Managing the supercell approximation for charged defects in semiconductors: Finite-size scaling, charge correction factors, the band-gap problem, and the *ab initio* dielectric constant

C. W. M. Castleton,[1,2,*] A. Höglund,[3] and S. Mirbt[3]

[1] *Material Physics, Materials and Semiconductor Physics Laboratory, Royal Institute of Technology (KTH), Electrum 229, SE-16440 Kista, Sweden*

[2] *Department of Physical Electronics/Photonics, ITM, Mid Sweden University, SE-85170 Sundsvall, Sweden*

[3] *Theory of Condensed Matter, Department of Physics, Uppsala University, Box 530, SE-75121 Uppsala, Sweden*

(Received 2 November 2005; published 25 January 2006)

The errors arising in *ab initio* density functional theory studies of semiconductor point defects using the supercell approximation are analyzed. It is demonstrated that (a) the leading finite size errors are inverse linear and inverse cubic in the supercell size and (b) finite size scaling over a series of supercells gives reliable isolated charged defect formation energies to around ±0.05 eV. The scaled results are used to test three correction methods. The Makov-Payne method is insufficient, but combined with the scaling parameters yields an *ab initio* dielectric constant of 11.6±4.1 for InP. Γ point corrections for defect level dispersion are completely incorrect, even for shallow levels, but realigning the total potential in real-space between defect and bulk cells actually corrects the electrostatic defect-defect interaction errors as well. Isolated defect energies to ±0.1 eV are then obtained using a 64 atom supercell, though this does not improve for larger cells. Finally, finite size scaling of known dopant levels shows how to treat the band gap problem: in ⩽200 atom supercells with no corrections, continuing to consider levels into the theoretical conduction band (extended gap) comes closest to experiment. However, for larger cells or when supercell approximation errors are removed, a scissors scheme stretching the theoretical band gap onto the experimental one is in fact correct.

DOI: 10.1103/PhysRevB.73.035215 PACS number(s): 61.72.Bb, 61.72.Ji, 71.15.Dx, 71.55.Eq

I. INTRODUCTION

Understanding the properties of point defects and dopants is of key importance in studying the electrical and optical properties of semiconductors. While various experimental techniques have been developed over the last half century it is only in recent years that they have started to be matched by accurate first principles computational techniques. Developments in computing power have now made *ab initio* density functional theory[1] (DFT) one of the most versatile atomic scale tools available for the investigation of defect properties in semiconductors and insulators. The key quantity to calculate is the defect formation energy

$$E_d^C = E_T^C(\text{defect}^q) - E_T^C(\text{no defect}) + \sum_i \mu_i n_i - q(\epsilon_v + \epsilon_F), \tag{1}$$

where $E_T^C(\text{defect})$ and $E_T^C(\text{no defect})$ are the total energy of the supercell C with and without the defect (of charge q) calculated using the same values of planewave cutoff, k-point grid, etc., to make use of the cancellation of errors. The defect is formed by adding/removing n_i atoms of chemical potential μ_i. ϵ_F is the Fermi level, measured from ϵ_v, the valence band edge (VBE). Almost all properties of a defect can be derived from variations in and differences between formation energies. The method is very powerful, but critical limitations remain, two of the most important being the relatively small number of atoms which can be treated and the effect of the approximations, such as the local density approximation (LDA) and generalized gradient approximation (GGA), required to solve the DFT itself. These treat quantum many body correlation and exchange effects incompletely, which in the case of semiconductors and insulators results in a roughly 50% underestimation of the bandgap. This in turn has severe consequences for the calculation of defect transfer levels, the values of ϵ_F at which the most stable charge state of the defect changes, given by the difference in E_d^C between the two states. The positions of the transfer levels govern whether the defect will be a single or multiple acceptor or donor, with levels deep inside the gap, or with shallow levels near to the band edges. Since the predicted band gap differs so strongly from the experimental one it is very hard to map the calculated transfer levels onto the experimental gap and hence predict the electrical properties of the material.

Meanwhile, the small number of atoms involved (100s or 1000s) means that the boundary conditions become very important. One of the most common approaches is to use periodic boundary conditions (PBCs) together with a plane wave basis set. [2] A supercell containing the defect in question is repeated periodically throughout space. The cell boundary thus looks bulklike, rather than being a vacuum as with open boundary conditions. However, it also means that the defect interacts with an infinite array of images of itself seen in the PBCs. This alters E_d^C, making it (and most other defect properties) supercell size-dependent. The "true" defect properties are only recovered in the limit of an infinitely large supercell, equivalent to the limit of an isolated defect. This problem is particularly severe in the case of charged defects, where the Madelung energy becomes infinite if the charge is not neutralized using a uniform jellium background.[3] Even with jellium, the calculated formation energies can be wrong by several eV in supercells of the order of 10s or 100s of atoms, and we have previously shown[4,5] that finite size errors on this scale can even arise for neutral defects. Various authors[6–8] have attempted to create correction schemes to estimate and remove these errors, the most widely known being that of Makov and Payne.[8] Although these corrections are often used their accuracy has been strongly questioned, with several studies suggesting that they are not reliable enough for regular use.[6,9,10]

We previously[4,5] suggested that the supercell size errors can instead be eliminated by calculating the same defect properties in a series of supercells of different sizes but the same symmetry and then finite size scaling the results to recover those of the infinite supercell. We found that E_d^C varies with the supercell size L as

$$E_d^C(L) = E_d^\infty + a_1 L^{-1} + a_n L^{-n}, \tag{2}$$

where a_1, a_n, and E_d^∞ are fitting parameters, E_d^∞ being the finite size-scaled formation energy corresponding to an infinitely large supercell. The linear term has been discussed many times previously, first by Leslie and Gillian.[3] For neutral defects we found the correct value for n to be 3. This is actually very intuitive: most sources of error should vary with either the supercell size L (the defect-defect image distance) or with the cell volume L^3 (proportional to the jellium charge density, the number of atoms, the number of electrons, etc.). Terms scaling with the surface area $6L^2$ seem unlikely to be dominant.

Here, two further sources of error must also be considered. First, since the electrostatic potential in a supercell with PBCs is only defined up to a constant, the zero on the energy scale must be chosen arbitrarily in each calculation. In the case of most pseudopotential codes (including the one we use) this occurs as an implicit average over values appropriate to each atom species in the supercell, weighted by the number of atoms of each species. This means that the numerical value of ϵ_F entering Eq. (1) changes with the contents of the cell, leading to an additional finite size error. If the number of defects per supercell is constant then this error decreases with the number of atoms in the cell—essentially with the volume of the cell L^3. Hence this error is completely taken care of in the infinite supercell limit of our finite size scaling scheme. For individual supercells, Van de Walle and Neugebauer[10] suggest correcting the error by realigning the potential in the defect cell to that of the bulk, using its real-space value at some chosen point in a bulklike region far from the defect. (We here use the point furthest from the defect in the unrelaxed cell.)

Secondly, additional errors come from the dispersion of the defect levels introduced by overlap between the defect state wave functions and their PBC images. It has been suggested[11] that this artificially raises E_d^C when k points other than just the Γ point are used. It is suggested[11] that E_d^C should then be shifted by $q\times(\epsilon_D^\Gamma - \epsilon_D^{\mathrm{KS}})$, where ϵ_D^Γ and ϵ_D^{KS} are the values of the Kohn-Sham level corresponding to the defect state calculated in the defect cell at the Γ point and averaged over the sampled k points respectively. The assumption is that the value of the defect level is correct at the Γ point, so the difference between that and the k-point averaged value should be removed. It has been shown by Höglund *et al.*[12] that this is completely incorrect for the example of the phosphorus antisite in GaP. By plotting the "band structure" of the defect level in different sized supercells it was shown that the defect level in the smaller cells is more or less correct when averaged over the sampled k points, but much too low at the Γ point. The same is also true for the As vacancy on the GaAs(110) surface, for example.[13] Van de Walle and Neugebauer[10] instead point out[10] that in this respect there is a fundamental difference between deep levels such as these and shallow defect levels. They suggest that the correction should only be applied when evaluating transfer levels for shallow donors and acceptors.

