Mémoire
d'Habilitation à Diriger des Recherches

***

Electronic Structure of Point Defects in Semiconductors

***

Fabien Bruneval

soutenue le 18 juin 2014
à l'Ecole Doctorale de l'université Claude Bernard Lyon I

Contents

Résumé

Ce mémoire d'habilitation à diriger des recherches présente la majeure partie de mon activité scientifique au cours de ces 7 dernières années, dans le domaine des calculs de structure électronique des défauts dans les solides.

Les défauts ponctuels (lacunes, interstitiels, impuretés) dans les matériaux fonctionnels jouent un rôle primordial pour déterminer si ces matériaux vont effectivement remplir le rôle qu'on leur a assigné ou pas. En effet, la présence de défauts est inévitable dès que la température s'élève ou que le matériau est soumis à des sollicitations externes comme l'irradiation dans les réacteurs nucléaires ou les satellites artificiels avec les rayonnements cosmiques. Cependant dans de nombreux cas, les défauts sont introduits dans le matériau de façon volontaire afin de contrôler les propriétés de transport électronique, optiques, ou même magnétiques. On parle alors du dopage des semiconducteurs, technique qui est à la base des transistors, des diodes ou des cellules photovoltaïques actuelles. Malheureusement, le dopage présente souvent des particularités inattendues, telles que les asymétries de dopage et l'épinglement du niveau de Fermi, qui ne peuvent s'expliquer que par des phénomènes complexes mettant en jeu différents types de défauts ou de complexes de défauts.

Dans ce contexte, les calculs de structure électroniques *ab initio* constituent un outil de choix pour compléter les observations expérimentales, pour affiner la compréhension des phénomènes au niveau atomique, et même pour prédire les propriétés des défauts. La force des calculs *ab initio* réside en ce qu'ils permettent en principe une description sans aucun ajustement spécifique de n'importe quel système d'électrons et de noyaux. Mais bien qu'il y ait un besoin fort de simulation numérique dans ce domaine, les calculs *ab initio* pour les défauts sont encore en développement à l'heure où ces mots sont écrits. Les travaux exposés dans ce mémoire résument ma contribution aux développements méthodologiques dans cette voie. Ces développements ont porté essentiellement sur deux pistes.

Le premier sujet d'étude est la meilleure compréhension des inévitables effets de taille finie. En effet, les défauts des semiconducteurs ou isolants sont généralement présents en concentration infimes (de l'ordre d'un pour un million). En revanche, étant donnée la lourdeur des calculs quantiques de structure électroniques qui croit très rapidement avec le nombre d'électrons, les systèmes simulés par ordinateur dépassent difficilement les quelques centaines d'atomes à l'heure actuelle. Ceci conduit à des concentrations de défaut effectives de l'ordre

du pourcent qui sont bien loin de la limite des défauts dilués de l'expérience. L'extrapolation des concentrations fortes vers les concentrations faibles est délicate car les défauts des semiconducteurs portent très souvent une charge électrique nette qui induit des interactions entre défauts chargés à longue portée. La première partie de mon travail expose les techniques disponibles dans ce domaine et quelques contributions à l'amélioration et à la compréhension des celles-ci.

Le second domaine de recherche présenté porte sur l'amélioration de la structure électronique des défauts dans les semiconducteurs et isolants. Les défauts dans ces matériaux introduisent des niveaux électroniques à l'intérieur de bande interdite du matériau parfait. Ces niveaux électroniques correspondent aux électrons participant au défaut. Leur fonction d'onde est plus ou moins localisée autour de la région du défaut et leur remplissage peut varier selon les conditions thermodynamiques. Ces niveaux à l'intérieur de la bande interdite gouvernent la modification des propriétés de transport électronique et optique. Malheureusement les techniques *ab initio* usuelles dans le cadre de la théorie de la fonctionnelle de la densité (DFT) sont incapables d'obtenir des largeurs correctes des bandes interdites des semiconducteurs et isolants. C'est pourquoi de nombreuses propriétés de défaut ne peuvent être prédites avec certitude avec cette approche. Cette deuxième partie de mon travail expose et démontre comment l'introduction de la théorie du problème à N corps dans l'approximation dite GW permet de résoudre le problème des bandes interdites et permet d'obtenir ainsi des propriétés des défauts plus fiables.

Bien entendu, le domaine de structure électronique *ab initio* des défauts est loin d'être un sujet de recherche épuisé, tant du point de vue des avancées théoriques, que des avancées expérimentales. L'avènement de calculateurs plus performants permettra d'utiliser des théories plus précises, de traiter des défauts plus dilués, et des défauts plus complexes. Nous pouvons aussi anticiper que dans un futur proche les besoins technologiques vont continuer à alimenter l'intérêt pour les défauts ponctuels: par exemple, l'informatique quantique se fonde en partie sur des bits élémentaires constitué de défauts préparés dans des états de spin bien déterminés; le développement de nouvelles cellules photovoltaïques nécessite la caractérisation des défauts qui limitent l'efficacité de la séparation de charge.

Summary

This "Habilitation à diriger des Recherches" memoir presents most of my scientific activities during the past 7 years, in the field of electronic structure calculations of defects in solids.

Point defects (vacancies, interstitials, impurities) in functional materials are a key parameter to determine if these materials will actually fill the role they have been assigned or not. Indeed, the presence of defects cannot be avoided when the temperature is increased or when the material is subjected to external stresses, such as irradiation in the nuclear reactors and in artificial satellites with solar radiations. However, in many cases, defects are introduced in the materials on purpose to tune the electronic transport, optical or even magnetic properties. This procedure is called the doping of semiconductors, which is the foundation technique for transistors, diodes, or photovoltaic cells. However, doping is not always straightforward and unexpected features may occur, such as doping asymmetry or Fermi level pinning, which can only be explained by complex phenomena involving different types of defects or complexes of defects.

In this context, the calculations of electronic structure *ab initio* is an ideal tool to complement the experimental observations, to gain the understanding of phenomena at the atomic level, and even to predict the properties of defects. The power of the *ab initio* calculations comes from their ability to describe any system of electrons and nuclei without any specific adjustment. But although there is a strong need for numerical simulations in this field, the *ab initio* calculations for defects are still under development as of today. The work presented in this memoir summarizes my contributions to methodological developments on this subject. These developments have followed two main tracks.

The first topic is the better understanding of the unavoidable finite size effects. Indeed, defects in semiconductors or insulators are generally present in trace concentration (of the order of one part per million). However, owing to the heavy burden of the quantum-mechanical electronic structure calculations, which grow very rapidly with the number of electrons, the present day simulations do not easily exceed a few hundred atoms nowadays. This induces effective defect concentrations of the order of one percent which are very far from the diluted defects observed in the experiments. The extrapolation of high concentrations to low concentrations is difficult because defects in semiconductors often bear a net electric charge which induces long-range interactions between the spuriously interacting charged defects. The first part of my work presents

the techniques available in this area, improvements in the techniques and some understanding of these spurious interactions.

The second topic addressed in this memoir focuses on improving the electronic structure of defects in semiconductors and insulators. Defects in these materials introduce discrete electronic levels within the band gap of the pristine bulk material. These electronic levels correspond to the electrons involved in the defect states. Their wavefunction is more or less localized around the defect region and the filling of the state may also vary with the thermodynamic conditions (Fermi level). These levels inside the band gap govern the modification of the properties of electronic and optical transport. Unfortunately the standard *ab initio* approaches, in the context of Density Functional Theory (DFT), are unable to get the correct band gaps of semiconductors and insulators. This is why many defect properties cannot be predicted with certainty within these approaches. This second part demonstrates how the introduction of the many-body perturbation theory in the so-called GW approximation solves the problem of band gaps and thus allows one to obtain more reliable defect properties.

Of course, the field of *ab initio* electronic structure for defects is far from being a finalized research subject, because of the theoretical advances, as well as the experimental progresses. The advent of more powerful computers allows the use of more accurate theories, the calculations of more diluted (more realistic) defects, and also the calculations of more complex defects. We can also anticipate that in the near future, technological applications will continue to feed the interest in point defects, e.g. with the field of quantum computing whose elementary bits are built from well-prepared defects in specific spin states; with the development of new solar cells that will require the fine characterization of defects that hinder the efficiency of the charge separation.

Part I

A short introduction to point defects in semiconductors

1 Why should we care about defects?

In crystalline solids, the presence of defects cannot be avoided. Point defects are disruptions in the perfect ordering of the atoms in a crystal. The defects can be caused by missing, substituted, or added atoms. Even the purest electronic grade silicon samples still contain impurities, e.g. other elements than silicon. Furthermore, the defect concentration is non zero at thermodynamical equilibrium because of entropy. Although point defects have a significant cost in terms of energy, named formation energy, they are ubiquitous in realistic samples.

The most common point defects are vacancies (a lattice site in the crystal remains empty), self-interstitials (an atom of the crystal lies off the lattice sites) and impurities. Other defect types are also possible (antisites, clusters etc...). The presence of defects is often thought as detrimental for the material properties. But it is not true in general: defects may induce many useful properties for technological applications.

The semiconductor electronics are entirely based on the tuning of the conduction properties of crystals with impurities. This case is referred to as "doping". With the introduction of the suitably chosen impurities, it is possible to transform intrinsic silicon, which is a poor charge conductor, into a reasonable conductor for holes (p-type doping) or electrons (n-type doping). Sandwiching the n and p conducting regions is the recipe for the production of semiconductors transistors, light-emitting diodes, etc. Photovoltaic cells also rely on p and n layers, which induce a permanent electric field ensuring the charge separation, once a photon has been absorbed.

Optical properties can also be tuned by defects. Think of the difference between the colorless alumina (Al_2O_3), the blue sapphire and red ruby gemstones. The three crystals share the same corundum matrix, however sapphire contains traces of iron and titanium and ruby contains impurities of chromium. A tiny change in the composition can induce magnificent changes in terms of optical transmission!

The mechanical properties also affected by the defects. The difference between the ductile iron and the brittle steel is explained by the carbon impurities.

However, the defects are not always desirable. In semiconductor doping, other impurities or self-defects can compensate the targeted doping. Many wide band gap semiconductors unfortunately experience a doping asymmetry: while one doping type is readily obtained, the opposite doping is extremely difficult. In photovoltaic cells, the defects are the location for the detrimental electron-hole recombinations, which produce useless photons instead of the desired electric current.

The characterization of defects is particularly relevant in the harsh environments encountered in nuclear plants or in satellites subjected to cosmic rays. When a crystal is subjected to irradiation (fast electrons, ions, or neutrons etc...), the atoms of the solid are regularly kicked off their perfect lattice site due to the interaction with the energetic impinging particles. When an atom is ejected from its site, this induces a defect pair called a Frenkel pair, which consists of an self-interstitial atom and a vacancy.

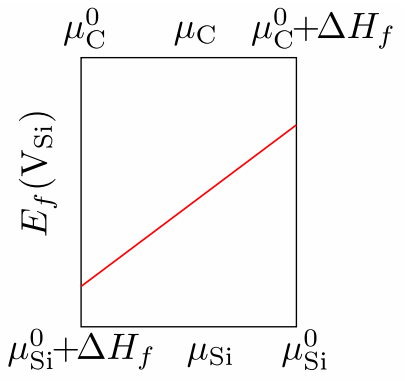

Figure 1: Schematic representation of the formation energy of a silicon vacancy $\mathrm{V_{Si}}$ in silicon carbide as a function of the silicon chemical potential μ_{Si} or alternatively as a function of the carbon chemical potential μ_{C}. The chemical potentials are bounded by the inequalities of Eq. (12a).

After this short walk through the zoo of the different defect types and their applications, I hope the interest of point defects for practical applications has been made clear. The remainder of this introductory chapter is organized as follows: the next section will introduce the important physical quantities for the characterization of point defects; the following section will describe in a few words the most important experimental techniques for measuring defects; then the final section will introduce the numerical techniques for the computer simulations of point defects and their limitations. This last section will outline the motivations for the research works presented in this memoir.

2 Point defect basic properties

2.1 Formation energy and charge transition levels

To characterize a specific defect, the very first question is the probability of occurrence of such a defect. This probability (or in other words its concentration) is a thermodynamical quantity, which depends on several intensive parameters (temperature, pressure, chemical potentials...). The correct thermodynamical quantity to obtain the concentration of defects would be then the formation Gibbs free energy. However, for most applications of *ab initio* calculations, the physical description remains at zero temperature with zero pressure. Thus the central quantity simply becomes the formation energy of a defect E_f. Neglecting the entropic contributions might pose problems when comparing the calculated quantities to high-temperature measurements. Obtaining the concentration of defects under pressure would require to introduce the enthalpy instead.

Let me introduce the expression for the formation energy for the simplest

case and progressively introduce complexity to finally reach the general case: an unspecified defect having a net charge in a compound solid. First consider the formation of a neutral vacancy in an elemental solid, say silicon,

$$\mathrm{Si}_N \rightleftharpoons \mathrm{Si}_{N-1} + \mathrm{Si}. \tag{1}$$

The perfect crystal with N silicon atoms is transformed into an defective crystal with $N-1$ atoms and a silicon atom placed in a reservoir. The corresponding formation energy reads

$$E_f(\mathrm{V_{Si}}) = E_{\mathrm{total}}(\mathrm{Si}_{N-1}) + \mu_{\mathrm{Si}} - E_{\mathrm{total}}(\mathrm{Si}_N), \tag{2}$$

where μ_{Si} is the chemical potential of the silicon atom fixed by a reservoir. In a pure elemental solid, the chemical potential of the element is imposed by the stability of the host material:

$$\mu_{\mathrm{Si}} = \mu^0_{\mathrm{Si}} = \frac{1}{N} E_{\mathrm{total}}(\mathrm{Si}_N). \tag{3}$$

Finally, Eq. (2) simply reads

$$E_f(\mathrm{V_{Si}}) = E_{\mathrm{total}}(\mathrm{Si}_{N-1}) - \frac{N-1}{N} E_{\mathrm{total}}(\mathrm{Si}_N) \tag{4}$$

and the formation energy is a number, independent from any external parameter.

When turning to a defect, in a compound, the situation becomes slightly more complex. Consider now a neutral silicon vacancy in silicon carbide to fix the ideas. The formation reaction is very similar:

$$\mathrm{Si}_N\mathrm{C}_N \rightleftharpoons \mathrm{Si}_{N-1}\mathrm{C}_N + \mathrm{Si} \tag{5}$$

and the formation energy also remains

$$E_f(\mathrm{V_{Si}}) = E_{\mathrm{total}}(\mathrm{Si}_{N-1}\mathrm{C}_N) + \mu_{\mathrm{Si}} - E_{\mathrm{total}}(\mathrm{Si}_N\mathrm{C}_N). \tag{6}$$

The only change comes from the chemical potential of silicon μ_{Si}. Since now the solid is silicon carbide, this reservoir only imposes the chemical potential of SiC:

$$\mu_{\mathrm{SiC}} = \mu_{\mathrm{Si}} + \mu_{\mathrm{C}} = \mu^0_{\mathrm{SiC}}. \tag{7}$$

Just the sum of carbon and silicon chemical potentials is fixed to μ^0_{SiC}, not their individual values. As a consequence, the formation energy is not any more a single number but it is rather a function of the chemical potentials. However, the thermodynamical conditions impose a finite range of variation for μ_{Si} and μ_{C}. The compound material should indeed be stable against the decomposition into separated phases. Here we introduce the zero temperature heat of formation, which reads in the case of silicon carbide:

$$\begin{aligned} \Delta H_f &= \mu^0_{\mathrm{SiC}} - \mu^0_{\mathrm{Si}} - \mu^0_{\mathrm{C}} && (8)\\ &= E_{\mathrm{total}}(\mathrm{SiC}) - E_{\mathrm{total}}(\mathrm{Si}) - E_{\mathrm{total}}(\mathrm{C}). && (9) \end{aligned}$$

Bulk silicon carbide is stable against phase separation into bulk silicon and bulk carbon (graphite) at zero temperature, since ΔH_f has a negative value.

