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

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 .h2d /geneo9c8tainy.fis dominated 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.

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-

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

ΔHf ge47,6798446Sf03of3069.ages-38f]7.99846 0 0 6.98465364TmOTj9.978 0 0 9.978 183.8755 99.079Tj704 07.1itsi38f

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

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

$$\tag{1}$$

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

1. ⸱ ⸱ ⸱ *t* ⸱ ⸱ ⸱ ⸱⸱ ⸱ ⸱ ⸱⸱ ⸱⸱

$E_V = E_H\left(0\right) - E_H\left(+1\right)$ is defined as the energy difference between the pure host $\left(q=0\right)$ and the host with one hole $\left(q=+1\right)$ 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

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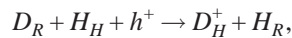
where D_R denotes the donor atom in its chemical reservoir before defect formation, H_H denotes a host atom at its native lattice site, and h^+ denotes a hole at the valence-band maximum VBM.

2. ⸱ ⸱ ⸱ ⸱

E_F is conventionally defined with respect to VBM, $E_F = E_V$

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 $_i$ is the energy gain during relaxation of the initial-state wave functions. Comparing with the Hartree-Fock HF

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. $\;$ $_{,\;}$ $^{t}_{,}$ ^{t} $_{,\;\cdots\;,\;\cdots\;\cdots}$ $_{,\;\cdots}$ $_{,\;\cdots}$ $^{t}_{,\;}$ ^{t} ( )

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 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. $\;$ $_{,\;}$ $^{t}_{,}$ $_{,\;\cdots\;\cdots}$ ^{t} ^{t} $_{,\;}$ $^{t}_{,\;}$ $^{t}_{,\;\cdots}$ ( )

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.

$\;$ $_{\cdot\;\;\cdots\cdots\;\cdots\;\;\cdots}$ $_{\cdot\cdots}$ ( )

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

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

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*, $e_{\text{KS}} \Gamma = e_C \Gamma - e_V \Gamma$, in the limit of a dilute gas of free electrons and holes. Thus, for approximate functionals such as LDA or GGA, the quasiparticle gap $E_{\text{QP}}^{\text{LDA}}$ shows the same band-gap error as the single-particle gap $e_{\text{KS}}^{\text{LDA}}$. 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,

$$e_g \text{ BZ av} = \sum_{\mathbf{k}} w_{\mathbf{k}} \, \eta_{C,\mathbf{k}} e_{C,\mathbf{k}}^* - \eta_{V,\mathbf{k}} e_{V,\mathbf{k}}^* \, , \qquad 13$$

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 $I = -e_V$

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.

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$_2$O,[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

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:

$$e_{n,\mathbf{k}}(\lambda) = \langle \psi_{n,\mathbf{k}} | \mathbf{H}_H | \psi_{n,\mathbf{k}} \rangle + \lambda \langle \psi_{n,\mathbf{k}} | \Delta \mathbf{H}_P | \psi_{n,\mathbf{k}} \rangle$$

$$= e_{n,\mathbf{k}}(0) + \lambda \frac{\partial e_{n,\mathbf{k}}(\lambda)}{\partial \lambda}. \qquad (16)$$

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

$$e_D(\lambda) = \langle \psi_D | \mathbf{H}_H + \Delta \mathbf{H}_D | \psi_D \rangle + \lambda \langle \psi_D | ,\mathbf{k} \rangle$$

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

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

1.

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

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		
$\varepsilon(2+/0)=E$	I3e5E				

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 oxy-gen 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 ε $2+/0 = E_V + 2.3$ eV, close to the result of Refs. [37] and

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

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

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_{As}^{3+} 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 bcc 141352944 8 s k0 Tc with 378.A5 pote5 J/pply data5 Jbofit 5 J upe5 Jalignment 312.5 and TJ/F4 1 Tf20.6732 0 ave

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. *······ t· ··· ··· t··· ··· ··· ··· t···*

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_{As}^{3+} the third-order correction ΔE_{MP}^3 (second term in Eq. (11)) as a function of the respective first correction ΔE_{MP}^1 (first term in Eq. (11)), we find a clear proportionality, shown in Fig. 8,

$$\Delta E_{MP}^3 = f \Delta E_{MP}^1, \qquad (23)$$

which strongly deviates from the behavior that would be expected from the nominal $1/L^3$ scaling of the third-order term ΔE_{MP}^3, 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_{MP}^3 scales effectively in the same way as ΔE_{MP}^1, i.e., as q^2/L, which indicates the implicit dependency $Q_r \propto qL^2$ 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_{MP}^3 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_{MP}^1, the remaining finite-size dependence scales solely as $1/L^3$.

In order to study the origin of the dependency $Q_r \propto qL^2$ and, hence, of the unexpected scaling behavior of the third-order term ΔE_{MP}^3, 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

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_{MP}^3 and ΔE_{MP}^1: The delocalized part of the defect density $\widetilde{\rho}_D(\mathbf{r})$ 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 $Q_r \propto qL^2$, which explains the observed $1/L$ scaling of ΔE_{MP}^3. Since, ΔE_{MP}^3 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

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_{As}^0 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 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_{As}^0 donor in GaAs is in stark contrast with the behavior of the deep V_O^0 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 ΔH V_O^0 from the cell size corroborates our argument see Sec. III E 1 that the donor state of V_O^0 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_O^0 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 ΔH V_O^0 . 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.

Regarding the convergence of *single-particle* defect states, we find pronounced finite-size effects for the a_1 gap level of V_O^0 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

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.

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 perturba-

---

[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]

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

[114] J. Vidal and F. Bruneval private communication .

[115] S. B. Zhang, S.-H. Wei, and A. Zunger, Phys. Rev. Lett. **84**, 1232 2000 .