Journal of Applied Physics PERSPECTIVE scitation.org/journal/jap

Optimized hybrid functionals for defect calculations in semiconductors

Cite as: J. Appl. Phys. **126**, 130901 (2019); doi: 10.1063/1.5110643
Submitted: 24 May 2019 · Accepted: 21 September 2019 ·
Published Online: 3 October 2019

Peter Deák,[a] Michael Lorke, Bálint Aradi, and Thomas Frauenheim

AFFILIATIONS

Bremen Center for Computational Materials Science, University of Bremen, P.O. Box 330440, D-28344 Bremen, Germany

[a]deak@uni-bremen.de

ABSTRACT

Defects influence the electronic and optical properties of crystals, so their identification is crucial to develop device technology for materials of micro-/optoelectronics and photovoltaics. The identification requires the accurate calculation of the electronic transitions and the paramagnetic properties of defects. The achievable accuracy is strongly limited in the case of the (semi)local approximations to density functional theory, because of the underestimation of the gap and of the degree of localization. In the past two decades, hybrid functionals, mixing semilocal and nonlocal exchange semiempirically, have emerged as an alternative. Very often, however, the parameters of such hybrids have to be tuned from material to material. In this paper, we describe the theoretical foundations for the proper tuning and show that if the relative positions of the band edge states are well reproduced, and the generalized Koopmans's theorem is fulfilled by the given parameterization, the calculated defect levels and localizations can be very accurate. As demonstrated here, this can be achieved with the two-parameter Heydt-Scuseria-Ernzerhof hybrid, HSE(α,μ) for diamond, Si, Ge, TiO_2, GaAs, $CuGaS(Se)_2$, GaSe, GaN, and Ga_2O_3. The paper describes details of the parameterization process and discusses the limitations of optimizing HSE functionals. Based on the gained experience, future directions for improving exchange functionals are also provided.

I. INTRODUCTION

Defect engineering in the materials of microelectronics and optoelectronics or photovoltaics is a critical part of technology development. It starts with the identification of the most important intrinsic and extrinsic point defects, which may influence device behavior. Mostly, this identification cannot be achieved by experiments alone, and electronic structure calculations, using density functional theory (DFT), have become an indispensable tool in the process. The underestimated gap and the tendency for delocalization with (semi)local approximations of DFT, however, cause uncertainties in predicting charge transition levels and paramagnetic resonance parameters of defects. These problems are aggravated in wide bandgap materials, where also polaronic self-trapping of carriers may occur. As the weight of research shifted from microelectronics to optoelectronics and photovoltaics, it has soon become clear that the local density and the semilocal generalized gradient approximations (LDA and GGA, respectively) to the exchange functional in DFT just will not do for the defects of the wide bandgap materials applied in these areas.[1,2] After some years of searching, the hybrid functionals (mixing semilocal and nonlocal exchange semiempirically) have emerged as a useful alternative.[3–5] (Note that presently, there is no *ab initio* total energy method which could be carried out accurately for supercell sizes required in defect calculations.) For years, the use of hybrid functionals has been controversial, so the first purpose of this paper is to explain why and when hybrids are appropriate. After describing the technical details of the calculations in Sec. II, it will be demonstrated in Sec. III on Group-IV semiconductors and TiO_2 that, if the hybrid functional satisfies certain conditions in the given material, defect levels can be calculated with unprecedented accuracy. It will then be shown in Sec. IV that in materials where this is not the case, the same criteria can be used to optimize the hybrid functional for defect calculations. The procedure for optimization and results for Ga-based semiconductors, Ga_2O_3, GaN, GaSe, $CuGaS(Se)_2$, and GaAs, will be given. The experience with optimized HSE functionals is summarized then in Sec. V, where also an analysis of their successes and failures is provided, and perspectives for further development are given.

II. COMPUTATIONAL FRAMEWORK

All calculations presented here were carried out with the Vienna *Ab initio* Simulation Package (VASP), using the projected augmented wave (PAW) method.[6] In the case of Ga atoms, the semicore d-orbitals were treated explicitly in Ga_2O_3, GaN, and GaAs, but—for practical reasons—excluded in the case of $CuGaS(Se)_2$ and GaSe. For titanium, the 3*p*, 3*d*, and 4*s* orbitals were treated explicitly. The cutoff energy for the expansion of the wave function was set to 420 eV if the material contained C, O, or N but lowered to 320 eV otherwise. The cutoff for expanding the charge density was set to be double of the wave function cutoff. The equilibrium lattice parameters have been determined either by constant volume relaxations of the primitive cell and fitting to Murnaghan's equation of state[7] or by brute force shape- and volume-relaxation with a 30% higher cutoff energy (as recommended by the VASP manual). For these calculations, a Γ-centered Monkhorst-Pack (MP) *k*-point set,[8] either $6\times6\times6$ or $7\times7\times7$, was applied for the Brillouin zone (BZ) integration. The supercell model was used for defect calculations, in the largest supercell allowed by our computer resources (at the time of doing the actual calculations), choosing supercell vectors of nearly equal length if possible, to ensure a directionally unbiased relaxation around the defect. The supercell calculations were carried out in the Γ-point approximation for BZ sampling. The force criterium for geometry optimization was 0.02 eV/Å.

For charged defects, the total energy was corrected *a posteriori*. In our earlier calculations on C, Si, Ge, and TiO_2, we have used the Lany–Zunger method (LZ),[9] which applies a cell size- and shape-dependent correction, taking into account a monopole and a quadrupole term in the multipole expansion of the defect charge. As it was shown, this method works well for sufficiently localized defect states.[10] Later, for Ga-based semiconductors, we have switched to the method of Freysoldt, Neugebauer, and Van de Walle (FNV),[11] which models the defect charge by a single, spherical Gaussian distribution. This method, as implemented in the SPHINX package,[12] and can also handle nonorthogonal and nonhexagonal supercells (as is the case for β-Ga_2O_3). Recently we have found, however, that defect charge distributions in layered compounds cannot be well described by a single spherical Gaussian, while on surfaces, the change of the dielectric constant has to be taken into account. Therefore, in the case of GaSe, we have used our own code, SLABCC,[13,14] which realizes the Komsa-Pasquarello method[15,16] to calculate the total energy correction and can be used for bulk, slabs, and monolayers alike. From the viewpoint of the last two, it is especially important that it allows the usage of arbitrary nonspherical and/or multiple Gaussian models of the defect charge, as well as an anisotropic dielectric constant in the solid part. The position of the interfaces (i.e., of the step in the dielectric constant normal to the interface) as well as the parameters of the Gaussians (position, width in the *x*, *y*, *z* directions, and charge fraction) are optimized to get the best fit to the calculated defect potential.