The stability of silicon and carbon in SiC imposes the two following inequalities:

$$\mu_{\mathrm{Si}} < \mu^0_{\mathrm{Si}} \tag{10a}$$
$$\mu_{\mathrm{C}} < \mu^0_{\mathrm{C}}. \tag{10b}$$

And as the sum of these two chemical potentials is imposed in Eq. (7), then two additional inequalities hold

$$\mu_{\mathrm{Si}} = \mu^0_{\mathrm{SiC}} - \mu_{\mathrm{C}} > \mu^0_{\mathrm{SiC}} - \mu^0_{\mathrm{C}} = \mu^0_{\mathrm{Si}} + \Delta H_f \tag{11a}$$
$$\mu_{\mathrm{C}} = \mu^0_{\mathrm{SiC}} - \mu_{\mathrm{Si}} > \mu^0_{\mathrm{SiC}} - \mu^0_{\mathrm{Si}} = \mu^0_{\mathrm{C}} + \Delta H_f. \tag{11b}$$

Summarizing all these inequalities, μ_{Si} is bounded as

$$\mu^0_{\mathrm{Si}} + \Delta H_f < \mu_{\mathrm{Si}} < \mu^0_{\mathrm{Si}} \tag{12a}$$

and μ_{C} is bounded in a similar way:

$$\mu^0_{\mathrm{C}} + \Delta H_f < \mu_{\mathrm{C}} < \mu^0_{\mathrm{C}}. \tag{12b}$$

The limit when μ_{Si} approaches to μ^0_{Si} is named silicon-rich thermodynamical conditions. This situation occurs when the silicon atoms are in excess compared to carbon atoms. The opposite limit when μ_{Si} approaches $\mu^0_{\mathrm{Si}} + \Delta H_f$ implies an excess of carbon atoms: this is the carbon-rich thermodynamical conditions.

The schematic representation of the silicon vacancy formation energy is provided in Figure 1. The formation energy is obtained from Eq. (6) and the chemical potentials are bounded with the inequalities of Eq. (12a). The formation energy of a defect is then a function of the chemical potential imposed by the equilibrium in certain thermodynamical conditions. It is not surprising that the higher the chemical potential of silicon, the more difficult the creation of a vacancy. Within silicon rich conditions ($\mu_{\mathrm{Si}} = \mu^0_{\mathrm{Si}}$), the silicon element is abundant and the vacancy formation is limited. At the opposite, in silicon poor conditions (equal to carbon rich), the silicon element is more rare and the vacancy concentration increases.

Some time has been spent in explaining in details the chemical potential dependence. This is useful to introduce the charged defects in equilibrium with an electron reservoir. Let me exemplify this with the charged silicon vacancy $\mathrm{V}^+_{\mathrm{Si}}$ in SiC. The corresponding formation reaction reads

$$\mathrm{Si}_N\mathrm{C}_N \rightleftharpoons \mathrm{Si}^+_{N-1}\mathrm{C}_N + \mathrm{Si} + e^-, \tag{13}$$

where e^- stands for an electron. The formation energy now reads

$$E_f(\mathrm{V}^+_{\mathrm{Si}}) = E_{\mathrm{total}}(\mathrm{Si}_{N-1}\mathrm{C}^+_N) + \mu_{\mathrm{Si}} + \mu_e - E_{\mathrm{total}}(\mathrm{Si}_N\mathrm{C}_N). \tag{14}$$

The chemical potential of the electrons μ_e is the Fermi level in the host material. As the chemical potential of the element was bounded by thermodynamical

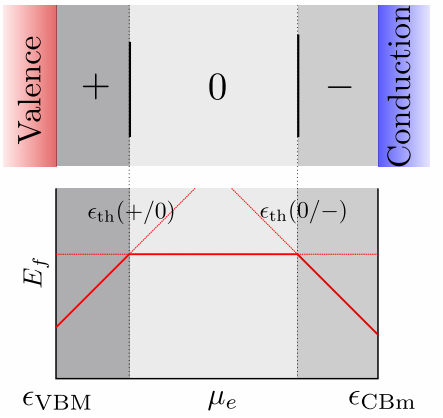

Figure 2: The two usual representations of the charged defect stability range as a function of the Fermi level μ_e. Upper panel shows the thermodynamically most stable charge state and the Fermi level for which the stability changes. Lower panel plots the formation energy as a function of the Fermi level for the different charge states. The energy of the most stable charge state is plotted with a thick red line. This representation provides more date than the other one.

considerations, the chemical potential μ_e is bounded to the band gap region of the host in the case of a non-degenerate semiconductor. If the zero of the Fermi levels is set at the valence band maximum $\epsilon_{\rm VBM}$, then the Fermi level is to vary between 0 and E_g, the band gap energy. As a consequence, the formation energy of a charged defect is a function of the Fermi level and it is generally plotted as a straight line with the slope determined by the number of electrons added or removed. Note that in principle, the Fermi level of a material could be calculated: it is not an external parameter. However this calculation would require to have the database of all the possible defects for all the possible charge states with the corresponding formation energies and also the concentration of the impurities with their charge state. Then for sake of comparison, the formation energies are preferably given in the literature as a function of the Fermi level.

The general expression of the formation energy of a generic defect X in charge state q reads

$$E_f(\mathrm{X}^q) = E_{\rm total}(\mathrm{X}^q) - E_{\rm total}(\mathrm{bulk}) - \sum_i n_i\mu_i + q(\epsilon_{\rm VBM} + \mu_e), \qquad (15)$$

where n_i counts the number of element i that were inserted $n_i > 0$ or extracted $n_i < 0$ to form the defect. The charge q counts the number of electrons that were extracted to form the defect.

With this definition, we are now able to understand all the peculiarities

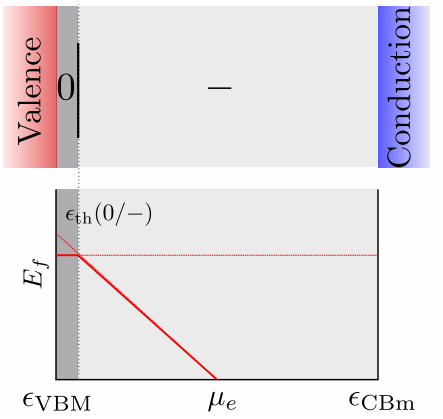

Figure 3: The two usual representations of the charged defect stability range as a function of the Fermi level μ_e exemplified with a good acceptor.

of charged defects in semiconductors and insulators. By plotting the different charge state formation energies in the same graphs, as in the lower panel of Figure 2, crossing points between formation energy lines may occur. The meaning of these intersections is clear: it accounts for a change of the thermodynamical most stable charge state. The thermodynamical charge transition level $\epsilon_{\rm th}(q+1/q)$ is the Fermi level value for which the formation energies of charge states q and $q+1$ match. The equality

$$E_f(\mathrm{X}^q) = E_f(\mathrm{X}^{q+1}) \tag{16}$$

implies the definition

$$\epsilon_{\rm th}(q+1/q) = E_{\rm total}(\mathrm{X}^q) - E_{\rm total}(\mathrm{X}^{q+1}) - \epsilon_{\rm VBM}. \tag{17}$$

When one focuses on the electrical properties induced by the charged defect, another handy representation exists as shown in the upper panel of Figure 2. This representation highlights the most stable charge state as a function of the Fermi level. It contains less information compared to the formation energy plot in the lower panel, however for many applications this is already sufficient. For instance, it allows one to appraise which defects will contribute to doping with electrons or holes.

2.2 Acceptors or donors? Shallow or deep?

As the defects in semiconductors and insulators may have a net charge, the overall charge compensation in a macroscopic sample is ensured either by other charged defects or by free charge carriers, such as electrons in the conduction band or holes in the valence band. When a charge transition level from neutral

to negative exists below the conduction band, then the defect is said to be an acceptor or to have an acceptor level. Since it can bear a negative charge, positive holes could be induced to ensure charge compensation. The defect has then some affinity to "accept" additional valence electrons. When a charge transition level from neutral to positive exists above the valence band, then the defect is said to be a donor or to have a donor level. As it may have a positive charge, conduction electrons may be "donated" by charge compensation. For instance, the defect represented in Figure 2 is both an acceptor and a donor.

In the semiconductors based electronics and in photovoltaic cells, donor and acceptor impurities are instrumental to engineer the transport properties of a semiconductor. To efficiently introduce holes in the valence band, an acceptor should have its charge transition level $\epsilon_{\rm th}(0/-)$ as close as possible from the valence band maximum. Figure 3 shows the formation energies and charge transition levels of such a good acceptor defect. If this accepting defect is predominant, the Fermi level is to be pinned in between the valence band maximum and the defect level. As a rule of thumbs, the charge transition level should be not farther than a few k_BT from the valence band maximum to have a significant probability being charged -1 and therefore to induce conducting holes. Obtaining a effective donor is exactly the reverse situation: the neutral to positive charge transition level $\epsilon_{\rm th}(+/0)$ should lie within a few k_BT to the conduction band minimum. At room temperature, the electrically active defects should be located within, say, 100 meV from the band edges. These active defects are called *shallow* defects; the other ones are categorized as *deep*. However, to address the possibility of carrier doping in a semiconductor, both shallow and deep defects should be considered. Indeed, the shallow defects are the desired one, but the deep defects can behave as charged compensators and thus hinder the doping capabilities.

The depth of the defect level inside the band gap is also in general a measure of the localization of the defect wavefunctions. The quantum electronic states introduced by a shallow defect can easily hybridize with the bulk electronic states. The corresponding wavefunction is in general very delocalized. A simple hydrogenoid model gives an effective radius a of (Grosso and Pastori Paravicini, 2000):

$$a = a_0 \frac{\varepsilon_\infty}{m^*}. \tag{18}$$

For instance, in silicon, the effective is around $m^* = 0.3$ for holes and the dielectric constant is around $\epsilon_\infty = 12$, and finally the effective hydrogenoid wavefunction radius is $a = 18$ Å. Of course, this is a very rough estimate which does not depend on the defect introduced, but it already shows that the numerical simulation of shallow defect will necessitate large systems with many atoms: the defect wavefunction needs to fit inside the supercell. A recent work calculated 64,000 atom supercell to obtain reliable shallow defect properties (Zhang et al., 2013). At the opposite, the wavefunctions of the states induced by a deep defect are generally well localized in the vicinity of the defect as exemplified in Figure 4 for a vacancy in silicon. The silicon vacancy in the neutral charge states experiences a Jahn-Teller distortion, which means that in the most en-

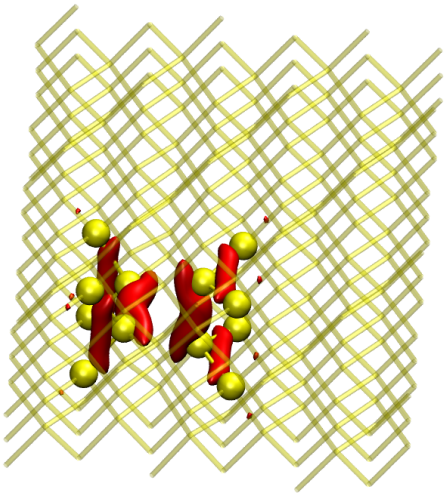

Figure 4: Isosurface of the defect wavefunction in the band gap of silicon (red surface), when a vacancy is introduced. The first and second nearest neighbors are drawn, whereas the other atoms of the 216 atom supercell are transparent.

ergetically favorable configuration, the neighboring atoms form new bonds (the symmetry is then lowered). The isodensity surface indeed shows an accumulation of electrons in between the atoms which had dangling bonds owing to the vacancy.

2.3 Migration, diffusion, clustering

Besides the tuning of the electronic transport properties in semiconductors, the point defects also play the prominent role in matter transport in crystals. The motion of impurities, as well as the motion of the crystal atoms, are mostly mediated by the defects. Indeed, the direct exchange between two atoms in neighboring crystalline sites would require to overcome exceedingly high energy barriers. The jumps of atoms from a crystalline site to another site is much more favorable when the final site is empty (vacancy mediated diffusion). Alternatively, a lattice atom may also swap with a nearby self-interstitial atom (self-interstitial mediated diffusion). Furthermore, one can intuitively imagine that the self-interstitial atoms do not easily fit in the host matrix: they are not trapped in deep energy wells and therefore are quite mobile in the crystal.

For these reasons, it is important to know the defect properties to characterize the matter transport in crystals. And conversely, the measure of the diffusion coefficient in a crystal gives insight about the defect properties. For a quantitative understanding of self-diffusion, it is time to introduce the driving energies.

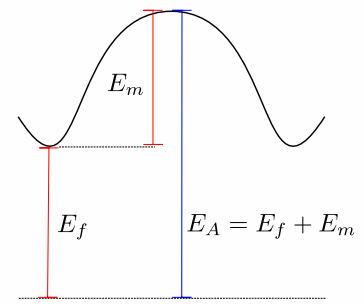

Figure 5: Schematic representation of the energy landscape in the diffusion process. The driving energies are shown: the formation energy in the stable configuration E_f, the migration energy of the defect E_m, and the activation energy of diffusion E_A. Note that the activation energy can hence be defined as the formation energy of the saddle point configuration.

The diffusion is mediated either through vacancies or through self-interstitials. As for the conductance in parallel electric circuits, the total diffusion coefficient D is then the sum of the two contributions:

$$D = D_V + D_I. \tag{19}$$

Each component of the diffusion is proportional to the number of diffusive defects, i.e. the concentration of vacancies or interstitials, and proportional to the probability to hop to the next site, measured by the energy barrier $E_m(\mathrm{X})$, as shown in Figure 5. Then the total diffusion can be expressed as

$$D(T) \propto f_V C_V e^{-E_m(V)/k_B T} + f_I C_I e^{-E_m(I)/k_B T}, \tag{20}$$

where the correlation factors f_V and f_I have been introduced. The correlation factors measure the probability that a diffusive specie comes back to in its previous position when it reaches the saddle point. In general, the correlation factors do not deviate much from 0.5. Since the concentration of a defect (at thermodynamical equilibrium) is measured by its formation energy, the diffusion coefficient finally reads

$$D(T) \propto f_V e^{-[E_f(V)+E_m(V)]/k_B T} + f_I e^{-[E_f(I)+E_m(I)]/k_B T}, \tag{21}$$

This leads to the definition of an diffusion activation energy for defect X:

$$E_A(\mathrm{X}) = E_f(\mathrm{X}) + E_m(\mathrm{X}). \tag{22}$$

These three energies are schematically represented in Figure 5. Note that the entropic effects have been disregarded in the previous equations. In principle, the free energies should be considered instead of the energies.

The measurement of the diffusion coefficient as a function of T in Arrhenius plots ($\ln D(1/T)$) gives access to the activation energy, not to the individual values of E_f and E_m. In the experiment the range of accessible temperature is limited unfortunately. Too high temperatures would melt the sample, too low temperatures would hinder so much the diffusion that no significant signal would be visible even waiting for several months. Furthermore, if the activation energies of vacancies and interstitials are not well separated, the Arrhenius plot would be very similar to a single straight line and only one total activation energy would be obtained. Owing to these limitations, the computer simulations of diffusion appear as a tool of choice to complement the experiment.