In the current paper we will show in Sec. III that Eq. (2) with $n=3$ also holds for charged defects, so that finite size scaling can be used to produce fully finite-size corrected defect formation and other energies, with well defined error bars and uncertainty. To do this we will study 11 example defects in the zinc-blende structured III-V semiconductor InP. These are chosen to include all types of native defects (vacancies, antisites, and interstitials) as well as some common dopants at both substitutional and interstitial sites. Each is studied in one charge state only, usually the one that previous studies[5,14] suggest it has over the majority of the bandgap. The specific choices have been made to include all nonzero values from -3 to $+3$.

These results will also enable us in Sec. IV to perform the most objective and comprehensive reliability test we are currently aware of on other, computationally cheaper, correction schemes. (Previous tests rely on only one or two—usually rather simple—examples, and do not generally have reliable isolated formation energies to compare with.) We will test the Makov-Payne scheme, potential realignment and dispersion corrections for shallow levels. In Sec. IV D we will also derive an *ab initio* dielectric constant for InP by combining the scaling results with those of the Makov-Payne correction scheme. Finally, in Sec. V we will use finite size scaling to provide the first clear-cut answer to the problem of mapping LDA or GGA transfer levels onto the experimental bandgap. Computational details are in the next section, and in Sec. VI we will conclude.

II. COMPUTATIONAL DETAILS

We use plane-wave *ab initio* DFT (Ref. 1) within the local density approximation (LDA) together with ultrasoft pseudopotentials[15] (USPP) using the VASP code.[16] Since we expect (at least) a three parameter fit we need at least four supercells. These must all be of the same symmetry since the errors scale differently for different symmetries. We choose simple cubic supercells containing 8, 64, 216, and 512 atoms. It would be preferable to replace the 8 atom cell by the 1000 atom one, but our computing resources are currently insufficient for k-point converged calculations with 1000 atoms. On the other hand, we previously found that, somewhat surprisingly, the 8 atom supercell is good enough in most cases: formation energies in this cell usually lie very close to the scaling curves, providing satisfactorily small error bars on the scaled values. Similarly, memory limitations force us to treat the indium $4d$ electrons as core, even though they are comparatively shallow: about 14.5 eV below the VBE. This leads[5] to errors of up to $\sim$0.5 eV, but these are essentially supercell size-independent. They can easily be estimated in, say, the 64 atom cell and added back onto the scaled E_d^{∞} at the end. Our optimized LDA lattice constant using these chosen pseudopotentials[17] is 5.827 Å and the band gap is 0.667 eV, compared to 5.869 Å and 1.344 eV in experiment. We use $\mu_{\mathrm{P}}=3.485$ eV and $\mu_{\mathrm{In}}=6.243$ eV, corresponding to stoichiometric conditions, together with $\mu_{\mathrm{Zn}}=1.891$ eV, $\mu_{\mathrm{Si}}=5.977$ eV, and $\mu_{\mathrm{S}}=4.600$ eV. For the 64 atom cell a plane-wave cutoff energy of 200 eV and a Monkhorst-Pack $4\times4\times4$ k-point grid[18] was previously found[17] sufficient to restrict errors to O(0.01 eV) or less. When analyzing the errors arising from the supercell approximation itself, nonfinite size-dependent errors[5] (from the In pseudopotential, plane-wave cutoff, etc.) are not a problem. However, we do need to keep the k-point sampling errors down to at least the meV scale, since this convergence rate varies with supercell size. This is a much higher convergence criterion than is normally practical, necessary or even meaningful, and it is the reason that we pick only a limited number of example defects for this study. This convergence level can be achieved[5] by using the average

$$\overline{E_d^C} = \frac{\sum_N N^3 E_d^C(N)}{\sum_N N^3} \tag{3}$$

weighted by the number of points in the full Brillouin zone, where $E_d^C(N)$ is the formation energy calculated using an $N\times N\times N$ Monkhorst-Pack k-point grid. The sum over N is taken up to 12 in the 8 atom cell, 8 in the 64 atom cell and 4, or for certain cases 6, in the 216 and 512 atom supercells in the unrelaxed geometries. [The weighted mean $\overline{E_d^C}$ converges much faster than the unweighted mean or the individual values $E_d^C(N)$ themselves.]

We present both nonrelaxed (ions at ideal lattice sites) and relaxed calculations. No restrictions are placed upon the symmetry of relaxations, but we do not allow atoms located on the surface of the cell to relax. The relaxation energy

$$\epsilon_{\mathrm{relax}}(N) = E_d^{C:\mathrm{Rx}}(N) - E_d^{C:\mathrm{Id}}(N), \tag{4}$$

where $E_d^{C:\mathrm{Rx}}(N)$ and $E_d^{C:\mathrm{Id}}(N)$ are $E_d^C(N)$ with atoms at relaxed and ideal positions, respectively, converges faster with N than either $E_d^{C:\mathrm{Rx}}(N)$ or $E_d^{C:\mathrm{Id}}(N)$. Hence we save computational time by approximating the relaxed formation energies $E_d^{C:\mathrm{Rx}}$ by

$$\overline{E_d^{C:\mathrm{Rx}}} \approx \overline{E_d^{C:\mathrm{Id}}} - \epsilon_{\mathrm{relax}}(N) = \overline{E_d^{C:\mathrm{Id}}} + E_d^{C:\mathrm{Rx}}(N) - E_d^{C:\mathrm{Id}}(N). \tag{5}$$

The relaxation energies used are weighted averages using $6\times6\times6$ and $8\times8\times8$ k-point grids in the 8 atom cell, $2\times2\times2$ and (if the convergence is uncertain) $4\times4\times4$ grids in the 64 atom cell and $2\times2\times2$ in the 216 and 512 atom supercells. For the latter cells we usually restrict the k-point grid to the irreducible Brillouin zone of the undisturbed bulk lattice. In other words, we use just the special k point (0.25,0.25,0.25): the first Chadi-Cohen k point.[19] This restriction is equivalent to assuming that the distortion in the band structure due to the presence of the defect is either localized (thus important only very near Γ) or symmetric. It introduces a small error whose significance again disappears in the large supercell limit.

III. FINITE SIZE SCALING OF DEFECT FORMATION ENERGIES

A. Scaled formation energies for the example defects

Figure 1 shows the formation energy scaling for the 11 example defects in InP. The scaling curves using the uncorrected, as-calculated values are shown as solid lines in the figures (black in the online color version). Their y-axis intersects give the E_d^{∞} values listed in Table I. The curves also serve to predict the formation energy which would be expected in any finite sized supercell: for example, the formation energies in the 8000 atom supercell are those at $1/L=0.0172$ in the figures. We can estimate how accurate the E_d^{∞} values are by adding the four dotted (black) curves shown for each example in Fig. 1, in each of which one of the four data points has been omitted. (Note that for some cases the errors are so small that the dotted lines are hard to pick out, but they are still present in the figure.) The spread in y-axis intersects gives the error bars listed in the table. This is one of the particular advantages of using finite size scaling: it is possible not only to correct the finite size errors themselves, but also to obtain a well defined uncertainty on the resulting energies—something other correction schemes cannot provide.