The need for charge correction in supercells arises from the spurious Coulomb-interaction of repeated charges and requires the dielectric constant of the material, for which we have used well established experimental data. In the following, we will need the ionized state of the defect with the geometry fixed at that of the neutral one (vertical ionization), or at the equilibrium geometry of the charged state (adiabatic ionization). In the first case, obviously, the high frequency dielectric constant ε_∞ must be used. In the second case, the static dielectric constant, ε_0, would be relevant in principle, because of the ionic screening. Earlier we have observed, however, that in cases where $e_0 \gg e_\infty$, much too shallow adiabatic charge transition levels are obtained this way, while using ε_∞ reproduces the experimental values better.[17–19] The explanation probably is that relatively large supercells describe a substantial part of the ionic screening explicitly, and using the bulk value of ε_0 amounts to double-counting.

III. WHEN AND WHY DO HYBRID FUNCTIONALS WORK FOR SOLIDS

The deficiencies of LDA/GGA can be traced back to the deviation from the correct piecewise linear behavior of the total energy as a function of the occupation numbers, as shown in Fig. 1. The (semi)local exchange functionals do not show a derivative discontinuity at integer occupation numbers at all, so that will be small in the total energy as well (some discontinuity coming from the kinetic energy). This is the reason for the underestimated gap, as shown in Fig. 1 by the LDA/GGA bands in red. In contrast, pure (unscreened) nonlocal exchange leads to a much too strong discontinuity and to the overestimation of the gap (blue bands in Fig. 1). The linearity condition directly implies that the Kohn-Sham (KS) level of the defect remains constant upon changing its occupation (see the green line in Fig. 1), and the KS-level of the occupied (unoccupied) frontier orbital supplies exactly the ionization energy (electron affinity) of the system.[20] This is called the generalized Koopman' theorem (gKT). For example, for the electron addition case shown in Fig. 1, it can be formulated as

$$\varepsilon_{KS} - E_C \equiv \Delta KS = \Delta SCF \equiv E(N+1) - E(N) - E_C, \qquad (1)$$

where ΔKS and ΔSCF are, respectively, the position of the KS-level and the self-consistently calculated electron affinity with respect to the conduction band minimum, E_C.

LDA and GGA exchange gives rise to a convex behavior of the total energy between integer occupation numbers, while unscreened nonlocal exchange to a concave one. Therefore, adding an electron to the defect level in the gap pushes it up in the case of LDA/GGA, and down in the case of the unscreened nonlocal exchange,[20] as shown by the red and blue lines, respectively, in Fig. 1. The farther the defect level is from the band edge, the more localized the corresponding one-electron state will be, so LDA/GGA favors delocalization, while unscreened nonlocal exchange overlocalizes.

Hybrid functionals employ a semiempirical mixture of nonlocal and LDA/GGA exchange,[21,22] and from Fig. 1 it is quite obvious that this can lead to error compensation between the pure functionals. Indeed, the improvement of the bandgap has been observed early on[3,4] in calculations using the original 3-parameter quantum chemical hybrid B3LYP.[21] Later, a one-parameter hybrid, the so-called PBE0 functional[23] was shown to be better for solids. The only parameter here is the mixing ratio between semilocal and (unscreened) nonlocal exchange, which was set to 0.75/0.25, in agreement with the suggestion in Ref. 24 for saturated bonds,

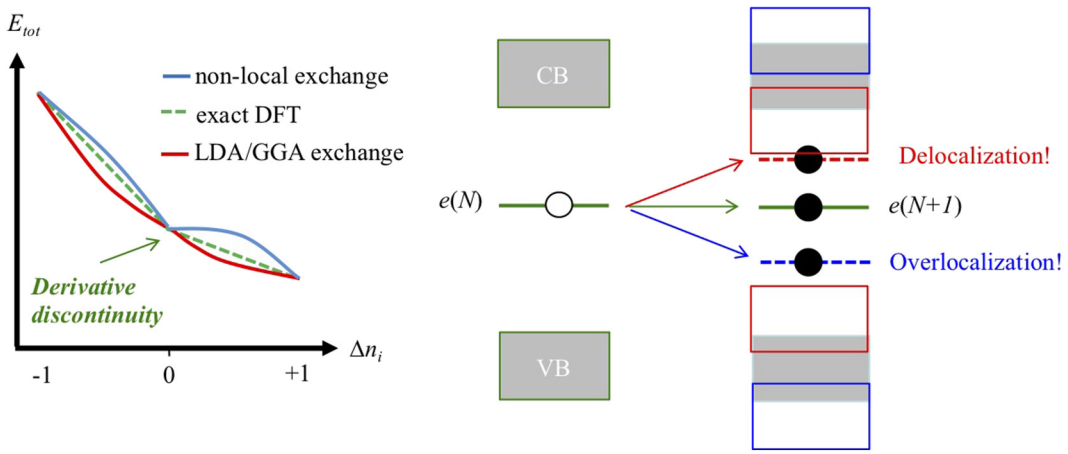

FIG. 1. Deviations from the proper piecewise linear behavior of the total energy as a function of occupation numbers and its consequences. (For details see the text.)

which is mostly the case in solids. Actually, this so-called α parameter was shown to be related to ε_∞ of the material,[25–29] so it ought to be chosen to be material-dependent.