As the defects are relatively mobile inside the crystalline matrix, the meeting of two defects is a possible event. In this case, the binding of two defects, says X and Y, that forms a complex XY, can be written as a chemical reaction

$$\mathrm{X} + \mathrm{Y} \rightleftharpoons \mathrm{XY}. \tag{23}$$

The binding energy $E_b(\mathrm{XY})$ is hence defined as

$$E_b(\mathrm{XY}) = E_f(\mathrm{XY}) - E_f(\mathrm{X}) - E_f(\mathrm{Y}). \tag{24}$$

Note that a negative binding energy is not sufficient to ensure the existence of the complex XY. As in chemistry, the chemical reaction is driven by the law of mass-action:

$$C_{\mathrm{XY}} = C_{\mathrm{X}} C_{\mathrm{Y}} e^{-E_b(\mathrm{XY})/k_B T}, \tag{25}$$

where the influence of the concentration of the reactants C_{X} and C_{Y} is made obvious.

When the two defects are a vacancy and a self-interstitial atom, the clustering of the two defects is a recombination which heals the crystalline structure. In this case, the binding energy of the recombination is simply the opposite of the formation energy of the so-called Frenkel pair.

3 Experimental characterization of defects

This section is a quick overview of a few selected experimental techniques that can help characterizing a defect in a solid. For more details, the reader is referred to the books Lannoo and Bourgoin (1981); Bourgoin and Lannoo (1983).

As detailed in the previous section, the measurement of the diffusion coefficient in a crystal gives access to the diffusion activation energy. The formation and migration have just been described earlier. In the experimental setup, mono-crystalline samples are grown with a controlled content of tracers: for instance for silicon ^{30}Si instead of the most common isotope ^{28}Si. The samples are then placed in an oven at a fixed temperature for a long period of time. Finally, secondary ion mass spectroscopy (SIMS) permits one to figure out the concentration profile of the isotopes after diffusion has occurred.

The photoluminescence technique is one the most used experimental technique to characterize the defect in semiconductors. It is based on the fluorescence phenomenon: a photo-excited system re-emits a photon at a different

energy. This photon is collected for analysis. By shining a laser on a sample, the photons are absorbed and create electron-hole pairs (excitons) in the crystal. If they were no defects in the sample, then the exciton would finally recombine and emit a photon at the free-exciton energy (labeled FX). However in the presence of defects, numerous process can happen. The exciton can bind to an acceptor or a donor and then re-emit light at a lower energy (the binding energy with the defect). These processes are labeled A^0X and D^0X. Light can be emitted from an ionized donor and an ionized acceptor (labeled DAP). The exciton can also decay into a free electron and a free hole, that may later on ionize the defects:

$$D^0 + h^+ \quad \rightarrow \quad D^+ + h\nu \tag{26}$$

$$A^0 + e^- \quad \rightarrow \quad A^- + h\nu. \tag{27}$$

These processes are labeled respectively (h,D^0) and (e,A^0). In addition to this possible phenomena, the coupling with optical phonons of the lattice adds some extra peaks. Finally, the photoluminescence spectra are rather messy, but the dependence of the peaks upon temperature, defect concentration, strain etc. can help sorting out the nature of each peak.

The infrared and the Raman spectroscopy is the inelastic diffusion of light involving loss through the phonons of the lattice. Since the defects break the translational invariance of the crystal, it is clear that additional modes will be induced by the defects. Detecting these vibrational modes help characterizing the nature and the concentration of defects.

An other prominent experimental technique involves the magnetic moment carried by a single defect. In Electron Paramagnetic Resonance (EPR), the possible unpaired spin introduced by the defect are measured by lifting the degeneracy between spin up and spin down thanks to a static magnetic field (Zeeman effect). Then transitions from the spin up and spin down states are produced with a second oscillatory magnetic field. This technique is very powerful. Since most pristine bulk materials consists only of spin-paired electrons, EPR is only sensitive to the defects. However it is limited to defects have a non-zero magnetic moment $S^2 \neq 0$. EPR technique furthermore gives information about the point group symmetry of the defect and its environment, as the magnetic moment are slightly modified by the spin states of the neighboring nuclei. EPR allows one to measure the total magnetic moment and then to have a measure of the concentration of the defect, for instance along with a thermal annealing.

There are several other important techniques that I will only mention here: Deep Level Transient Spectroscopy (DLTS), which plays with the occupation of the defect states with the temperature; resistivity recovery measurements, which measures the resistivity during a thermal annealing; Hall current measurement, which characterizes the nature of the conduction in a semiconductor (p-type or n-type).

All these experimental techniques provide us with some knowledge about the defects involved. However many of them need theoretical model for further interpretation. In the recent years, *ab initio* calculations have become the tool

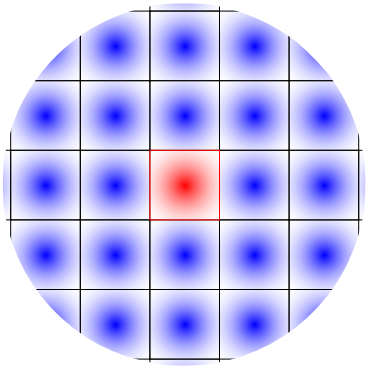

Figure 6: Supercell technique exemplified in two-dimensions: the red cell containing the defect of interest is replicated (blue cells). The blurs are simplified (artistic?) representation of the defect wavefunction.

of choice to complement the experimental findings.

4 Challenges for the *ab initio* calculation of defects

As stated above, the *ab initio* calculation are highly relevant for defects, since they are able to provide in a consistent manner the atomistic and electronic description of point defects. This piece of information is crucial for the interpretation of experiment and for the prediction of material properties. The coming-of-age of *ab initio* calculations in condensed matter can be witnessed nowadays: Calculations are no longer used to back-up the experimental findings, they are also used to predict the properties of never synthesized materials (Van Noorden, 2014). Unfortunately, whereas for many ground-state properties the *ab initio* calculations can be considered as reliable enough for materials' prediction, the situation is more contrasted when studying point defects. There are two main issues that prevents the defect calculations from being considered as predictive: the image interactions and the band gap problem.

4.1 Supercell technique and the spurious image interactions

The *ab initio* calculations in condensed matter are all based on periodic boundary conditions. There is no alternative nowadays. Of course, the real samples are finite; however calculating a real size sample is far beyond the computational capabilities. Try to imagine the calculation of a crystal of a length of 100 atoms: this would require a 1 million atom calculations, with no certainty that the bulk effects will prevail over the surface effects. Indeed, even with a cube of 1 million atoms, 6 % of the atoms are located on the surface! On the contrary, the periodic boundary conditions are perfectly suited to evaluate infinite crystal properties just calculating the unit cell. This method matches our needs for pristine bulk properties. But for defects, the unit cell description is not suitable anymore. In order to mimic an isolated defect, one may then introduce the defect in a supercell consisting of several unit cells. The larger the number of unit cells in the supercell, the more diluted the defect concentration.

The experimental concentration of defects is generally very low. They are a few examples in which the dopant concentration is of the order of a few percents, as for transparent-conductive oxide ZnO doped with 3 % of Al. This is case the defects cannot be considered as isolated and the semiconductors becomes degenerate: it behaves as a metal (Grosso and Pastori Paravicini, 2000). However in general, the concentration is way lower: at thermodynamical equilibrium it is given by $\exp(-E_f/k_BT)$. In the calculations instead, the concentration of defects is driven by the number of atoms in the supercell. Unfortunately, nowadays it is not possible to consider supercell sizes with the correct defect concentration. The typical *ab initio* can consider 1500-3000 electrons, which translates into 500 atoms for silicon, but only 100 atoms for ZnO.

The too-small supercell size has several uncomfortable consequences as depicted in Figure 6. If the defect is too shallow, it may happen that the defect wavefunction simply does not fit inside the supercell. These defects cannot be addressed as of today, unless the supercell is increased facing the computational scaling problem. Even for deeper defect, the defect wavefunction may have a significant overlap with the image defect wavefunctions. The overlap can be appreciated from the width of the defect band. An isolated defect should have be described dispersiveless states in the Brillouin zone. In practical applications, this is rarely true. See for instance Figure 6.2 for a carbon vacancy in silicon carbide in a cubic 216 atom supercell. Finally, the supercell technique poses deep problem when combining it with long-range interactions, such as the Coulomb interaction between the charges. Due to the long-range nature of the Coulomb interaction, the mathematical treatment in the periodic boundary conditions is very delicate. Many subtle points deserve a special care. In particular, charge interactions between the defect and its images through periodic boundary conditions add a spurious Coulomb energy, which decays very slow. Hence, correcting schemes have been designed, however it is not yet clear which correcting framework is the most reliable.

This point will be extensively discussed in Part II of the memoir.

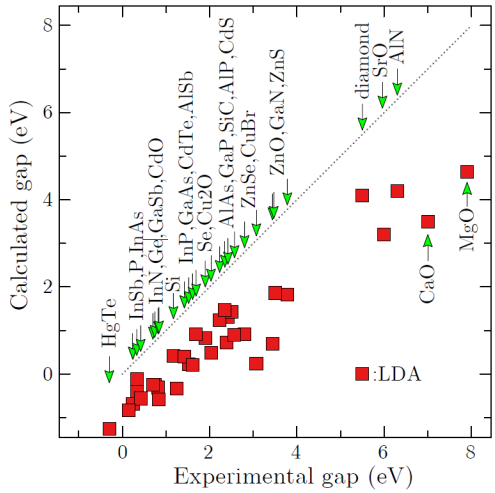

Figure 7: LDA band gaps for a large family of semiconductors and insulators compared to the experimental value. A perfect agreement would place all the points along the diagonal. Figure is adapted from van Schilfgaarde et al. (2006).

4.2 Band gap problem and electronic structure of point defects

Point defects are also problematic for the electronic structure description itself. Intuitively, one can imagine that these systems will push the theoretical approaches close to their limit. The most used practical framework in the context of electronic structure is Density-Functional Theory (DFT) (Hohenberg and Kohn, 1964; Kohn and Sham, 1965; Parr and Yang, 1989). This theory is in principle exact, however the practical usage heavily relies on approximation for the unknown exchange-correlation (xc) term. The reliable approximations have been obtained and validated either in the chemistry context, with finite isolated atoms and molecules having localized wavefunctions, or in the physics context, with periodic crystal unit cells having delocalized wavefunctions. Point defects in crystalline structure will introduce both situations in the same simulation supercell. Localized defect states will co-exist with the delocalized Bloch states of the crystalline matrix. In such a complex situation, the conception flaws of the practical xc approximations may appear to open light.

A prominent problem of DFT approximations for solid is the systematic error for the band gaps of semiconductors and insulators. As shown in Figure 7, the Local-Density Approximation (LDA) severely underestimates the band gaps. The same conclusion holds for all the semi-local approximations (the Generalized Gradient Approximation family). Whereas many properties of solids are properly obtained even within an approximation that does catch the correct

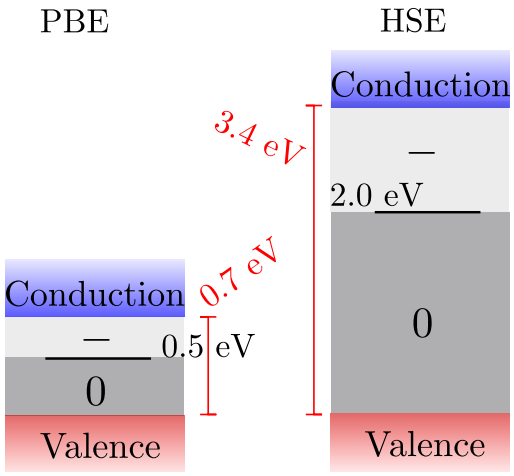

Figure 8: Nitrogen acceptor charge transition level in ZnO as calculated from PBE and from HSE06.

band gap, the band gap error is particularly relevant for point defects. Indeed, the point defect in semiconductors and insulators introduce electronic state inside the band gap, as explained above. The correct positioning of these defect levels is crucial for the charge transition levels, for the formation energy etc. Remember that the nature of the defect wavefunctions strongly depends on the depth of the defect electronic level inside the band gap.

Having a too small band gap will spuriously produce shallow defects. Let me exemplify this problem with the quest of p-type ZnO. n-type zinc oxide samples are easily obtained and are already being used as transparent conductive oxide in thin film solar cells. It would be highly desirable to obtain also the p-type ZnO to produce cheap p-n junctions for light-emitting diodes for instance. The band gap of ZnO (3.4 eV) matches the technological needs for this application. Unfortunately p-type ZnO has remained elusive and controversial. Nitrogen doping (in substitution for oxygen) is the most intuitive candidate to create a shallow acceptor, after a quick glance at the periodic table. When calculating the nitrogen acceptor level from PBE (Perdew et al., 1996), one of the GGA flavors, an acceptor level indeed exists and is located 0.5 eV above the valence band maximum. This acceptor level is not, strictly speaking, shallow enough to induce hole doping, however it seems to point towards the right direction. But PBE failure with the band gap of ZnO is even more pronounced as the usual underestimation (0.7 eV instead of 3.4 eV). Then the acceptor level lies only 0.2 eV below the PBE conduction band! Based on the sole PBE results, it is then impossible to conclude about the depth of the charge transition level and thus impossible to conclude about the potentialities of N doping in PBE.

In the recent years, new approximations for the xc energy have emerged.

The so-called hybrid functionals (Becke, 1993) come from the quantum chemistry community. They use a mixture of Hartree-Fock exact exchange and of semi-local exchange from the usual DFT approximations. Whereas these mixtures have been designed and validated for small molecules, their results for crystalline solids are also improving a lot over the semi-local approximations. In particular, the most widely used hybrid functionals in the context of solids are PBE0 (Adamo and Barone, 1999) and HSE06 (Heyd et al., 2006). Figure 8 shows the nitrogen acceptor level as obtained from HSE06, with a further fitting of the exact exchange content (38 %) (Petretto and Bruneval, 2014). Now that the band gap of ZnO is realistic, the depth of the acceptor level is obtained close to the middle of the band gap. There is as a consequence no hope that the nitrogen doping could induce the p-type conduction in ZnO by substitution of a lattice oxygen atom. This example is a very convincing reason for which the band gap must be correctly calculated to perform predictive electronic structure calculations.

Though attractive, the hybrid functionals have still some deficiencies. First, there is a problem of principle: the mixture of exact exchange and DFT xc functionals is based on empirical grounds. The precise recipe is usually designed to minimize the error with respect to a database of experimental values for molecules. This is procedure is no longer truly *ab initio*. Furthermore, the fitting on the specific subset of molecules can introduce biases. There is no guarantee that the systems, such as crystals, which are far from the ones included in the database are reliably described. For instance, most of hybrid functionals have underestimation problems with the wide band gap semiconductors and insulators. Second, the hybrid functionals require the evaluation of the exact exchange contribution, which is non local in space. This evaluation is much cumbersome in periodic systems. Then the supercell sizes must be decreased when using hybrid functionals, and the finite-size effects discussed in the previous subsection are increased again. Third, there was a transient problem when I started with point defect calculations: the availability of hybrid functional codes with periodic boundary conditions was scarce at that time. Only VASP (Kresse and Furthmüller, 1996; Paier et al., 2006) had the hybrid funtionals fully operational back in 2007. Nowadays the situation has improved much, since Quantum Espresso (Giannozzi et al., 2009) and Abinit (Gonze et al., 2009) have implemented them, however only for norm-conserving pseudopotentials.

