electron (hole) delocalizes over the array of F atoms, E_{QP} converges to e_{KS} in the limit of large N, as expected from the considerations above. Note that this delocalization occurs due to energy minimization and is not an artifact of periodic boundary conditions.

As discussed in recent literature,[108,109] the delocalization of electrons and holes in an array of distant atoms is a result of the incorrect convex behavior of LDA and GGA energies between integer occupation numbers in separated open systems. Whereas $I-A$ should be invariant with respect to the system size N, LDA or GGA incorrectly finds the delocalized electron (hole) state that is distributed over all N atoms lower in energy than the state in which the wave function is localized on just one atom. Comparing the results in Figs. 2(a) and 2(b), it appears that the band-gap problem has a different character in the ZnO solid-state system and in the F atom separated open system: Whereas in the free-atom case, the band-gap error can easily be avoided by enforcing the "correct" localization (one electron or hole per atom), the amount of localization needed to significantly increase the quasiparticle band gap in the case of a semiconductor [cf. Fig. 2(a)] is incompatible with the physically correct delocalized nature of the free-electron and free-hole quasiparticle states. Thus, in contrast to atomic systems, the correction of the band gap E_g (quasiparticle gap) in solid-state systems requires generally information from outside the LDA or GGA, such as, e.g., from post-LDA methods or from experiment.

B. Host band-edge corrections and their effect on defect energies

Figure 3 illustrates for the case of a single donor the effect of the band-gap correction on the defect formation energy ΔH. For the ionized donor D^+, where the donor state is unoccupied, the primary effect of the band-gap correction is to extend the range of Fermi levels within the gap to $E_V^{LDA}+\Delta E_V<E_F<E_C^{LDA}+\Delta E_C$. Here ΔE_V and ΔE_C are the individual band-edge corrections for the VBM and the CBM, respectively, needed to recover the experimental band gap, $E_g(\mathrm{expt})=(E_C^{LDA}+\Delta E_C)-(E_V^{LDA}+\Delta E_V)$. As seen in Fig. 3 and in Eq. (1), the lowering of the VBM by ΔE_V results in lower formation energies of D^+ for Fermi levels low in the gap,

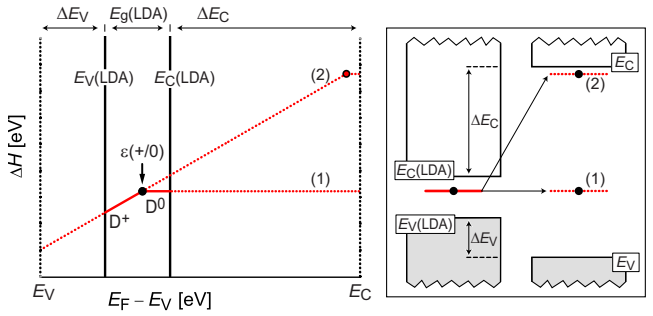


FIG. 3. (Color online) Schematic illustration of the effect on the formation energies (left) and single-particle energies (right) when the valence- and conduction-band edges in LDA are corrected by ΔE_V and ΔE_C toward the experimental gap $E_g = E_C - E_V$. Solid lines correspond to situation before correction and dotted lines to the situation after correction. The scale is chosen so as to illustrate the magnitude of corrections needed in ZnO. In general, the defect levels can be affected by the correction in varying degrees, as illustrated by examples (1) and (2).

whereas the upward shift of the CBM by ΔE_C causes increased $\Delta H(D^+)$ for Fermi levels high in the gap, thereby increasing the range of possible formation energies. One component in the VBM shift ΔE_V is due to self-interaction effects in occupied metal d shells whose orbital energies are generally too high in LDA and GGA. If such d states occur in the LDA calculation close to the VBM, e.g., in the case of Zn in ZnO (Refs. 6, 7, 39, and 40) or Cu in $CuInSe_2$ (Refs. 5, 8, and 79) or Cu_2O,[83,86,110] the VBM energy can be expected to lie too high in energy as well, due to p-d repulsion[80] between the metal d states and the anion p states in the valence band. Therefore, we used in Refs. 6, 7, 79, 83, and 110 the LDA+U (or GGA+U) method for the metal d states to determine the correction ΔE_V for the VBM.

We emphasize here that the individual band-edge shifts ΔE_V and ΔE_C, which determine the corrections for the charged-defect formation energies (see Fig. 3), need to be determined with respect to a *bulk-internal* potential reference. When ΔE_V and ΔE_C are determined, a *constant shift* of the potential due to a *bulk-external* source (e.g., the capacitorlike potential step due to a surface or interface dipole) needs to be avoided. This is because a constant shift of the external potential does not only shift the band-edge energies E_V and E_C but also the electrostatic energy of the charged defect, e.g., $E(D^+)$ in Eq. (1), so that the charged-defect formation energy $\Delta H(D^+)$, in fact, remains invariant. Such an undesired shift of the external potential can occur when, as done in Ref. 111, a ZnO(LDA)/ZnO(LDA+U) interface is constructed that can lead to the development of an interface dipole. Since the ensuing potential step causes a contribution to the VBM shift, it affects, e.g., $\Delta H(D^+)$ in Eq. (1) via the term E_V, and an error in $\Delta H(D^+)$ is introduced when the corresponding change in $E(D^+)$ due to the potential step is neglected.

Using a self-consistent method such as LDA+U, it is difficult, in principle, to determine the change in the band-edge energies with respect to an internal potential reference, since this reference can change during the self-consistent calculation as well. Therefore, we determined in Refs. 6 and 7 the change in the VBM energy in ZnO relative to the deep anion s state Γ_{1v}, which has a_1 symmetry (in zinc-blende notation[112]) and does not directly couple to the e_g and t_2 symmetries of the metal d states on which the LDA+U method was used. Also, we confirmed that the results were very similar when the anion-site average potentials (see Sec. II D 4) were used as a potential reference. Our finding of ΔE_V between −0.8 eV (LDA) and −0.7 eV (GGA) for U =7 eV (J=0) (Refs. 6 and 7) is consistent with the GW result of ΔE_V=−0.5 eV.[12] (Note that in this GW calculation the band gap and the too shallow Zn d band energies were not completely corrected.) Indeed, even though the GW method is presently not applicable in calculation of total energies for host+defect systems, it can be a useful augmentation to LDA for the purpose of determining ΔE_V and ΔE_C.[11,12,29] Given the observation[113] that, with respect to an internal potential reference, defect levels are often invariant under the band gap correction, the relative contributions of ΔE_V and ΔE_C to the band gap correction are very important, as they largely determine the position of the defect levels inside the gap after the gap correction [cf. Fig. 3, example (1)]. Our previously published VBM shifts of ΔE_V= −0.37 eV in $CuInSe_2$ and $CuGaSe_2$ (Refs. 7 and 79) and ΔE_V=−0.32 eV in Cu_2O (Ref. 83) have been shown to be consistent with GW calculations, which yield ΔE_V= −0.38 eV and ΔE_V=−0.22 eV for $CuInSe_2$ and Cu_2O, respectively.[114]

C. Determining defect level shifts by the perturbation-extrapolation method

In the case of the charge-neutral donor D^0 in Fig. 3, where the donor state is occupied, there is a contribution to ΔH if the donor level shifts to higher energies during the band-gap correction. The degree to which the donor level follows the band-edge correction ΔE_C determines the amount of correction in ΔH, as illustrated in Fig. 3. While shallow conduction-band-like donor states can be expected to fully track the CBM during band-gap correction (Sec. II D 2),[5,9] the magnitude of correction needed for deeper donor states is not generally known.

One way to determine the shift of a donor state during band-gap correction is the extrapolation of a band-gap-opening perturbation toward the experimental band gap.[34,37,39,40,115] This method employs the notion that the defect (donor) state $\psi_D(\mathbf{r})$ can be expanded in terms of the Bloch functions $\psi_{n,\mathbf{k}}(\mathbf{r})$ of the defect-free host, which form a complete basis:

$$\psi_{n,\mathbf{k}}(\mathbf{r}) = e^{i\mathbf{k}\mathbf{r}} u_{n,\mathbf{k}}(\mathbf{r}), \tag{14}$$

$$\psi_D(\mathbf{r}) = \sum_{n,\mathbf{k}} A_{n,\mathbf{k}} \psi_{n,\mathbf{k}}(\mathbf{r}). \tag{15}$$

Here, n is the band index, $\mathbf{k}$ is a wave vector in the Brillouin zone corresponding to the host unit cell, and $u_{n,\mathbf{k}}(\mathbf{r})$ is the lattice-periodic part of $\psi_{n,\mathbf{k}}(\mathbf{r})$, where the periodicity is that of the underlying host lattice, not that of the supercell in which the defect is calculated. If a (physical) band-gap-opening perturbation $\Delta\mathbf{H}_P$ is added to the pure-host Hamil-

tonian $\mathbf{H}_H$ via a multiplier $0 \leq \lambda \leq 1$, the host-band (single-particle) energies $e_{n,\mathbf{k}}$ are moved into direction of the correct experimental energies:

$$e_{n,\mathbf{k}}(\lambda) = \langle \psi_{n,\mathbf{k}} | \mathbf{H}_H | \psi_{n,\mathbf{k}} \rangle + \lambda \langle \psi_{n,\mathbf{k}} | \Delta \mathbf{H}_P | \psi_{n,\mathbf{k}} \rangle$$
$$= e_{n,\mathbf{k}}(0) + \lambda \frac{\partial e_{n,\mathbf{k}}(\lambda)}{\partial \lambda}. \quad (16)$$

If the same perturbation is applied to the "host+defect" system ($\mathbf{H}_H + \Delta \mathbf{H}_D$), one finds the (single-particle) energy e_D of the defect state as

$$e_D(\lambda) = \langle \psi_D | \mathbf{H}_H + \Delta \mathbf{H}_D | \psi_D \rangle + \lambda \langle \psi_D | \Delta \mathbf{H}_P | \psi_D \rangle$$
$$= e_D(0) + \lambda \frac{\partial e_D(\lambda)}{\partial \lambda}. \quad (17)$$

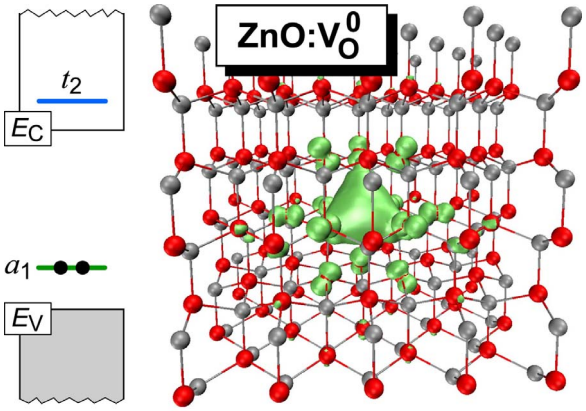
Within first-order perturbation theory (unchanged wave functions ψ_D upon application of $\Delta \mathbf{H}_P$), Eq. (17) is equivalent to

$$e_D(\lambda) = e_D(0) + \lambda \sum_{n,\mathbf{k}} A_{n,\mathbf{k}}^2 \frac{\partial e_{n,\mathbf{k}}(\lambda)}{\partial \lambda} \quad (18)$$

by the definition of $\psi_D(\mathbf{r})$ [Eq. (15)]. Thus, upon phasing in the perturbation, the energy of the defect state follows the corrections of the host states $e_{n,\mathbf{k}}$ in proportion to the respective squared coefficients $A_{n,\mathbf{k}}^2$ of expansion (15).

Once a value of $\lambda = \lambda_{ex}$ has been determined for the pure-host case from Eq. (16) such that the experimentally correct *host*-band energies $e_{n,\mathbf{k}}$ are recovered, the corrected *defect* energy e_D can be obtained in two ways: either from Eq. (17) when $\partial e_D / \partial \lambda$ is determined by calculation of e_D in the presence of the perturbation or from Eq. (18) when the coefficients $A_{n,\mathbf{k}}$ of expansion (15) are determined by projection of the defect state onto host bands.[116] Note that the defect energy e_D in Eq. (18) depends on the type of perturbation only through the respective shifts of the host-band energies $e_{n,\mathbf{k}}$. Thus, all perturbations that extrapolate the host states $e_{n,\mathbf{k}}$ to correct experimental band energies will give the same result for the extrapolated defect energy $e_D(\lambda_{ex})$. Under the assumption that the expansion of ψ_D in Eq. (15) is unaffected by the perturbation, the defect energy e_D is then also extrapolated correctly. Of course, the success of the perturbation-extrapolation scheme requires two conditions: (a) All host bands $(n,\mathbf{k})$ which significantly contribute to expansion (15) of the defect state must be "correctly corrected" by the perturbation $\Delta \mathbf{H}_P$ when λ is extrapolated to λ_{ex}. (b) The first-order perturbation theory arguments applied must be valid; i.e., the coefficients $A_{n,\mathbf{k}}$ must not change significantly during the band-gap correction.

Since a *physical* perturbation that corrects all host bands simultaneously toward their experimental energies is not known, practical perturbation-extrapolation methods use rather technical perturbation parameters for band-gap opening, such as, e.g., the energy cutoff of the plane-wave basis[34] or the parameter U in LDA+U,[37,39,40] which are then extrapolated to generally *unphysical* values such that the extrapolated band gap coincides with the experimental gap. In such schemes, only the difference $e_g = e_C - e_V$ is, in fact, corrected, whereas host-band states other than the CBM and the VBM are extrapolated to uncontrolled energies.



	R [Å]	s [%]	p [%]	d [%]
V_O (empty sphere)	1.80	29.1	0.1	0.0
NN-Zn (4 atoms)	1.22	4.4	7.5	4.8
NNN-O (12 atoms)	0.81	0.4	26.7	0.0

FIG. 4. (Color online) Left: The charge-neutral O vacancy V_O^0 in ZnO introduces two defect-localized states: An occupied a_1 symmetric state deep in the gap and an unoccupied t_2 level resonant inside the conduction band. Right: Isosurface plot (green) of the wave-function square of the a_1 gap state. Bottom: Angular momentum decomposed contributions to the occupied a_1 level from the vacancy site, the NN Zn shell, and the next-nearest-neighbor (NNN) O shell. R is the atomic-sphere radius used for the projection (touching spheres).