The errors obtained are on the 0.01–0.1 eV range or below (smaller errors are rounded to 0.01 eV) and can doubtless be further improved if still larger supercells are used. Note that, by construction, the errors which arise if only the 8, 64, and 216 atom supercells are used for the scaling are also on this 0.01–0.1 eV level (see Table I). The fact that such small error bars can be obtained indicates that (a) scaling is a viable and practical approach to supercell approximation errors, (b) the k-point convergence is sufficient for our current purpose, and (c) our enforced use of the 8 atom supercell is actually reasonable, for the same reasons described above and previously.[5]

B. Form of the scaling

The choice of $n=3$ again provides the best overall fit to the data, both for relaxed and nonrelaxed calculations. Normalized χ^2 tests[5] show that on average $n=2$ provides fits 2.9 times worse than $n=3$ while $n=4$ is 2.2 times worse. We note, however, that there are additional small [probably O(0.1) eV or less] short-ranged errors present which decay exponentially with supercell size. These arise chiefly from the direct overlap of bound defect states with their PBC images and the resulting dispersion of the defect levels. In the case of relaxed energies some additional short ranged errors can appear because defects in the 8 atom cell are only surrounded by 1 shell of relaxable atoms. The effect of this upon the form of the scaling can be seen in Fig. 2. Here we show the scaling of the elastic contribution to the finite size errors. This is done by calculating formation energies in the

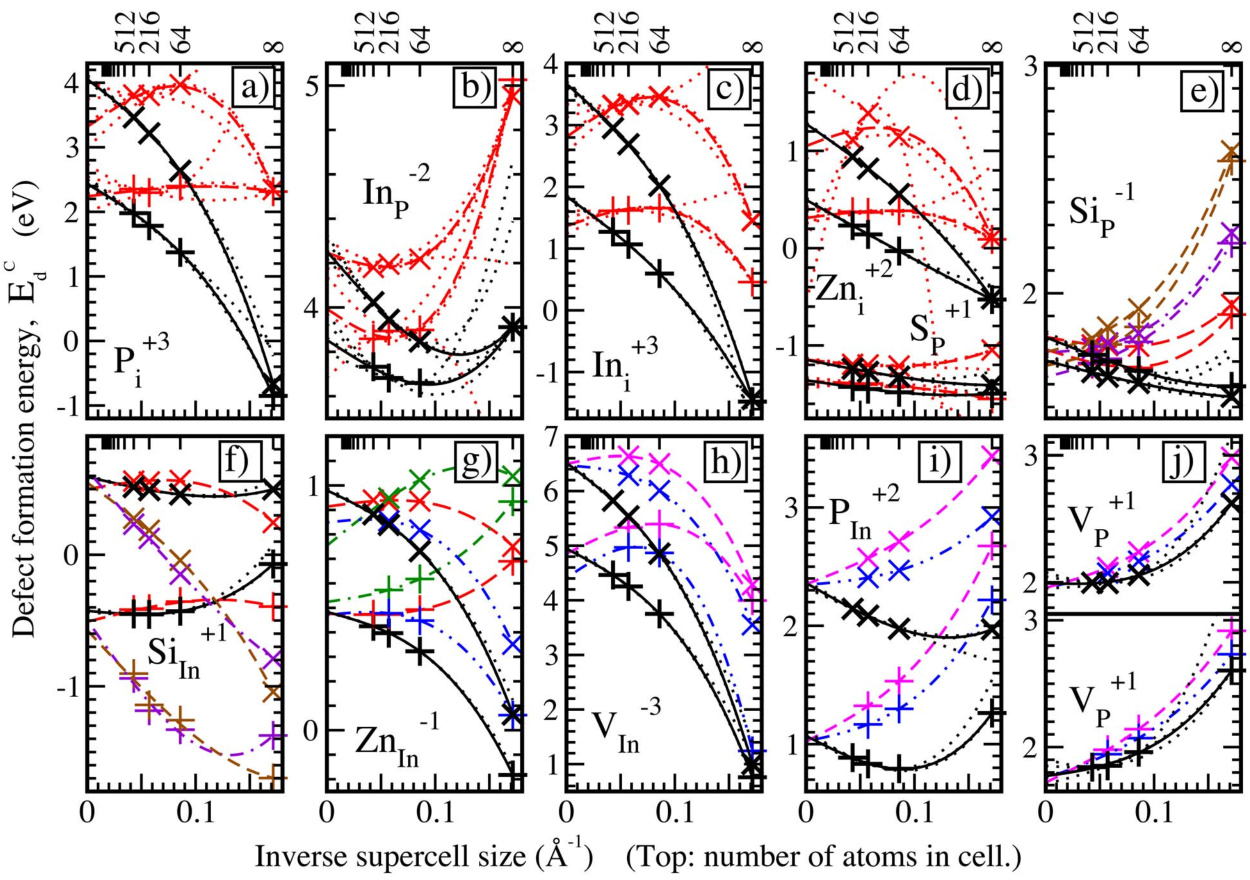

FIG. 1. (Color online) Scaling of (×) unrelaxed and (+) relaxed formation energies. Curves are fits to Eq. (2) with $n=3$. Solid (black) curves are fits to the four points as calculated (no corrections.) Dotted (black) lines each have one cell omitted for accuracy assessment. Scaling of the calculated values with various correction factors are shown for certain examples, as follows. Potential realignment: long dashed (red) lines in panels (a) to (g), and accuracy assessment for them: dotted (red) lines in panels (a) to (d). Dispersion corrections: short dashed (brown) lines in (e) and (f). Dispersion+potential corrections combined: dot-dashed (purple) lines in (e) and (f). First order (L^{-1}) Makov-Payne corrections as dot-dot-dashed (blue) lines in panels (g) to (j). First+third order ($L^{-1}+L^{-3}$) Makov-Payne corrections: short dashed (pink) lines in (h) to (j). First order Makov-Payne+potential corrections combined: dash-dash-dot (green) lines in panel (g).

216 atom cell only, so that the electrostatic errors are essentially constant. The number of shells of atoms permitted to relax around the defect is varied and the resulting formation energies are plotted against the inverse of the radius of the outermost relaxed atom shell. Hence the y intersect corresponds to the formation energy expected if an infinite number of shells are relaxed around the defect, but with the electrostatic errors inherent for the 216 atom supercell. As expected, and as for the neutral defects,[5] the elastic errors are predominantly linear. Indeed, if the "one shell only" point from each curve is omitted then a linear fit works perfectly. (Solid lines in Fig. 2.) The one shell only point corresponds to relaxations in the 8 atom cell, so we expect that the elastic contribution to the supercell approximation errors scales linearly with supercell size apart from some additional short range errors essentially only affecting the 8 atom cell.

These various short ranged errors have nevertheless only a very limited impact upon the final results, introducing only some additional scatter in the curves in Fig. 1, and hence leading to larger error bars in some cases. They also lead to $n=2$ or $n=4$ actually providing the best fit for some individual defects. However, in these latter cases the fitting with $n=3$ is almost always a very close second. These problems can be overcome in a few years time once improved computing resources allow the study to be repeated using the 1000 atom supercell. For now we can still conclude that the elastic errors are essentially inverse linear in supercell size, while the total formation energy errors (relaxed or unrelaxed) do indeed scale with the inverse-linear dimension and the inverse volume of the supercell.

IV. ASSESSMENT OF CORRECTION SCHEMES

In addition to the as-calculated formation energies, Fig. 1 also shows the formation energy scaling using various correction schemes. For clarity and space we do not show all possible corrections for all example defects, but results for all schemes are listed in Tables I and II. All schemes recover the correct formation energy in the infinite supercell limit, but not all produce improvements over the uncorrected formation energies for smaller supercells. This is shown in Table 2, which lists the residual errors (relative to the infinite supercell limit) when the corrections are applied in the 64 atom cell. The uncorrected 64 atom cell formation energies

TABLE I. Scaled relaxed and unrelaxed (ideal lattice sites) formation energies with ($E_d^{\infty,\phi}$) and without (E_d^{∞}) potential corrections, for various example defects in InP. Note that the error bars are not actually symmetric: the widest has been listed in each case. $\varepsilon(L^{-1})$ and $\varepsilon(L^{-3})$ are *ab initio* values of the dielectric constant ε for InP, calculated by comparing the Makov-Payne corrections of order L^{-1} and L^{-3} with the coefficients a_1 and a_3 obtained from the scaling. $\varepsilon^{\phi}(L^{-1})$ is the same thing calculated from the potential-corrected formation energies. All energies in eV, dielectric constants in units of the free space dielectric constant ε_0.