Later functionals have taken screening into account in a more explicit manner. The sX method applies screened exact (nonlocal) exchange,[30] while the HSE functional[31] of Heyd, Scuseria, and Ernzerhof takes electronic screening also into account in the nonlocal part of the exchange by phasing out its admixture beyond a given screening length $2/\mu$. The screening parameter μ was originally used to fine-tune the bandgaps. Nowadays, the HSE functional is mostly used with the parameters $\alpha = 0.25$ and $\mu = 0.20$, resulting in the best bandgap for a large number of semiconductors. This choice of the parameters is often referred to as the HSE06 functional.[32]

A single parameter in the functional, or a purely empirical fitting of the second, does not guarantee the restoration of both important attributes of the exact functional: the reproduction of the bandgap (proper derivative discontinuity) and the fulfillment of the gKT (linearity). (In fact, it has been shown that both cannot be achieved by tuning α alone in a PBE0-type hybrid.[28,29]) As it happens, however, the standard HSE06 = HSE(0.25,0.20) functional satisfies these conditions to a good approximation in Group-IV semiconductors[33] and also in TiO_2,[34] as shown in Table I. The applied supercell sizes were 512-atoms for C, Si, and Ge, 96-atoms for anatase-TiO_2, and 72-atoms for rutile-TiO_2. It must be noted that electronic structure calculations within the Born–Oppenheimer approximation result in the single-particle (SP) gap at 0 K, which may differ substantially from the room-, or even from the low-temperature optical bandgap, which has to be renormalized for electron-phonon interactions.[35] (Obviously, the exciton binding energy should also be taken into account.) Alternatively, low-temperature photoelectron spectroscopy data have to be used for comparison. (For a discussion about that, see Ref. 18.)

Consequently, in these materials, HSE06 is a good approximation of the exact DFT functional. This can be seen in the prediction of the adiabatic (or thermal) charge transition levels of defects. Table II shows examples for diamond.

TABLE I. Comparison of the single-particle (SP) bandgaps obtained by the HSE06 = HSE(0.25,0.20) hybrid with experiment, and the fulfillment of the gKT for Group-IV semiconductors and for TiO_2. (The charge corrections were calculated by the LZ method and we note that in the case of TiO_2, a check with the FNV correction resulted in the deviation from the gKT to be less than 0.1 eV.) All values in electron volt. Donor and acceptor levels are given with respect to the conduction band minimum (E_C) and valence band maximum (E_V), respectively.

	SP-gap			With respect to E_C			With respect to E_V	
Material	Calc.	Exptl.	Donor	ΔKS	ΔSCF	Acceptor	ΔKS	ΔSCF
Diamond	5.42	5.48[a]	P_C	−0.6	−0.6	B_C	0.3	0.4
Silicon	1.17	1.23[a]	C_i	−1.0	−0.9	O_{Si}	0.9	1.0
Germanium	0.84	0.79[a]	S_{Ge}	−0.4	−0.3	O_{Ge}	0.4	0.4
Anatase	3.58	3.6[b]	Nb_{Ti}	−0.5	−0.4	Al_{Ti}	2.2	2.2
Rutile	3.37	3.4[b]	Nb_{Ti}	−1.0	−0.9	Al_{Ti}	0.5	0.4

[a]For reference, see Ref. 33.
[b]For reference, see Ref. 34.

TABLE II. Adiabatic (thermal) charge transition levels of defects in diamond (in electron volt). For references, see Refs. 33 and 36. (V_C denotes a vacancy.)

Defects levels in diamond	HSE06 in 512-atom cell	Exptl.
P_C (+/0)	$E_C - 0.5$	$E_C - 0.6$
B_C (0/−)	$E_V + 0.4$	$E_V + 0.4$
N_C (+/0)	$E_C - 1.8$	$E_C - 1.7$
V_C (+/0)	$E_C - 4.4$	$E_C - 4.3$
N_CV_C (0/−)	$E_V + 2.7$	$E_V + 2.6$
$(N_2)_C$ (+/0)	$E_C - 4.0$	$E_C - 4.0$
N_CV_CH (0/−)	$E_V + 2.4$	$E_V + 2.4$

The accuracy of predicting the charge transition levels is connected to the accuracy of the predicted localization of the defect state. This is especially important in the case of TiO_2, where it is well known that the same Nb_{Ti} substitutional has a deep donor level in rutile and a relatively shallow one in anatase, allowing the latter to be used as a transparent electrode (see Ref. 37 and references therein). HSE06 provides the donor level of Nb_{Ti} in a rutile supercell, corresponding to a concentration of ∼1%, at $E_C - 1.0$ eV,[34] in good agreement with the infrared transition at 0.8 eV, observed in samples with 3% Nb_{Ti}.[38] At the same time, Nb_{Ti} in anatase was found to be relatively shallow ($E_C - 0.4$ eV in the Γ-point approximation) with HSE06. In contrast, Al_{Ti} in anatase was found to be a deep acceptor, and the calculated recombination energy of an electron with the hole trapped at Al_{Ti}, 2.3 eV,[34] was in very good agreement with the observed 2.2 eV.[39] The reason for the difference between rutile and anatase can also be explained based on the HSE06 calculations. Rutile can trap free electrons in small polaron states, while anatase—in the bulk at least—cannot. In contrast, anatase in the bulk can capture free holes in small polaron states. As a consequence, *n*-type doping of rutile results in an electron-polaron with considerable binding energy, while it produces an effective mass like state in anatase. The situation is reversed for *p*-type doping, which produces hole-polarons in anatase and shallower states in rutile.[34]

The fine difference in carrier localization between rutile and anatase is actually a powerful test for any theoretical method. The well-localized small polaron states cannot be reproduced by pure ("convex") GGA calculations, which result in shallow effective mass like states for Nb_{Ti} and Al_{Ti} in both rutile and anatase, whereas the still "concave" hybrids, PBE0 and B3LYP localize both carriers in both modifications into small polaron states. (Note that GGA + U[40] with carefully chosen U values can lead to qualitatively correct results but cannot provide quantitative level positions because of the underestimation of the gap. Fitting U to reproduce the bandgap is physically unjustified and leads to grossly overlocalized defect states.)

In summary of this section, we may conclude that the two criteria of reproducing the 0 K single-particle bandgap and fulfilling the gKT, in other words the proper piecewise linear nature of the total energy as a function of the occupation numbers, allow the very accurate reproduction of the defect levels in the gap and of the localization of the defect state.