Employing the notion that a defect level is constructed in some relative proportions from both conduction- and valence-band states [cf. Eq. (15)], one can expect,[40] subject to condition (b) above, that by applying a perturbation that shifts the valence and conductions bands relative to each other, the defect level shifts according to these proportions. For illustration, consider a defect state that is constructed from only the VBM and CBM,

$$\psi_D(\mathbf{r}) = A_V \psi_V(\mathbf{r}) + A_C \psi_C(\mathbf{r}). \quad (19)$$

In this case, the extrapolated defect level $e_D(\lambda_{ex})$ follows the band-edge corrections ΔE_V and ΔE_C in proportion to the squares of the coefficients A_V and A_C [cf. Eq. (18)],

$$\varepsilon_D(\lambda_{ex}) = \varepsilon_D(0) + A_V^2 \Delta E_V + A_C^2 \Delta E_C. \quad (20)$$

In this rather idealized case, the use of an unphysical perturbation parameter appears to be justified, as all band-gap-opening perturbations would lead to the same result for the defect energy relative to the VBM because the gap correction $\Delta E_g = \Delta E_C - \Delta E_V$ is the same by construction. However, if the defect state is constructed mostly from states other than the VBM and the CBM, the use of an unphysical extrapolation parameter can lead to uncontrolled errors in the defect level energy because in this case the correction of the defect level in Eq. (18) is based mostly on the behavior of host states $e_{n,\mathbf{k}}$, which are extrapolated to uncontrolled and possibly incorrect energies. As we show in Sec. III E 2 for the example of V_O in ZnO, two different perturbations parameters, which superficially may seem equally justified, lead to

entirely different predictions for the defect level of V_O. Thus, it is essential to choose a perturbation that reproduces a physically correct *band structure* after extrapolation, not just a correct band gap [cf. condition (a) above]. Indeed, the limited expansion in Eq. (19) is severely incomplete for localized states such as that of V_O in ZnO, shown in Fig. 4, because the construction of localized defect states from host-band states generally requires the superposition of a very large number of host Bloch functions [cf. Eq. (15)] from the *entire* Brillouin zone of the host.[117] Since the projections A_V and A_C of a localized state onto the delocalized band-edge states ψ_V and ψ_C [cf. Eq. (19)] vanishes in the limit of large supercells (due to negligible overlap between defect and host wave functions), we conclude that the shift of the localized defect level e_D upon gap correction is not directly connected with the shifts of the band-edge states e_V and e_C, in contrast to the assertion made in Ref. 40.

D. Self-consistent band-gap corrections via LDA+U

It was recognized by Christensen[118] that a band-gap correction can be achieved through the addition of empirical external potentials in the self-consistent LDA calculation. This method has been applied, e.g., to improve the description of the *p-d* coupling in Mn chalcogenides[119] and to predict band-structure properties of ordered III-V alloys.[120] A modification of this method with nonlocal (angular momentum–dependent) external potentials was used to describe defect states due to substitutional nitrogen in the conduction band of GaAs.[121] Such *l*-dependent potentials can also be created by the LDA+U method when applied to cation *s* orbitals (Ref. 44 and present work) or to anion *s* orbitals.[82,83] The latter requires a negative value for U and opens the gap essentially by reducing the dispersion of the conduction band, which, however, leads often to an overestimation of the electron effective mass.

LDA+U creates an attractive potential for occupied orbitals (e.g., LDA+U_d for Zn *d*) but a repulsive potential for unoccupied orbitals (e.g., LDA+U_s for Zn *s*), which becomes apparent when considering the simplified LDA+U description of Dudarev *et al.*,[85] in which the "+U" part of the potential is

$$V_{m,\sigma}^{+\mathrm{U}} = (U - J)\left(\frac{1}{2} - n_{m,\sigma}\right) \tag{21}$$

in the diagonal representation. Here, $n_{m,\sigma}$ is the occupation (partial charge) $0 \leq n_{m,\sigma} \leq 1$ of the m sublevel of spin σ. Thus, LDA+U_s moves the conduction-band states toward higher energies when applied to the Zn *s* states that mainly contribute to empty conduction-band states in ZnO.[44] Due to hybridization, however, there is also a Zn *s* contribution to the valence-band states. leading to a nonzero Zn *s* partial charge of 0.48*e* in LDA ($n_{\mathrm{Zn}\,s}=0.24$). In contrast, LDA+U_d lowers the energy of the VBM (see Sec. III B), but after application of LDA+U_d ($U_{\mathrm{Zn}\,d}=7$ eV) alone the ZnO gap is still underestimated by ~50%. In a combined LDA+$U_{s/d}$ approach, the full band-gap correction is achieved with a smaller (downward) correction of the VBM ($\Delta E_V \approx -0.7$ eV) due to the lowered Zn *d* energy, and a larger (upward) correction of the CBM ($\Delta E_C \approx +2.0$ eV), due to the increased Zn *s* energy, in accord with the general finding of GW calculations that the band gap is usually corrected mostly by an upshift of the conduction bands.[11,12,29] Using the LDA+$U_{s/d}$ method, the experimental band gap of ZnO is recovered for the values $U_s=38$ eV and $U_d=4$ eV for Zn *s* and Zn *d*, respectively. Here, the value for U_d required to achieve agreement with photoemission data is smaller than in the conventional LDA+U_d method[7] because LDA+U_s additionally lowers the Zn *d* energy. The reason for the lowering of the Zn *d* energy is that the *s*-repulsive potential reduces the Zn *s* partial charge (0.18*e* in LDA+$U_{s/d}$) and, hence, reduces the screening of the Zn ionic charge. In contrast to U_d, which, within some bounds, has a physical meaning as the Coulomb-interaction parameter, U_s should be regarded instead as an empirical parameter for the repulsive Zn *s* potential chosen such that the experimental band gap is reproduced.

A considerable shortcoming of the LDA+$U_{s/d}$ method is that it does not reproduce the experimental wurtzite lattice structure of ZnO. When the cell-external lattice parameters are relaxed, we obtain the hexagonal BN structure with fivefold coordination and nearest-neighbor distances of 2.07 Å (in plane) and 2.17 Å (along the *c* axis), similar to the metastable hexagonal modification of MgO found in Ref. 122. The relaxation to a structure with higher coordination number is probably caused by the exaggerated depletion of the Zn *s* partial charge from 0.48*e* in LDA to only 0.18*e* in LDA+$U_{s/d}$, which renders ZnO more ionic than it should be (for comparison, the Mg *s* partial charge in MgO is 0.14*e*). This spurious increase in the ionicity can in some cases influence the energy of defect levels (see below). Considering these shortcomings of LDA+U applied to *s* states, we conclude that LDA+$U_{s/d}$ may not generally be a quantitative total-energy method for band-gap-corrected defect calculation. We recently developed an alternative method for correcting the band gap within the self-consistent calculation which is based on nonlocal external potentials (NLEPs).[123] In contrast to the LDA+U potential, the NLEPs do not depend on the partial charge (sublevel occupation) and provide a general and flexible method for empirically fitting the band structure as well as structural parameters for targeting values, e.g., taken from experiment. Applied to ZnO, we achieved with the NLEP method a good description of the ZnO band structure while reproducing the correct wurtzite ground-state structure.[123]

E. Comparison of different correction schemes for the case of V_O in ZnO

In this section, we discuss and compare different band-gap correction schemes for a specific example, namely, the oxygen vacancy in ZnO, a case which has been subject of considerable controversy and debate in literature.[6,7,33–40,42,44] A wide range of formation energies have been proposed for V_O, as illustrated in Fig. 1. Since, $\Delta H(V_O^0)$ in the uncorrected GGA calculation is hardly affected by supercell-size effects, as shown in Fig. 5, the spread of literature results originates mainly from the way the band gap was corrected. Therefore,

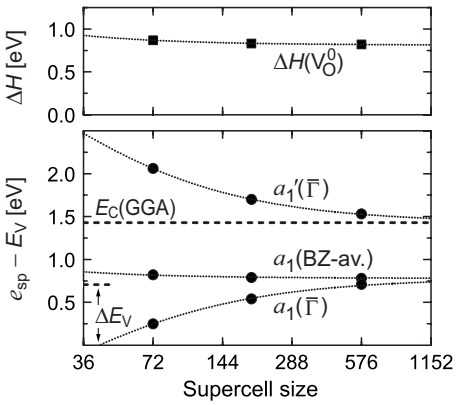


FIG. 5. Formation energy (squares) and single-particle energies e_{sp} (circles) of the neutral O vacancy in ZnO as a function of supercell size, determined in GGA. Shown is e_{sp} for the a_1 symmetric defect state of the charge-neutral V_O, obtained either at $\bar{\Gamma}$, the center of the Brillouin zone of the supercell, or as an average over the Brillouin zone (BZ av) with appropriate weights. Also shown is $a_1'(\bar{\Gamma})$, the first unoccupied conduction-band state at the zone-center. The single-particle energies are given relative to the corrected energy of the VBM, $E_V(\mathrm{GGA})+\Delta E_V$.

we now compare and assess the results of different band-gap correction schemes.

1. "Band-edge-only" correction

Considering that the appropriate correction for the energy levels of localized defect states is generally independent of the band-edge corrections ΔE_V and ΔE_C of the host, as discussed in Sec. III C, and taking into account the general observation that localized defect states do not respond as strongly to external perturbations (e.g., pressure) as do the host band-edge states,[124] one can expect that the strongly localized defect state of V_O (cf. Fig. 4) is much less affected by the band-gap problem than the band edges of the ZnO host. Accordingly, the most conservative correction approach is to avoid an explicit correction to the defect state of V_O and to correct only the band edges E_C and E_V which exhibit the band-gap error. This "band-edge-only" approach, as employed in our previous works,[6,7] corresponds to example (1) in Fig. 3. Here, the LDA or GGA *formation energy* of the charge-neutral V_O^0 state is not changed by the gap correction. In contrast, the position of the corrected *defect level* depends strongly on the relative contributions of ΔE_V and ΔE_C to the gap correction. For example, the single-particle energy of the a_1 level of V_O^0 lies at $E_V+0.1$ eV in the uncorrected GGA calculation and moves to $E_V+0.8$ eV after application of the VBM shift $\Delta E_V=-0.7$ eV (Fig. 5). Here it is important that the band-edge shifts ΔE_V and ΔE_C are determined with respect to the (internal) potential reference of the GGA calculation, so as to maintain consistency between the position of the defect level in the supercell calculation and the band-edge energies E_V and E_C which are obtained by applying corrections in the defect-free pure-host calculation (see Sec. III B). Additional justification for the band-edge-only correction method can be deduced from the recent observation that many defect levels, in particular those with localized wave functions, align between a GGA and an approximately gap-corrected hybrid-DFT calculation.[113]

The band-edge-only correction method was very successful in describing the anion-vacancy-related optical transition levels in ZnO (Refs. 6 and 7) and other II-VI systems.[79] The low formation energy $\Delta H(V_O^0)=0.8$ eV in GGA (Table I and Fig. 5) explains the formation of high V_O concentrations under Zn-rich conditions, which is further supported by recent experiments (see Sec. V). [Note that under the decisive Zn-rich growth condition, LDA and GGA yield similar formation energies for V_O (see Table I).] Very recently, additional theoretical support for low V_O formation energies was drawn from the self-interaction-corrected calculations by Pemmaraju *et al.*,[42] and by Oba,[43] both finding a formation energy close to that in the GGA.[6]

2. Ambiguity of the perturbation-extrapolation scheme

Before the band-gap correction, the occupied a_1 symmetric defect level of V_O^0 in ZnO lies just above the VBM, e.g., at $E_V^{\mathrm{GGA}}+0.1$ eV in GGA (Fig. 5), when determined from the appropriate Brillouin-zone average (see Sec. IV B 2). If, upon band-gap correction this level follows the upward shift of the CBM, then it can be expected that the formation energy will increase (Fig. 3), since, on an absolute scale, most of the band-gap correction occurs in the conduction band (cf. Secs. III B and III D). As described above, the perturbation-extrapolation method provides a prediction of the extent to which a defect state follows the respective band edges during the correction. However, as we pointed out in Sec. III C, the result of the extrapolation may not be unique if a perturbation parameter is used that yields only the correct band gap after extrapolation but not the correct full band structure. In order to test whether the prediction for the V_O level in ZnO is affected by this uncertainty, we calculate the equilibrium transition levels of V_O in ZnO with an extrapolation based on two different perturbation parameters: First, when the LDA $+U$ method is applied on the Zn d orbitals (U_d), the band gap is increased mainly via the reduced p-d repulsion lowering the VBM energy. Second, LDA$+U$ applied on Zn s orbitals (U_s) mainly causes an upshift of the conduction band (see Sec. III D).

While both methods open the band gap and, by construction, give the same band-gap correction $\Delta E_g=\Delta E_C-\Delta E_V$, the extrapolated transition levels of V_O lead to entirely different predictions, as shown in Table I: In the U_d-extrapolation scheme, the equilibrium transition energy $\varepsilon(2+/0)=E_V+1.9$ eV lies in the upper part of the band gap, whereas in the U_s-extrapolation scheme, it moves to $E_V-0.6$ eV even below the VBM, which indicates that the 2+ state is unstable for all Fermi levels within the ZnO band gap. Also, the extrapolated single-particle energies of the occupied a_1 level of V_O^0 differ considerably, by ~ 1 eV. Similarly, large differences occur when comparing the extrapolated results based on LDA$+U_d$ (Ref. 37) with those using the energy cutoff of the plane-wave basis as perturbation parameter.[34] As noted by Vlasenko and Watkins,[125,126] the two extrapolation schemes in Refs. 34 and 37 favor different possible interpretations about the $\varepsilon(0/+)$ level of V_O in optically detected

TABLE I. Comparison of different methods for determining the band-gap-corrected transition levels and formation energies of V_O in ZnO: LDA+$U_{Zn\,d}$+extrapolation (U_d extr.), LDA+$U_{Zn\,s}$+extrapolation (U_s extr.), LDA+$U_{Zn\,s}$+$U_{Zn\,d}$ ($U_{s/d}$), and LDA or GGA+band-edge-only correction ($\Delta E_V+\Delta E_C$), as published in Refs. 6 and 7. The U_d- and U_s-extrapolation methods are based on calculations with $U_s=U_d=0$ and $U_d=4.7$ eV, as in Refs. 37 and 40, and $U_s=10$ eV, which then are extrapolated. Given are the U parameters, the $\varepsilon(2+/0)$ and $\varepsilon(+/0)$ equilibrium transition energies, the single-particle energy e_{sp} of the a_1 gap state of V_O^0, and the formation energy of V_O^0 under Zn-rich ($\Delta\mu_{Zn}=0$) and O-rich ($\Delta\mu_O=0$) conditions. All numbers are in eV and correspond to the band-gap-corrected situation, $E_g=E_g$(expt). Image charge and potential-alignment corrections were applied to charged supercell energies. Single-particle energies were determined from the appropriate Brillouin-zone average (cf. Fig. 5 and Sec. IV B 2).