	Ideal structures					Relaxed structures				
Defect	E_d^{∞}	$E_d^{\infty,\phi}$	$\varepsilon(L^{-1})$	$\varepsilon(L^{-3})$	$\varepsilon^{\phi}(L^{-1})$	E_d^{∞}	$E_d^{\infty,\phi}$	$\varepsilon(L^{-1})$	$\varepsilon(L^{-3})$	$\varepsilon^{\phi}(L^{-1})$
V_P^{+1}	1.95±0.09	2.03±0.01	77.21	8.37	64.17	1.77±0.14	1.81±0.03	−16.38	4.32	−17.41
V_{In}^{-3}	6.52±0.06	5.75±0.78	12.12	−12.86	−14.43	4.95±0.05	4.63±0.52	17.96	−16.91	−39.93
P_{In}^{+2}	2.36±0.04	2.04±0.31	15.63	37.37	−14.88	1.07±0.04	0.83±0.27	19.83	17.96	−18.21
In_P^{-2}	4.25±0.08	4.25±0.12	14.33	29.16	46.67	3.85±0.13	4.00±0.31	25.53	30.12	37.69
$P_{i(P)}^{+3}$	4.05±0.07	3.32±0.71	14.22	15.49	−15.83	2.43±0.11	2.24±0.50	18.08	−26.71	−90.00
$In_{i(P)}^{+3}$	3.67±0.08	2.80±0.54	8.18	−15.78	−14.79	1.85±0.04	1.36±0.25	14.18	215.60	−29.86
$Zn_{i(P)}^{+2}$	1.28±0.01	1.05±0.31	10.57	−22.52	−18.98	0.50±0.02	0.31±0.12	13.58	472.79	−54.67
Zn_{In}^{-1}	0.98±0.01	0.91±0.07	9.78	−13.23	−32.65	0.48±0.01	0.47±0.03	16.59	−9.36	117.20
Si_P^{-1}	1.82±0.03	1.85±0.03	10.85	42.10	22.63	1.71±0.04	1.75±0.13	16.65	87.59	8.78
S_P^{+1}	−1.17±0.02	−1.15±0.03	9.24	5.38	16.91	−1.34±0.01	−1.37±0.07	12.28	6.71	39.50
Si_{In}^{+1}	0.62±0.01	0.50±0.11	10.81	20.48	−13.90	−0.36±0.03	−0.51±0.11	27.56	8.89	−6.82
Average	±0.05	±0.27	17.54	8.54	2.27	±0.05	±0.21	15.08	71.91	−4.88
Average over both relaxed and unrelaxed structures:						±0.05	±0.24	16.31	40.23	−1.31

contain average errors of about 0.5–0.6 eV, while using the potential realignment scheme produces errors of around 0.1 eV. The Makov-Payne scheme does much worse (average errors around 0.1–0.4 eV, but often much larger) and the dispersion "corrections" produce errors which can be even larger than those in the uncorrected formation energies.

A. Potential realignment

The potential realignment scheme is illustrated by the long-dashed (red) curves and points in Figs. 1(a)–1(g). Even by the 64 atom supercell the values are very good indeed. However, a lot of additional scatter is introduced into the corrected formation energies $E_d^{C,\phi}$. Indeed, the average errors relative to (the uncorrected) E_d^{∞} do not shrink at all with increasing supercell size: 0.07 eV in the 64 atom cell, 0.10 eV in the 216 atom cell, and 0.09 eV in the 512 atom cell (see Table II). This leads to wide error bars if the $E_d^{C,\phi}$ values are scaled to give the infinite supercell limit $E_d^{\infty,\phi}$. We have derived scaling error bars by the same technique described above. The resulting $E_d^{\infty,\phi}$ values and error bars are listed in Table I, although the (red) dotted curves with data points omitted are only shown in Figs. 1(a)–1(d). We find error bars of up to ±0.78 eV, average ±0.24 eV. This means that potential realignment is a useful correction for the results from individual supercells, but should *not* be used if more accurate results or defined error bars on results are required. In that case nonrealigned values should be scaled. These error bars are certainly too large to provide a basis for analysis of other correction schemes. The reason is that the correction scheme, good though it is, is not actually complete or correct. Even in the largest supercells, the point furthest from the defect is *not* bulklike, as the scheme assumes, resulting in either an over estimate or an under estimate, depending upon the specific conditions.

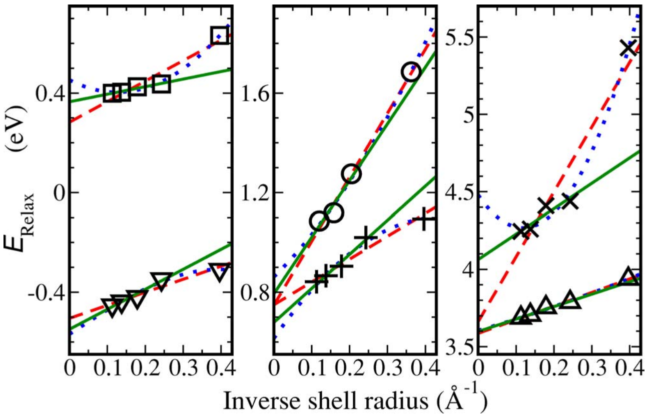

FIG. 2. (Color online) Scaling of the elastic contribution to the finite size errors in defect formation energies. Formation energies in the 216 atom shell are plotted vs the inverse of the radius of the outermost shell of atoms permitted to relax. Zn_{In}^{-1} (□), Si_{In}^{+1} (▽), $In_{i(P)}^{+3}$ (○), P_{In}^{+2} (+), V_{In}^{-3} (×), and In_P^{-2} (△). Solid (green) and dashed (red) lines: linear fits with the 1 shell only point omitted and included respectively. Dotted (blue) lines: quadratic fits to all points.

B. Dispersion corrections

The dispersion correction scheme is illustrated for shallow donors and acceptors in Figs. 1(e) and 1(f) and in Table I, both with (short dashed, brown curves) and without (dot-dashed, purple curves) potential alignment. Although the acceptor states fare better than the donors, the "corrected" val-

TABLE II. Assessment of correction schemes. Finite size errors (relative to the scaled values) are shown for the 64 atom supercell: δ_E is the error in the as-calculated formation energy and δ_{E+1} and δ_{E+1+3} are the errors when Makov-Payne corrections are used to order L^{-1} and L^{-3}, respectively. $\delta_{E+1}^{\mathrm{LDA}}$ is the error when order L^{-1} corrections are used, calculated with the *ab initio* dielectric constant evaluated from the results themselves (see text). $\delta_{E+\mathrm{k}}$ is the error when defect level dispersion is corrected for in the ionized states of the shallow donors and acceptors. Columns δ_E^{ϕ} etc. are the same as δ_E etc. but electrostatic potential realignments added. The +averages are of the absolute error values $|\delta_E|$. All energies in eV.