IV. OPTIMIZATION OF HSE(α,μ) HYBRIDS

Generally, the HSE06 hybrid has produced a substantial improvement over LDA/GGA in the predicted gaps of semiconductors, however, very wide gaps are underestimated,[41,42] while HSE fails in metals completely.[43] Obviously, its success is connected to the description of electronic screening.[44] However, the performance of the HSE06 parameterization is uneven also in the medium gap range. As shown in Table III for a series of Ga-based semiconductors, standard HSE06 consequently underestimates the gap all over the 1.2–4.9 eV range spanned by these materials. For cases like that, the general practice in the literature has become to increase the mixing parameter α to reproduce the room-temperature optical gap. The increased admixture of nonlocal exchange leads also to a more concave total energy as a function of the fractional occupation number and to stronger localization. However, as long as the linearity condition is not met, the accuracy in the predicted position of defect levels and in the degree of localization remains uncertain.

In a series of papers,[18,19,48] we have shown that, by tuning both parameters of a HSE(α,μ) functional simultaneously, one can achieve both the reproduction of the SP-gap and the fulfillment of the gKT in the semiconductors shown in Table III. The optimal α values are close to 0.25, as expected. The small variation of μ is significant for fulfilling the gKT.

TABLE III. Single-particle (SP) bandgaps calculated by standard HSE06 and by an optimized HSE(α,μ) hybrid. For the latter, the requirement of the gKT, i.e., the equality between t $\Delta KS = \Delta SCF$, is satisfied within <0.05 eV in every case, if using the FNV method for charge correction.

Semiconductor	SP-gap (eV) from experiment	SP-gap (eV) with HSE(0.25,0.20)	SP-gap (eV) with optimal HSE(α,μ)	Optimal HSE parameters α	μ
GaAs	1.6[a]	1.2	1.5	0.25	0.15
$CuGaS(Se)_2$	∼2.6(∼1.8)[b]	2.2	2.6 (1.8)	0.26	0.08
GaN	3.6[a]	3.2	3.6	0.26	0.10
Ga_2O_3	4.9[c]	4.3	5.0	0.26	0.00

[a]Zero temperature value, renormalized for electron-phonon interactions.[35]
[b]Values estimated from the optical gaps of 2.53 eV and 1.73 eV, respectively, measured at 2 K for $CuGaS_2$ in Ref. 45 and at 77 K for $CuGaSe_2$ in Ref. 46.
[c]Angle resolved photoemission spectroscopy at 20 K.[47]

TABLE IV. Relative energy of the band edges in GaN, as obtained by the GW_0 method and by the optimized HSE(0.26,010) hybrid.[19] (The zero of the energy scale is set to the valence band edge. All values in electron volt.)

k-point in the BZ	GW_0	HSE(0.26,0.10)
Valence band Γ	0.00	0.00
A	−0.61	−0.59
M	−1.13	−1.09
K	−3.09	−3.00
Conduction band Γ	3.63	3.61
A	6.23	6.14
M	6.85	6.79
K	6.72	6.70

TABLE V. The minimum bandgap of the β-Ga_2O_3 supercell (in the Γ-point approximation, at the experimental geometry of the supercell) and the fulfillment of the generalized Koopmans's theorem for V_O, as a function of the HSE(α,μ) parameters. The target value for the SP-gap was chosen to be 5.0 ± 0.1 eV.[18] (All values in electron volt. Starting and end values in bold.)

α/μ	E_g (perfect)	ΔKS−ΔSCF
0.35/0.10	5.5	0.1
0.35/0.20	**5.0**	**0.3**
0.45/0.20	5.8	0.2
0.26/0.00	**5.1**	**0.0**

Considering the first criterion, we emphasize again that the optical gap, which is most often available experimentally (especially at room temperature), is an underestimation of the SP-gap (at 0 K), which is expected from a DFT calculation in the case of the exact functional. Using the room-temperature (RT) optical gap as the target might seem very convenient (because, experimentally, the defect levels are most often determined at RT, too, and related to the optical gap), but it corrupts the method. Experimentally, the SP-gap can be obtained from photoelectron spectroscopy; however, those experiments are rarely carried out at low temperatures, and the experimental uncertainty may often be several tenths of an electron volt. Therefore, it is better to exclude experimental data and use first-principle many-body methods, like GW[49] (on the primitive unit cell), to determine the 0 K SP-gap. Generally, partially self-consistent GW_0 calculations give a better agreement with the experiment than the single shot G_0W_0 or fully self-consistent GW without vertex corrections;[50,51] therefore, such calculations are recommended, making sure that the result is converged with respect to the empty states involved.

Actually such GW_0 calculations are unavoidable anyway, as the reproduction of the minimum bandgap alone is not sufficient. Since the wave function of a defect can be derived from a superposition of *all* band edge states, there is little chance to reproduce the correct defect level position if the shape of the band edges is in error. Therefore, the optimal HSE(α,μ) has to reproduce the relative position of the band edge states over the BZ, as obtained from the reference GW_0 calculation. In the semiconductors listed in Table III, this condition has been met, as shown, e.g., for GaN in Table IV.

Checking the second condition, the fulfillment of the gKT, requires tests on a defect with a sufficiently localized gap state, since the accuracy of the charge correction in calculating Δ*SCF* increases with localization. (Note that for practical supercell sizes, the use of a dispersion correction may be necessary.)

It is also important to note that the gKT only applies if the change of the occupation does not influence much the other single-particle states. This requirement is posed by the fact that the derivation of the gKT relies on the existence of a connecting state between the N and N + 1 particle Hilbert spaces, in analogy to a Slater transition state.[52] However, for a strongly correlated N particle state, such a transition state will not exist in general. It has been shown[53] that an analogous expression to Janak's theorem can be formulated in the general case, but for populating a many-body state between the N and N + 1 Hilbert spaces. However in the (generalized) KS scheme, DFT is not able to do this, and hence the fulfillment of the gKT is not guaranteed for strongly correlated systems. Therefore, defects with strongly correlated states, e.g., in Jahn-Teller degenerate systems, or with high spin multiplicity, should be avoided. The best choice is to find a defect which produces a single dangling bond state on one of the host atoms. Actually, in binary semiconductors, it is recommended to check the gKT on orbitals of both host atoms.