For all these reasons, I address in Part III of the present memoir, the problem of going beyond DFT for the electronic structure of point defects. The method I analyze arises from a different theoretical framework: Many-Body Perturbation Theory or Green's function theory (Fetter and Walecka, 1971; Mahan, 2000). In this field, the most prominent approximation for solids is certainly the so-called GW approximation (Hedin, 1965; Aryasetiawan and Gunnarsson, 1998). Part III is fully devoted to application of the GW approximation to the point defect electronic structure. The Random-Phase Approximation, which is the total energy expression corresponding to GW, is also examined. In contrast with hybrid functionals, GW approximation is fully *ab initio*, it is known to perform extremely well for many solids, from narrow band gap semiconductors

to insulators and, last but not least, I did have a working code, Abinit, that implements these calculations.

Part II

Supercell induced artifacts in point defect calculations

Chapter 1

Introduction to the corrections to the formation energy in supercells

The formation energy of a charged defect was introduced as the central quantity to measure the stability of a defect. Calculating it accurately is thus of the utmost interest. As long as the calculations are treated on the paper, there is no conceptual problem with the procedure. However when we turn to the practical calculations with the use of supercells, then the extraction of the formation energy value becomes a real challenge. We would like to extract data for isolated defects, but the supercells technique provides us with a lattice of interacting defects.

The combination of periodic supercells and of charged systems induces artifacts which prevents the brute force supercell convergence. First of all, it is easily understood that the defect can interact with its periodic images and, as the Coulomb interaction $1/r$ is long ranged, the convergence with supercell size is not surprisingly extremely slow.

However this is not the only problem induced by the use of charged supercell. In periodic systems, the Coulomb energy is only finite for charge neutral systems. Simulating charged defects requires then the addition of a compensating background density. Even with this compensating background density, the electrostatic potentials in a periodic cell remain defined up to a constant. In practice, the average electrostatic potential is generally set to zero, however we should keep in mind that this is a conventional choice. The undefined value of the average electrostatic potentials is already encountered in basic solid-state physics when introducing the Madelung constant for ionic solids. The Madelung constant is defined only within its summation scheme (spherical summation, cubic summation, etc...), as the Coulomb interactions in $1/r$ induce conditionally convergent sums.

The slow rate of convergence induced by the aforementioned issues is exem-

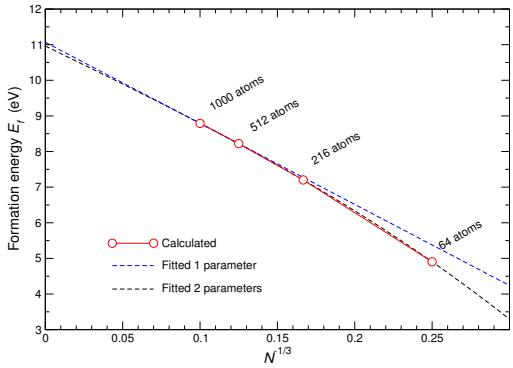

Figure 1.1: Convergence of the formation energy of a silicon interstitial in 3C-SiC, Si^{4+}_{TC}, as a function of N, the number of atoms in the supercell. Red line is the formation energy as obtained from Quantum-Espresso code within LDA (without atomic relaxation). Blue dashed line is a fit to the calculated value with function $\gamma_1 N^{-1/3}$ Black dashed line is a fit with function $\gamma_1 N^{-1/3} + \gamma_3 N^{-1}$. Thermodynamical conditions are Si-rich and μ_e set to the valence band maximum.

plified in Figure 1.1 for a quite heavily charged defects (4+). Even the largest supercell I could afford (1000 atoms) is 2 eV off the converged value. Reversely, one can estimate that the supercell size to obtain a realistic formation energy within 0.3 eV (which is not very accurate anyway) would require a 125 000 atom supercell! It would take years before the computational power would reach this level. There is therefore a stringent need to cure the sources of error in the supercell approach, since the brute force approach is doomed to fail.

With these two artifacts (image charge interactions and undefinition of the absolute electrostatic potential), the formation energy from Eq. (15) has to be complemented with two correcting terms $\Delta E_{e.s.}$ and ΔV:

$$E_f(X^q) = E_{tot}(X^q) - E_{tot}(\text{bulk}) - \sum_i n_i \mu_i + \Delta E_{e.s.} + q(\epsilon_{\text{VBM}} + \mu_e + \Delta V). \quad (1.1)$$

1 Electrostatic correction $\Delta E_{e.s.}$

The first correcting term $\Delta E_{e.s.}$ affects directly the energy. It is meant to correct the spurious electrostatic interaction between the charge defects and its periodic images. This spurious interaction is present in the total energy of the charged supercell $E_{tot}(X^q)$ (the first term in Eq. (1.1)).

The historically first expression for this term was given by Leslie and Gillan (1985), assuming the charged defect induces a point charge distribution. In the supercell calculation, the density induced by the defect is of course "background

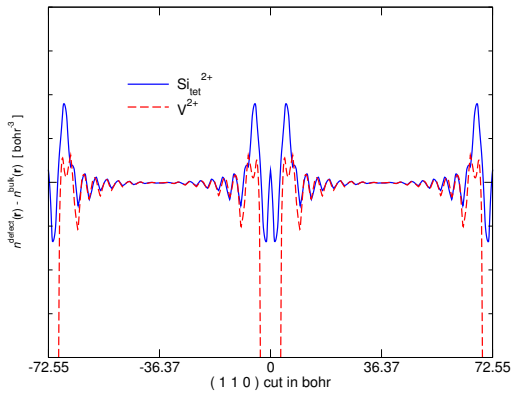

Figure 1.2: Cut of the electronic density difference $n^{\mathrm{defect}}(\mathbf{r}) - n^{\mathrm{bulk}}(\mathbf{r})$ along the (110) direction passing through the bond centers for 1000 atoms cubic supercell for two prototypical defects in silicon, an interstitial $\mathrm{Si}^{2+}_{\mathrm{Tet}}$ and a vacancy $\mathrm{V}^{2+}_{\mathrm{Si}}$. Note that the (110) direction was intentionally selected since this is the most favorable direction for density fluctuations in the diamond structure.

compensated":

$$n(\mathbf{r}) = q\left[\sum_{\mathbf{R}} \delta(\mathbf{r}-\mathbf{R}) - \frac{1}{\Omega}\right], \tag{1.2}$$

where Ω is the volume of the supercell and $\mathbf{R}$ are the lattice vector. With this notation, the periodicity of the defect density in the supercell calculations has been emphasized. Then the difference between the isolated point charge Coulomb energy and the point charge lattice Coulomb energy is given by the Madelung energy,

$$\Delta E_{e.s.} = \frac{\alpha_M q^2}{2\varepsilon\Omega^{1/3}}, \tag{1.3}$$

with α_M the Madelung constant of the lattice. The only refinement compared to the usual Madelung expression is the introduction of the dielectric constant ε that screens the bare Coulomb interaction in condensed matter. This very simple expression, also named monopole correction, yields the leading effect inducing the slow convergence with respect to the supercell size ($\Omega^{-1/3}$ behavior). This correction explains what we observed for the charged defect provided in Figure 1.1. To further prove that the monopole correction captures most of the error, let us evaluate the dielectric constant from the fit in Figure 1.1. Using Eq. (1.3), we obtain a value of dielectric constant of 7.13. This value should be compared not to the experiment, but rather to the calculated LDA clamped-ion dielectric constant, as the atoms have not been allowed to relax in the calculations of Figure 1.1. According to Karch et al. (1996), this dielectric constant $\varepsilon_\infty^{\mathrm{LDA}}$ is 7.02, which is in very close agreement with the previous evaluation from the monopole formula (~ 1.5 %).

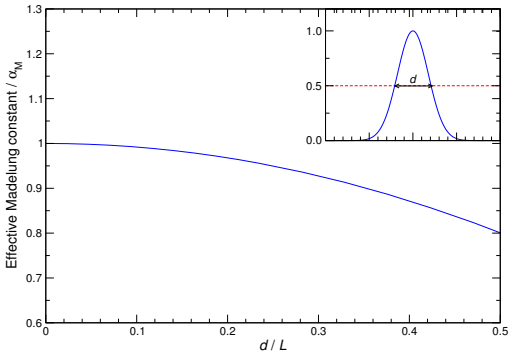

Figure 1.3: Reduction of the Madelung constant induced by the finite extension of the charge density in units of the Madelung constant. The results are obtained using a Gaussian charge distribution in a cubic lattice with edge L. The Gaussian distribution with a width at half maximum d is shown in the inset.

The defect induced charge density significantly departs from the point charge distribution, as shown for two defects of silicon in Figure 1.2. There is a zone around the defect, where the electronic density is vastly reorganized and farther there are some oscillating tails, the so-called Friedel oscillations, which propagates in the polarizable density. However, why does the point charge model perform well in capturing the global trend? This is mainly driven by the defect extension d compared to the supercell dimension L. If the extension of the defect is small compared to the supercell, then the point charge modeling is to work properly. If it is not so, then the point charge model will be insufficient. A strong point supporting the monopole correction is the observation that for any defect extension d there always exists a supercell dimension large enough so that the monopole correction will be eventually justified. However, the shallow defects may have a very large extension that no supercell affordable in DFT would contain.

In principle, the point charge model could be complemented with higher terms in the multipole expansion. This is the rationale behind the corrections of Makov and Payne (1995). However, these terms are difficult to calculate for solids, where one has to consider the dielectric screening. Furthermore, the monopole term already captures the majority of the corrections, as shown in Figure 1.3. In the Figure, I have evaluated numerically the complete correction induced by the Gaussian charge distribution displayed in the inset. This calculation is a simplified version of the Ewald technique (Ewald, 1921; Martin, 2004), in which the short-range terms have been dropped. In other words, the curve in the Figure is the difference between the Coulomb energy of a single isolated Gaussian with a width at half maximum d and a cubic lattice of Gaussian distributions with a compensating background. The reduction of the

Madelung constant induced by the finite extension of the distribution is only marginal: consider a Gaussian whose width is 25 % the supercell dimension, then the Madelung constant is reduced by only 5 %! The inclusion of higher multipoles can be then considered as a correction to the correction.

However, the Madelung correction is not sufficient to correct the calculation in practice. In realistic cases, often due to the limited computer resources or due to the complexity of the calculations (think of hybrid functionals, think of crystal containing f electron elements) the supercell one can afford are not in the limit where the Madelung correction is sufficient. Having a new glance at Figure 1.1, the calculations we perform in practical cases are most often in the regime where the two fits do differ. Remember that Figure 1.1 was obtained within LDA for the simple crystal of SiC without relaxing the atomic positions. That is why the second correction type, namely the potential alignment, could be also significant.

2 Potential alignment correction ΔV

The potential alignment, labeled ΔV in the corrected formation energy expression (1.1), arises from the necessity to define an electron reservoir in order to balance the formation equation. The chemical potential of the electron reservoir μ_e, or in other words the Fermi level, is easily defined for the bulk system. As written in Part I, this study focuses on non-degenerate semiconductors for which the Fermi level is bound in the range $[\epsilon_{\mathrm{VBM}}, \epsilon_{\mathrm{CBm}}]$.

Unfortunately, when performing the total energy calculation with the charged supercell $E_{tot}(X^q)$ term in Eq. (1.1), the absolute position of the eigenvalues is lost. Indeed, in the periodic solids the electrostatic potentials are defined only up to a constant (and also possibly to a surface dipole term). It is therefore impossible in theory to compare straightforwardly the eigenvalue position in the defective charged supercell with the eigenvalue position in the pristine neutral bulk system. That is why the eigenvalues of the bulk system have to be shifted with a constant ΔV in the correction formation energy.

The origin of this shift in the eigenvalues is two-fold: in the charged defective cell, both the number of atoms has been changed and the charge has been changed. First, when inserting a defect in the supercell (removing/adding an atom), the electrostatic potential induced by the ions is modified. In the DFT language, the external potential v_{ext} has been modified. Since this electron-ion electrostatic potential is defined up to a constant, a potential alignment is necessary. Second, when considering a charged defect, the more or less localized charge around the defect will induce a modification in the electron-electron electrostatic potential. Away from the charged defect, the added electrostatic potential should behave as $1/\varepsilon r$. However, with the supercell technique, this behavior is not allowed by the periodic boundary conditions. The charged defect induced potential will be automatically altered by the periodicity and once again, the induced potential will be defined up to a constant.

Disentangling the two origins (atomic and electronic) is not straightforward.

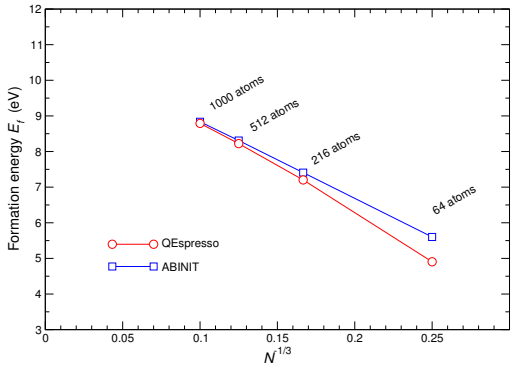

Figure 1.4: Convergence of the formation energy of a silicon interstitial in 3C-SiC, Si_{TC}^{4+}, as a function of N, the number of atoms in the supercell. Red circles is the unrelaxed formation energy as obtained from Quantum-Espresso, and blue squares from Abinit. Both calculations use the same norm-conserving pseudopotential, the same k-point grid, the same energy cutoff, and the same Fast Fourier Transform grids. Thermodynamical conditions are Si-rich and μ_e set to the valence band maximum.

The potential alignment is sometimes believed to be small in practice. However it is not always true. Re-considering the example of the silicon interstitial with charge $4+$ in SiC, Figure 1.4 shows the uncorrected formation energy as obtained from two plane-wave codes (Abinit and QEspresso) with the very same inputs. The two code yield the same self-consistent charge density, so the electron-electron potential modification should be identical. However these two codes use different conventions for the ion-electron potentials. The difference for the formation energy can be as large as 0.7 eV for the 64 atom supercell! Fortunately the difference decreases with the supercell size, as fast as $1/N$.

3 Motivation for the following sections

Still nowadays the two corrections $\Delta E_{e.s.}$ and ΔV are not completely understood. The following sections will summarize the research lines I have been following in order to improve the understanding and the quality of these two corrections for charged systems. In particular, it should be noted that the meaning of the potential alignment ΔV is far less understood than the spurious charge-charge interactions. Furthermore, as the magnitude of ΔV is in general smaller, it is often difficult to appreciate it from the numerical results. The following section are mainly devoted to the definition of ΔV, both for the electron-ion part and the electron-electron part.

Chapter 2

Ionic potential alignment in projector augmented wave method and in norm-conserving pseudopotentials

This section summarizes, further explains, and exemplifies the article by Bruneval et al. (2014) printed in Appendix C.

1 Conventions in periodic codes vary

The starting point of this study was the apparent discrepancy between plane-wave codes I mentioned in the previous chapter in Figure 1.4. Since the potential alignment is a subtle correction that consists of two different origins, it appeared valuable to us to investigate on the origin of the difference between codes. The proper potential alignment definition should of course reconcile the different codes. However it is nevertheless interesting to understand what tiny differences in the implementation would induce those sizable changes (0.7 eV for the example in Figure 1.4) and how the potential alignment procedure would fix the errors.

As the aforementioned calculations using Abinit or QEspresso are rigorously identical, the difference could only arise from a difference in conventions for the electrostatic potential. It is usually said that all the codes use the same convention for the ill-defined electrostatic potential: the "zero average" convention. However, this statement is not true in general. For the electron-ion potential, codes use diverse conventions, which in turn produce diverse formation energies

when they are uncorrected. Due to the use of norm-conserving pseudopotentials, which can be though as created by a pseudo-charge, it is not surprising that different conventions exist.