	U_d extr. (LDA)[a,b]	U_s extr. (LDA)[b]	$U_{s/d}$ (LDA)[c,d]	$\Delta E_V+\Delta E_C$ [e] (LDA)[d]	$\Delta E_V+\Delta E_C$ [f] (GGA)[b]
U_d	17.2		4.0	7.0	7.0
U_s		37.3	38.0		
$\varepsilon(2+/0)=E_V+$	1.93	−0.61	0.61	1.60	1.30
$\varepsilon(+/0)=E_V+$	n/a[g]	n/a[g]	0.79	0.94[g]	n/a[g]
$e_{sp}(V_O^0-a_1)=E_V+$	1.3	0.3	1.0	0.9	0.8
$\Delta H_f(V_O^0)$, Zn rich	3.23[h]	1.16[h]	1.98[h]	1.14	0.83
$\Delta H_f(V_O^0)$, O rich	6.86	4.79	5.61	4.67	3.76

[a]Calculation comparable to that in Refs. 37 and 40.
[b]Theoretical (relaxed) host lattice constants.
[c]Calculation comparable to that in Ref. 44.
[d]Experimental host lattice constants.
[e]Reference 7.
[f]Reference 6.
[g]The energy of V_O^+ cannot reliably be determined in LDA, GGA, or their $+U_d$ extensions because the a_1 level lies above the calculated CBM in this charge state (see Sec. III E 3). In Ref. 7, the $\varepsilon(+/0)$ transition energy was deduced from an interpolation in the configuration coordinate diagram.
[h]Since the elemental Zn-metal reference energy is not well defined in LDA+U, the experimental heat of formation $\Delta H_f=-3.63$ eV of ZnO was used for the Zn-rich condition.

magnetic-resonance (ODMR) experiments (see also below, Sec. III E 3).

According to the discussion in Sec. III C, the reason for the large differences between the U_d- and U_s-extrapolation schemes lies in the fact that in either case only the difference e_C-e_V is corrected, whereas the energies of states other than the CBM and the VBM, including those with large contributions to expansion (15), are extrapolated to different and not necessarily physically correct energies. Indeed, in the U_d-extrapolation scheme, the Coulomb parameter for Zn d was extrapolated to an unphysically large value of $U=17.4$ eV in Refs. 37 and 40, leading to an extrapolated energy of the Zn d band at E_V-10 eV, considerably deeper than the experimental position between $E_V-7.5$ eV (Ref. 127) and $E_V-8.8$ eV.[128,129] Apart from the d-band energies, other materials properties are also extrapolated to unphysical values; e.g., the extrapolated lattice constant is more than 7% smaller than the experimental one.[37,40]

In Ref. 40, Janotti and van de Walle suggested that defect levels follow the corrections of the VBM and of the CBM in proportion of the fractions VB and $CB=1-VB$, respectively [cf. the coefficients A_C^2 and A_V^2 in Eq. (20)], which are interpreted as measures of the valence- and conduction-band characters. This assumption led them to speculate,[40] "The assumption that the transition levels associated with the oxygen vacancy do not shift when the conduction band is corrected is equivalent to saying that the a_1 state has purely valence-band character." However, as we showed in Sec. III C, localized defect states cannot be decomposed into just two contributions from the valence- and conduction-band states. Due to the incompleteness of the limited expansion in Eq. (19), the correction of the defect state cannot be directly linked to the corrections of either the VBM or the CBM, or a combination thereof in fractions VB and $CB=1-VB$. Since localized deep levels are constructed from valence- and/or conduction-band states from throughout the Brillouin zone (see Sec. III C), their behavior during the band-gap correction depends on the detailed behavior of the entire band structure upon applying the perturbation. Indeed, it is the different behavior of states other than the VBM and the CBM that leads to the extremely different predictions of different perturbations (e.g., LDA+U_d versus LDA+U_s) for the extrapolated energy of the V_O defect level in ZnO (see Table I).

In the U_d-extrapolation method, our calculated transition energy of the $\varepsilon(2+/0)=E_V+1.9$ eV level lies somewhat lower in the gap than in Ref. 37, despite the nominally identical parameters used here (see Table I). This difference is mainly due to the application of image charge corrections in the present work, which, as we show in Sec. IV A, yield

important contributions to the defect formation energy. Omitting these corrections,[40] we would extrapolate the transition level to $\varepsilon(2+/0)=E_V+2.3$ eV, close to the result of Refs. 37 and 40. Thus, the present theoretical analysis suggests that the high energy of the $\varepsilon(2+/0)$ donor transition in the U_d-extrapolation scheme of Janotti and van de Walle[37,40] results mostly from an exaggerated lowering of the VBM energy relative to the V_O level, due to extrapolation of U_d to an unphysically large value of $U_d=17$ eV (cf. Table I), but, to a lesser extent, also from the neglect of image charge corrections.

We further give in Table I the absolute formation energies $\Delta H(V_O^0)$ obtained by application of the U_d- and U_s-extrapolation schemes directly to ΔH (analogous to the extrapolation of defect levels, as discussed above). Since the use of the LDA+U method for Zn precludes the use the calculated total energy of Zn metal as a reference state (see Sec. II C 2), we used in these cases the experimental heat of formation $\Delta H_f=-3.63$ eV (Ref. 130) of ZnO to determine ΔH under Zn-rich conditions ($\Delta\mu_O=\Delta H_f$). The formation energy $\Delta H=3.2$ eV in the U_d-extrapolation method is close to the value of 3.7 eV found by Janotti and van de Walle.[40] [Note, however, that they did not extrapolate ΔH directly but rather reconstructed $\Delta H(V_O^0)$ and $\Delta H(V_O^+)$ by using the extrapolated transition levels and assuming that $\Delta H(V_O^{2+})$ is correct for $E_F=E_V$ in the nonextrapolated situation for $U=4.7$ eV in the LDA+U calculation.] However, similarly as for the transition levels, we again observe large differences between the two choices of the perturbation parameter, U_d versus U_s, the latter predicting a much lower formation energy of only $\Delta H=1.2$ eV (Table I). These findings corroborate our conclusion that perturbation-extrapolation schemes that require the extrapolation of the perturbation parameter toward unphysical values to sufficiently open the band gap are arbitrary as to the choice of the perturbation parameter and cannot reliably be used to predict corrected defect levels.

3. V_O in the self-consistently gap-corrected LDA+$U_{s/d}$ approach

In the combined LDA+$U_{s/d}$ approach, the corrected V_O defect levels can be determined within the self-consistent calculation, which is particularly advantageous for the singly charged V_O^+ state, which we discuss now in more detail: Experimentally, the g factor observed for V_O^+ in electron paramagnetic resonance (EPR) is close to the free-electron value, implying a localized deep gap state.[131] In contrast, the energy of the a_1 state lies above the CBM in LDA and in LDA+U_d, which implies a shallow effective-mass-like state.[7] While in small supercells, band-filling effects (cf. Secs. II D and IV B) lead to a partial occupation of the a_1 resonance in the conduction band, in the limit of large cells the electron will eventually occupy the CBM; i.e., increasing the cell size leads to the gradual transition $V_O^+ \rightarrow V_O^{2+}+e$ ($a_1^1 \rightarrow a_1^0+e$). As a result, the calculated $\Delta H(V_O^+)$ is not unambiguously defined and depends on the actual supercell size used (cf. Sec. IV B). Thus, the V_O^+ state cannot be realistically described in LDA or LDA+U_d. In order to avoid the problem of spurious hybridization of the a_1 level with the conduction band, we determined in Ref. 7 the formation energy of V_O^+ by an interpolation of the configuration coordinate diagram (see Sec. III F). These difficulties can be avoided in the self-consistently band-gap-corrected LDA+$U_{s/d}$ method where the singly occupied a_1 level of V_O^+ occurs correctly inside the band gap.

In Refs. 125 and 126, Vlasenko and Watkins gave two possible interpretations of their ODMR experiments, one of which placed the $\varepsilon(+/0)$ transition energy of V_O at $E_V+0.9$ eV, whereas the second would require an energy much higher in the gap at $E_V+2.5$ eV. Notwithstanding the above discussed ambiguous energy of the V_O^+ state, Janotti and van de Walle[40] interpreted their result of $\varepsilon(+/0)=E_V+2.0$ eV in the U_d-extrapolation scheme as supporting the second possibility. In contrast, both our published[7] value of $\varepsilon(+/0)$ and the value obtained here with the $U_{s/d}$ method (see Table I) are close to the experimental $\varepsilon(+/0)$ level in the first interpretation, which was also supported by the recent $U_{s/d}$ calculation of Paudel and Lambrecht,[44] as well as by the band-gap-corrected hybrid-DFT calculations of Patterson[23] using the B3LYP (Refs. 21 and 132) functional.

Regarding the $(2+/0)$ equilibrium transition level, LDA+$U_{s/d}$ finds $\varepsilon(2+/0)=E_V+0.6$ eV ($E_V+0.8$ eV in Ref. 44), notably even lower in the gap than in the band-edge-only correction method ($E_V+1.3$ eV, see Table I). Comparing with the results of self-interaction correction ($E_V+1.1$ eV, Ref. 42) and hybrid-DFT ($E_V+2.2$ eV, Ref. 43), we see that different self-consistently gap-corrected methods still give a large range of predictions for the position of the V_O defect level even though they do not rely on *a posteriori* corrections. Also, the formation energies under O-poor conditions, $\Delta H(V_O^0)=0.7$ eV in SIC (Ref. 42) and 1.0 eV in hybrid-DFT (Ref. 43), are close to the GGA result of 0.8 eV (Table I), but $\Delta H(V_O^0)=2.0$ eV in LDA+$U_{s/d}$ is considerably larger than in GGA. Considering the expectation that the relative contributions of the band edge corrections ΔE_V and ΔE_C to the gap correction will be decisive for the position of the defect levels in the corrected gap (see Sec. III E 1), a particular concern is the relative magnitude of ΔE_V and ΔE_C in different self-consistently gap-corrected approaches. Indeed, it was found in Ref. 29 that screened exchange [which, similar as some hybrid-DFT variations (Refs. 43 and 133) includes HF-exchange via a model screening function] increases the band gap primarily by lowering E_V, whereas GW primarily raises E_C.[11,12,29] Thus, the accuracy of the shifts ΔE_V and ΔE_C in e.g., the different variations of hybrid-DFT should be addressed in future band-gap corrected defect calculations.

4. Prediction of persistent photoconductivity due to V_O in different correction methods

Finally, we consider the optical transitions that were predicted to lead to persistent photoconductivity (PPC) in Refs. 6 and 7. After two successive optical excitations, $V_O^0 \rightarrow V_O^+ +e$ ($a_1^2 \rightarrow a_1^1+e$) and $V_O^+ \rightarrow V_O^{2+}+e$ ($a_1^1 \rightarrow a_1^0+e$), the doubly charged vacancy is formed with two electrons being excited to the CBM. This $V_O^{2+}+2e$ state was predicted[7] to exist as a metastable state; i.e., the electron capture process that leads back to the V_O^0 ground state requires the activation of an energy barrier. We here address two related issues that remain unsettled or controversial, comparing the trends of dif-

ferent band-gap correction methods including also recent hybrid-DFT results by Patterson[23] and the self-interaction-correction calculations by Pemmaraju *et al.*[42] First, the experimental assignment of the excitation energy for the $V_O^0 \rightarrow V_O^+ + e$ transition remains ambiguous, as discussed in detail in Sec. V. Theoretically such optical excitation energies can be calculated from total-energy differences according to Eqs. (3) and (4). For the sake of comparability with the data given for hybrid-DFT and SIC in Refs. 23 and 42, however, we here regard the respective single-particle energies. Indeed, we found before[6,7] that the difference [see Eq. (5)] is small in this specific case; i.e., the optical transition energy is only about 0.2 eV larger compared to the estimate based on the single-particle energy of the a_1 state of V_O in the gap (the difference is expected to be even smaller in hybrid-DFT or SIC calculations; cf. Sec. II B). Notably, the result of our previous band-edge-only correction method of an excitation energy $\varepsilon_O(0/+;e)=2.8$ eV,[7] corresponding to a single-particle energy of the a_1 state of V_O^0 at $E_V+0.9$ eV (Table I), is very closely reproduced by all the self-consistently band-gap-corrected methods. That is, the a_1 state lies at $E_V+1.0$ eV in the present LDA$+U_{s/d}$ result (Table I), in agreement with the finding of Paudel and Lambrecht,[44] at $E_V+0.7$ eV and 1.0 eV in the hybrid-DFT calculations of Refs. 23 and 43, respectively,[134], and at $E_V+0.8$ eV in the SIC calculation.[42] (The extrapolation methods based on U_d and U_s give somewhat higher and lower energies for the a_1 state, respectively; see Table I). Thus, the agreement among these different theories strongly supports the assignment[7] of the photoluminescence excitation threshold for the green luminescence, observed at 3.1 eV (see also Sec. V), to the first excitation of the O vacancy, $V_O^0 \rightarrow V_O^+ + e$.