The actual optimization process can be made very simple (with some sacrifice of accuracy). We have carried out calculations

TABLE VI. Adiabatic charge transition levels (in electron volt), obtained by the optimized hybrids,[18,19] in comparison with the experiment.

Material	Defect	Level	HSE(α,μ)	Experiment
GaN	C_N	(0/−)	E_V + 1.1	E_V + (0.95–1.06)[a,b]
	$N_i + xC_N$	(2+/+)	E_V + (0.2–0.3)	E_V + 0.2[c]
β-Ga_2O_3	$V_{Ga(1)}$	(2−/3−)	E_C − 0.7	E_C − 0.8[d,e]
	Hole-polaron	(+/0)	E_C − 4.50/4.61	E_C − 4.5[b,f]

[a]Deep Level Transient Spectroscopy (DLTS) from Ref. 55.
[b]The room-temperature experimental values were corrected with half the increase of the gap between RT and 0 K.
[c]DLTS from Ref. 56.
[d]DLTS from Ref. 57.
[e]It should be noted that in the case of the (2−/3−) transition, this good agreement can only be achieved by using ε_∞ instead of ε_0.
[f]Deep level optical spectroscopy (DLOS) from Ref. 58.

for a medium-size supercell with experimental lattice parameters and checked the gKT for the chosen defect in the same supercell. (Note: for ε_∞ in the charge correction we have used the experimental value mostly, to circumvent the need for a self-consistent fitting, but this empiricism can be avoided by using ε_0 from the reference GW_0 calculation.) Since both the value of the gap and the difference $\Delta KS - \Delta SCF$ (which should be zero if the gKT is fulfilled) change nearly linearly for small changes of both α and μ, extrapolation is easy and usually less than half a dozen calculations are needed. This is demonstrated for β-Ga_2O_3 in Table V. In that given case, we have used the donor level of the oxygen vacancy (V_O) and started from the tuned α parameter found in the literature to reproduce the bandgap. As can be seen, the deviation from the gKT is substantial. For a donor, the positive value of $\Delta KS - \Delta SCF$ means an upshift of the defect level upon electron removal, i.e., a convex behavior. Therefore, the weight of nonlocal exchange has to be increased. This can be achieved either by increasing α or decreasing μ. However, as shown in Table V, both increase the bandgap beyond the target value. So, the only way to satisfy both the bandgap and the gKT criterion is to decrease both μ and α by different amounts. After some trials, we have arrived[18] at the combination of $\alpha = 0.26$ and $\mu = 0.00$ (which, in this case is very near to the PBE0 functional). Optimizing the lattice constants with these parameters has resulted in $a = 12.25$ Å, $b = 3.04$ Å, $c = 5.79$ Å, and $\beta = 103.8°$, in comparison to the experimental values,[54] 12.23 Å, 3.04 Å, 5.81 Å, and 103.8°, respectively. At this geometry, the minimum bandgap is 5.02 eV.

As a result of similar optimizations, we have obtained reliable data for defects in all Ga-based semiconductors listed in Table III. Table VI shows adiabatic charge transition levels and Table VII photoluminescence (PL) energies, calculated by the optimized hybrids, in comparison with the experiments.

In addition to the defect levels, the localization of the defect state is also well described. This can be seen in Table VIII, where the calculated hyperfine parameters for a hole-polaron, trapped at a Mg_{Ga}^{-} substitutional in β-Ga_2O_3,[63] are compared to the experiment. It should be noted that due to the anisotropy of the trigonal supercell, the wave function is not equally well accommodated in all directions, leading to some anisotropy in the calculated results. The dielectric constant is very low in this material, so even a 160 atom unit cell is somewhat small, therefore, the localization is also slightly overestimated.

TABLE VII. PL energies (in electron volt), obtained by the optimized hybrids,[48,59] in comparison with the experiment.

Material	Defect	HSE(α,μ)	Experiment
$CuGaSe_2$	Ga_{Cu}	1.3	1.2[a]
β-Ga_2O_3	Hole-polaron	3.7/3.5	3.6/3.3[b]
	$V_{Ga} + V_O$	2.9	3.0[b]
	O_i	2.3	2.4[b]
	N_O	1.8	1.8[c]

[a]Reference 60.
[b]Reference 61.
[c]Reference 62.

TABLE VIII. Superhyperfine interaction of the trapped-hole in Mg-dopes β-Ga_2O_3 along the crystal axes (*a*, *b*, *c*) from the optimal HSE(0.26,0.0) calculation and from the experiment.[64]

	HSE(0.26,0.0) (MHz)			Experiment (MHz)		
Site	*a*	*b*	*c*	*A*	*b*	*c*
Ga1	76.32	81.17	72.14	73.15	71.74	71.46
Ga2	40.43	41.21	39.78	33.07	33.35	31.67

To summarize this section we can conclude that tuning the parameters of HSE(α,μ) to fulfill the two criteria (gap and gKT) leads to quantitatively accurate prediction of defect levels and localizations in all the investigated cases.

V. SUMMARY AND OUTLOOK

The characteristic fingerprints of defects in the materials of micro-/optoelectronics and photonics are the charge transition levels, measured, e.g., by deep level spectroscopies, the defect related electronic transitions, measured, e.g., in photoluminescence, and the electron spin resonance signals related to the paramagnetic state of the defect. To identify specific defects based on these fingerprints, electronic structure calculations are needed which can reproduce the bandgap and predict the correct position of the defect levels in it, as well as the localization of the related wave function. The fulfillment of these requirements is connected in exact DFT to the fact that the total energy is a piecewise linear function of the occupation numbers. The proper amount of derivative discontinuity at integer numbers ensures the correct gap and the linearity for fractional numbers leads to the fulfillment of the generalized Koopman's theorem, i.e., to accurate level positions and to the accurate degree of localization. We have found that, for C, Si, Ge, TiO_2, GaAs, $CuGaS(Se)_2$, GaSe, GaN, Ga_2O_3 and h-BN, the parameters of a HSE(α,μ) hybrid functional can be adjusted to achieve the correct piecewise linear behavior. For calculating defect properties in bulk semiconductors, such optimized hybrids allow unprecedented accuracy, which at present cannot be matched by any other method (either due to inherent theoretical limits or due to computational costs).