	Ideal structures								Relaxed structures							
Defect	δ_E	δ_{E+1}	δ_{E+1+3}	$\delta_{E+1}^{\mathrm{LDA}}$	δ_E^{ϕ}	δ_{E+1}^{ϕ}	δ_{E+k}	δ_{E+k}^{ϕ}	δ_E	δ_{E+1}	δ_{E+1+3}	$\delta_{E+1}^{\mathrm{LDA}}$	δ_E^{ϕ}	δ_{E+1}^{ϕ}	δ_{E+k}	δ_{E+k}^{ϕ}
$\mathrm{V}_{\mathrm{P}}^{+1}$	0.11	0.29	0.18	0.22	0.02	0.20			0.19	0.37	0.30	0.34	0.11	0.29		
$\mathrm{V}_{\mathrm{In}}^{-3}$	−1.67	−0.02	−0.51	−0.31	−0.13	1.51			−1.20	0.44	−0.08	0.16	−0.01	1.63		
$\mathrm{P}_{\mathrm{In}}^{+2}$	−0.39	0.35	0.11	0.23	−0.02	0.72			−0.28	0.46	0.23	0.34	0.04	0.78		
$\mathrm{In}_{\mathrm{P}}^{-2}$	−0.40	0.33	0.08	0.21	−0.03	0.70			−0.19	0.54	0.30	0.42	0.05	1.16		
$\mathrm{P}_{\mathrm{i(P)}}^{+3}$	−1.41	0.23	−0.28	−0.05	−0.08	1.56			−1.06	0.58	0.05	0.30	−0.03	1.61		
$\mathrm{In}_{\mathrm{i(P)}}^{+3}$	−1.65	−0.01	−0.50	−0.29	−0.21	1.44			−1.25	0.39	3.48	0.11	−0.19	1.45		
$\mathrm{Zn}_{\mathrm{i(P)}}^{+2}$	−0.72	0.01	−0.16	−0.11	−0.13	0.60			−0.53	0.21	−0.03	0.08	−0.12	0.62		
$\mathrm{Zn}_{\mathrm{In}}^{-1}$	−0.25	−0.07	−0.16	−0.10	−0.04	0.14	0.19	0.40	−0.17	0.02	−0.03	−0.01	0.01	0.19	0.05	0.21
$\mathrm{Si}_{\mathrm{P}}^{-1}$	−0.16	0.02	−0.03	0.00	−0.04	0.14	0.16	0.27	−0.10	0.08	0.03	0.06	−0.03	0.09	0.09	0.5
$\mathrm{S}_{\mathrm{P}}^{+1}$	−0.15	0.03	0.00	−0.02	−0.06	0.12	−0.89	−0.78	−0.14	0.04	0.05	0.03	−0.07	0.12	−1.07	−1.02
$\mathrm{Si}_{\mathrm{In}}^{+1}$	−0.16	0.02	−0.04	0.02	−0.02	0.16	−0.61	−0.50	−0.07	0.11	0.06	0.15	0.07	0.26	−0.90	−0.83
Average	0.64	0.13	0.19	0.14	0.07	0.66	0.46	0.49	0.47	0.29	0.42	0.18	0.07	0.75	0.53	0.64
	Average over both relaxed and unrelaxed structures:								0.56	0.21	0.31	0.16	0.07	0.71	0.50	0.57

ues are always worse than those using only potential realignment, and usually worse than even the uncorrected formation energies. Clearly, even for shallow defect levels, which closely follow[10] the VBE or conduction band edge (CBE), ϵ_D^{Γ} still produces worse formation energies than ϵ_D^{KS}.

C. Makov-Payne corrections

Figure 1(g) shows the first order L^{-1} Makov-Payne corrections, with (dash-dash-dot, green) and without (dot-dot-dash, blue) potential realignment. When used together the two schemes usually produce a large overestimate of the required correction (see columns 7 and 15 of Table II) almost as if using both corrections actually makes the *same* correction twice. Since the combination does so much worse than either technique alone there is no point going further with it. Instead, Figs. 1(h)–1(j) show Makov-Payne corrections only, with formation energies including both the order L^{-1} corrections (short dashed, magenta) and the order L^{-1} plus order L^{-3} corrections (dot-dot-dashed, blue). The order L^{-1} corrections work well in some cases (such as $\mathrm{In}_{i(\mathrm{P})}^{+3}$ when relaxations are omitted), but in most cases they are too large by a factor of about $1\frac{1}{2}$ to 2 (as also noted by others[6,9]) so that the "corrected" formation energies are a little better than the uncorrected ones. When the order L^{-3} corrections are added the correct formation energies are obtained in some cases, such as $\mathrm{V}_{\mathrm{In}}^{-3}$, but in other cases, such as $\mathrm{P}_{\mathrm{In}}^{+2}$, they help but are not sufficient. For other cases, such as $\mathrm{V}_{\mathrm{P}}^{+1}$, the corrections actually move the formation energies in the wrong direction.

Table II shows that the corrections are generally more likely to succeed for unrelaxed formation energies which is to be expected since the electrostatic monopole terms are not the only ones to scale as L^{-1}: the elastic errors do too. This means that even in principle the Makov-Payne corrections are only useful for nonrelaxed formation energies, which are rarely the interesting ones. In addition to this, the corrections also do better for more highly charged defects. This confirms that one of the problems is that they do not take into account the various other error terms which depend upon supercell size but not on charge state. These errors mostly have to do with the spurious defect level dispersion introduced by the PBCs. Although the direct contributions of these are exponentially decaying,[5] their effects can still be seen in supercells on the scale of 10–100 atoms. Indeed the actual band width can be on the order of, for example, 0.5 and 2 eV in the 64 atom and 8 atom supercells[12] and remain significant even beyond that. Indirect dispersion effects can also be very important: for example, in a partially filled, erroneously dispersed defect level only the lower part will be filled, leading to too low a value for E_d^C. Worse happens if the defect level lies outside the band gap, either because it genuinely does or because the supercell is too small. This can lead to strong linear terms in the supercell size errors even for neutral defects:[5] a neutral defect can behave as, say a −1 charged defect with (to a first approximation) a +1 charged jellium background. This is not limited to neutral defects: a calculation for a defect anticipated to be in a +2 charge state (with a −2 charged jellium) could end up behaving more like a +3 charge defect with a −3 charged jellium. If the defect level moves outside the bandgap at certain k points only it can lead to a linear error term which is not even proportional to the square of an integer charge. Overall, even the leading linear error term may be very different from that predicted by Makov and Payne's corrections. Unfortunately, beyond noting that things get better on average for larger charge states and for nonrelaxed calculations there seems to be no *a priori* method for determining whether the corrections will make things better or worse in a specific case. They are thus of little practical help, since they do not take into account enough of the specific behavior of individual defects and materials. Indeed, it seems unlikely that any such highly generalized model for prediction of finite size error correction factors will ever fully succeed.

D. Calculating the *ab initio* dielectric constant

Makov and Payne predicted that the two leading terms in the errors should be linear and cubic and our results show that they were correct in that respect. Their "corrected" formation energy takes the form

$$E^{C}_{d:\mathrm{MP}}(L) = E^{\infty}_{d:\mathrm{MP}} - k_1(\epsilon L)^{-1} - k_3(\epsilon L)^{-3}, \qquad (6)$$

where ϵ is the dielectric constant, $k_1 = q^2\alpha/2$, and $k_3 = 2\pi q Q/3$. (q is the charge of the defect, α is the Madelung constant for the supercell and Q is the quadrupole moment of the defect.) Comparing this with Eq. (2) we find $a_n = -k_n/\epsilon$. If we assume that the scheme is correct after all, then, q being known and Q having been calculated from the charge density, the only variable is the dielectric constant ϵ. We can then use the correction scheme together with the scaling results to derive an *ab initio* value of ϵ. This can be done twice for each defect, as shown in Table I. We find a wide scatter in the results, reflecting the wide variations in the effectiveness of the corrections. Indeed, the values of ϵ obtained are completely crazy when order L^{-3} corrections are used, as these are much more sensitive to short range effects and other errors. This again reflects the fact that the situation described by Makov and Payne was highly idealized and ignores too many of the details of the charge distribution around specific defects. Nevertheless, the averaged values ϵ from the order L^{-1} corrections are reasonably good. The most physically correct approach is to use the unrelaxed formation energies only (with no elastic effects); indeed the values derived using relaxed values, third order corrections or Makov-Payne plus potential realignment make little sense (see Sec. I). From the first order nonrelaxed curves we obtain a dielectric constant of 17.5 ± 19.0. (The error bar is the standard deviation from the average.) Numerical problems with the V_P value have made it rather unreliable, and very different to the the others. Omitting it gives the perhaps more consistent value of 11.6 ± 4.1. These values compare to 9.6 in experiment or 10.7 calculated[20] using more traditional *ab initio* DFT-LDA techniques.[21] We thus obtain a fairly reasonable estimate of ϵ as a free side-effect of performing accurate defect calculations—an interesting alternative to the traditional calculations methods. The uncertainty in the value obtained is obviously rather large, but should improve if more defects in more charge states are included in the average.