However we observed the same difference within the projector-augmented wave (PAW) formalism (Blöchl, 1994; Kresse and Joubert, 1999). In PAW, the all-electron quantities can be re-construct from pseudo-quantities, indicated with a tilde sign. In particular, the valence electrons of the system experience the effect of the true nucleus potential Z/r, not a smooth pseudo-potential which is supposed to mimic the core plus nucleus potential. Of course, in the PAW framework, a working pseudopotential, labeled $v_H[\tilde{n}_{Zc}]$, is introduced. However, the pseudopotential is simply an intermediate quantity, chosen for numerical convenience. At the end of the calculation, the effect of the pseudopotential is compensated in spheres around each atoms and the true physical potential $v_H[n_{Zc}]$ is recovered. In the previous notations $v_H[n]$ stands for the electrostatic potential induced by the charge density n.

If all the codes were using the same zero average convention for the electrostatic potentials, the absolute position of the Kohn-Sham eigenvalues should be identical. We have tested this against published absolute eigenvalues within full-potential augmented plane wave (FLAPW) from Ishii et al. (2010). The results provided in Table 2.1 demonstrate that none of the tested codes (Abinit and QEspresso) implement the zero average convention. Note that VASP code (Kresse and Furthmüller, 1996) would yield the same result as Abinit, since Abinit implementation closely follows the article of Kresse and Joubert (1999).

2 Deriving the PAW formalism with a zero average potential

In the landmark paper of Kresse and Joubert (1999), the PAW formalism is derived in a manner that emphasizes the similarities with the usual pseudopotential approach. In contrast with pseudopotential, the PAW framework requires to evaluate the quantities partly on the plane-wave basis and partly on radial grids in spheres around each atoms. One constantly needs to map plane-wave description onto radial grid description. In the original paper of Kresse and Joubert (1999), the electronic and nuclear densities are not background compensated (see Eq. (1.2)). The background compensation is then manually imposed at the

Table 2.1: Absolute value of the LDA valence band maximum of diamond at lattice constant 3.564 Å from different calculations (eV).

code Approach	FLAPW[a]	Abinit PAW	QEspresso PAW	Abinit PAW Corr.	Abinit NC	QEspresso NC	Abinit NC Corr.
ϵ_{VBM}	13.39	12.25	13.08	13.40	11.06	12.95	13.25

[a]Ref. (Ishii et al., 2010)

end of the derivation by setting some potential averages to zero.

In our article (Bruneval et al., 2014), we re-derived the PAW equations taking into account the background compensated densities from the very beginning. The considered valence electron density is

$$n'(\mathbf{r}) = n(\mathbf{r}) - \frac{N}{\Omega} \tag{2.1}$$

where N is the number of valence electron per unit cell of volume Ω. The core plus nucleus density reads

$$n'_{Zc}(\mathbf{r}) = n_{Zc}(\mathbf{r}) - \frac{N_c - Z}{\Omega}, \tag{2.2}$$

where Z is the number of proton and N_c the number of core electrons in the unit cell. The total background compensated density is then

$$n'_T(\mathbf{r}) = n(\mathbf{r}) + n_{Zc}(\mathbf{r}) - \frac{N_c + N - Z}{\Omega}. \tag{2.3}$$

For charge neutral calculations, the number of electrons $N_c + N$ equals the number of protons Z and the background automatically disappears.

With the introduction of these densities, we realized that when mapping a density from plane-waves to the PAW augmentation sphere, a background contribution in the sphere had to be incorporated for the calculation of the Coulomb energy of the system:

$$E_{\text{PAW bg}} = \frac{Z - N_c - N}{\Omega} \int d\mathbf{r} v_H[n_T^1 - \tilde{n}_T^1](\mathbf{r}). \tag{2.4}$$

The total densities n_T^1 and $\tilde{n}_T^1$ have a superscript 1, which in the PAW convention indicates quantities that vanish outside the PAW augmentation spheres. The integration in the previous equation is hence limited to the spheres of the unit cell. This term was not obtained previously in the PAW framework to the best of my knowledge.

The detailed analysis of the new term $E_{\text{PAW bg}}$ is given in Bruneval et al. (2014). However let me insist on a few intriguing characteristics. The background energy gives a non-zero contribution to the energy only for charged systems where $N_c + N \neq Z$. The trace of the stress tensor is also modified for charged systems, since the term has a $1/\Omega$ dependence. However the most surprising feature (to me) is its non-zero contribution to the Kohn-Sham potential, irrespective to the charge of the system. Indeed, the Kohn-Sham Hamiltonian is obtained by derivation of the energy with respect to the electron density $n(\mathbf{r})$. The derivative of $E_{\text{PAW bg}}$ with respect to n is not null, mainly due to the number of valence electrons N as a prefactor. N is of course a (simple) functional of n

$$N = \int d\mathbf{r} n(\mathbf{r}). \tag{2.5}$$

In other words, this is not because a function is zero for a given x-coordinate that its derivative is zero too! When calculating the contributions to the potential

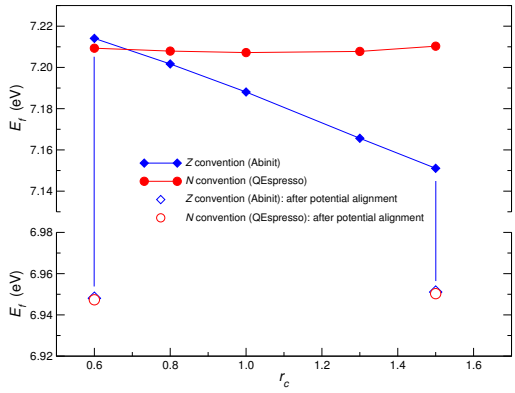

Figure 2.1: Formation energy of the unrelaxed vacancy in diamond with charge 2+, as obtained from LDA with Abinit code (blue diamond) or with QEspresso (red circles), with the same norm conserving pseudopotentials. The cutoff radius of the local component (d channel) of the pseudopotential has been varied from 0.6 bohr to 1.5 bohr. The uncorrected results are given with solid symbols, whereas the potential aligned results are given with open symbols.

induced by $E_{\mathrm{PAW\,bg}}$, we obtained the revised value for the diamond valence band maximum given in the column labeled "PAW Corr." in Table 2.1. This absolute eigenvalue nicely reproduces the FLAPW result. We could adapt the PAW to the norm-conserving case. Obviously, the norm-conserving pseudopotentials are farther from the all-electron results. However, we could introduce the effect of the finite extent of the core electron density in the framework to obtain the result "NC Corr." in Table 2.1.

Obtaining the same absolute numbers in FLAPW and PAW is not a goal *per se*. Of course, the physical properties are never extracted from the absolute position of the Kohn-Sham eigenvalues. However, for benchmarking and accuracy checks, it appears to me that comparing the absolute Kohn-Sham eigenvalues could help in the future. Furthermore, turning back to our original issue of the potential alignment, the previous development have allowed us to highlight the pseudopotential effect on the average potential.

3 Consequences for charged defects

As noted above, different codes may use different conventions for the average electrostatic potentials. This is the case for the ionic potential in Abinit and in QEspresso. With two different choices of average potential, the proper potential alignment technique should be able to reconcile any convention. Let me follow this idea in order to further precise the potential alignment technique.

The difference between Abinit and QEspresso arises from the expression cho-

sen for the so-called "$Z\alpha$" term in the total energy (Ihm et al., 1979; Martin, 2004). The term arises from the ion-ion repulsion, which is usually calculated from the Ewald summation technique (Ewald, 1921). As the Ewald sums accounts for the repulsion between point charges, the energy has to be corrected, since the real positive charges used in the pseudopotential calculations are instead a smooth pseudo-charge (remember $\tilde{n}_{Zc}$ in the PAW language). Abinit (Gonze, 1997) implements the original formula of Ihm et al. (1979):

$$E_{Z\alpha} = \frac{Z_{ion}}{\Omega} \sum_a \alpha_a, \tag{2.6}$$

where $Z_{ion} = Z - N_c$ and the coefficients α_a are integrals calculated for each atom a in the unit cell. The expression of α_a can be found in the Appendix C. QEspresso implements another version of the formula

$$E_{N\alpha} = \frac{N}{\Omega} \sum_a \alpha_a, \tag{2.7}$$

that can be found in many text books (Martin, 2004; Payne et al., 1992). Once again, as long as the system is charged neutral, $N = Z_{ion}$ and the total energy is identical within the two conventions. However the absolute Kohn-Sham potentials differ even in the charge neutral case:

$$v_{Z\alpha}(\mathbf{r}) = \tfrac{\delta E_{Z\alpha}}{\delta n(\mathbf{r})} = 0 \tag{2.8a}$$

$$v_{N\alpha}(\mathbf{r}) = \tfrac{\delta E_{N\alpha}}{\delta n(\mathbf{r})} = \frac{1}{\Omega} \sum_a \alpha_a. \tag{2.8b}$$

Let me compare the formation energy a charged defect would have using the two above mentioned conventions. Figure 2.1 shows the formation energy of a 2+ vacancy using Abinit (Z convention) and using QEspresso (N convention). I have modified the cutoff radius of the local component of the norm-conserving pseudopotential to see how the formation energy behaves as a function of the pseudization details. I expect a weak dependence of the physical formation energy on the d component of the pseudopotential in diamond, since the valence electrons have mainly a character sp. Obviously, the Z convention is much more sensitive to the details of the pseudization, whereas the N convention is rather insensitive.

With this, let me try to visualize the effect of changing the local component of the pseudopotential. In Figure 2.2, one can appreciate how the introduction of a charged vacancy changes the Kohn-Sham potential (electrostatic part only) far from the defect, using the different conventions and different local pseudopotentials. The proper potential alignment should compensate the spurious dependence on the pseudization details observed in Figure 2.1:

$$\Delta V_1 = v_H[n_T^{\text{Defect}}](\mathbf{r}_{\text{far}}) - v_H[n_T^{\text{Pristine}}](\mathbf{r}_{\text{far}}). \tag{2.9}$$

$\mathbf{r}_{\text{far}}$ denotes a point in the supercell "far enough" from the defect. Note that I work here with the electronic potential and not the electrostatic potentials

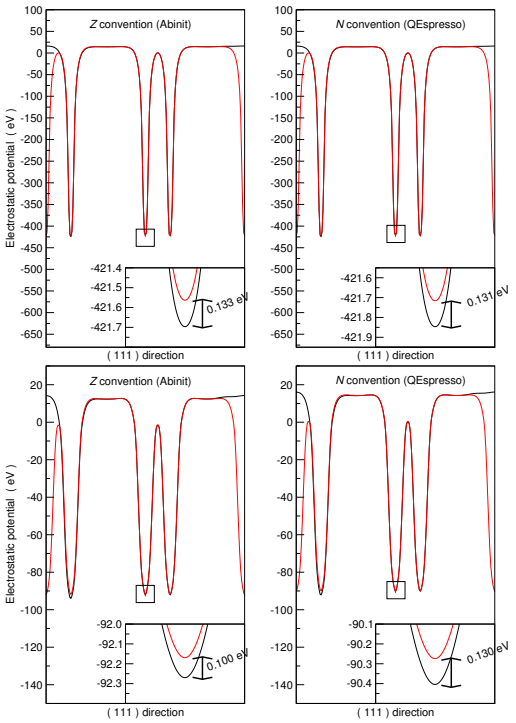

Figure 2.2: Cuts along the (111) direction of the Kohn-Sham potential arising from electrostatic interactions in 64 atom supercell of diamond, with a vacancy at the origin (black line) or without (red line). A close-up view of the effect of the vacancy on the potential is given in the inset. The left-hand column graphs employ the Z convention, whereas the right-hand column graphs employ the N convention. The local pseudopotential has a cutoff radius of 1.5 bohr in the top graphs and a cutoff radius of 0.6 bohr in the bottom graphs.

that differ in sign, owing to the conventional negative charge of the electron. The negative charge of the electron is the cause of many misunderstandings that this convention should avoid. Applying this expression to the uncorrected results in Figure 2.1 will, as announced, reconcile the results arising in the different conventions and also from the different local pseudopotentials (see the open symbols in the Figure).

However, the story has not reached an end yet. Note that the potential alignment in Eq. (2.9) has been labeled with a subscript 1. In fact, this simple definition is sufficient to capture the atomic potential alignment, as just demonstrated above. But it will unfortunately induce a double counting of the electron-electron interactions. This double counting is to be discussed in the next section.

Chapter 3

Electronic potential alignment: Relation between ΔV and $\Delta E_{e.s.}$

This section explains the central argument of the article by Taylor and Bruneval (2011), printed in Appendix C. It also provides updated conclusions, which go beyond the original paper.

In the previous chapter, I have analyzed the potential alignment part that aroused from the electrostatic potential induced by the ions. I have deliberately disregarded the potential alignment produced by the electron-electron interaction. The present section is dedicated to this issue.

1 Is there a need for further potential alignment when the electrostatic correction $\Delta E_{e.s.}$ is used?

The electron-electron electrostatic interactions (spurious Hartree contribution) are indeed problematic for the definition of the potential alignment. First of all, one may think that the spurious image interactions are fully corrected by the other correction term $\Delta E_{e.s.}$ in Eq. (1.1) and then no further term is required. Then if one thinks that the correction term $\Delta E_{e.s.}$ only partially removes the spurious interactions, a further potential alignment $q\Delta V$ may be needed. However, this potential alignment term should not double count the interactions that have already been taken into account with the previous term. Furthermore, when charging the defect (i.e. adding or removing electrons to the neutral defect), the extra charge-charge interaction we spuriously introduce is screened by the other charges present in the system. As a consequence, the spurious interaction are divided by the dielectric constant of the system, which can be approximated by the dielectric constant of the pure crystal as a first guess. If only the electrons have been relaxed in the presence of the added charge, it is

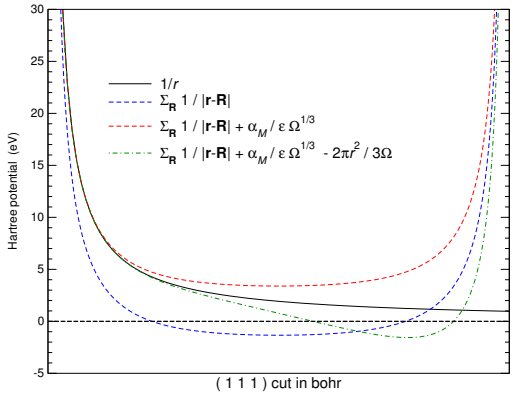

Figure 3.1: Hartree potential cut in eV created by a single point charge $q = -1$ (thin black line) compared to a cubic lattice of point charges (thick dashed blue line). The periodic potential can be corrected by adding increasing multipole orders in the electrostatic correction $\Delta E_{e.s.}$ (thin dashed red line and thin dashed-dotted green line). Note that the Hartree potential differs from the usual electrostatic potential by a minus sign.

logically to employ the ion-clamped dielectric constant ε_∞, whereas if also the nuclei have been allowed to relax, then the corresponding dielectric constant would be the static one ε_0.

The relation between $\Delta E_{e.s.}$ and ΔV is not obvious at first sight. However, considering a practical example, the situation becomes much clearer. In Figure 3.1, I compared the Hartree potential of a single point charge $q = -1$ to the Hartree potential created by a cubic lattice of the same point charge (with a compensating background). The point charge is placed at the origin. The periodic electrostatic potential has been evaluated by the usual Ewald summation (Ewald, 1921; Martin, 2004), with a zero average convention. Due to the presence of an arbitrary constant in the periodic potential, there is no simple way to compare the two potentials. The periodic potential presents a plateau shape far the defect. This plateau is induced by the presence of equidistant charges when the potential is evaluate around the cell boundary: this plateau should not be considered as the zero of the potentials! Indeed, adding a constant shift of $2\Delta E_{e.s.}^{\mathrm{Mad}}/q$ to periodic potential makes more sense: the obtained periodic potential nicely matches the isolated potential in the vicinity of the defects. This demonstrates a tight link between the electrostatic correction $\Delta E_{e.s.}$ and the potential alignment ΔV. But where does this link come from?