The question arises, whether the optimization procedure outlined here can always be carried out successfully. Here, we show exceptions where this is not the case.

As mentioned earlier, the key to success is the proper description of electronic screening.[65] This is calculated microscopically in the GW method, while HSE supplies only a simple approximation. For example, direction- or orbital-dependent screening effects cannot be mimicked by it. This can be seen, e.g., in GaSe, where the conduction band edge states do not contain a contribution from Ga 3*d* orbitals, so their inclusion in a HSE(α,μ) calculation does not influence the relative position of the band edge states, while the screening effect of the 3*d* states in GW_0 changes them substantially. Another example is ZnO, where the convergence of the GW calculation is very difficult to achieve.[66] This indicates that screening is quite complicated in this material. Consequently, (α,μ) parameters, optimized to fulfill the gKT and reproduce the minimum gap, do

not provide a correct description for the relative position of the band edge states, and the predicted defect levels are in error.[18]

Obviously, the optimized (α,μ) parameters are materials specific. The case of $CuGaS_2$ and $CuGaSe_2$, where the same parameters work in both (see Table III), seem to be rather exceptional. When, e.g., In is substituted for Ga in these compounds, new optimization becomes necessary. Therefore, alloys like $CuGa_{1-x}In_xS(Se)_2$ cannot be treated well by a HSE(α,μ) functional. The problem is similar at solid interfaces, with the optimal parameters mostly being different in the two phases.

Actually, surfaces are also problematic, since screening in the vicinity of the solid/vacuum interface is different from that in the bulk.[67] This can also be seen in layered materials, like GaSe. The dielectric constant in bulk GaSe is still quite isotropic, so we have been able to find an optimized parameter set, which fulfills the gKT and reproduces the gap as well as the relative position of the band edge states across the BZ.[68] However, the values $\alpha = 0.40$ and $\mu = 0.25$ are quite different from the ones found for the other Ga-based semiconductors in Table III. (We note that the optimal parameters of layered bulk hexagonal-BN are similar in value.[68]) In addition, the parameters optimized for bulk GaSe do not work for a monolayer (ML). It has been noticed earlier, that HSE06 breaks down in quasitwo-dimensional (2D) layers.[69] An ML of GaSe is about 5 Å thick and has nearly the same ε_∞ as in the bulk, so we have been able to retune the HSE parameters.[68] However, the simplistic description of screening becomes untenable in the case of 2D hexagonal boron nitride, where the criteria of reproducing the gap and fulfilling the gKT cannot be achieved simultaneously.[68]

The experience, described here, allows us to sketch the expected road to further development. The original idea of hybrid functionals was based on the adiabatic connection principle, or coupling strength integration,[21] which, however, cannot be used to deduce a simple but general two-point approximation in terms of mixing GGA and HF exchange (V^X) in some given ratio. Besides Becke's B3LYP functional[21] (which has helped DFT to earn a Nobel prize in chemistry), the one-parameter PBEα functionals

$$\begin{aligned} V^X(q) &= \alpha V^X_{HF}(q) + (1-\alpha)V^X_{GGA}(q) \\ &= \varepsilon^{-1}(q)V^X_{HF}(q) + [1-\varepsilon^{-1}(q)]V^X_{GGA}(q) \end{aligned} \tag{2}$$

have initially been applied in solids, where α is the mixing ratio. The use of α = ¼ can be justified based on perturbation theory[24] and is used also in the PBE0 functional.[23] It should be noted that the mixing can also be interpreted as a choice for an effective screening function of the quasimomentum q, $\varepsilon^{-1}(q)$, as shown in Eq. (1). However, obviously, screening in solids is more complicated than that, and an $\varepsilon^{-1}(q) = konst$ approximation is insufficient. Therefore, several approaches introduce a "screening length" on which the mixing is phased out.[22,31,65] The ones favored mostly in the solid state community today are the HSE(α, μ) functionals,[31,32] which are also employed in this work. The screening parameter μ provides an additional leverage to adjust the simple and phenomenological description of screening, however, at the end of day, one can only achieve a semiempirical error compensation between the two types of exchange, leading to nontransferable parameters.

An open issue with HSE is that its distance dependent screened potential

$$W^{HSE}(r) = \frac{erfc(\mu r)}{r} \tag{3}$$

does not obey the correct limits, i.e., it does not approach the unscreened Coulomb potential in the $r \to 0$ limit (see below). This means, on the one hand, that the parameters α and μ cannot be regarded as "physical" parameters, determined directly from $\varepsilon^{-1}(q)$. On the other hand, it is also theoretically unclear how details of the band structure can be reproduced by tuning α and μ. (As mentioned above on the example of ZnO, the reproduction of the minimum gap and fulfillment of the gKT does not guarantee that.) An attempt to remedy this situation has recently been suggested,[29] using a model dielectric function to mix GGA and HF exchange. An alternative is the screened exchange (sX) approximation,[30,70] where HF exchange is screened with a model dielectric function, but no mixing is used. The advantage lies in the fact that the screening tries to emulate the microscopic dielectric function. The momentum dependence of the effective dielectric functions used by the HSE and sX functionals is shown in Fig. 2. HSE has the wrong long wave-number limit, $\varepsilon^{-1}(q \to \infty) = \alpha = \frac{1}{4}$, while the sX method obeys the correct one, $\varepsilon^{-1}(q \to \infty) = 1$, expressing the fact that two electrons, that come arbitrarily close, do not feel any screening. Missing in both approaches, however, is the limit $\varepsilon^{-1}(q \to 0) = \epsilon_\infty^{-1}$, that states that two electrons very far from each other should feel the background screening of the crystal. In the sX approach, this problem frequently leads to an underestimated bandgap. In HSE, the behavior for $q \to 0$ and $q \to \infty$ do not obey the correct physical limits. In order to nevertheless approximate the total exchange potential, HF and GGA exchange are mixed. However, to get reasonable agreement in the quasimomentum range most important for the material in question, both α and μ need to be tuned and will change from material to material.