The second open issue related to the PPC phenomenon is the energy of the empty a_1^0 level in the doubly charged V_O^{2+} state. The emergence of an energy barrier, which protects the conductive metastable state from spontaneous relaxation into the nonconductive V_O^0 ground state by electron capture, requires the empty a_1^0 level of V_O^{2+} to be located as a resonance inside the conduction band, rather than to occur as a gap state. In the uncorrected LDA calculation, the a_1 resonance of V_O^{2+} lies far inside the LDA conduction band, suggesting that it would still be resonant when the CBM is corrected by a rigid shift.[7] In contrast, as pointed out by Paudel and Lambrecht,[44] the a_1^0 level of V_O^{2+} lies inside the band gap around $E_V+2.0$ eV in the LDA$+U_{s/d}$ method, which contradicts the PPC model for O vacancies in ZnO. However, this behavior is not present in hybrid DFT (Refs. 23 and 43) and SIC,[42] which do not show a (single-particle) defect state inside the band gap for the V_O^{2+} state. There, the lowest unoccupied state in the V_O^{2+} defect calculation is a strongly dispersive band, which has to be identified with the perturbed-host state that is formed due to the interaction between the host conduction band and the resonant a_1 state at higher energies (see Sec. IV B 2). Similarly, no gap state is created when the upshift of the CBM in LDA$+U_{s/d}$ is achieved by applying U_s on the O s orbitals instead of the Zn s orbitals (cf. Secs. II C 2 and III D). Thus, the appearance of the a_1 defect level inside the gap for the V_O^{2+} state appears to be a peculiarity of the LDA$+U$ method when applied to Zn s orbitals, which is probably related to the artificially increased ionicity of ZnO

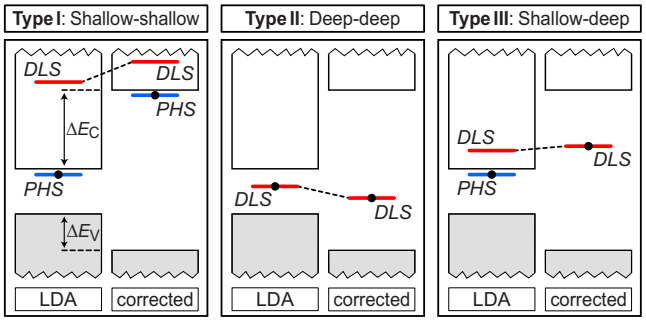


FIG. 6. (Color online) Schematic illustration three qualitatively different behaviors of defect level during band-gap correction. If the primary defect level, i.e., the defect-localized state (DLS, red) is resonant inside the conduction band, the electron is released to a secondary, conduction-band-like perturbed-host state (PHS, blue). In this case, the defect exhibits shallow behavior. If the DLS lies inside the gap, the defect exhibits deep behavior. Type I: Shallow in LDA, shallow after correction. Type II: Deep in LDA, deep after correction. Type III: Shallow in LDA, deep after correction.

in LDA$+U_{s/d}$ (cf. Sec. III D). Indeed, we found before[135,136] that in the more ionic oxides, such as MgO, CaO, and TiO_2, the V_O gap level has a much less pronounced tendency to shift to higher energies upon ionization, compared to the oxides ZnO and In_2O_3, which have a significant covalent character. Accordingly, O vacancies are expected to cause PPC in ZnO and In_2O_3 (Refs. 6 and 7) but not in MgO, CaO, or TiO_2.[135,136]

F. Classification scheme for defect level corrections

Since, as discussed in Sec. III E, current schemes for LDA correction do not represent a universal cure for the band-gap problem in defect calculations, we now turn to developing a general classification of different defect behaviors that require different treatments when the band edges are corrected. These defect behaviors are characterized by the energies of the single-particle defect levels relative to the band edges: When a defect is introduced into a semiconductor host lattice, it generally creates defect-localized states (DLS), which can either fall inside the gap or occur as resonances inside the host bands.[7] These DLS typically originate from the orbital interaction between the atomic orbitals of the impurity atom and the dangling bonds of the host.[137] In case of vacancies, these DLS are created due to the interaction between the dangling-bond orbitals. For example, the four Zn-site-centered dangling bonds around V_O in ZnO originate from the Zn (s/p)-O p atomic orbital interaction (Fig. 4) and combine to form a lower-energy a_1 symmetric level and a triply degenerate t_2 symmetric level resonant inside the conduction band (Fig. 4).[112] As discussed in Ref. 7, the DLS can, as a function of the defect potential, either anticross with the host-band edge (in case the DLS and band edge have the same symmetry representation) or cross the band edge (in case the DLS and band edge have different symmetry representations). Thus, the energy of the defect-induced DLS relative to the host bands determines whether a defect behaves as deep or shallow[7] and, hence, in which way

corrections should be applied to the defect formation energy. For example, a shallow-donor state is formed when the primary DLS occurs as a resonance inside the conduction band and releases the electron into the host conduction band. The resulting unoccupied (ionized) charged donor leads to the formation of a shallow, effective-mass-like secondary state, i.e., the perturbed-host state (PHS).[7] Since, however, the donor concentrations corresponding to usual supercell sizes on the order of ~100 atoms correspond normally to the case of degenerate doping, the PHS does not appear as a gap state in the calculation. Instead, Moss-Burstein-type band-filling effects raise the Fermi level above the CBM. These band-filling effects are associated with a considerable and strongly supercell size-dependent increase in the donor formation energy, and need to be corrected to obtain the formation energy for the situation of dilute doping. In the case of deep defects where the primary defect state (i.e., the DLS) occurs inside the band gap, such size-dependent band-filling effects do not exist (see Sec. IV B 1).

In Fig. 6, we distinguish three general types of defect behaviors which require different treatments when the band gap is corrected. The V_O defect in ZnO is a remarkable defect in the sense that it assumes all three behaviors when the charge states is increased from 0 to 2+.

Type I: Shallow behavior before gap correction and shallow behavior after gap correction (Fig. 6, left). If the primary DLS lies so high in energy that it exceeds the energy of the CBM even after the shift of the CBM by ΔE_C (Fig. 6, left), it can be expected that it is still resonant inside the conduction band after the band-gap correction. Since, the PHS that carries the donor electron is derived from the host-band structure, it can be expected that PHS-like donor levels follow the correction of the CBM. Therefore, the "shallow-donor correction"[5] (see Sec. II D 2) should be applied; i.e., the formation energy is to be corrected by $z_e \Delta E_C$, where z_e is the number of electrons occupying the PHS (e.g., $z_e=1$ for the shallow Te_{As} donor in GaAs in its charge-neutral state; see Sec. IV B 1). This situation corresponds to example (2) in Fig. 3. Due to the defect-to-host (DLS-to-PHS) charge transfer, the occupied donor state ψ_D has a large overlap with the CBM state of the pure host, i.e., $A_C^2 \approx 1$ in Eq. (19), and the prediction in Eq. (20) of the extrapolation scheme (Sec. III C) yields the same result for the donor transition energy as the shallow-donor correction $z_e \Delta E_C$. Type-I behavior is assumed by V_O in ZnO in the metastable shallow state ($V_O^{2+}+2e$), where the atomic configuration is that of the ionized $q=+2$ state and where two electrons occupy the PHS (see Sec. III E 4 above).

Type II: Deep behavior before gap correction and deep behavior after gap correction (Fig. 6, center). If the primary DLS occurs inside the uncorrected gap, the defect state exhibits deep and localized behavior. In this case, the formation energy is not affected by the Moss-Burstein shift, and the band-filling correction Eq. (6) vanishes. Thus, the formation energy of type-II defects such as V_O^0 in ZnO (see Sec. III E 2 above) converges typically very quickly with supercell size (Fig. 5), with no further finite-size corrections needed in the case of the neutral charge state (see also Sec. IV B 1).

Due to the localized nature of type-II defects, it cannot be expected that the donor state follows the CBM correction ΔE_C, and practical perturbation-extrapolation schemes that correct only the band gap but not the entity of all host-band states will generally fail to correctly predict how the energy of the defect level changes upon band-gap correction (see Sec. III C). Fortunately, however, the defect states of type-II defects are often rather well described in LDA, so that no defect-specific LDA correction is needed, i.e., the band-edge-only correction is often sufficient (see Sec. III E 1). This situation corresponds to example (1) in Fig. 3. It is still important to correctly determine the correct band-edge shifts ΔE_V and ΔE_C with respect to the original LDA potential reference (see Sec. III B) and, hence, with respect to the defect level. Some defects may, however, require additional corrections. For example, the DLS of transition-metal impurities lies usually at too high energy[138] due to self-interaction effects in strongly localized occupied levels. In this case, defect-specific corrections can often be applied via the LDA+U method.

Type III: Shallow behavior before gap correction and deep behavior after gap correction (Fig. 6, right). The most difficult situation occurs when the DLS is located above the LDA-calculated CBM but inside the gap after the correction of the CBM (Fig. 6, right). In this case, the character of the donor state ψ_D changes from hostlike and delocalized to defectlike and localized. Therefore, the first-order perturbation requirement [condition (b) in Sec. III C] for the perturbation-extrapolation method is violated. Thus, the extrapolation scheme would fail even if a perturbation were found that corrects all host states and not just the VBM and the CBM (Sec. III C). The formation energy in LDA is generally too low for type-III defects because the electron erroneously relaxes into the host conduction-band states located at too low energy. Ironically, the LDA description of type-III defects may worsen with increasing supercell size (see also Sec. IV B), since a larger fraction of the donor electron relaxes into these host states as the host conduction-band density of states is increased with supercell size. On the other hand, the application of the shallow-donor correction $z_e \Delta E_C$ (see Sec. II D 2) would also lead to a wrong result, i.e., to overcorrection, since defect level shift is smaller than that of the CBM (see Fig. 6, right).

An example for type-III behavior is the singly charged state V_O^+ of the O vacancy in ZnO (see Sec. III E 3 above). Since, in this specific case, V_O transforms from type I to type II along the reaction coordinate connecting the equilibrium configurations of V_O^{2+} and V_O^0, respectively, we could obtain $\Delta H(V_O^+)$ in Ref. 7 by interpolation of the configuration coordinate diagram. More generally, one needs to modify the host-band structure *during the self-consistent calculation* such that the position of the DLS within the gap is uncovered and the correct orbital configuration is obtained.[123] Here, the LDA+U_s method (see Sec. III D) may be helpful in achieving at least a qualitative picture about the energy of the defect level in the corrected gap.

IV. CORRECTION OF FINITE-SIZE EFFECTS IN SUPERCELL CALCULATIONS

There are several distinct causes of finite-size effects in supercell calculations, i.e., in particular, those concerning the

TABLE II. Supercell-size dependence of the charged-defect formation energies $\Delta H(E_F=E_V)$ of V_{As}^{3+}, V_{Ga}^{3-}, As_{Ga}^{2+}, and Te_{As}^{+} in GaAs under As-rich conditions ($\Delta\mu_{As}=0$). Given are ΔH for the 1728- and 64-atom supercells at different levels of corrections for finite-size effects: our present method of "potential alignment+image charge" corrections (PA+MP), uncorrected (UC) supercell energies, and "potential-alignment-only" (PA) corrected energies. The formation energies are in eV and include only the NN relaxation as determined in the 64-atom cell.

	$\Delta H(1728)$ (PA+MP)	$\Delta H(64)$ (PA+MP)	$\Delta H(1728)$ (UC)	$\Delta H(64)$ (UC)	$\Delta H(1728)$ (PA)	$\Delta H(64)$ (PA)
V_{As}^{3+}	4.15	4.15	3.88	3.68	3.86	3.30
V_{Ga}^{3-}	4.32	4.36	4.01	3.06	4.03	3.51
As_{Ga}^{2+}	1.28	1.29	1.15	0.73	1.16	0.92
Te_{As}^{+}	0.20	0.19	0.17	−0.03	0.17	0.09

convergence of elastic energies, potential-alignment effects, image charge interactions, and band-filling effects. Defects with large lattice relaxations have a considerable contribution to their formation energy due to elastic energies, which depend on the supercell size. However, such elastic energies can usually be explicitly converged in supercells of affordable sizes. For example, ΔH of the fully relaxed neutral O vacancy in ZnO differs by only 0.05 eV between a 72-atom and a 576-atom supercell (Fig. 5), despite the large lattice relaxation of this defect. Similarly, in the case of the triply charged V_{As}^{3+} defect in GaAs (see below) which also exhibits large lattice relaxations,[139] we find convergence of the elastic energy within 0.06 eV for 128-atom and larger supercells. Therefore, we focus here on the slower-converging size-dependent effects that in general cannot be converged by simply calculating large enough cells. These slow-converging finite-size effects are, in particular, the electrostatic image charge interaction in the case of charged defects (Sec. IV A) and the Moss-Burstein-type band-filling effects in the case of occupied shallow levels that are caused, e.g., by charge-neutral shallow donors (Sec. IV B).

A. Image charge interactions

The treatment of charged supercells and the question of whether or not the image charge corrections proposed by Makov and Payne[31] are appropriate have been subjects of considerable discussion and debate in literature.[5,9,39,40,44–57] In particular, concerns were raised[44,48,57] that the "defect charge," i.e., the charge difference between the "host+defect" and "pure-host" systems, may be too delocalized, so that the point-charge model underlying the (first-order) image charge correction in Ref. 31 may not hold. Therefore, we assess here the validity of the image charge correction [Eq. (11)] by calculating a dense series of supercell sizes for a highly charged test case, i.e., the 3+ state of the As vacancy in GaAs, and compare the results of the image charge correction with the prediction of the finite-size scaling method.[50,54,56]

1. Image charge correction versus finite-size scaling

Finite-size scaling methods provide, in principle, accurate formation energies for the limit of infinite dilution provided that sufficiently many terms are taken into account in the expansion and that sufficiently many and sufficiently large supercells are calculated, so as to be able to accurately fit the expansion coefficients. It was found[50,56] that within an expansion in powers of the reciprocal linear supercell dimension ($1/L$, where $L=V_{SC}^{1/3}$ is the cubic root of the supercell volume V_{SC}), the first- and third-order terms dominate,

$$\Delta H_D(L) = \Delta H_D(\infty) + \frac{\gamma_1}{L} + \frac{\gamma_3}{L^3}. \tag{22}$$

The finite-size scaling method requires the calculation of a set of different supercell sizes with the largest the size of at least a few hundred atoms, preferably on the order of a thousand atoms.[54,56] After the fit according to Eq. (22), the extrapolated formation energy $\Delta H_D(\infty)$ for infinite supercell size should be accurate, irrespective of the validity of the analytic form for γ_1 and γ_3 as given by Makov and Payne,[31] according to Eq. (11). More simplified scaling methods consider only either the $1/L$ term (e.g., Refs. 34 and 55) or the inverse volume $1/L^3$ term.[44,48,51] Another modification of the finite-size scaling method was proposed by Erhart *et al.*,[39] who assumed that the analytic expression for the factor γ_1 according to Eq. (11) fully accounts for the $1/L$ contribution to the scaling, and fitted the remaining size dependence only through the parameter γ_3 in the $1/L^3$ term in Eq. (22).