The order L^{-1} Makov-Payne corrections do improve using this new value for the dielectric constant, see columns 5 and 13 of Table I. However, some individual values still have errors of up to 0.3–0.4 eV, and there is still no way to know when the corrections are making things better and when they are making things worse, so from a practical point of view the Makov-Payne scheme is still not reliable enough for accurate calculations.

V. THE BANDGAP PROBLEM

We now turn to the band gap problem and the issue of how to map calculated transfer levels onto the experimental gap. In practice several alternative—and essentially incompatible—methods are normally used.

(1) *The extended gap scheme*. Align the theoretical and experimental VBEs and start plotting defect transfer levels from there, continuing *past* the theoretical CBE until one reaches the experimental one. In the section of the thus plotted "band gap" which lies above the theoretical CBE one automatically includes calculations in which supposedly localized, defect-bound electrons are in reality located in delocalized conduction band states. The properties of the defect itself (transfer levels and local relaxed structure, etc.) reenter primarily via hybridization of the conduction band states with the localized defect states, though this hybridization becomes smaller as the supercell size grows.

(2) *The scissors scheme*. Align both the theoretical VBE *and* CBE with their experimental counterparts, performing a "scissors" operation to stretched out the theoretical gap states over the experimental gap. The manner in which this scissors operation should be done is not uniquely defined. A common option is to place acceptor levels the same distance above the experimental VBE that they appear above the theoretical VBE in calculations, and donor levels the same distance below the experimental CBE that they appear below the theoretical CBE in calculations. A better alternative is to actually examine the form and symmetry of the defect states themselves, and see whether they hybridize more strongly with host states near the CBE or with states near the VBE. If they hybridize most strongly with VBE states then they should be plotted the calculated distance above the (experimental) VBE, and if they hybridize most strongly with CBE states they should be placed the calculated distance below the CBE.

(3) *The reference level scheme*. The basis of this scheme is rather different: the transfer level for the defect of interest is calculated, together with that of a similar reference defect for which the experimental value of the transfer level is well known, both done to the same level of accuracy. The difference between the experimental and calculated levels of the known defect is subtracted from the calculated value of the new defect, so that the new level is only found relative to the old one. This idea is not without practical merit, but is very empirical. Its accuracy depends critically upon the choice of an appropriate reference defect, which must be as similar to the new one as possible, so it will not be discussed further here. However, it has an occasionally used *ab initio* variant, which will be discussed.

(4) *The charged bulk reference scheme*. The reference state is not that of another defect, but is either the VBE or the CBE, meaning that a charged bulk total energy appears in Eq. (1), rather than a neutral one. In principle this provides

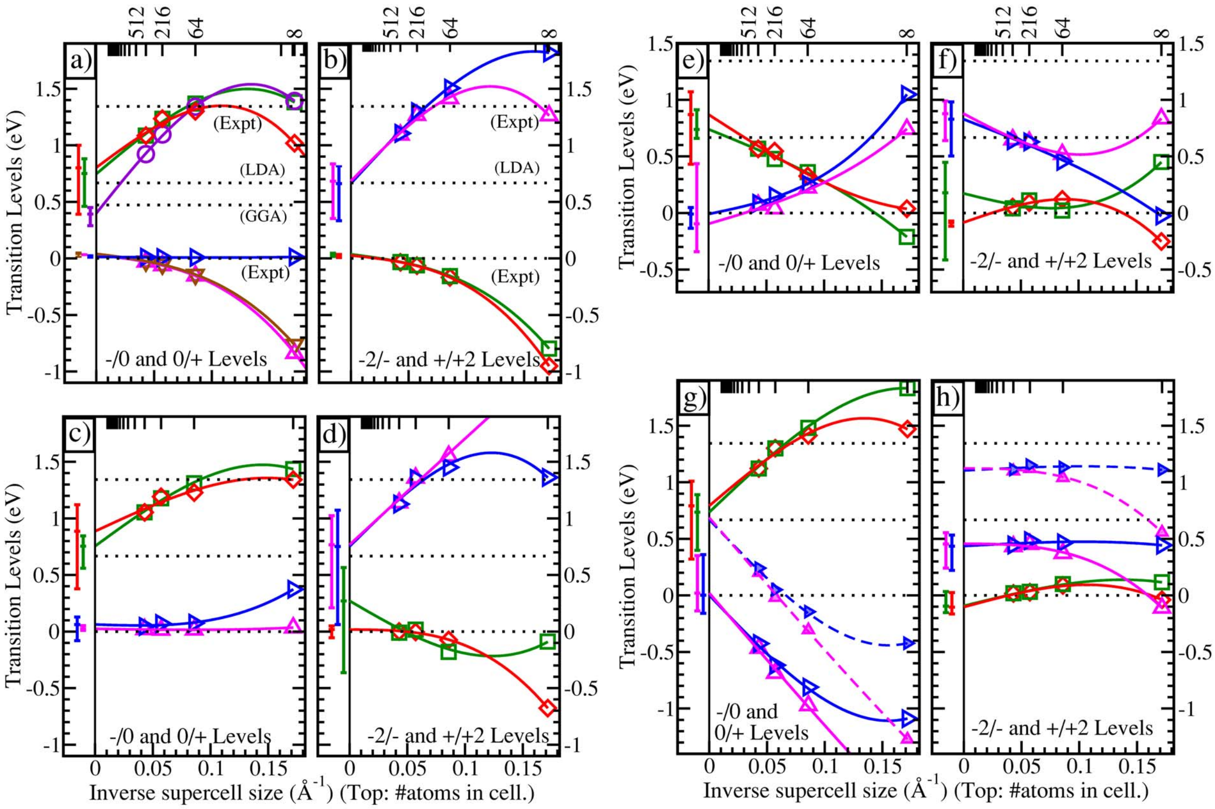

FIG. 3. (Color online) Scaling of transfer levels for simple donors and acceptors calculated using (a) & (b) neutral bulk as reference, with no corrections. (c) & (d) neutral bulk reference with potential corrections. (e) & (f) neutral bulk reference with dispersion corrections. (g) & (h) charged bulk as reference, with no corrections. Left panels: the dopant levels themselves. Right panels: the double donor or double acceptor levels, which should lie outside the bandgap. Using LDA: $S_P^{+/0}$ and $S_P^{+2/+}$ (□, green), $Si_{In}^{+/0}$ and $Si_{In}^{+2/+}$ (◇, red), $Si_P^{0/-}$ and $Si_P^{-/-2}$ (▷, blue), $Zn_{In}^{0/-}$ and $Zn_{In}^{-/-2}$ (△, pink). Using GGA: $S_P^{+/0}$ (○, purple) and $Zn_{In}^{0/-}$ (▽, brown). In (g) & (h) the smaller symbols with dashed lines show the acceptor-type levels relative to the experimental CBE rather than the LDA one. On all panels: the dotted lines are (in order of increasing energy) the VBE and CBE from GGA [panel (a)], LDA, and experiment. The error bars shown have been constructed as described in Sec. III though the dotted lines are omitted for clarity.

an alternative route around the band gap problem. (Details are given below.)

Obviously, none of these schemes is fully correct, since the LDA/GGA bandgap problem is a fundamental one, but the important practical question of which approach comes closer to giving the correct physical picture remains unanswered. In principle, it can be answered by examining various experimentally well known defect levels. The exact location of most native defect levels is rather hard to measure to a sufficiently high accuracy to answer this question, but many simple donor and acceptor levels are known very accurately. We will use the 0/− acceptor level of Zn_{In}, which in experiment lies 0.035 eV from the VBE, and the +/0 donor levels of S_P and Si_{In}, which experiment finds about 0.006 eV from the CBE. We will also add in the 0/− transfer level of Si_P, which would be a simple acceptor if Si_{In} had not been the more stable site for Si in InP. This gives us an example of a donor and an acceptor on each sublattice, so that all bonding and band hybridization possibilities are represented. Unfortunately, calculations of these levels in finite sized supercells in the 100–200 atom range have never produced a clear answer to the question, so we will use finite size scaling to correct for the supercell approximation errors. The results are shown in Fig. 3(a) using as calculated values, (c) adding in potential corrections, and (e) using dispersion corrections. (Van der Walle and Neugebauer[10] suggested that dispersion corrections should still be correct for shallow transfer levels.) The results using as calculated transfer levels and potential corrected ones are very similar. The dispersion corrections, on the other hand, are clearly completely incorrect: they place both acceptor and donor levels in the midgap for most practical supercell sizes, whether the potential corrections are added (not shown) or omitted (as here). Meanwhile, in Figs. 3(b), 3(d), and 3(f) we also show the second donor/acceptor levels $Zn_{In}^{-/-2}$, $S_P^{+2/+}$, etc., calculated using the same correction schemes. Since these levels are never observed experimentally they must lie outside the band gap. Hence the VBE should lie between the double donor levels (right panels) and the single acceptor levels (left panels). Similarly the CBE should lie between the single donor and double acceptor levels. In practice, these pairs of levels more or less coincide, doubtless a result of the remaining limitations in the use of DFT-LDA for semiconductor defects. Fortunately this still leaves us with a clear view of how to treat the band gap problem.