2 Demonstration of the link between the potential alignment and the electrostatic correction

This question is answered in the article by Taylor and Bruneval (2011) in an elementary way that I would like to reproduce here. The argument developed in this article is based on the comparison of the Kohn-Sham potential obtained in a truly isolated defect and in a periodic system subjected to spurious charge-charge interactions. Starting from the total energy expressions, the electrostatic correction term $\Delta E_{e.s.}$ is precisely meant to link these two systems' total energies:

$$E_{\text{iso}} = E_{\text{per}} + \Delta E_{e.s.}. \tag{3.1}$$

The Kohn-Sham potential(Kohn and Sham, 1965; Parr and Yang, 1989) is defined as the functional derivative of the total energy without the kinetic energy T with respect to the electron density $n(\mathbf{r})$. Hence, the isolated and periodic Kohn-Sham potentials v are linked through

$$v_{\text{iso}} = v_{\text{per}} + \frac{\delta \Delta E_{e.s.}}{\delta n(\mathbf{r})}. \tag{3.2}$$

At first sight, $\Delta E_{e.s.}$ may look independent from $n(\mathbf{r})$ However this is not true, following the same argument I developed in the previous chapter. Let us specify the expression of $\Delta E_{e.s.}$ using the simplest expression available, the Madelung correction of Leslie and Gillan (1985). Note that one could use more advanced expressions, such as the correction of Makov and Payne (1995), as did Komsa et al. (2012). The Madelung correction introduced in Eq. (1.3) does have a dependence with the charge of the supercell q. Indeed, the charge of the supercell can be expressed (re-using the notation introduced in the previous chapter) as

$$q = Z_{ion} - N = Z_{ion} - \int d\mathbf{r} n(\mathbf{r}). \tag{3.3}$$

Here arrives the dependence on the electron density $n(\mathbf{r})$. And the final results reads

$$v_{\text{iso}} = v_{\text{per}} - \frac{\alpha_M q}{\varepsilon \Omega^{1/3}}. \tag{3.4}$$

This conclusion justifies the shift applied to the periodic potential in the Figure 3.1. As written above, it is possible to go beyond the simplest monopole correction and introduce the next term, the quadrupole term, following Komsa et al. (2012):

$$\Delta E_{e.s.}^{\text{MP}} = \Delta E_{e.s.}^{\text{Mad}} - \frac{2\pi q Q}{3\varepsilon\Omega} \tag{3.5}$$

where the quadrupole Q reads

$$Q = \int d\mathbf{r} r^2 n(\mathbf{r}). \tag{3.6}$$

Adding this term in the derivation improves the description. Besides a quadrupolar term, it also introduces a position dependent term which is also proportional to q:

$$v_{\rm iso}(\mathbf{r}) = v_{\rm per} - \frac{\alpha_M q}{2\varepsilon\Omega^{1/3}} + \frac{2\pi q}{3\varepsilon\Omega}. \tag{3.7}$$

With this term, the obtained potential is not any more periodic as shown in Figure 3.1. The comparison between the isolated potential and the periodic one is then valid in a broader range. As the Madelung correction is exact for the energy of a point charge, I can conjecture that the isolated potential should be also exact if all the higher order terms in the multipole expansion are included. The surprise, in my opinion, is the presence of terms that arises from all the momenta of the distribution. For a point charge, the higher momenta beyond the monopole are all going to zero. As a consequecne, their contribution to the energy is zero, but their contribution to the potential (which is a derivative) remains finite!

3 Consequences for the potential alignment definition

Turning back to the potential alignment definition, I have just shown that the Madelung corrections already accounts for a transformation in the Kohn-Sham potential, rigorously transforming the periodic electrostatic potential into the isolated potential. The challenge is then to find out the charge distribution $n_d(\mathbf{r})$ induced by the charged defect. In realistic case, $n_d(\mathbf{r})$ is not a simple point charge and finding the corresponding $\Delta E_{e.s.}$ may become challenging. Recently, Freysoldt et al. (2009) have shown that the potential alignment procedure may be indeed useful to address the details of the charge distribution which have not been captured by the charge distribution selected to calculate $\Delta E_{e.s.}$. As further explained by Komsa et al. (2012), when doing so, the potential alignment precisely designed for that purpose.

These difficult concepts are better understood with a practical example. Let me explain the ideas of the electronic potential alignment with Gaussian defect distributions, normalized to -1:

$$n_d^\sigma(\mathbf{r}) = \frac{-1}{(\sqrt{2\pi}\sigma)^3} e^{-r^2/2\sigma^2}, \tag{3.8}$$

where the length σ measures the spread of the distribution. Gaussian defect distributions are particularly convenient, since all the calculations, both for isolated charge distribution and for periodic distributions, can be straightforwardly evaluated numerically. The Gaussian distributions are meant to be a more realistic model for realistic defects that spread over a finite volume in the supercell. Figure 3.2 shows the periodic potential created by the periodic Gaussian distributions

$$n_d^{\sigma'}(\mathbf{r}) = \sum_{\mathbf{R}} n_d^\sigma(\mathbf{r}-\mathbf{R}) + \frac{1}{\Omega}, \tag{3.9}$$

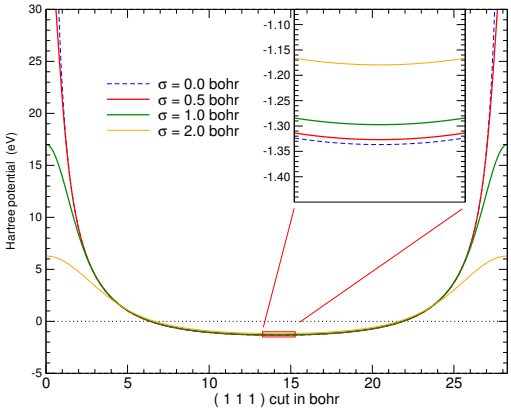

Figure 3.2: Periodic Hartree potential cut in eV created by Gaussian distributions with different widths. $\sigma = 0$ corresponds to a lattice of point charges. The inset shows a close-up view of the cell boundary region. Note that the Hartree potential differs from the usual electrostatic potential by a minus sign.

with the standard zero average convention. I employ here the prime notation introduced in the previous chapter for periodic background compensated densities. The effect of the zero average convention is made obvious: wider spread Gaussian distributions induce a smaller potential in the region of the defect and therefore the potential is less negative far away from the defect (see the inset).

Imagine now that one had use a point charge modeling to obtain the value of the electrostatic correction $\Delta E_{e.s.}$, but the true defect distribution is a Gaussian with a finite width σ. The argument of Freysoldt et al. (2009) is that the potential shift far from the defect can further correct for an incorrect modeling of the charge distribution $n_d(\mathbf{r})$. For Gaussian distributions, one can calculate the exact electrostatic correction (this is the Ewald energy without the short range part). Let me compare the electrostatic correction and the Hartree potential far from the defect for several values of σ as shown in Figure 3.3. There is a perfect linear correspondence between the electrostatic correction and the potential at cell boundary. At least for Gaussian distribution, it is then justified to use

$$\Delta E^{\sigma_1}_{e.s.} = \Delta E^{\sigma_2}_{e.s.} + q\left(v^{\sigma_1}_H(\mathbf{r}_{\text{far}}) - v^{\sigma_2}_H(\mathbf{r}_{\text{far}})\right). \tag{3.10}$$

Remember that the example in Figure 3.3 used $q = -1$. As a consequence in a practical case, where the Madelung correction was used for the electrostatic correction ($\Delta E^{\text{Mad}}_{e.s.} = \Delta E^{\sigma=0}_{e.s.}$), the correction for the true σ can be obtained with

$$\Delta E^{\sigma}_{e.s.} = \Delta E^{\text{Mad}}_{e.s.} + q\left(v^{\sigma}_H(\mathbf{r}_{\text{far}}) - v^{\sigma=0}_H(\mathbf{r}_{\text{far}})\right). \tag{3.11}$$

This formula obtained empirically for a Gaussian charge distribution in a cubic lattice was also checked successfully for linear combination of Gaussian in non-cubic lattices. Komsa et al. (2012) obtained analytically the same result from

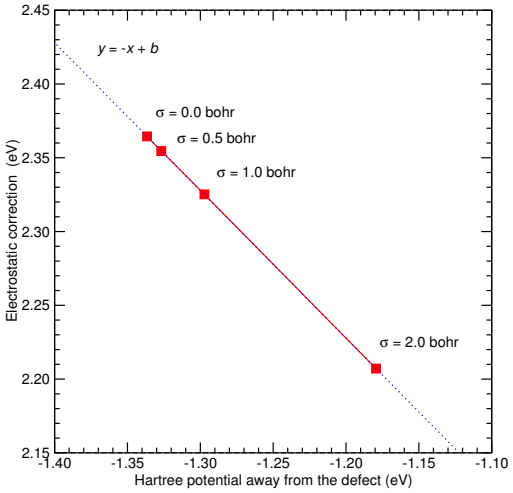

Figure 3.3: Electrostatic correction $\Delta E_{e.s.}$ as function of the Hartree potential away from the defect as extracted from the previous Figure for several Gaussian distributions normalized to $q = -1$. The blue dotted line is a linear fit of the data with slope -1.

the expansion of the electrostatic corrections up to the quadrupole. It appears, according to my numerical results, that higher order momenta give a vanishing contribution.

4 Concluding remarks

In summary, the spurious charge interaction between the images could completely be considered through the electrostatic correction $\Delta E_{e.s.}$ or completely be considered through the potential alignment $q\Delta V$. In practical cases, anyone can choose the part of the correction included in $\Delta E_{e.s.}$ or in $q\Delta V$. In my opinion, it appears useless to refine the modeling of the defect charge density beyond a point charge or beyond a single Gaussian distribution with a fixed width, since anyway the potential alignment will be present to fix the inadequacy of the model charge selected. This final statement goes beyond the idea present in the original paper of Freysoldt et al. (2009).

Summing up the atomic and electronic potential alignement contributions as described in the present and previous chapter, I recommend the expression for the total finite size effect corrections:

$$\begin{aligned}\Delta E_{e.s.} + q\Delta V = \Delta E_{e.s.}^{\mathrm{Mad}} \\ + q\left[v_H[n_T^{\mathrm{Defect}}](\mathbf{r}_{\mathrm{far}}) - v_H[n_T^{\mathrm{Pristine}}](\mathbf{r}_{\mathrm{far}}) - v_H^{\sigma=0}(\mathbf{r}_{\mathrm{far}})\right]. \quad (3.12)\end{aligned}$$

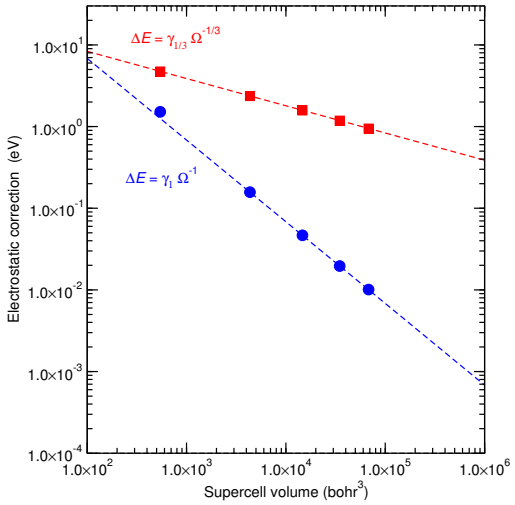

Figure 3.4: Electrostatic corrections evaluated for a unity Gaussian distribution with width $\sigma = 2.0$ bohr in a cubic lattice as a function of the supercell volume Ω. The Madelung correction is given with red square symbols. The correction on top of the Gaussian is plotted with blue circles. The dashed lines highlight the slope in the log-log scale.

This expression contains both the atomic and the electronic potential alignments contributions with no distinction.

It is worth noting that, while these results agree with the developments of Freysoldt et al. (2009) and Komsa et al. (2012), they contradict the highly cited paper of Lany and Zunger (2008). Indeed, Lany and Zunger argued that the quadrupole term in the Makov-Payne expansion scales with the supercell edge $\Omega^{-1/3}$. They proposed therefore to merge the monopole and the quadrupole terms or, in other words, to rescale the monopole with an empirical factor $2/3$. In Figure 3.4, both monopole correction and the difference between the monopole and the exact electrostatic correction are plotted for a Gaussian distribution with a finite width as a function of the supercell volume Ω. In log-log scale, it becomes obvious that the monopole indeed scales with $\Omega^{-1/3}$. However, the remaining correction does scale with Ω^{-1} almost perfectly, in contradiction with Lany and Zunger (2008).

Part III

Many-body Perturbation Theory for point defects

Chapter 4

Introduction to the GW approximation

Before starting with the application of the GW approximation, it is worthwhile recapping the basics underlying the GW theory. Since the GW approximation is not so widely spread, I adapted the review chapter I coauthored with Matteo Gatti on the topic (Bruneval and Gatti, 2014).

1 The Green's function G and the self-energy Σ

The single-particle Green's function G is the most basic ingredient of MBPT. The time-ordered Green's function describes the propagation of an extra electron in an electronic system for positive times and the propagation of a missing electron (i.e. a hole) for negative times:[1]

$$\begin{aligned} iG(\mathbf{r}t,\mathbf{r}'t') &= \theta(t-t')\langle N0|\psi(\mathbf{r}t)\psi^\dagger(\mathbf{r}'t')|N0\rangle \\ &\quad -\theta(t'-t)\langle N0|\psi^\dagger(\mathbf{r}'t')\psi(\mathbf{r}t)|N0\rangle, \end{aligned} \tag{4.1}$$

where $|N0\rangle$ denotes the exact ground-state wavefunction of an N electron system, ψ and $\psi^\dagger$ are the annihilation/creation field operators in the Heisenberg picture, and θ is the step function.

The physical meaning of G becomes clear when inserting the closure relation in between the two field operators and taking a Fourier transform in time. The so-called Lehmann representation reads

$$G(\mathbf{r},\mathbf{r}',\omega) = \sum_i \frac{f_i(\mathbf{r})f_i^*(\mathbf{r}')}{\omega - E_i}. \tag{4.2}$$

[1] Here the spin degrees of freedom are omitted for simplicity. The generalization is however straightforward.

The poles of G are located at the energies E_i

$$\begin{aligned} E_i &= E_{N+1i} - E_{N0} - i\eta \text{ when } E_i > \mu \\ &= E_{N0} - E_{N-1i} + i\eta \text{ when } E_i < \mu, \end{aligned} \tag{4.3}$$

where the energy $E_{N\pm1i}$ are the exact eigenenergies of the $N \pm 1$ electron system and i is the index labeling the exact eigenvectors of both the $N-1$ and $N+1$ electron systems. The ubiquitous vanishing positive η has naturally arisen from the Fourier transform of the step functions. In a solid, the discrete set of poles in Eq. (4.2) merges into a branch-cut. The so-called Lehmann amplitudes f_i are then defined as

$$\begin{aligned} f_i(\mathbf{r}) &= \langle N0|\psi(\mathbf{r}0)|N+1i\rangle \text{ when } E_i > \mu \\ &= \langle N-1i|\psi(\mathbf{r}0)|N0\rangle \text{ when } E_i < \mu. \end{aligned} \tag{4.4}$$

Note that the Lehmann amplitudes f_i are not mutually orthogonal. From this representation, we see that the poles E_i carry the exact ionization energies of electrons in the system or the exact affinity energies. The analytic structure of G is also made clear: the poles lie slightly above the real axis for $E_i < \mu$ and slightly below for $E_i > \mu$. The poles can be directly compared to the peaks obtained from a photoemission or inverse photoemission experiment.