We here present several showcase examples of charged defects in GaAs which are calculated in very large supercells of up to 1728 atoms (Table II) and which are chosen such that there exist no size-dependent band-filling effects (Sec. II D 3) that may convolute the finite-size scaling. In order to eliminate the size dependence of the elastic energies, the formation energies are calculated with the same nearest-neighbor atomic relaxation determined in a 64-atom cell, while all other atoms remain at their ideal lattice positions. (In separate calculations where we relaxed all cell-internal degrees of freedom, we determined for V_{As}^{3+} the residual relaxation energy as 0.51 eV in the limit of large cells, which was reached within 0.15 eV in the fully relaxed 64-atom cell and within 0.06 eV in cells of 128 atoms or larger.) In order to accurately determine the exact scaling behavior for the highly charged test case V_{As}^{3+}, we use a dense series of 15 different cells representing the GaAs zinc-blende lattice in simple cubic (sc), face-centered-cubic (fcc), or body-

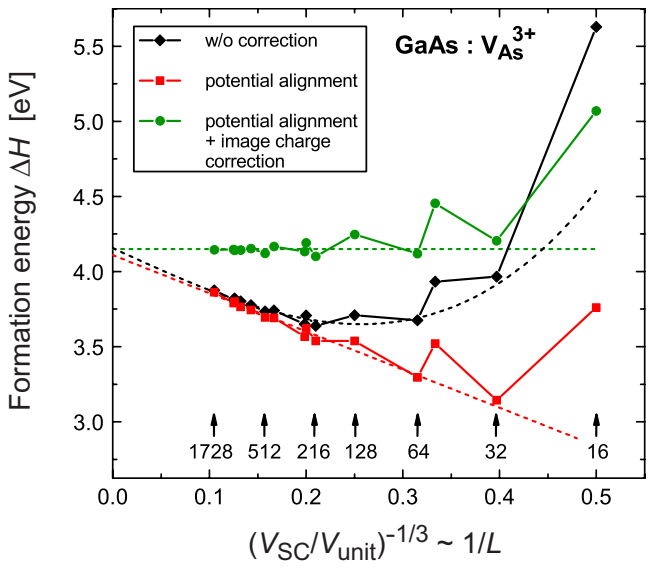


FIG. 7. (Color online) Scaling of the formation energy $\Delta H(E_F = E_V, \mu_{As} = \mu^0_{As})$, of the V^{3+}_{As} defect in GaAs as a function of $(V_{SC}/V_{unit})^{-1/3}$, which is proportional to the reciprocal linear supercell dimension $1/L$ (V_{SC}: supercell volume; V_{unit}: volume of the two-atom GaAs unit cell). Shown are different levels of total-energy corrections and their respective finite-size scalings (dashed lines), as determined by a fit using the data points between 64 and 1728 atoms. Diamonds: No correction, $1/L+1/L^3$ scaling. Squares: Potential-alignment correction, $1/L$ scaling. Circles: Potential alignment+image charge correction, no size dependence.

centered-cubic (bcc) supercells (see Ref. 140 for the justification of combining the different cell symmetries in the finite-size scaling).

Figure 7 shows the size dependence of ΔH of V^{3+}_{As} as a function of the inverse linear supercell dimension $1/L$ for three different levels of corrections, along with a respective fit according to Eq. (22) (the fit includes all supercells with 64 or more atoms): (1) (diamonds) uncorrected supercell energies, with fit of γ_1, γ_3, and $\Delta H(\infty)$; (2) (squares) supercell energies after the potential-alignment correction [Eq. (7)], with fit of γ_1 and $\Delta H(\infty)$; and (3) (circles) supercell energies after potential-alignment and image charge corrections [Eqs. (7) and (11)],[141] with fit of $\Delta H(\infty)$ only [i.e., $\Delta H(\infty)$ is the average ΔH for the cell sizes between 64 and 1728 atoms]. Comparing the three different finite-size scaling data sets, we see that very similar extrapolations to infinite cell size are obtained; i.e., the values of $\Delta H(\infty)$ obtained by the three fits agree within 0.04 eV. It is notable that even the result of the 32-atom cell (bcc) is rather well converged within only 0.06 eV, whereas the supercells in fcc symmetries (e.g., 16, 54, and 128 atoms) yield somewhat slower convergence (see Fig. 7).[142]

After applying both the potential-alignment correction [Eq. (7)] and the analytic form of image charge correction [Eq. (11)] including the first- and the third-order terms (circles in Fig. 7), we find that the formation energy of V^{3+}_{As} becomes *size independent* for cells of 64 atoms and larger, and the $\Delta H(\infty)$ obtained by the fit of the corrected energies (circles) agrees within 0.01 eV with the respective $\Delta H(\infty)$ obtained from the finite-size scaling of the uncorrected energies (diamonds). Also, we find that the corrected $\Delta H(V^{3+}_{As})$ for the relatively small 64-atom cell is converged within 0.03 eV compared to $\Delta H(\infty)$. Thus, the *combination* of potential alignment plus image charge correction including the third-order term enables the calculation of formation energies with essentially the *same accuracy as finite-size scaling*, but requires the calculation of only one relatively small supercell, e.g., of 64 atoms, thereby reducing the computational effort dramatically. Our results demonstrate that finite-size errors can be corrected to high accuracy if finite-size effects of different physical origins are deconvoluted and treated separately, despite the expectation[54,56] that it would be difficult to obtain a general analytic correction method for finite-size errors. Also, we find no indication of significant contributions to the finite-size scaling from terms other than the $1/L$ and $1/L^3$ terms in Eq. (22) (at least when excluding the 16- and 54-atom fcc supercells), which account for electrostatic interaction and potential-alignment effects. The sporadic observation[50,56] of scaling behaviors in deviation from Eq. (22) are possibly related to band-filling effects (cf. Secs. II D 3 and IV B) which do not follow the functional form of Eq. (22).

Some previous works[44,48,51] used an approximate $1/L^3$ scaling, found empirically within some range of supercell sizes, to extrapolate uncorrected supercell energies to the infinite limit. We emphasize here that the potential-alignment correction does scale as $1/L^3$. Therefore, the uncorrected formation energies may give the impression of inverse-volume scaling if potential-alignment effects are pronounced. Indeed, when we plot the uncorrected energies in Fig. 7 as a function of inverse volume, we find approximate linearity with a positive slope (lower ΔH for larger cells) up to cell sizes of 216 atoms. For even larger cells, however, the image charge interaction becomes more important and changes the sign of the slope (cf. the nonmonotonic behavior of the uncorrected energies in Fig. 7). Note that this change in slope may not be noticed if cells up to only few hundred atoms are considered. Thus, when we use the formation energies calculated for the 216-atom and smaller cells for linear extrapolation to infinity in the $1/L^3$ plot, we obtain an error as large as 0.9 eV compared to the converged formation energy of V^{3+}_{As}. This example shows that the extrapolation based on inverse-volume scaling[44,48,51] can lead to errors that are even larger than the errors of the uncorrected energies (excluding the small 16-atom cell, the largest error of about 0.5 eV occurs at the fairly large cell size of 216 atoms; see Fig. 7).

We further tested the present correction scheme for few additional defects in GaAs, i.e., the 3− charge state V^{3-}_{Ga} of the Ga vacancy, the 2+ state of the EL2-related As^{2+}_{Ga} antisite defect,[143] and the singly charged Te^{+}_{As} as a prototypical ionized shallow donor (see also Sec. IV B). Here, we did not repeat the full series of supercells calculated for V_{As} but simply compare in Table II the corrected and uncorrected formation energies for 64-atom and 1728-atom supercells. In order to exclude the cell-size dependence of elastic energies, we again consider nearest-neighbor (NN) relaxation only. As seen in Table II, our "potential alignment+image charge" correction method consistently removes the supercell-size dependence of ΔH, whereas uncorrected energies show large discrepancies of up to ~1 eV between the 1728- and 64-

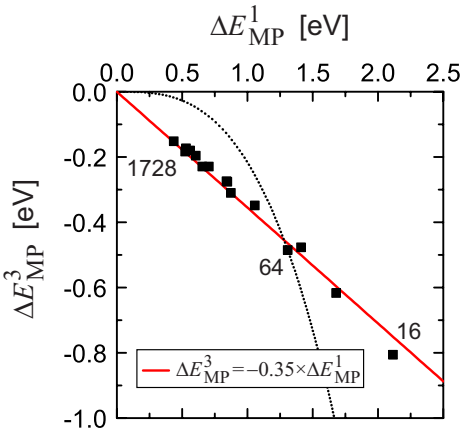


FIG. 8. (Color online). The third-order term ΔE^3_{MP} of the image charge correction as a function of the respective first-order term ΔE^1_{MP} [cf. Eq. (11)], calculated for V^{3+}_{As} in GaAs in supercells with between 16 and 1728 atoms. The observed proportionality $\Delta E^3_{MP} \sim \Delta E^1_{MP}$ demonstrates the unexpected effective $1/L$ scaling of ΔE^3_{MP}. For comparison, the dashed line illustrates the behavior that would be expected from the nominal $1/L^3$ scaling of ΔE^3_{MP} relative to the calculated image charge energy for the 64-atom supercell.

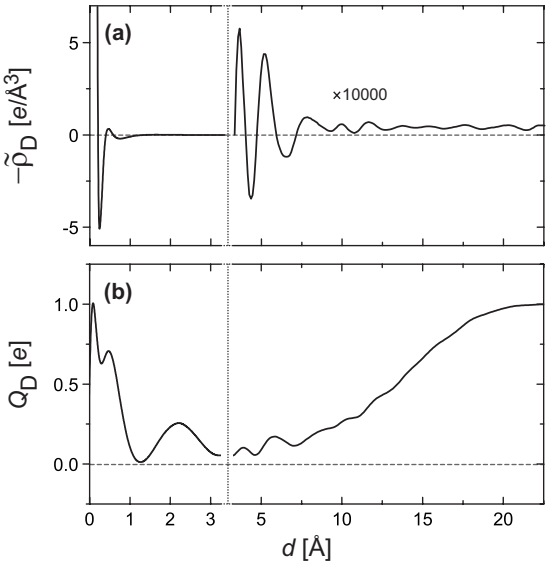


FIG. 9. (a) The spherically averaged defect-charge density $\tilde{\rho}_D(r)$ caused by the ionized Se^+_{As} donor in GaAs, obtained in a 1000-atom supercell, shown as a function of the distance d from the donor (jellium background charge is not included). For graphical clarity, the density is amplified by a factor of 10 000 beyond $d=3.3$ Å. The contribution due to the different nuclear charges of Se and As is represented by a Gaussian distribution with width $\sigma=0.05$ Å. (b) Integration $Q_D(d)=-\int_0^d \rho_D(r)dr$ of the defect charge shown in (a).

atom cells. *Large errors occur also if potential-alignment effects are considered but no image charge corrections are applied*, as done, e.g., in the recent ZnO defect calculations by Janotti and van de Walle.[40] In the examples shown in Table II, such potential-alignment-only corrected energies for the typical cell size of 64 atoms deviate from the converged, i.e., the fully corrected energies by up to 0.9 eV, highlighting the importance of taking into account image-charge and potential-alignment corrections simultaneously.

2. Unexpected scaling of the image-charge correction

A surprising observation in Fig. 7 is that the data set including only the potential-alignment corrections (squares) but not the image-charge correction can be well fitted with only the first-order parameter γ_1, i.e., with the setting $\gamma_3=0$. This means that after the potential alignment (which scales as $1/L^3$), no significant third-order contribution remains, despite the nominal $1/L^3$ scaling of the second term in Eq. (11), and that the image-charge correction effectively scales as $1/L$. Indeed, when we plot for the case of V^{3+}_{As} the third-order correction ΔE^3_{MP} [second term in Eq. (11)] as a function of the respective first correction ΔE^1_{MP} [first term in Eq. (11)], we find a clear proportionality, shown in Fig. 8,

$$\Delta E^3_{MP} = f \Delta E^1_{MP}, \tag{23}$$

which strongly deviates from the behavior that would be expected from the nominal $1/L^3$ scaling of the third-order term ΔE^3_{MP}, as illustrated by the dashed line in Fig. 8. Additionally, from calculation of defects with different charge states in GaAs (Table II) we find that the proportionality factor $f=-0.35$ is essentially independent of q. Thus, ΔE^3_{MP} scales effectively in the same way as ΔE^1_{MP}, i.e., as q^2/L, which indicates the implicit dependency $Q_r \sim qL^2$ for the second moment of the defect density $\tilde{\rho}_D(\mathbf{r})$ [cf. Eqs. (11) and (12)]. Notice that the effective $1/L$ scaling of ΔE^3_{MP} implies that a significant error can be introduced in the scaling method of Erhart *et al.*,[39] where it is assumed that after applying the first order correction ΔE^1_{MP}, the remaining finite-size dependence scales solely as $1/L^3$.

In order to study the origin of the dependency $Q_r \sim qL^2$ and, hence, of the unexpected scaling behavior of the third-order term ΔE^3_{MP}, we calculated the (all-electron) defect-induced electron-density difference $\tilde{\rho}_D(\mathbf{r})$ (cf. Sec. II D 6) due to the ionized Se^+_{As} donor in a 1000-atom supercell of GaAs. Thus, Fig. 9(a) shows the defect-induced charge density $-\tilde{\rho}_D$, which is the negative of the electron-density difference $\tilde{\rho}_D$ (due to the negative charge of electrons). Se_{As} is a shallow donor, similar to Te_{As} (cf. Sec. IV B) but is particularly suited to studying the defect charge, since there is no contribution due to additional core electrons. Also, in order to avoid the large positive and negative contributions to the defect charge due to lattice relaxation, we use here the ideal lattice positions for all atoms (atomic relaxation increases the Se-Ga NN distance only moderately by 0.09 Å compared to the As-Ga distance). As seen in the integrated defect charge $Q_D(r)$ [Fig. 9(b)], the net $q=+1$ charge stemming from the nuclear charge difference between Se and As is screened to about $1/\varepsilon$ at a distance $d \approx 3$ Å from the site of the donor. The charge $-q(1-1/\varepsilon)$ needed for this screening is drawn more or less uniformly from throughout the supercell, as evident from the approximately homogeneous positive defect charge $-\tilde{\rho}_D(d)$ beyond $d>7$ Å in Fig. 9(a) (which is not to be confused with the compensating jellium background that is not explicitly introduced in the calculation; see Sec. II D 4).

Considering that the third-order term ΔE^3_{MP} of the image charge interaction is designed to describe the interaction of

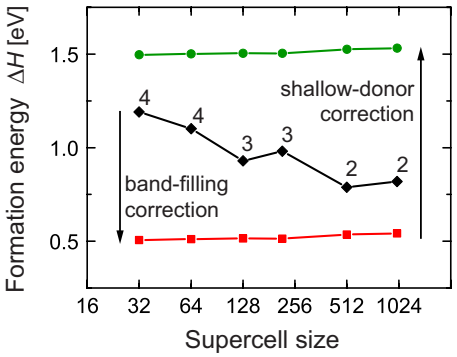


FIG. 10. (Color online) Scaling of the formation energy $\Delta H(\mu_{As}=\mu^0_{As}, \mu_{Te}=\mu^0_{Te})$ of the Te^0_{As} defect in GaAs as a function of the supercell size. Shown are the uncorrected ΔH (diamonds), ΔH after the band-filling corrections (squares), and ΔH after the band-filling and shallow-donor corrections (circles). The numbers n next to the uncorrected energies indicate the use of a $n\times n\times n$ k mesh in the respective supercell calculation.

the (background-compensated) point charge with the delocalized part of the defect charge,[31] the following physical picture emerges for the unexpected proportionality between ΔE^3_{MP} and ΔE^1_{MP}: The delocalized part of the defect density $\tilde{\rho}_D(\mathbf{r})$ arises due to the dielectric screening response of the host upon introduction of a defect with charge q. This delocalized defect density is proportional to q and is essentially constant in the regions farther away from the defect, i.e., those regions that primarily contribute to the second radial moment Q_r of the defect charge. Thus, by the definition of Q_r [Eq. (12)] it follows the proportionality $Q_r \sim qL^2$, which explains the observed $1/L$ scaling of ΔE^3_{MP}. Since, ΔE^3_{MP} is determined by the screening response of the host, rather than by any defect-specific property, the proportionality factor f in Eq. (23) should be independent of the specific defect, so that Eq. (11) can be simplified to

$$\Delta E_{MP} = (1+f)\frac{q^2\alpha_M}{2\varepsilon L}, \tag{24}$$

where $f=-0.35$ is determined from the data shown in Fig. 8, in agreement with the previous observation that the third 30% of the order term is about monopole correction.[5] We can now explain this empirical observation by considering that in a semiconductor material with typically $\varepsilon \gg 1$, the screening charge which accumulates close to the defect is approximately $-q$, leading to a change of the charge density by q/V_{SC} throughout the supercell, except the region close to the defect [which hardly contributes to the second radial moment, cf. Eq. (12)]. Using $-\tilde{\rho}_{D,q}=q/V_{SC}$ in Eq. (12), we obtain $f=-0.37$, -0.34, and -0.34 for the sc, fcc, and bcc supercell geometries, respectively. Thus, our present analysis suggests that Eq. (24) with $(1+f)\approx 2/3$ can serve as a general, simple and reasonably accurate correction formula for the image charge interaction.