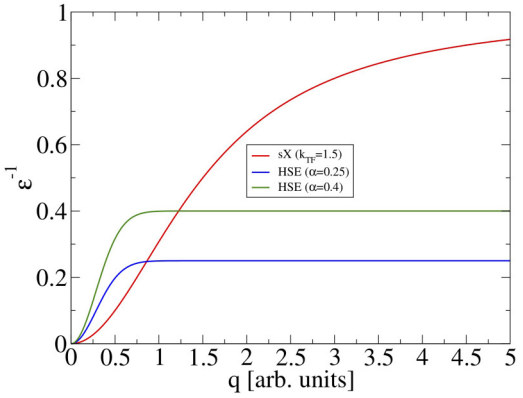

FIG. 2. The effective dielectric functions in the sX and in the HSE method at two different mixing parameters. (Note: the effect of the screening length μ is not considered here.)

To develop the field further and to come to a more general method for accurately describing defects in semiconductors, a better description of screening is needed, adhering to the microscopic limits of the dielectric function. Optimally, the parameters of such a functional should be physically motivated and transferable between different materials to describe alloys. While such a general functional could work in the bulk if the screening is not orbital-dependent, the presence of an interface introduces spatial dependence into the screening function. To accommodate that in the functional would be a next step. Work in both directions is in progress.

ACKNOWLEDGMENTS

Funding by the DFG research project (No. FR2833/63-1) and the support of the Supercomputer Center of Northern Germany (HLRN Grant No. hbc00027) is acknowledged. P.D. is indebted to S. Lany and G. Kresse for very fruitful discussions.