In the 64 atom cell the donor (and double acceptor) levels lie roughly the experimental band gap (1.3 eV) above the

VBE, while the acceptor (and double donor) levels lie on average a little below the VBE. However, coming to the larger cells the donor levels fall and the acceptor levels rise. Finite size scaling places the acceptor levels $\mathrm{Zn}_{\mathrm{In}}^{0/-}$ and $\mathrm{Si}_{\mathrm{P}}^{0/-}$ 0.03 and 0.01 eV above the VBE, respectively, in rather good agreement with experiment. The single donor (and double acceptor) levels all scale to the theoretical CBE.[22] To be more specific, transfer levels calculated using LDA scale to the LDA band edges, while the $\mathrm{Zn}_{\mathrm{In}}^{0/-}$ and $\mathrm{S}_{\mathrm{P}}^{+/0}$ transfer levels calculated using the Perdew-Wang GGA (Ref. 23) scale to the edges of the GGA band gap—Fig. 3(a). (The GGA CBE lies 0.2 eV below the LDA one when the lattice parameter has been optimized.)

Hence, scheme 1, the extended gap scheme, is seen to be the most appropriate when only reporting uncorrected results from supercells of about 50–100 atoms. However, when the finite size errors are removed (by scaling or by some other technique) it becomes clear that the scissors scheme, scheme 2, is physically far more correct. Unscaled LDA or GGA results in supercells over a few 1000 atoms would also be best reported using the scissors scheme. For intermediate (100–1000 atom) supercells some kind of hybrid approach is required. The result also indicates why the controversy has lasted so long: ultimately the scissors scheme is correct, but this only shows up for very large supercells or with scaling.[24]

We now return to scheme (4), the charged bulk reference. This amounts to replacing the terms $-E_T^C(\text{no defect})$ and $-q\epsilon_F$ in Eq. (1) by the term $-E_T^C(\text{no defect}^q)$, which is the total energy of the bulk supercell C with $-q$ extra electrons and neutralizing jellium. Figures 3(g) and 3(h) show the transfer levels calculated like this, with no correction terms. The donor levels behave in the same way as using Eq. (1) in Fig. 3(a), but the acceptor levels are less straightforward. Using a charged bulk reference the levels come out relative to the CBE, rather than the VBE: they implicitly *include* the bandgap, which must be subtracted off again to place them on the same overall scale as the donor levels. This gives a "choice" for the value for the bandgap to subtract, which is how the potential route around the band gap problem enters. Namely, if the 0/− transfer level emerges as, say, −0.5 eV, one could place it 0.5 eV below the experimental CBE, thus plotting the transfer levels over the experimental bandgap. (Small symbols and dashed scaling curves in Figs. 3(g) and 3(h).) For the single acceptor levels this clearly does not work: although they land accidentally close to the VBE for smaller supercells they actually scale to the theoretical CBE, which is completely wrong. Instead, they should be placed below the theoretically CBE (large symbols, solid curves), where they scale to the VBE. Unfortunately the opposite is true for the double acceptor levels. These work out moderately well if plotted relative to the experimental CBE—lying outside the theoretical band gap, even if still inside the experimental one—but using the theoretical CBE (as required for the single acceptors) they lie inside the theoretical band gap, disagreeing with experiment. Hence using a charged bulk total energy as the reference for charged defect calculations is not even internally consistent and the scheme is thus fundamentally incorrect.

VI. CONCLUSIONS

In this paper we have shown that finite size errors in the supercell approximation scale with the linear dimension and with the volume of the supercell, and that finite size scaling the results from a series of supercells removes the supercell approximation errors, leaving accurate information on isolated semiconductor defects, without the need for corrections. We also obtain error bars defining the uncertainly on the results obtained, and as far as we are aware this is the only method which is able to remove these errors in a controlled manner with defined uncertainty. We have demonstrated this using a variety of different types of defects with charge states ranging from −3 to +3 and find that it is possible to reduce formation energy errors from the 0.1–2 or so eV range of practical supercells down to the 0.01−0.1 eV range or below—doubtless much lower if still larger supercells are used. By construction, errors on this scale also occur if only the 8, 64, and 216 atom supercells are used.

We then used the scaled results for the first full reliability test of three correction schemes. We found that dispersion corrections are incorrect and Makov-Payne corrections are poor (with both the experimental and LDA dielectric constants), though they did allow us to obtain a reasonable *ab initio* LDA dielectric constant of $\epsilon = 11.6 \pm 4.1$ for InP. On the other hand, the potential realignment scheme was found to be remarkably successful, removing much of the electrostatic defect-defect error as well, to leave average residual errors of about 0.1 eV, from single calculations with supercells in the 64–512 atom range.

This obviously raises the question of why the potential re-alignment scheme is *so* successful, when it does not set out to correct defect-image interaction errors at all. The fact that it produces similar (but more reliable) corrections to the Makov-Payne scheme suggests that it is some how dealing with the electrostatic errors anyway. We noted in Sec. IV A that the scheme assumes that the real-space potential at some point in the cell far from the defect is bulklike, even though for practical cell sizes it is not bulklike at all. The resulting additional shift in this local real-space potential reflects the effects of the electrostatic defect-image interactions. Doing the potential realignment in this way therefore fails to properly correct the mismatch in the zeros of the energy scales between the bulk and defect cells, but the "error" in the realignment more or less corrects for the electrostatic errors arising from the PBCs.

Finally, we have given the long awaited answer to the dilemma of how best to map LDA and GGA calculated defect transfer levels onto the experimental gap, and indicated why the issue was previously so hard to settle. The key result is that the scissors method is physically more correct, though the extended gap scheme is best when reporting results from single supercells on the 1–200 atom scale without finite size corrections. For uncorrected results from supercells over a few 1000 atoms the scissors method is best, with a hybrid method needed in between. The best, of course, is to use the scissors scheme, with either scaled or corrected results, regardless of supercell size. The apparent success of the essentially incorrect extended gap scheme for uncorrected results in manageably sized supercells is the basic reason for the debate lasting so long.