Finally, the present showcase of Se^+_{As} explains also why the delocalization of the defect charge cannot serve as an argument against the image charge correction [Eqs. (11) and (24)]: The total defect charge $\tilde{\rho}_D(\mathbf{r})$ of Se^+_{As} stems from two distinct contributions, i.e., first, the localized point-charge-like contribution due to the ionic substitution (solely the difference in the nuclear charge in case of Se^+_{As}) and, second, the delocalized contribution due to the screening response of the host. Thus, the underlying physical picture of the leading term of ΔE_{MP}, i.e., that of a point charge in a dielectric medium, is essentially correct. In contrast, the model of Segev and Wei[48] for the Coulomb interaction between delocalized periodic test charges, which is sometimes cited as evidence against the Makov-Payne correction,[44,57] ignores the point-charge-like ionic contribution to the defect charge and neglects dielectric screening. This simplified model, therefore, does not capture the underlying physics of the electrostatic interaction between charged point defects in a semiconductor.

B. Finite-size effects due to host and impurity band dispersion

1. Total-energy contributions due to band filling

In charge-neutral defect calculations, strong finite-size effects usually occur only if occupied shallow defect levels are present, leading to band-filling effects (see Sec. II D 3). (This situation may, of course, also occur for charged states, e.g., for the shallow Zn interstitial Zn^+_i in ZnO. In this case the charge-related and band-filling-related corrections have to be treated separately.) In the case of type-I behavior (see Fig. 6, Sec. III F), the DLS remains unoccupied, as it releases the electron into the hostlike delocalized PHS, which usually has a large dispersion. Due to the high defect concentration implied by typical supercell sizes on the order of 100 atoms, the occupation of the PHS can lead to large band-filling effects,[5,64] in particular when the respective host band has a small effective mass (large dispersion). These band-filling effects are illustrated in Fig. 10 for the case of the Te_{As} in GaAs (full atomic relaxation is included), which is a well-established shallow donor.[144] We find that the uncorrected formation energy is strongly cell-size dependent and the data points are somewhat jagged, despite the relatively dense k-mesh used in the calculations shown in Fig. 10 (smooth behavior is expected in the limit of a dense k mesh.) After application of the band-filling correction, Eq. (6), the formation energies are smooth and almost independent of the cell size. Note that a fairly large band-filling correction of 0.6 eV is necessary for a typical supercell size of 64 atoms, and considerable corrections are still needed even for large supercells, e.g., 0.3 eV at a cell size of 1000 atoms.

The successful removal of the size dependence of the formation energy of the shallow Te^0_{As} by the band-filling correction contrasts with the finding in Ref. 54 that the "dispersion correction" studied there leads to worse ΔH even for shallow defects. [Note that this dispersion correction (cf. also Ref. 52) should have a similar effect as our band-filling correction, Eq. (6).] The origin of this discrepancy may be the convolution of finite-size effects of different physical origins (and different scaling behaviors; see Sec. IV A) and/or the convolution of finite-size and band-gap errors. When we also apply the shallow-donor correction originating from the band-gap error (Sec. II D 2), which increases the formation energy of the Te^0_{As} donor by ΔE_C ($z_e=1$ due to the singly

occupied donor state; see also Sec. III F), we see in Fig. 10 that at small cell sizes the band-filling and shallow-donor corrections partly cancel each other. As a result, the uncorrected formation energy is closer to the corrected ΔH at small cell sizes than at large cell sizes. This cancellation effect is exploited in a method to calculate formation energies and transition levels by determining the band-edge energies of the host as the (supercell) Brillouin-zone average instead as the band energy at the extremal points (e.g., Γ),[9,94] while no band-filling corrections are applied to the defect state. Of course, correcting band-gap errors and band-filling effects separately yields more accurate energies and does not depend on the actual supercell size used.

Notably, a slight increase in the formation energy with cell size is still observed in Fig. 10 after application of the band-filling and shallow-donor corrections. This increase can be explained by a residual image charge interaction, considering that an ionic +1 quasipoint charge is created by replacing the As^{+5} ionic core with a Te^{+6} ionic core, which is compensated by the donor electron in a shallow, delocalized state (PHS). Since the shallow-donor states overlap with their periodic images, even the formally charge-neutral Te^0_{As} donor can be regarded as a (screened) point charge in a compensating background. Since, however, the compensation charge, i.e., the electron in the shallow-donor state, is not strictly homogeneous as the compensation jellium background in the case of the ionized Te^+_{As} donor, the effect is smaller, i.e., only $\sim$40% of the magnitude expected by the respective correction ΔE_{MP} for Te^+_{As} according to Eq. (11).

The strong supercell-size dependence of the uncorrected ΔH of the shallow Te^0_{As} donor in GaAs is in stark contrast with the behavior of the deep V^0_O donor in ZnO, in which case the formation energy is practically independent of the size of the supercell (see Fig. 5) and there is no need for correction of size effects. Due to the deep and localized donor state of V_O (cf. Fig. 6, center), the electrons occupy the DLS, i.e., the primary defect state, and not the host-band-derived PHS (cf. Sec. III F). Accordingly, no finite-size effects associated with band filling in the strongly dispersive host conduction band occur. Thus, the independence of $\Delta H(V^0_O)$ from the cell size corroborates our argument (see Sec. III E 1) that the donor state of V^0_O does not have the character of the host conduction band and should not experience a shift with the CBM during band-gap correction. Since the deep level of V^0_O is formed below the CBM of LDA or GGA, the band-filling correction, as formulated in Eq. (6), automatically vanishes despite the large dispersion of the impurity band within the LDA or GGA band gap, thereby correctly reflecting the size independence of $\Delta H(V^0_O)$. Thus, we agree with the conclusion by Castleton *et al.*[54] that the dispersion correction is not appropriate for deep defects.

A more difficult situation arises if a deep-donor level occurs below the experimental CBM energy but above the CBM in the LDA calculation (type-III behavior; see Sec. III F and Fig. 6). In this case, the simultaneous application of the band-filling and shallow-donor corrections would incorrectly predict a shallow level after correction. On the other hand, in the limit of large supercells, the introduced donor electron would relax to the energy of the LDA-calculated CBM which is lower than the appropriate defect level energy (Fig. 6, right). Thus, type-III behavior can lead to the unsuspected situation that the uncorrected energies are more accurate for small cell sizes than for large sizes (cf. Sec. III F) because the band-filling effect causes the (correct) occupation of the defect level inside the LDA conduction band. Such convolutions of band-gap errors due to LDA and finite-size errors may be the origin of the conclusion obtained in Ref. 54 that the appropriate band-gap correction method depends on the supercell size used in the respective calculation, whereas, in principle, band-gap and finite-size errors are of fundamentally different origins. In order to avoid the convolution between both types of errors, it can be very useful to correct the band edges within the self-consistent calculation through additional potentials,[123] which, at the same time, removes the spurious hybridization between the defect state and the host-band states, and enables the calculation of transition levels inside the corrected band gap.

2. Convergence of single-particle energies

Regarding the convergence of *single-particle* defect states, we find pronounced finite-size effects for the a_1 gap level of V^0_O (Fig. 4) if it is determined at $\bar{\Gamma}$, i.e., the center of the Brillouin zone corresponding to the supercell; see "$a_1(\bar{\Gamma})$" in Fig. 5. A similar observation was recently made by Li and Wei,[145] who calculated the (single-particle) gap level of the isovalent O_{Te} defect in ZnTe, and found slow convergence with the size of the supercell considering cells up to 4096 atoms in non-self-consistent calculations. We find here, however, that the Brillouin-zone average "a_1(BZ av)" (Fig. 5) of V_O is already well converged in a 72-atom cell, which explains the insensitivity of the formation energy on the cell size, considering the small relative weight of the $\bar{\Gamma}$ k vector in case of the 72-atom supercell. Thus, deep (i.e., type-II) defect states (Fig. 6, center) in small supercells are generally much more accurately described by an average over the Brillouin zone than at the zone center $\bar{\Gamma}$, as was also shown for the P antisite defect in GaP.[53] Figure 5 shows that the slow convergence of the $a_1(\bar{\Gamma})$ state of V_O is caused by the concentration-dependent repulsive interaction with the host conduction band $a'_1(\bar{\Gamma})$ at the zone center, as evident by the fact that $a'_1(\bar{\Gamma})$ converges to the CBM of the defect-free bulk in the limit of large supercells (small V_O concentrations), whereas $a_1(\bar{\Gamma})$ converges toward the Brillouin-zone average a_1(BZ av) of the V_O defect state. Thus, the dispersion of the impurity band is not due to direct defect-defect interaction between different images, which should decay exponentially with the defect-defect distance,[50] but arises from a *concentration-dependent hybridization* between the defect state and the host conduction band. This hybridization is strongest at $\bar{\Gamma}$ because at this k vector the energy difference between the defect state and the conduction band is minimal.

A different situation arises for shallow defect states (type-I behavior), in which case the DLS forms a resonance inside the host-band continuum. This situation occurs for the doubly charged vacancy state V^{2+}_O in ZnO, where the order of the a_1 (DLS) and a'_1 (PHS) states in Fig. 5 is inverted[7] (see also Sec. III E 4). Thus, the interaction between the higher-

energy DLS and the lower-energy PHS reduces the energy of the $a_1'(\bar{\Gamma})$ state, i.e., of the CBM-derived PHS, which occurs below the CBM at the zone center $\bar{\Gamma}$. However, the Brillouin-zone average of the dispersive PHS a_1' remains above the CBM for typical cell sizes, as observed, e.g., in the gap-corrected defect-bandstructure for V_O^{2+} in Ref. 43. (Note that the Brillouin-zone average of a PHS is size dependent, whereas that of a DLS is essentially size independent, because the number of available host states increases with the cell size, whereas that of the defect states does not.) With increasing supercell size, the PHS a_1' of V_O^{2+} converges toward the host-conduction-band-like shallow effective-mass level just below the CBM.[7]

V. EXPERIMENTAL SITUATION ON O VACANCIES IN ZnO

Due to the theoretical controversy about O deficiency in ZnO,[6,7,37,40] we summarize here briefly the experimental situation on this issue: While O vacancies are frequently considered to be an abundant defect in ZnO, a conclusive analysis of the ZnO stoichiometry is still lacking in ZnO, unlike the case of the related oxides In_2O_3 (Ref. 61) and SnO_2,[62] for which O deficiency up to the percent range has been rather directly shown by thermogravimetric analysis. Note that in the case of In_2O_3, we showed in Ref. 6 that our method of band-gap correction gives good agreement with these thermogravimetric experiments, when the V_O concentration is calculated with first-principles thermodynamics methods. Significant equilibrium concentrations of F-centers (O vacancies) have also been measured in MgO after thermochemical reduction (i.e., Mg-rich treatment).[63] Recently, positron annihilation experiments provided the first step toward an unambiguous experimental proof of the existence of V_O in as-grown ZnO.[58,146] Also, an $S=1$ spin-triplet signal observed by optically detected magnetic resonance in the ubiquitous green luminescence of ZnO, has tentatively been assigned to V_O.[147] Further support for the existence of V_O in O-poor grown or treated ZnO can be drawn from the pronounced coloration effects, i.e., the fact that ZnO becomes successively yellow and red under Zn-rich treatment:[60,146] Such effects are typical for F centers in various II-VI materials[59,63,148] and are explained by our calculated absorption levels of anion vacancies in ZnO, ZnS, and ZnSe.[7,79] The assignment of these coloration effects to O vacancies was recently further corroborated by Evans *et al.*,[149] who found that the respective absorption bands are identical with those observed after high-energy electron irradiation, where the existence of V_O can be directly observed in electron paramagnetic resonance (EPR).

An argument against the existence of O vacancies in ZnO, brought forward in Ref. 40, is the fact that the well-known EPR signal of the spin-singlet ($S=1/2$) V_O^+ state[131] is not observed in as-grown ZnO but only after artificial creation of vacancies by high-energy electron irradiation. However, the classic experiments of Soriano and Galland,[150] as well as those of Locker and Meese,[151] already showed that the $S=1/2$ EPR is not observed in insufficiently compensated samples, even though O vacancies were created by the electron irradiation. Thus, the $S=1/2$ EPR of V_O^+ is not observed in n-type ZnO even if O vacancies exist. Indeed, one can expect[7] that a photoexcited V_O^+ binds free electrons in n-type ZnO, so that the excited $S=1$ triplet state of V_O^0 is observed[147] instead of the $S=1/2$ state. (V_O^+ and the electron can also couple to an excited $S=0$ state, which is not observed and decays through optical recombination very quickly into the ground state of V_O^0.[7])

Based on our calculated optical absorption energies,[7] we suggested that in ZnO the first excitation $V_O^0 \rightarrow V_O^+ + e$ occurs at 2.8 eV, consistent with the yellow color, and that the second excitation $V_O^+ \rightarrow V_O^{2+} + e$ occurs at 2.4 eV, which could explain the red coloration observed under strongly reducing conditions (high V_O concentrations). No theoretical explanation besides that of V_O exists to date for these coloration effects. Note that Soriano and Galland[150] already observed yellow coloration in their classic experiments after they selectively produced O vacancies by electron irradiation below the Zn-displacement threshold. After an early suggestion[34] that O vacancies may be involved with the phenomenon of persistent photoconductivity (PPC), we developed in Ref. 7 a detailed configuration coordinate model for V_O and predicted that PPC should arise from a double excitation mechanism. While PPC in ZnO is indeed frequently observed, it is often associated with oxygen adsorption and desorption at the surface.[152] Interestingly, however, pronounced PPC effects were recently observed in ZnO that were correlated with two optical absorption levels at 2.15 and 2.5 eV,[153] close to our calculated optical levels of V_O (Ref. 7) (which, however, correspond to the vertical excitation threshold, not to the maximum). Considering that the concentration of extrinsic donors, such as hydrogen, should not dramatically increase during the high-temperature thermochemical reduction process,[59,60,146] PPC caused by oxygen vacancies is presently the only model[6] that can explain the apparent paradoxical coexistence of coloration (indicating a deep level in the optical range) and high electron concentrations above 10^{18} cm^{-3} (requiring a shallow level) observed after such reduction treatment. Further experimental verification is, however, desirable.