REFERENCES

[1]A. Alkauskas, P. Deák, J. Neugebauer, A. Pasquarello, and C. G. Van de Walle, *Advanced Calculations for Defects in Materials: Electronic Structure Methods* (Wiley-VCH, Berlin, 2011).
[2]C. Freysoldt, B. Grabowski, T. Hickel, J. Neugebauer, G. Kresse, A. Janotti, and C. G. Van de Walle, Rev. Mod. Phys. **86**, 253 (2014).
[3]R. Dovesi, R. Orlando, C. Roetti, C. Pisani, and V. R. Saunders, Phys. Status Solidi (B) **217**, 63 (2000).
[4]J. Muscat, A. Wander, and N. M. Harrison, Chem. Phys. Lett. **342**, 397 (2001).
[5]P. Deák, A. Gali, A. Sólyom, A. Buruzs, and T. Frauenheim, J. Phys. Condens. Matter **17**, S2141 (2005).
[6]G. Kresse and J. Hafner, Phys. Rev. B **49**, 14251 (1994). ;G. Kresse and J. Furthmüller, ibid. **54**, 11169 (1996);G. Kresse and D. Joubert, ibid. **59**, 1758 (1999).
[7]F. D. Murnaghan, Proc. Nat. Acad. Sci. U.S.A. **30**, 244 (1944).
[8]H. J. Monkhorst and J. D. Pack, Phys. Rev. B **13**, 5188 (1976).
[9]S. Lany and A. Zunger, Modell. Simul. Mater. Sci. Eng. **17**, 084002 (2009).
[10]H.-P. Komsa, T. T. Rantala, and A. Pasquarello, Phys. Rev. B **86**, 045112 (2012).
[11]C. Freysoldt, J. Neugebauer, and C. G. Van de Walle, Phys. Rev. Lett. **102**, 016402 (2009).
[12]Source code can be downloaded from http://www.sphinxlib.de
[13]M. Farzalipour Tabriz, B. Aradi, T. Frauenheim, and P. Deák, Comput. Phys. Commun. **240**, 101 (2019).
[14]Source code can be downloaded from https://github.com/MFTabriz/slabcc
[15]H.-P. Komsa and A. Pasquarello, Phys. Rev. Lett. **110**, 095505 (2013).
[16]H.-P. Komsa, N. Berseneva, A. V. Krashennikov, and R. Nieminen, Phys. Rev. X **4**, 031044 (2014);H.-P. Komsa, N. Berseneva, A. V. Krashennikov, and R. Nieminen, Errata: *ibid.* . **8**, 039902(E) (2018).
[17]P. Deák, B. Aradi, and T. Frauenheim, Phys. Rev. B **86**, 195206 (2012).
[18]P. Deák, Q. D. Ho, F. Seemann, B. Aradi, M. Lorke, and T. Frauenheim, Phys. Rev. B **95**, 075208 (2017).
[19]P. Deák, M. Lorke, B. Aradi, and T. Frauenheim, Phys. Rev. B **99**, 085206 (2019).
[20]S. Lany and A. Zunger, Phys. Rev. B **80**, 085202 (2009).
[21]A. D. Becke, J. Chem. Phys. **98**, 5648 (1993).
[22]A. D. Becke, J. Chem. Phys. **104**, 1040 (1996).
[23]C. Adamo and V. Barone, J. Chem. Phys. **110**, 6158 (1999).
[24]J. P. Perdew, M. Ernzerhof, and K. Burke, J. Chem. Phys. **105**, 9982 (1996).
[25]A. Alkauskas, P. Broqvist, and A. Pasquarello, Phys. Status Solidi (B) **248**, 775 (2011).
[26]M. A. L. Marques, J. Vidal, M. J. T. Oliveira, L. Reining, and S. Botti, Phys. Rev. B **83**, 035119 (2011).
[27]M. Gerosa, C. E. Bottani, L. Caramella, G. Onida, C. Di Valentin, and G. Pacchioni, Phys. Rev. B **91**, 155201 (2015).
[28]G. Miceli, W. Chen, I. Reshetnyak, and A. Pasquarello, Phys. Rev. B **97**, 121112(R) (2018).
[29]W. Chen, G. Miceli, G.-M. Rignanese, and A. Pasquarello, Phys. Rev. Mater. **2**, 073803 (2018).
[30]S. J. Clark and J. Robertson, Phys. Rev. B **82**, 085208 (2010).
[31]J. Heyd, G. E. Scuseria, and M. Ernzerhof, J. Chem. Phys. **118**, 8207 (2003).
[32]A. V. Krukau, O. A. Vydrov, A. F. Izmaylov, and G. E. Scuseria, J. Chem. Phys. **125**, 224106 (2006).
[33]P. Deák, B. Aradi, T. Frauenheim, E. Janzén, and A. Gali, Phys. Rev. B **81**, 153203 (2010).
[34]P. Deák, B. Aradi, and T. Frauenheim, Phys. Rev. B **83**, 155207 (2011).
[35]M. Cardona and M. L. W. Thewalt, Rev. Mod. Phys. **77**, 1173 (2005).
[36]P. Deák, B. Aradi, M. Kaviani, T. Frauenheim, and A. Gali, Phys. Rev. B **89**, 075203 (2014).
[37]H. A. Huy, B. Aradi, T. Frauenheim, and P. Deák, J. Appl. Phys. **112**, 016103 (2012).
[38]D. Morris, Y. Dou, J. Rebane, C. E. J. Mitchell, R. G. Egdell, D. S. L. Law, A. Vittadini, and M. Casarin, Phys. Rev. B **61**, 13445 (2000).
[39]T. Sekiya, S. Kamei, and S. Kurita, J. Lumin. **87–89**, 1140 (2000).
[40]A. I. Liechtenstein, V. I. Anisimov, and J. Zaane, Phys. Rev. B **52**, R5467 (1995);S. L. Dudarev, G. A. Botton, S. Y. Savrasov, C. J. Humphreys, and A. P. Sutton, Phys. Rev. B **57**, 1505 (1998).
[41]J. Paier, M. Marsman, K. Hummer, G. Kresse, I. C. Gerber, and J. G. Ángyán, J. Chem. Phys. **124**, 154709 (2006).
[42]M. Marsman, J. Paier, A. Stroppa, and G. Kresse, J. Phys. Condens. Matter **20**, 064201 (2008).
[43]W. Gao, T. A. Abtew, T. Cai, Y.-Y. Sun, S. Zhang, and P. Zhang, Solid State Commun. **234–235**, 10 (2016).
[44]F. Fuchs, J. Furthmüller, F. Bechstedt, M. Siskin, and G. Kresse, Phys. Rev. B **76**, 115109 (2007).
[45]C. Bellabarba, J. González, and C. Rincón, Phys. Rev. B **53**, 7792 (1996).
[46]B. Tell and P. M. Bridenbaugh, Phys. Rev. B **12**, 3330 (1975).
[47]M. Mohamed, I. Unger, C. Janowitz, R. Manzke, Z. Galazka, R. Uecker, and R. Fornari, J. Phys. Conf. Ser. **286**, 012027 (2011).
[48]M. Han, Z. Zeng, T. Frauenheim, and P. Deák, Phys. Rev. B **96**, 165204 (2017).
[49]L. Hedin, Phys. Rev. **139**, A796 (1965).
[50]U. von Barth and B. Holm, Phys. Rev. B **54**, 8411 (1996).
[51]M. Shishkin, M. Marsman, and G. Kresse, Phys. Rev. Lett. **99**, 246403 (2007).
[52]J. C. Slater, Adv. Quantum Chem. **6**, 1 (1972).
[53]A. Gonis, World J. Condens. Matter Phys. **4**, 78 (2014).
[54]K. E. Lipinska-Kalita, P. E. Kalita, O. A. Hemmers, and T. Hartmann, Phys. Rev. B **77**, 094123 (2008).
[55]A. Armstrong, A. R. Arehart, D. Green, U. K. Mishra, J. S. Speck, and S. A. Ringel, J. Appl. Phys. **98**, 053704 (2005).
[56]U. Honda, Y. Yamada, Y. Tokuda, and K. Shiojima, Jpn. J. Appl. Phys. **51**, 04DF04 (2012).
[57]M. E. Ingebrigtsen, J. B. Varley, A. Yu. Kuznetsov, B. G. Svensson, G. Alfieri, A. Mihaila, U. Badstübner, and L. Vines, Appl. Phys. Lett. **112**, 042104 (2018).
[58]Z. Zhang, E. Farzana, A. R. Arehart, and S. A. Ringel, Appl. Phys. Lett. **108**, 052105 (2016).
[59]Q. D. Ho, T. Frauenheim, and P. Deák, Phys. Rev. B **97**, 115163 (2018).
[60]C. Spindler, D. Regesch, and S. Siebentritt, Appl. Phys. Lett. **109**, 032105 (2016).
[61]T. Onuma, S. Fujioka, T. Yamaguchi, M. Higashiwaki, K. Sasaki, T. Masui, and T. Honda, Appl. Phys. Lett. **103**, 041910 (2013).

[62]T. Zhang, J. Lin, X. Zhang, Y. Huang, X. Wu, Y. Xue, J. Zou, and C. Tang, J. Lumin. **140**, 30 (2013).
[63]Q. D. Ho, T. Frauenheim, and P. Deák, J. Appl. Phys. **124**, 145702 (2018).
[64]B. E. Kananen, L. E. Halliburton, E. M. Scherrer, K. T. Stevens, G. K. Foundos, K. B. Chang, and N. C. Giles, Appl. Phys. Lett. **111**, 072102 (2017).
[65]X. Zheng, A. J. Cohen, P. Mori-Sanchez, X. Hu, and W. Yang, Phys. Rev. Lett. **107**, 026403 (2011).
[66]J. Klimeš, M. Kaltak, and G. Kresse, Phys. Rev. B **90**, 075125 (2014).
[67]M. Farzalipour Tabriz, B. Aradi, T. Frauenheim, and P. Deák, J. Phys. Condens. Matter **29**, 394001 (2017).
[68]P. Deák, E. Khorasani, M. Lorke, M. Farzalipour Tabriz, B. Aradi, and T. Frauenheim, "Defect calculations with hybrid functionals in layered compounds and in slab models," Phys. Rev. B (submitted).
[69]M. Jain, J. R. Chelikowsky, and S. G. Louie, Phys. Rev. Lett. **107**, 216806 (2011).
[70]Y. Guo, J. Robertson, and S. J. Clark, J. Phys. Condens. Matter **27**, 025501 (2015).