This leads to another issue which is also apparent from our results. It is very dangerous to report calculations from single supercells without trying to estimate the errors contained. Quantitatively these can be ~1–2 eV or more, but we have cases here where conclusions are even qualitatively wrong in supercells up to and even including the 512 atom cell. For example,[25] comparing these results for P_{In}^{+2} with those[5] for P_{In}^{+0} we find that even at the CBE, the +2 charge state appears more stable than the +0 in all four supercells. The fact that it is actually 0.19 eV *less* stable only emerges when the finite size errors are removed, either by scaling or (leaving residual errors from 0.05–0.13 eV) by using potential realignment. Similarly, at the VBE, V_{In}^{-3} and $In_{i(P)}^{+3}$ are more stable than V_{In}^{+0} and $In_{i(P)}^{+0}$, respectively, in both the 64 and 216 atom supercells. The correct stability order only appears in the 512 atom cell (neutrals more stable by 0.21 and 0.16 eV, respectively), and the correct order of magnitude for the difference (0.68 and 1.17 eV) is only obtained by scaling. Another striking example is that, according to LDA in cells ⩽512 atoms, *p*-type Zn-doped InP—a material upon which much of current optoelectronics depends—should not be *p* type at all. For the roughly stoichiometric conditions of, say, Czochralski growth, LDA in the 64 and 216 atom cells places Zn not as the shallow acceptor Zn_{In} but as the interstitial $Zn_{i(P)}$, where it is a deep double donor. Even in the 512 atom cell the two are degenerate, suggesting at best semi-insulating material. According to this Zn is only an acceptor for InP grown under strongly nonequilibrium conditions, such as with molecular beam epitaxy. However, Zn *is* a *p* type dopant, even grown from the melt, and this fact *can* be predicted using LDA, but only for supercells of the order of 1000 s of atoms, or if the supercell size errors are removed—by scaling or otherwise. Doing this using potential realignment works for all these examples: even in the 64 atom supercell reasonable results can be obtained. However, caution should still be used: First, for our examples it worked much better for the formation energies than for the shallow dopant transfer levels. Secondly, potential realignment makes P_{In}^{+0} (correctly) more stable than P_{In}^{+2} at the CBE in all but the 8 atom supercell, but if those corrected results are then scaled the wrong answer returns, with P_{In}^{+2} more stable than P_{In}^{+0} because of the large error bars found when scaling potential realigned energies.

In short, it is essential, to estimate and report the finite size errors for each specific case when reporting supercell defect calculations. This is often omitted, or is only done using the unreliable Makov-Payne scheme. When it is done this is usually by doing most calculations in a cell of, say, 50–200 atoms, and then repeating a few of them in a slightly larger cell. If the calculated results do not change much then they are considered converged. However, even this should be done with extreme caution. Even with only a linear contribution, the finite size errors in the 64 atom supercell are three times the difference between the 64 and 216 atom cell energies, the 216 atom cell errors are still twice this estimate. Even the errors in the 512 atom cell are three times the difference between the 216 and 512 atom energies.

So, how *should* finite size errors within the supercell approximation be treated? Ideally, using finite size scaling of otherwise *un*corrected energies. This is, of course, costly in both human and computer time. The best alternative is simply to use potential realignment in as large a supercell as time and resources permit. However, one should be aware that (a) this does not help the elastic errors, (b) potential realignment should *not* be combined with finite size scaling, and (c) there is no way to estimate the remaining errors or the reliability of the results. For our examples, the average errors using this method are ~0.10 eV, but with some examples up to 0.21 eV, and nothing to say that much larger errors will never occur. If the conclusions being drawn from a calculation are not adversely affected by uncontrolled errors of 0.1–0.2+ eV then this method is reasonably good. Otherwise, the only truly reliable method of controlling the errors in the supercell approximation, and defining the uncertainly in the results, is finite size scaling.

The calculations in this paper were performed at Uppsala University and at the Parallel Computing Centre (PDC) Stockholm, Sweden. The authors would also like to thank the Göran Gustafsson Foundation, the Swedish Foundation for Strategic Research (SSF), and the Swedish Research Council (VR) for financial support.

*Present address: Materials Chemistry, Box 538, SE-75121 Uppsala, Sweden. Email address: Christopher.Castleton@mkem.uu.se

[1] W. Kohn and L. Sham, Phys. Rev. **140**, A1133 (1965).

[2] M. C. Payne, M. P. Teter, D. C. Allan, T. A. Arias, and J. D. Joannopoulos, Rev. Mod. Phys. **64**, 1045 (1992).

[3] M. Leslie and M. J. Gillian, J. Phys. C **18**, 973 (1985).

[4] C. W. M. Castleton and S. Mirbt, Physica B **340-342**, 407 (2003).

[5] C. W. M. Castleton and S. Mirbt, Phys. Rev. B **70**, 195202 (2004).

[6] P. A. Schultz, Phys. Rev. Lett. **84**, 1942 (2000); U. Gerstmann, P. Deák, R. Rurali, B. Aradi, Th. Frauenheim, and H. Overhof, Physica B **340-342**, 190 (2003); B. Aradi, P. Deák, A. Gali, N. T. Son, and E. Janzén, Phys. Rev. B **69**, 233202 (2004).

[7] L. N. Kantorovich, Phys. Rev. B **60**, 15 476 (1999); L. N. Kantorovich and I. I. Tupitsyn, J. Phys.: Condens. Matter **11**, 6159 (1999); P. A. Schultz, Phys. Rev. B **60**, 1551 (1999); H. Nozaki and S. Itoh, Phys. Rev. E **62**, 1390 (2000); A. Castro, A. Rubio, and M. J. Stott, Can. J. Phys. **81**, 1151 (2003).

[8] G. Makov and M. C. Payne, Phys. Rev. B **51**, 4014 (1995).

[9] J. Lento, J.-L. Mozos, and R. M. Nieminen, J. Phys.: Condens. Matter **14**, 2637 (2002).

[10] C. G. Van de Walle and J. Neugebauer, J. Appl. Phys. **95**, 3851 (2004).

[11] S.-H. Wei, Comput. Mater. Sci. **30**, 337 (2004).

[12] A. Höglund, C. W. M. Castleton, and S. Mirbt, Phys. Rev. B **72**, 195213 (2005).

[13] M. Hedström, A. Schindlmayr, and M. Scheffler, Phys. Status Solidi B **234**, 346 (2002).

[14] R. W. Jansen, Phys. Rev. B **41**, 7666 (1990); A. P. Seitsonen, R. Virkkunen, M. J. Puska, and R. M. Nieminen, *ibid.* **49**, 5253 (1994); A. Höglund, M. C. Qian, C. W. M. Castleton, M. Göthelid, B. Johansson, and S. Mirbt (unpublished).

[15] D. Vanderbilt, Phys. Rev. B **41**, R7892 (1990); G. Kresse and J. Hafner, J. Phys.: Condens. Matter **6**, 8245 (1994).

[16] G. Kresse and J. Furthmüller, Comput. Mater. Sci. **6**, 15 (1996).

[17] C. W. M. Castleton and S. Mirbt, Phys. Rev. B **68**, 085203 (2003).

[18] H. Monkhorst and P. Pack, Phys. Rev. B **13**, 5188 (1976).

[19] D. J. Chadi and M. L. Cohen, Phys. Rev. B **8**, 5747 (1973).

[20] B. Arnaud and M. Alouani, Phys. Rev. B **63**, 085208 (2001).

[21] The dielectric constant calculation quoted here used plain LDA with no added local correlation effects, GW quasi-particle corrections or scissors shifts of the bandstructure, and so comes from a similar level of method to ours.

[22] The scaling of the single donor and double acceptor levels produces rather wide error bars. The reason is that the particular Kohn-Sham level which is half filled in the 0 charge state but empty in the +1 charge state lies above the CBE in all of these (finite) supercells, though it will reappear inside the gap for large enough cells. As we showed previously (Ref. 5) this tends to produce poor scaling in the 0 charge state.

[23] J. P. Perdew and Y. Wang, Phys. Rev. B **13**, 5188 (1976).

[24] The data shown in Fig. 3 were calculated using the In 4*d* electrons as core, but the conclusions are completely unaffected: correcting this error using In 4*d* valence calculations in the 64 atom cell, the curve for $\mathrm{Zn}_{\mathrm{In}}^{0/-}$ moves down by about 0.003 eV, that for $\mathrm{Si}_{\mathrm{P}}^{0/-}$ moves up about 0.04 eV and those for $\mathrm{Si}_{\mathrm{In}}^{+/0}$ and $\mathrm{S}_{\mathrm{P}}^{+/0}$ move down about 0.08 eV. We also note here that the present results are for relaxed formation energies, but that we can reach exactly the same conclusions using nonrelaxed formation energies.

[25] The numerical results reported in this paragraph include corrections for the use of the In 4*d* core pseudopotential. They are thus considered accurate to around 0.01–0.02 eV, not counting errors in the LDA itself.