Regarding the optical excitation threshold for excitation of V_O^0, it should be noted that there appears to be a discrepancy between high-energy electron-irradiated ZnO and samples where O vacancies are formed during the growth process: In high-energy electron-irradiated ZnO, where O and Zn vacancies are generated simultaneously, the V_O^+ state can be excited by photon energies of around 2 eV.[149,151] In contrast, in as-grown samples, the photoexcitation threshold for the creation of the excited spin-triplet ($S=1$) state of V_O,[147] which should correspond to the $V_O^0 \rightarrow V_O^+ + e$ excitation,[7] is much larger at 3.1 eV.[154] As we discussed in Ref. 7, the generation of multiple defects due to high-energy irradiation may create different channels for the photoexcitation of V_O^+, so that the excitations energies around 2 eV may not necessarily correspond to the $V_O^0 \rightarrow V_O^+ + e$ transition. We here suggest that further experiments with electron irradiation below the Zn-displacement threshold[150] should provide an opportunity to more unambiguously determine the optical excitation energies of V_O.

VI. SUMMARY AND CONCLUSIONS

A. Band-gap correction

By calculating the quasiparticle band gap from total-energy differences, we demonstrated that the well-known band-gap problem is a real deficiency of the approximate LDA and GGA functionals, not just a fallacy caused by the nonphysical meaning of the Kohn-Sham single-particle energies. Given that accurate self-consistently band-gap-corrected total-energy calculations for large-scale defect systems remain challenging, we assessed current schemes for *ex post facto* band-gap corrections for the conventional LDA and GGA functionals. We demonstrated that extrapolation schemes, in which a band-gap-opening perturbation is extrapolated toward the experimental gap, depend in general very sensitively on the type of perturbation applied. Thus, such methods are arbitrary as to the choice of the perturbation parameter except in the case of a perturbation that corrects all host bands and not just the energy gap (of course, if such a perturbation were known, it could be directly applied without the need for extrapolation). A direct band-gap correction can be achieved by means of the LDA+U method when applied to s states in addition to the more conventional application to cation d orbitals. While this method may convey some qualitative insight on how defect levels react on band-gap correction, we find that it is not a general quantitative method for band-gap-corrected defect calculation. For example, the application of LDA+U on the Zn s orbitals for the purpose of band-gap correction leads to a wrong ground-state structure of ZnO.

Since a universal method that would avoid the band-gap problem and at the same time could accurately predict defect formation energies and transition levels remains elusive at the present, we defined a general qualitative classification scheme, describing different defect behaviors with regard to the energy of their (single-particle) defect states relative to the band edges. This classification scheme provides general guidelines on how band-gap corrections should be applied for the different behaviors. Since the primary defect levels, which can occur either in the gap or as resonances inside the host-band continuum, are typically less affected by LDA deficiencies than the band-edge energies and the gap, it is a well-defined and often reasonably accurate assumption to correct only the band-edge energies, but to retain the LDA description of the genuine defect states. In order to ensure consistency between the band-gap correction and the defect calculation, it is important that the shifts in the individual band-edge energies are determined with respect to an internal potential reference in the respective uncorrected LDA calculation. Further improvements upon this band-edge-only correction can be achieved through physically motivated defect-specific corrections, if such are known, e.g., by the LDA +U method for transition-metal impurities.

B. Correction of supercell finite-size effects

In order to assess finite-size effects in defect supercell calculations, we calculated highly charged test cases in GaAs supercells of up to 1728 atoms. Our results highlight the benefit of treating finite-size effects of different physical origins separately, whereas in finite-size scaling methods, simple functional forms for the scaling may not accurately account for all size-dependent effects that may exist simultaneously. In particular, we find that when error sources other than the electrostatic image charge interaction are eliminated, the image charge correction up to the third-order term, as proposed by Makov and Payne,[31] affords fast convergence and excellent formation energies even for small supercell sizes, such as the typical 64-atom cell of the zinc-blende structure, whereas uncorrected energies can show large errors on the order of 1 eV. Thus, the individual treatment of size effects of different physical origins should generally serve to dramatically reduce the computational effort associated with finite-size scaling methods, without sacrificing the accuracy. Specifically, we demonstrated that both image-charge interactions and potential-alignment effects need to be corrected simultaneously.

Notably, we found that the third-order term in the Makov-Payne correction is proportional to the first-order term and, thus, effectively scales as $\sim q^2/L$ despite its nominal $\sim q/L^3$ dependence. This behavior results from implicit q and L dependencies in the second radial moment of the defect charge that also enters the third-order term. We explained the success of the image charge correction and the unexpected scaling behavior based on the observation that the defect charge has two distinct contributions: First, the difference in the ionic charge (nuclear charge+core electrons) upon atomic substitution creates an essentially point-charge-like localized contribution to the defect charge, which equals the nominal defect charge q. Second, the dielectric screening response of the semiconductor host produces a contribution to the defect charge which is delocalized throughout the supercell. Thus, the underlying physical picture of the image charge correction, i.e., a point charge in a dielectric medium, is essentially correct, and previous claims that the Makov-Payne formula leads to overcorrection due to the delocalization of the defect charge are not substantiated.

We further demonstrated that, in addition to the image charge interaction in the case of charged impurities, large and slowly convergent size-dependent energy contributions also occur when shallow donors or acceptors release their carriers into the dispersive conduction- or valence-band states. The resulting Moss-Burstein-type band-filling effects due to the high dopant concentration implied by the supercell formalism necessitate corrections reaching into the order of 1 eV for common supercell sizes. Regarding the single-particle defect energies, we demonstrated for the example of the V_O defect in ZnO the fast convergence of the Brillouin-zone-averaged defect level energy with cell size, whereas the energy at the zone center converges very slowly. It is important, however, to correctly discriminate between the primary defect state (which may occur either inside the gap or as a resonance inside the continuum of host band) and the secondary perturbed-host state. The impurity band dispersion arises from a concentration-dependent hybridization between defect and host states. Thus, the band-filling effect and the deep-level impurity band dispersion which both are slowly converging with cell size are distinctly different from the effect of direct defect-defect interaction (wave-function overlap), which would be expected to decay exponentially with supercell size.

ACKNOWLEDGMENTS

This work was funded by the U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy, under Contract No. DE-AC36-08GO28308 to NREL. S.L. acknowledges discussions with Hannes Raebiger on the topic of supercell defect calculations.

[1] F. A. Kröger, *The Chemistry of Imperfect Crystals* (North-Holland, Amsterdam, 1974).
[2] *Identification of Defects in Semiconductors*, Semiconductors and Semimetals, Vol. 51A, edited by M. Stavola (Academic, Boston, 1998); *Identification of Defects in Semiconductors*, Semiconductors and Semimetals, Vol. 51B, edited by M. Stavola (Academic, Boston, 1999).
[3] G. A. Baraff and M. Schlüter, Phys. Rev. Lett. **55**, 1327 (1985).
[4] S. B. Zhang and J. E. Northrup, Phys. Rev. Lett. **67**, 2339 (1991).
[5] C. Persson, Y. J. Zhao, S. Lany, and A. Zunger, Phys. Rev. B **72**, 035211 (2005).
[6] S. Lany and A. Zunger, Phys. Rev. Lett. **98**, 045501 (2007).
[7] S. Lany and A. Zunger, Phys. Rev. B **72**, 035215 (2005).
[8] S. Lany, Y. J. Zhao, C. Persson, and A. Zunger, Appl. Phys. Lett. **86**, 042109 (2005).
[9] S. B. Zhang, J. Phys.: Condens. Matter **14**, R881 (2002).
[10] L. Hedin, Phys. Rev. **139**, A796 (1965).
[11] X. Zhu and S. G. Louie, Phys. Rev. B **43**, 14142 (1991).
[12] M. Usuda, N. Hamada, T. Kotani, and M. van Schilfgaarde, Phys. Rev. B **66**, 125101 (2002).
[13] F. Gygi and A. Baldereschi, Phys. Rev. Lett. **62**, 2160 (1989).
[14] D. M. Bylander and L. Kleinman, Phys. Rev. B **41**, 7868 (1990).
[15] R. Asahi, W. Mannstadt, and A. J. Freeman, Phys. Rev. B **59**, 7486 (1999).
[16] J. Robertson, K. Xiong, and S. J. Clark, Thin Solid Films **496**, 1 (2006).
[17] J. D. Talman and W. F. Shadwick, Phys. Rev. A **14**, 36 (1976).
[18] R. W. Godby, M. Schlüter, and L. J. Sham, Phys. Rev. Lett. **56**, 2415 (1986).
[19] A. Görling and M. Levy, Phys. Rev. A **50**, 196 (1994).
[20] P. Rinke, A. Qteish, J. Neugebauer, C. Freysoldt, and M. Scheffler, New J. Phys. **7**, 126 (2005).
[21] A. D. Becke, J. Chem. Phys. **98**, 1372 (1993).
[22] J. Robertson, P. W. Peacock, M. D. Towler, and R. Needs, Thin Solid Films **411**, 96 (2002).
[23] C. H. Patterson, Phys. Rev. B **74**, 144432 (2006).
[24] J. P. Perdew and A. Zunger, Phys. Rev. B **23**, 5048 (1981).
[25] A. Svane and O. Gunnarsson, Phys. Rev. Lett. **65**, 1148 (1990); W. M. Temmermann, A. Svane, Z. Szotek, and H. Winter, in *Electronic Density Functional Theory*, edited by J. F. Dobson, G. Vignale, and M. P. Das (Plenum, New York, 1996).
[26] D. Vogel, P. Krüger, and J. Pollmann, Phys. Rev. B **54**, 5495 (1996).
[27] A. Filippetti and N. A. Spaldin, Phys. Rev. B **67**, 125109 (2003).
[28] C. D. Pemmaraju, T. Archer, D. Sánchez-Portal, and S. Sanvito, Phys. Rev. B **75**, 045101 (2007).
[29] B. Lee, L.-W. Wang, C. D. Spataru, and S. G. Louie, Phys. Rev. B **76**, 245114 (2007).
[30] S. Kümmel and L. Kronik, Rev. Mod. Phys. **80**, 3 (2008).
[31] G. Makov and M. C. Payne, Phys. Rev. B **51**, 4014 (1995).
[32] T. S. Moss, Proc. Phys. Soc. London, Sect. B **67**, 775 (1954); E. Burstein, Phys. Rev. **93**, 632 (1954).
[33] A. F. Kohan, G. Ceder, D. Morgan, and Chris G. Van de Walle, Phys. Rev. B **61**, 15019 (2000).
[34] S. B. Zhang, S.-H. Wei, and A. Zunger, Phys. Rev. B **63**, 075205 (2001).
[35] F. Oba, S. R. Nishitani, S. Isotani, H. Adachi, and I. Tanaka, J. Appl. Phys. **90**, 824 (2001).
[36] E.-C. Lee, Y.-S. Kim, Y.-G. Jin, and K. J. Chang, Phys. Rev. B **64**, 085120 (2001).
[37] A. Janotti and Chris G. Van de Walle, Appl. Phys. Lett. **87**, 122102 (2005).
[38] W.-J. Lee, J. Kang, and K. J. Chang, Phys. Rev. B **73**, 024117 (2006).
[39] P. Erhart, K. Albe, and A. Klein, Phys. Rev. B **73**, 205203 (2006).
[40] A. Janotti and Chris G. Van de Walle, Phys. Rev. B **76**, 165202 (2007).
[41] H. Takenaka and D. J. Singh, Phys. Rev. B **75**, 241102(R) (2007).
[42] C. D. Pemmaraju, R. Hanafin, T. Archer, H. B. Braun, and S. Sanvito, Phys. Rev. B **78**, 054428 (2008).
[43] F. Oba, A. Togo, I. Tanaka, J. Paier, and G. Kresse, Phys. Rev. B **77**, 245202 (2008).
[44] T. R. Paudel and W. R. L. Lambrecht, Phys. Rev. B **77**, 205202 (2008).
[45] P. A. Schultz, Phys. Rev. Lett. **84**, 1942 (2000).
[46] J. Lento, J.-L. Mozos, and R. M. Nieminen, J. Phys.: Condens. Matter **14**, 2637 (2002).
[47] U. Gerstmann, P. Deák, R. Rurali, B. Aradi, T. Frauenheim, and H. Overhof, Physica B (Amsterdam) **340-342**, 190 (2003).
[48] D. Segev and S.-H. Wei, Phys. Rev. Lett. **91**, 126406 (2003).
[49] M. Bockstedte, A. Mattausch, and O. Pankratov, Phys. Rev. B **68**, 205201 (2003).
[50] C. W. M. Castleton and S. Mirbt, Phys. Rev. B **70**, 195202 (2004).
[51] S. Limpijumnong, S. B. Zhang, S.-H. Wei, and C. H. Park, Phys. Rev. Lett. **92**, 155504 (2004).
[52] C. G. van de Walle and J. Neugebauer, J. Appl. Phys. **95**, 3851 (2004).
[53] A. Höglund, C. W. M. Castleton, and S. Mirbt, Phys. Rev. B **72**, 195213 (2005).
[54] C. W. M. Castleton, A. Höglund, and S. Mirbt, Phys. Rev. B **73**, 035215 (2006).
[55] J. Shim, E.-K. Lee, Y. J. Lee, and R. M. Nieminen, Phys. Rev. B **71**, 035206 (2005).
[56] A. F. Wright and N. A. Modine, Phys. Rev. B **74**, 235209 (2006).
[57] A. Gali, T. Hornos, N. T. Son, E. Janzén, and W. J. Choyke, Phys. Rev. B **75**, 045211 (2007).
[58] F. Tuomisto, K. Saarinen, K. Grasza, and A. Mycielski, Phys.