Because G is the fundamental quantity, a great deal of effort has been put in to its evaluation in a many-body context. This poses a very large challenge since equation of motion for G involves the two-particle Green's function. Its equation of motion in turn involves the three-particle Green's function, and so on. The standard remedy in MBPT is to break this hierarchy by introducing an effective operator, the self-energy Σ. As Schwinger showed, by introducing an auxiliary external field $U(\mathbf{r}t)$ that is set to zero at the end, it is possible to formally express the two-particle Green's function as a function of the one-particle Green's function (Strinati, 1988). This results in a equation of motion for G alone:

$$\int d\mathbf{r}' \left\{ [\omega - h_0(\mathbf{r}) - V_H(\mathbf{r})]\, \delta(\mathbf{r}-\mathbf{r}') - \Sigma(\mathbf{r},\mathbf{r}',\omega) \right\} G(\mathbf{r}',\mathbf{r}'',\omega) = \delta(\mathbf{r}-\mathbf{r}''). \tag{4.5}$$

Here h_0 is the non-interacting Hamiltonian and V_H the Hartree potential. Note that the self-energy Σ hides all the complexity of the original problem and thus is a non-local, dynamical and non-Hermitian operator. When $\Sigma = 0$ the Green's function G_0 is simply the resolvent of the Hartree Hamiltonian: $G_0^{-1} = \omega - h_0 - V_H$. We refer the reader to the review articles of Strinati (1988) or of Hedin and Lundqvist (1970) for further details.

Dyson's equation results by multiplying Eq. (4.5) by G_0:

$$G(\mathbf{r},\mathbf{r}',\omega) = G_0(\mathbf{r},\mathbf{r}',\omega) + \int d\mathbf{r}_1 d\mathbf{r}_2 G_0(\mathbf{r},\mathbf{r}_1,\omega)\Sigma(\mathbf{r}_1,\mathbf{r}_2,\omega)G(\mathbf{r}_2,\mathbf{r}',\omega). \tag{4.6}$$

This equation establishes the link between the Hartree Green's function G_0 (easily calculated) and the fully interacting Green's function G (very hard to calculate) through the self-energy Σ.

The purpose of MBPT is then to provide approximations with increasing accuracy for the self-energy. The Coulomb interaction between electrons

$$v(\mathbf{r}-\mathbf{r}') = \frac{1}{|\mathbf{r}-\mathbf{r}'|} \tag{4.7}$$

is considered as the perturbation with respect to the independent-particle case. The first-order contribution to the self-energy is nothing else but the Fock exchange operator (the Hartree potential is already taken into account by G_0). This level of approximation is widely used for atoms and molecules, and in quantum chemistry perturbative methods in v with respect to Hartree-Fock are known as Møller-Plesset perturbation theory (Møller and Plesset, 1934). However for the homogeneous electron gas Hartree-Fock yields an anomalous zero density of states at the Fermi level. There is therefore a stringent need for higher order terms for periodic systems. Unfortunately, the analytic evaluation of one of the two second-order contributions is not finite in the case of the homogeneous electron gas (Fetter and Walecka, 1971; Mahan, 2000). Perturbation theory is thus not justified. How should one proceed to circumvent this problem, especially for periodic systems?

2 The screened Coulomb interaction W

This divergence can be addressed in an effective manner by introducing a screened counterpart to the Coulomb interaction v. Other electrons act as a dielectric medium that reduces the interaction between any pair. It is common sense that the interaction between charges is not the same in vacuum as in a dielectric medium. At the macroscopic scale, this is measured by the dielectric constant of the medium. At the microscopic scale, the screening of the Coulomb interaction is given by

$$W(\mathbf{r},\mathbf{r}',\omega) = \int d\mathbf{r}_1 \varepsilon^{-1}(\mathbf{r},\mathbf{r}_1,\omega) v(\mathbf{r}_1-\mathbf{r}'), \tag{4.8}$$

where the microscopic dielectric matrix ε^{-1} has been introduced. ε is linked to the macroscopic dielectric function ε_M (Adler, 1963; Wiser, 1963), which is a measurable quantity. For instance, $-\mathrm{Im}\varepsilon_M^{-1}$ is called the loss function and can be measured by electron energy loss spectroscopy (EELS) or inelastic X ray scattering (IXS).

So far, the expression of the dielectric matrix is not specified. Nevertheless, one can still analyze the physical meaning of the dynamically screened Coulomb interaction $W(\mathbf{r},\mathbf{r}',\omega)$. The effective interaction between electrons in a medium is decreased from v, the bare Coulomb interaction, to W the screened interaction. A perturbation theory based on W rather than on v makes then much more sense. However there is a price to pay: the screened interaction W is dynamical, meaning that the screening is more effective for some frequencies than for others. For metals, the static dielectric constant is infinite and consequently

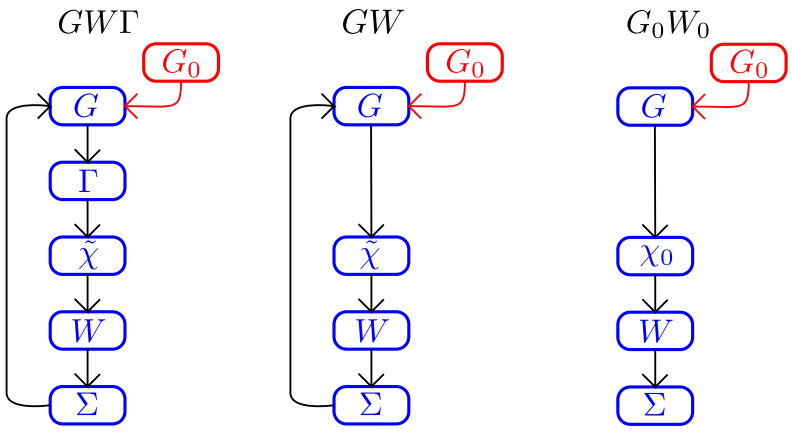

Figure 4.1: Schematic view of the three frameworks described in this review. The exact expression ($GW\Gamma$, left panel) can be obtained after an initialization with a guessed Green's function G_0 by iteration of all the Hedin's equations. The self-consistent GW approximation (GW, central panel) arises from the iteration of the equations keeping the vertex function $\Gamma = 1$. The perturbative GW approximation (G_0W_0, right panel) is the one-shot evaluation of the GW self-energy based on the guessed Green's function G_0.

the long-range component of W vanishes. This fixes the problem of the vanishing density of states at the Fermi level predicted by Hartree-Fock theory for the homogeneous electron gas. Conversely, in the high frequency limit, the screening by the electrons becomes completely ineffective and the screened Coulomb interaction is simply the bare Coulomb interaction.

The additional complexity contained in W compared to v bears the hope that the perturbation theory is to be rapidly convergent. Maybe, as W already contains an infinite sum of interactions v, just the first order in W will be sufficient, as proposed by Hedin (1965) ...

3 Hedin's equations and the GW approximation

Employing W instead of v in the MBPT allowed Hedin to reformulate the exact equations of the solution of the many-electron problem for the calculation of G

(Hedin, 1965). They read:

$$G(1,2) = G_0(1,2) + \int d(34) G_0(1,3)\Sigma(3,4)G(4,2) \tag{4.9a}$$

$$\Sigma(1,2) = i\int d(34) G(1,3)W(1^+,4)\Gamma(3,2,4) \tag{4.9b}$$

$$W(1,2) = \int d(3)\epsilon^{-1}(1,3)v(3,2) \tag{4.9c}$$

$$\epsilon(1,2) = \delta(1,2) - \int d(3)v(1,3)\tilde{\chi}(3,2) \tag{4.9d}$$

$$\tilde{\chi}(1,2) = -i\int d(34) G(1,3)G(4,1)\Gamma(3,4,2) \tag{4.9e}$$

$$\Gamma(1,2,3) = \delta(1,2)\delta(1,3) + \int d(4567)\frac{\delta\Sigma(1,2)}{\delta G(4,5)}G(4,6)G(7,5)\Gamma(6,7,3) \tag{4.9f}$$

Contracted indexes $(1) = (\mathbf{r}_1, t_1, \sigma_1)$ have been used for simplification. The index 1^+ denotes the times $t_1 + \eta$ for a vanishing positive η. Most of the quantities have been introduced earlier. $\tilde{\chi}$ is the irreducible polarizability and Γ is the three-point vertex function. These non linear equations are coupled. If solved self-consistently, these equations form an exact scheme to obtain the solution of the many-body problem. This process is pictured in the left panel of Figure 4.1. Of course, in practice even in the simplest cases, it is not feasible mainly due to the presence of the three-point vertex function Γ. It is then natural that an approximated scheme begins by simplifying this particular term.

The second term in the right-hand side of Eq. (4.9f) involves the self-energy which is of first-order in W according to Eq. (4.9b). Retaining only the zero-order terms in Eq. (4.9f) (i.e. the δ functions), Hedin's equations are greatly simplified:

$$G(1,2) = G_0(1,2) + \int d(34) G_0(1,3)\Sigma(3,4)G(4,2) \tag{4.10a}$$

$$\Sigma(1,2) = iG(1,2)W(1^+,2) \tag{4.10b}$$

$$W(1,2) = \int d(3)\epsilon^{-1}(1,3)v(3,2) \tag{4.10c}$$

$$\epsilon(1,2) = \delta(1,2) - \int d(3)v(1,3)\tilde{\chi}(3,2) \tag{4.10d}$$

$$\tilde{\chi}(1,2) = -iG(1,2)G(2,1). \tag{4.10e}$$

The irreducible polarizability $\tilde{\chi}$ is then a simple product of two Green's functions. This is the well known Random-Phase Approximation (RPA) to the dielectric matrix. Also the self-energy is much simplified: this is just the simple product of G and W, giving the name to the GW approximation. It is of first-order in W. The missing terms (second and higher orders in W) are commonly named the "vertex corrections".

The set of equations (4.10a-4.10e) still requires a self-consistent treatment since W and Σ depend on G, which is the quantity one needs to obtain. This

is pictured in the central panel of Figure 4.1. The practical implementation of these equations is still far from obvious. This is the reason why for many years the GW self-energy has been evaluated non self-consistently.

4 Practical calculation of the GW self-energy: the G_0W_0 approach

It is most often impossible to evaluate the Green's function self-consistently from Eqs. (4.10a-4.10e). However let us imagine that mean-field theories such as Hartree-Fock or Kohn-Sham would provide a good description of the electronic system under study. In a mean-field theory, the one-electron wavefunctions $\phi_i(\mathbf{r})$ and eigenvalues ϵ_i allow one to evaluate the independent-particle Green's function[2] G_0

$$G_0(\mathbf{r},\mathbf{r}',\omega) = \sum_i \frac{\phi_i(\mathbf{r})\phi_i^*(\mathbf{r}')}{\omega - \epsilon_i + i\eta \mathrm{sign}(\epsilon_i - \mu)}. \tag{4.11}$$

The location of the poles of G_0 are above the real axis for occupied states and below for empty states. As a consequence, $\tilde{\chi}$ and then W can be readily evaluated from this expression of G_0. Let us label this evaluation of the irreducible polarizability, χ_0, and of the screened Coulomb interaction, W_0. Finally, the GW self-energy is obtained as the convolution G_0W_0.

The so-called G_0W_0 approach consists in stopping the procedure immediately after the first evaluation of the self-energy, as shown in the right hand panel of Figure 4.1. This "one-shot" procedure is justified when the starting mean-field theory used for G_0 is accurate enough for the targeted property. The vast majority of the GW applications for almost 50 years have been obtained with the G_0W_0 procedure. Of course, the choice of the starting point is material dependent. The seminal paper of Hedin (1965) simply employed the free electron model to calculate the GW self-energy for the homogeneous electron gas. The first application of GW to real solids used either the Hartree-Fock approximation (Strinati et al., 1980) or the local density approximation (Hybertsen and Louie, 1985). For atoms, Shirley and Martin chose Hartree-Fock (Shirley and Martin, 1993). The rationale underlying the choice is the selection of the most accurate mean-field theory for the specific system under scrutiny. This strategy is sometimes referred to as the "best G, best W" approach.

In the quasiparticle approximation, the Dyson equation (4.6) becomes

$$\left(-\frac{\nabla^2}{2} + V_{ext}(\mathbf{r}) + V_H(\mathbf{r})\right)\psi_i(\mathbf{r}) + \int d\mathbf{r}'\Sigma(\mathbf{r},\mathbf{r}',E_i)\psi_i(\mathbf{r}') = E_i\psi_i(\mathbf{r}) \tag{4.12}$$

In the G_0W_0 framework one assumes that the quasiparticle wavefunctions ψ_i can be approximated by the Kohn-Sham orbitals ϕ_i. By comparing Eqs. (4.12) and the Kohn-Sham equation, one finds that the quasiparticle energies E_i can

[2] G_0 here can be understood as a generalization of the Hartree Green's function introduced in Eq. 4.6, and thus we keep the same notation for a distinct quantity.

be thus calculated as a first-order correction with respect to the underlying mean-field starting point from

$$E_i = \epsilon_i + \langle\phi_i|\Sigma(E_i) - V_{xc}|\phi_i\rangle, \tag{4.13}$$

where Σ is the G_0W_0 self-energy. From a linearization of the frequency dependence of Σ, one finally obtains

$$E_i = \epsilon_i + Z_i\langle\phi_i|\Sigma(\epsilon_i) - V_{xc}|\phi_i\rangle, \tag{4.14}$$

where the renormalization factors Z_i are

$$Z_i = \left[1 - \langle\phi_i| \left.\frac{\partial\Sigma(\omega)}{\partial\omega}\right|_{\omega=\epsilon_i} |\phi_i\rangle\right]^{-1}. \tag{4.15}$$

In most G_0W_0 calculations the band structures are obtained using Eq. (4.14). One can also calculate the spectral function (the imaginary part of the Green's function) from

$$A_{ii}(\omega) = \frac{1}{\pi}\frac{|\langle\phi_i|\mathrm{Im}\Sigma(\omega)|\phi_i\rangle|}{[\omega - \epsilon_i - \langle\phi_i|\mathrm{Re}\Sigma(\omega) - V_{xc}|\phi_i\rangle]^2 + [\langle\phi_i|\mathrm{Im}\Sigma(\omega)|\phi_i\rangle]^2}. \tag{4.16}$$

The spectral function has poles in correspondence to the quasiparticle energies, i.e. when $\omega - \epsilon_i - \langle\phi_i|\mathrm{Re}\Sigma(\omega) - V_{xc}|\phi_i\rangle = 0$ [compare with Eq. (4.13)]. The width of the quasiparticle peak is given by $\mathrm{Im}\Sigma(\omega)$, which is hence linked to the lifetime of the excitation (defined as the inverse of its width). The spectral function can have other peaks, the satellites, that originate from structures in $\mathrm{Im}\Sigma(\omega)$. Also $\omega - \epsilon_i - \langle\phi_i|\mathrm{Re}\Sigma(\omega) - V_{xc}|\phi_i\rangle$ can have additional zeroes, giving rise to satellites. Within the GWA this latter kind of satellites has been called plasmarons (Hedin et al., 1967; Lundqvist, 1967), but later they have been shown to be an artifact of the GWA (Blomberg and Bergersen, 1972; Bergersen et al., 1973; Guzzo et al., 2011). In Hartree-Fock the self-energy is Hermitian: $\mathrm{Im}\Sigma(\omega) = 0$. Therefore quasiparticle peaks become delta functions (i.e. the lifetime of quasiparticle becomes infinite). Moreover, since the self-energy is static, no other structures (e.g. satellites) can appear in the spectral function.