Status Solidi B **243**, 794 (2006).
[59] R. M. de la Cruz, R. Pareja, R. González, L. A. Boatner, and Y. Chen, Phys. Rev. B **45**, 6581 (1992).
[60] L. E. Halliburton, N. C. Giles, N. Y. Garces, M. Luo, C. Xu, L. Baic, and L. A. Boatner, Appl. Phys. Lett. **87**, 172108 (2005).
[61] J. H. W. de Wit, J. Solid State Chem. **13**, 192 (1975); **20**, 143 (1977); J. H. W. de Wit, G. van Unen, and M. Lahey, J. Phys. Chem. Solids **38**, 819 (1977).
[62] J. Mizusaki, H. Koinuma, J. I. Shimoyama, M. Kawasaki, and K. Fueki, J. Solid State Chem. **88**, 443 (1990).
[63] G. H. Rosenblatt, M. W. Rowe, G. P. Williams, Jr., R. T. Williams, and Y. Chen, Phys. Rev. B **39**, 10309 (1989).
[64] S. Lany, H. Wolf, and Th. Wichert, Phys. Rev. Lett. **92**, 225504 (2004).
[65] J. C. Slater, *Quantum Theory of Molecules and Solids* (McGraw-Hill, New York, 1974), Vol. 4.
[66] A. Zunger and A. J. Freeman, Phys. Rev. B **16**, 2901 (1977).
[67] T. C. Koopmans, Physica (Amsterdam) **1**, 104 (1934).
[68] S. Lany and A. Zunger, Phys. Rev. Lett. **100**, 016401 (2008).
[69] J. Ihm, A. Zunger, and M. L. Cohen, J. Phys. C **12**, 4409 (1979).
[70] P. E. Blöchl, Phys. Rev. B **50**, 17953 (1994).
[71] G. Kresse and D. Joubert, Phys. Rev. B **59**, 1758 (1999).
[72] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).
[73] B. Hammer, L. B. Hansen, and J. K. Norskov, Phys. Rev. B **59**, 7413 (1999).
[74] Note, however, that the O_2 molecule is not well described with the soft pseudopotential.
[75] R. Cherian and P. Mahadevan, Phys. Rev. B **76**, 075205 (2007).
[76] V. I. Anisimov, J. Zaanen, and O. K. Andersen, Phys. Rev. B **44**, 943 (1991).
[77] V. I. Anisimov, I. V. Solovyev, M. A. Korotin, M. T. Czyzyk, and G. A. Sawatzky, Phys. Rev. B **48**, 16929 (1993).
[78] A. I. Liechtenstein, V. I. Anisimov, and J. Zaanen, Phys. Rev. B **52**, R5467 (1995).
[79] S. Lany and A. Zunger, Phys. Rev. Lett. **93**, 156404 (2004).
[80] J. E. Jaffe and A. Zunger, Phys. Rev. B **29**, 1882 (1984).
[81] S. Lany, J. Osorio-Guillén, and A. Zunger, Phys. Rev. B **75**, 241203(R) (2007).
[82] C. Persson and A. Zunger, Appl. Phys. Lett. **87**, 211904 (2005).
[83] H. Raebiger, S. Lany, and A. Zunger, Phys. Rev. B **76**, 045209 (2007).
[84] In order to apply the LDA+$U_{s/d}$ method in Sec. III to ZnO, we modified the VASP code so as to treat more than one angular momentum number in LDA+U.
[85] S. L. Dudarev, G. A. Botton, S. Y. Savrasov, C. J. Humphreys, and A. P. Sutton, Phys. Rev. B **57**, 1505 (1998).
[86] A. Önsten, M. Månsson, T. Claesson, T. Muro, T. Matsushita, T. Nakamura, T. Kinoshita, U. O. Karlsson, and Oscar Tjernberg, Phys. Rev. B **76**, 115127 (2007).
[87] O. Gunnarsson, O. K. Andersen, O. Jepsen, and J. Zaanen, Phys. Rev. B **39**, 1708 (1989).
[88] L. Wang, T. Maxisch, and G. Ceder, Phys. Rev. B **73**, 195107 (2006).
[89] M. Cococcioni and S. de Gironcoli, Phys. Rev. B **71**, 035105 (2005).
[90] H. J. Kulik, M. Cococcioni, D. A. Scherlis, and N. Marzari, Phys. Rev. Lett. **97**, 103001 (2006).
[91] W. E. Pickett, S. C. Erwin, and E. C. Ethridge, Phys. Rev. B **58**, 1201 (1998).
[92] V. I. Anisimov and O. Gunnarsson, Phys. Rev. B **43**, 7570 (1991).
[93] C. Franchini, R. Podloucky, J. Paier, M. Marsman, and G. Kresse, Phys. Rev. B **75**, 195128 (2007).
[94] S. H. Wei, Comput. Mater. Sci. **30**, 337 (2004).
[95] D. B. Laks, C. G. Van de Walle, G. F. Neumark, P. E. Blöchl, and S. T. Pantelides, Phys. Rev. B **45**, 10965 (1992).
[96] T. Mattila and A. Zunger, Phys. Rev. B **58**, 1367 (1998).
[97] S. Lany and A. Zunger, J. Appl. Phys. **100**, 113725 (2006).
[98] J. P. Perdew and M. Levy, Phys. Rev. Lett. **51**, 1884 (1983).
[99] L. J. Sham and M. Schlüter, Phys. Rev. Lett. **51**, 1888 (1983).
[100] R. W. Godby, M. Schlüter, and L. J. Sham, Phys. Rev. B **37**, 10159 (1988).
[101] O. Gunnarsson and K. Schönhammer, Phys. Rev. Lett. **56**, 1968 (1986).
[102] W. G. Aulbur, L. Jönsson, and J. W. Wilkins, Solid State Phys. **54**, 1 (2000).
[103] Since ideally delocalized, i.e., bandlike, free electrons and holes do not break translational symmetry, $E_H(+1)$ and $E_H(-1)$ are calculated using a four-atom ZnO unit cell with the corresponding carrier density. A $18\times18\times12$ k mesh and the linear tetrahedron method are used for Brillouin-zone integrations. The interaction of the electronic charge with the jellium background is included (see Sec. II D).
[104] A. M. Stoneham, J. Gavartin, A. L. Shluger, A. V. Kimmel, D. Muñoz Ramo, H. M. Rønnow, G. Aeppli, and C. Renner, J. Phys.: Condens. Matter **19**, 255208 (2007).
[105] J. Lægsgaard and K. Stokbro, Phys. Rev. Lett. **86**, 2834 (2001).
[106] R. O. Jones and O. Gunnarsson, Rev. Mod. Phys. **61**, 689 (1989).
[107] Calculated in a slightly asymmetric cubic box of 15 Å side length and with the symmetry-broken solutions for F^0 and F^+. The monopole correction of Eq. (11) is included for F^+ and F^-. (The third-order term is small, on the order of 0.01 eV; omitting the monopole correction yields $I-A=11.25$ eV for this cell size.)
[108] J. P. Perdew, A. Ruzsinszky, G. I. Csonka, O. A. Vydrov, G. E. Scuseria, V. N. Staroverov, and J. Tao, Phys. Rev. A **76**, 040501(R) (2007).
[109] P. Mori-Sánchez, A. J. Cohen, and W. Yang, Phys. Rev. Lett. **100**, 146401 (2008).
[110] H. Raebiger, S. Lany, and A. Zunger, Phys. Rev. Lett. **99**, 167203 (2007).
[111] A. Janotti, D. Segev, and C. G. Van de Walle, Phys. Rev. B **74**, 045202 (2006).
[112] Despite the (global) C_{3v} symmetry of the wurtzite ZnO lattice, the local symmetry is close to T_d with a near-tetrahedral coordination. In principle, a triply degenerate t_2 state splits into a nondegenerate a_1 and a doubly degenerate e_g level. We find, however, that this effect is small. For example, the splitting of the t_2 state of V_O in ZnO (Fig. 4) is only 0.2 eV and is very small compared to the a_1-t_2 splitting of ~4 eV. Further, we find that the splitting of the Zn d-like t_2 level into a_1 and e_g sublevels does not introduce significant coupling with the a_1 symmetric deep O s state or the a_1 symmetric CBM, as evident from comparison with LDA+U calculations in zinc-blende ZnO, where the d-t_2 level of Zn does not couple to the a_1 levels by symmetry.
[113] A. Alkauskas, P. Broqvist, and A. Pasquarello, Phys. Rev. Lett. **101**, 046405 (2008).

[114] J. Vidal and F. Bruneval (private communication).
[115] S. B. Zhang, S.-H. Wei, and A. Zunger, Phys. Rev. Lett. **84**, 1232 (2000).
[116] Note that the underlying theory of extrapolation methods described in Sec. III C concerns single-particle energies, whereas in practice such extrapolation schemes are usually applied to $\varepsilon(q/q')$ transition energies. Thus, the perturbation-extrapolation is implicitly also imposed on electronic and atomic relaxation effects (see Sec. II B).
[117] U. Lindefelt and A. Zunger, Phys. Rev. B **26**, 846 (1982).
[118] N. E. Christensen, Phys. Rev. B **30**, 5753 (1984).
[119] S. H. Wei and A. Zunger, Phys. Rev. B **48**, 6111 (1993).
[120] S. H. Wei and A. Zunger, Phys. Rev. B **57**, 8983 (1998).
[121] L.-W. Wang, Appl. Phys. Lett. **78**, 1565 (2001).
[122] S. Limpijumnong and W. R. L. Lambrecht, Phys. Rev. B **63**, 104103 (2001).
[123] S. Lany, H. Raebiger, and A. Zunger, Phys. Rev. B **77**, 241201(R) (2008).
[124] S. Y. Ren, J. D. Dow, and D. J. Wolford, Phys. Rev. B **25**, 7661 (1982).
[125] L. S. Vlasenko and G. D. Watkins, Phys. Rev. B **71**, 125210 (2005).
[126] L. S. Vlasenko and G. D. Watkins, Physica B (Amsterdam) **376-377**, 677 (2006).
[127] R. A. Powell, W. E. Spicer, and J. C. McMenamin, Phys. Rev. Lett. **27**, 97 (1971).
[128] C. J. Vesely, R. L. Hengehold, and D. W. Langer, Phys. Rev. B **5**, 2296 (1972).
[129] L. Ley, R. A. Pollak, F. R. McFeely, S. P. Kowalczyk, and D. A. Shirley, Phys. Rev. B **9**, 600 (1974).
[130] *CRC Handbook of Chemistry and Physics*, 87th ed. (Taylor & Francis, Boca Raton, 2007).
[131] J. M. Smith and W. E. Vehse, Phys. Lett. **31A**, 147 (1970).
[132] P. J. Stephens, F. J. Devlin, C. F. Chabalowski, and M. J. Frisch, J. Phys. Chem. **98**, 11623 (1994).
[133] J. Heyd, G. E. Scuseria, and M. Ernzerhof, J. Chem. Phys. **118**, 8207 (2003).
[134] Due to finite-size effects in the supercell used in Refs. 23 and 43, the energy of the a_1 state of V_O appears at somewhat too low energy at the Brillouin-zone center (cf. Fig. 5 and discussion in Sec. IV B 2). Therefore, we take here the Brillouin-zone average of the a_1 state as shown in Figs. 2 and 4 of Ref. 23 and in Fig. 3 of Ref. 43.
[135] J. Osorio-Guillén, S. Lany, S. V. Barabash, and A. Zunger, Phys. Rev. Lett. **96**, 107203 (2006).
[136] J. Osorio-Guillén, S. Lany, and A. Zunger, Phys. Rev. Lett. **100**, 036601 (2008).
[137] H. P. Hjalmarson, P. Vogl, D. J. Wolford, and J. D. Dow, Phys. Rev. Lett. **44**, 810 (1980).
[138] L. M. Sandratskii, P. Bruno, and J. Kudrnovsky, Phys. Rev. B **69**, 195203 (2004).
[139] J. E. Northrup and S. B. Zhang, Phys. Rev. B **50**, 4962(R) (1994).
[140] In the analytic form of the first-order correction term in Eq. (11), according to Makov and Payne (Ref. 31), the corresponding scaling prefactor γ_1 [cf. Eq. (22)] is slightly different for the different supercell geometries, i.e. $\gamma_1(\text{fcc}) \approx \gamma_1(\text{bcc}) = 1.018\gamma_1(\text{sc})$. Since the difference (less than 2%) is so small, we combine the data in one graph, thereby taking the advantage to use more data points for the fit of Eq. (22).
[141] In order to avoid the large (positive and negative) local charge differences caused by a shift of atoms, we determine the second-order moment Q_r in Eq. (11) with the same atomic relaxation pattern in the defect and host-reference calculations.
[142] This slow convergence is likely related to the fact that in the fcc supercell symmetry, the defect images are connected along the cation-anion "zigzag" chains in the [110] direction of the zinc-blende lattice, which apparently causes larger residual defect-defect interactions than in case of the sc or bcc symmetries, where the defect images are oriented along the [001] and [111] lattice directions, respectively.
[143] J. Dabrowski and M. Scheffler, Phys. Rev. Lett. **60**, 2183 (1988).
[144] J. Gebauer, E. R. Weber, N. D. Jäger, K. Urban, and Ph. Ebert, Appl. Phys. Lett. **82**, 2059 (2003).
[145] J. Li and S.-H. Wei, Phys. Rev. B **73**, 041201(R) (2006).
[146] F. A. Selim, M. H. Weber, D. Solodovnikov, and K. G. Lynn, Phys. Rev. Lett. **99**, 085502 (2007).
[147] F. H. Leiter, H. R. Alves, A. Hofstaetter, D. M. Hofmann, and B. K. Meyer, Phys. Status Solidi A **226**, R4 (2001).
[148] J. Schneider and A. Räuber, Solid State Commun. **5**, 779 (1967); K. Leutwein, A. Räuber, and J. Schneider, *ibid.* **5**, 783 (1967).
[149] S. M. Evans, N. C. Giles, L. E. Halliburton, and L. A. Kappers, J. Appl. Phys. **103**, 043710 (2008).
[150] V. Soriano and D. Galland, Phys. Status Solidi B **77**, 739 (1976).
[151] D. R. Locker and J. M. Meese, IEEE Trans. Nucl. Sci. **19**, 237 (1972).
[152] Y. Takahashi, M. Kanamori, A. Kondoh, H. Minoura, and Y. Ohya, Jpn. J. Appl. Phys., Part 1 **33**, 6611 (1994).
[153] D. Seghier and H. P. Gislason, Physica B (Amsterdam) **401-402**, 404 (2007).
[154] F. H. Leiter, H. Alves, D. Pfisterer, N. G. Romanov, D. M. Hofmann, and B. K. Meyer, Physica B (Amsterdam) **340-342**, 201 (2003).