The GW self-energy can be split into a Fock exchange term Σ_x and a correlation term $\Sigma_c(\omega)$: $\Sigma(\omega) = \Sigma_x + \Sigma_c(\omega)$. While $\Sigma_x = iGv$ is static, the evaluation of $\Sigma_c(\omega)$ requires the calculation of the convolution integral of G and $W_p = W - v$:

$$\Sigma_c(\mathbf{r}_1, \mathbf{r}_2, \omega) = \frac{i}{2\pi}\int d\omega' e^{i\eta\omega'} G(\mathbf{r}_1, \mathbf{r}_2, \omega + \omega')W_p(\mathbf{r}_1, \mathbf{r}_2, \omega'). \tag{4.17}$$

Since Σ_c is obtained through the frequency integration (4.17), the fine details of the energy dependence of W_p are often not important. In these cases one can approximate the imaginary part of the inverse dielectric function ϵ^{-1} as a single-pole function in ω (plasmon-pole model) (Hybertsen and Louie, 1986;

Godby and Needs, 1989). Plasmon-pole models can be used for calculating quasiparticle energies, but should be avoided for spectral functions, because for example they don't describe correctly ImΣ. In these cases the full-frequency dependence of Σ is required and the frequency integration has to be performed with care (Lebègue et al., 2003).

The spectral representation of W_p is given by (Hedin, 1999):

$$W_p(\mathbf{r}_1, \mathbf{r}_2, \omega) = \sum_s W_s(\mathbf{r}_1, \mathbf{r}_2) \left[\frac{1}{\omega - (\omega_s - i\eta)} - \frac{1}{\omega + (\omega_s - i\eta)} \right]. \qquad (4.18)$$

The poles of W_p are the energies ω_s that correspond to neutral excitations (electron-hole transitions and plasmons). By combining (4.18) with (4.11) and performing the frequency integration (4.17), one finds that the G_0W_0 self-energy is given by the sum of two terms:

$$\Sigma_{\mathrm{SEX}}(\mathbf{r}_1, \mathbf{r}_2, \omega) = -\sum_i \theta(\mu - \epsilon_i) \phi_i(\mathbf{r}_1) \phi_i^*(\mathbf{r}_2) W(\mathbf{r}_1, \mathbf{r}_2, \omega - \epsilon_i), \qquad (4.19a)$$

$$\Sigma_{\mathrm{COH}}(\mathbf{r}_1, \mathbf{r}_2, \omega) = \sum_i \phi_i(\mathbf{r}_1) \phi_i^*(\mathbf{r}_2) \sum_s \frac{W_s(\mathbf{r}_1, \mathbf{r}_2)}{\omega - (\omega_s - i\eta) - \epsilon_i}. \qquad (4.19b)$$

The first term arises from the poles in G and the second from the poles in W. Owing to the similarity of the first term with the Fock exchange, it is usually called the "screened exchange" term. The second term is referred to as the "Coulomb-hole" term (Hedin, 1965). If a further static approximation is carried out, this decomposition gives rise to the so-called COHSEX (Coulomb hole plus screened exchange), first introduced by Hedin (1965); Hedin et al. (1967). This static and Hermitian self-energy is obtained by setting $\omega - \epsilon_i = 0$ in $\Sigma_{\mathrm{SEX}}(\omega)$ and $\Sigma_{\mathrm{COH}}(\omega)$. This corresponds to assume that the main contribution to the self-energy $\Sigma(\omega)$ stems from the states ϵ_i close to ω. So $\omega - \epsilon_i$ is small compared to main excitations in W which are at the plasmon energies ω_s (Hedin, 1965; Hedin et al., 1967).

In the majority of the practical cases, the G_0W_0 scheme is largely sufficient to evaluate successfully the GW self-energy. However for some cases it is not. As long as we are interested in defects, the computational burden restrains us to G_0W_0.

Chapter 5

GW calculations for large supercells

The first problem before applying GW corrections to defective systems is the computational burden involved by the use of supercells. Reasonable supercell sizes consist of a minimum of, say, 50-100 atoms. Even with the constant increase of computational power, the GW calculations are most often limited to crystal unit cells. The reasons why the GW calculations are so cumbersome are the truly non-local nature of the screened Coulomb interaction W, see Eq. (4.8), and the dependence of the Green's function G upon empty states. The latter can really be problematic in actual calculations as documented in Aulbur et al. (2000); Tiago et al. (2004); van Schilfgaarde et al. (2006). For instance, it is not uncommon to have GW calculations a unit cell of semiconductor requiring 4 occupied states, but as many as 200 empty states.

In 2006, there already exist a few methods to remove (Reining et al., 1997) or limit (Tiago and Chelikowsky, 2006) the dependence on empty states. However, these methods have not been applied extensively, because of difficulties in the implementation or because of limited accuracy. I and Xavier Gonze decided then to cope with the slow convergence with respect to empty states in a simpler way (Bruneval and Gonze, 2008).

1 Dependence on empty states in GW calculations

In the expression of the non-interacting Green's function G_0 from Eq. (4.11), the sum over states runs over all the states, occupied and empty. In the GW framework, G_0 is then used in two different places: in non-interacting polarizability $\chi_0 = -G_0 G_0$ and in the self-energy $\Sigma = iG_0 W$. Therefore, the empty state dependence occurs both in χ_0 and in Σ. In Figure 5.1, the solid line shows how large this dependence is in the case of cubic silicon carbide (3C-SiC) in

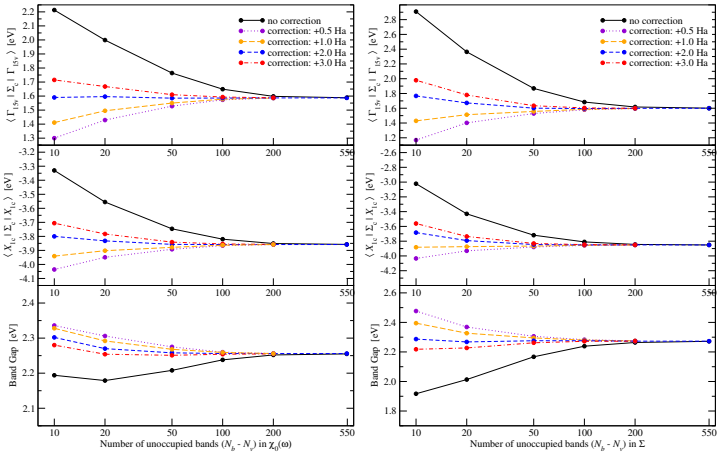

Figure 5.1: Convergence study of the correlation part of the self-energy at top valence (upper panel), at bottom conduction (middle panel), and of the band gap (lower panel) of 3C-SiC as a function of the number of unoccupied states explicitly included in the calculation of the polarizability (left) or in the calculation of the self-energy (right). The solid (black) curve shows the usual *GW* result with no correction. The other curves include the correction with different values for the energy parameter $\bar{\epsilon}_{\chi_0}$: 0.5 Ha (dotted violet), 1.0 Ha (long-dashed orange), 2.0 Ha (short-dashed blue), or 3.0 Ha (dot-dashed red) above the last explicitly calculated band.

both χ_0 (left-hand panel) and in Σ (right-hand panel). As far as the band gap is concerned, the slow convergence is visible, but not so problematic, because it is an energy difference. However, when we turn to the absolute position of a quasiparticle energy (referred to the mean electrostatic potential), the dependence becomes visible: 200 empty states are required for a 0.1 eV accuracy. Remember that the case of SiC is relatively simple. In transition metal oxides in which the nature of the valence band maximum differs from the nature of the conduction band minimum (e.g. Cu_2O), even the band gap evaluation requires a huge number of empty states.

2 Extrapolar idea

The starting point of our method was the extrapolar approximation used by Anglade and Gonze (2008), which owes much similarity to the common-energy denominator approximation (CEDA) used in quantum-chemistry, in particular for optimized effective potential (OEP) generation (Sharp and Horton, 1953; Krieger et al., 1990; Kümmel and Kronik, 2008). The expression of χ_0 in frequency is obtained by the convolution of G_0 with itself:

$$\chi_0(\mathbf{r},\mathbf{r}',\omega) = \sum_{ij} \phi_i(\mathbf{r})\phi_i^*(\mathbf{r}')\phi_j(\mathbf{r}')\phi_j^*(\mathbf{r}) \left[\frac{f_j(1-f_i)}{\omega - (\epsilon_i - \epsilon_j) + i\eta} - \frac{f_i(1-f_j)}{\omega - (\epsilon_i - \epsilon_j) - i\eta} \right], \tag{5.1}$$

where f_i is the occupation (from 0 to 1) and η is a vanishing positive real number. It would be tempting to get rid of the sum over i or j by using the closure relation:

$$\sum_{i>N_b} \phi_i(\mathbf{r})\phi_i^*(\mathbf{r}') = \delta(\mathbf{r}-\mathbf{r}') - \sum_{i\leq N_b} \phi_i(\mathbf{r})\phi_i^*(\mathbf{r}'). \tag{5.2}$$

Unfortunately, the dependence upon i is also present in the denominator through ϵ_i. Here comes the extrapolar idea: if beyond a given index $i > N_b$ one assumes that all the energy have the same common energy $\bar{\epsilon}$, then the closure relation could be applied. For example, the first term in χ_0 could be simplified from

$$\sum_{\substack{i \text{ occ} \\ N_v<j\leq N_b}} \frac{\phi_i(\mathbf{r})\phi_i^*(\mathbf{r}')\phi_j(\mathbf{r}')\phi_j^*(\mathbf{r})}{\omega - (\epsilon_i - \epsilon_j) + i\eta} + \sum_{\substack{i \text{ occ} \\ j>N_b}} \frac{\phi_i(\mathbf{r})\phi_i^*(\mathbf{r}')\phi_j(\mathbf{r}')\phi_j^*(\mathbf{r})}{\omega - (\bar{\epsilon} - \epsilon_j) + i\eta} \tag{5.3}$$

into

$$\sum_{\substack{i \text{ occ} \\ N_v<j\leq N_b}} \phi_i(\mathbf{r})\phi_i^*(\mathbf{r}')\phi_j(\mathbf{r}')\phi_j^*(\mathbf{r}) \left[\frac{1}{\omega - (\epsilon_i - \epsilon_j) + i\eta} - \frac{1}{\omega - (\bar{\epsilon} - \epsilon_j) + i\eta} \right] + \sum_{i \text{ occ}} \frac{\phi_i(\mathbf{r})\phi_i^*(\mathbf{r})\delta(\mathbf{r}-\mathbf{r}')}{\omega - (\bar{\epsilon} - \epsilon_j) + i\eta}. \tag{5.4}$$

In this last expression, all the states beyond index N_b are included in the calculation in a very approximate manner: they are assumed to have the same energy $\bar{\epsilon}$! However these states are taken into account in the calculation without the need to actually calculate them! The same procedure can be applied to the self-energy (see (Bruneval and Gonze, 2008) for further details).

3 Setting the position of the extra-pole

In practical calculations, the extrapolar approximation is used to complement the explicitly calculated χ_0 and Σ with an estimation of the remainder due to the empty states which have not been included in the calculation. In the limiting case for which all the states have been included, the remainder is strictly zero whatever the choice for the parameter $\bar{\epsilon}$. At this stage, a practical question rises: how do you set the energy of the extra pole $\bar{\epsilon}$? Admittedly, in most of the applications $\bar{\epsilon}$ has been considered as a parameter. If one refers the $\bar{\epsilon}$ with respect to the highest energy explicitly included in the calculation,

$$\bar{\epsilon} = \epsilon_{N_b} + \Delta, \tag{5.5}$$

the convergence with a fixed Δ value is very smooth, as shown in Figure 5.1. If Δ is too small, the extrapolar correction is too large and the self-energy converges from below, whereas if Δ is too large, the extrapolar correction is too small and the self-energy converges from above. A good strategy is to find the value of Δ that would make the convergence flat. I have no formal proof for this procedure. But it has always worked nicely for all the systems studied so far, crystals, supercells, nanowires (Peelaers et al., 2011), or isolated molecules (Bruneval, 2009). I have observed that the value of Δ is generally larger for crystalline systems (~ 2 Ha) than for finite systems (~ 0 Ha).

In the original article, we rather advocated for the fulfillment of a sum rule to set the most appropriate value for $\bar{\epsilon}$. This procedure is also a try-and-error procedure since one needs to perform the calculation of χ_0 to evaluate the fulfillment of the sum rule (in reciprocal space) (Mahan, 2000; Taut, 1985):

$$\int_0^{+\infty} d\omega \omega \mathrm{Im}\left[\varepsilon_{\mathbf{GG}}(\mathbf{q},\omega)\right] = \frac{\pi}{2}\omega_p^2, \tag{5.6}$$

where $\omega_p = \sqrt{4\pi n}$ is the classical plasma frequency (n being the average electronic density). $\mathbf{G}$ stands for a reciprocal lattice vector and $\mathbf{q}$ for a Brillouin zone vector. Though this procedure was not much used besides in our original work, it is still quite instructive about how to improve the approximation in the future. In Figure 5.2, I show in the upper how the usual χ_0 slowly fulfills the sum-rule when increasing the number of empty states. When the extrapolar scheme is switched on, the convergence behavior of the sum-rule becomes drastically different. In the lower panel of Figure 5.2, the fulfillment of the sum-rule strongly depends on the transferred momentum $|\mathbf{q}+\mathbf{G}|$. For instance, for large values $|\mathbf{q}+\mathbf{G}|$, the integral in Eq. (5.6) is clearly too small, which indicates a

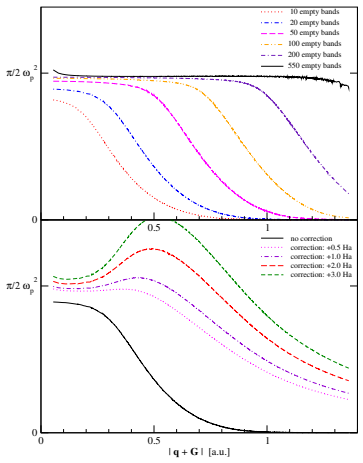

Figure 5.2: Upper panel: Value of the integral in Eq. (5.6) as a function of the transferred momentum $|\mathbf{q}+\mathbf{G}|$ without any correction using 10, 20, 50, 100, 200, 550 empty bands in 3C-SiC. Lower panel: Value of the integral in Eq. (5.6) as a function of the transferred momentum $|\mathbf{q}+\mathbf{G}|$ with 20 empty bands using no correction, or a correction with an average energy Δ of 0.5 Ha, 1.0 Ha, 2.0 Ha, or 3.0 Ha.

too low value for $\bar{\epsilon}$. An obvious improvement of the extrapolar method would be to have a common energy $\bar{\epsilon}$ that varies with $|\mathbf{q}+\mathbf{G}|$. In this case, the value of $\bar{\epsilon}$ would still be independent from the state i and therefore the closure relation could still be applied. The implementation would only marginally be more complicated. Note that the $|\mathbf{q}+\mathbf{G}|$-dependent energy is also obtained within the scheme of Berger et al. (2010).

Finally, I would like to conclude with the application of the extrapolar method to realistic system size of defect calculations. Figure 5.3 shows the convergence behavior with respect to empty states with and without the extrapolar method. It shows that the absolute position of the states are converged with "only" 600 empty states. If ones compares to the 128 occupied states, this is a 1:5 ratio, which is much lower than the 1:50 ratio a standard *GW* calculation would require.

In summary, the extrapolar method has greatly helped the application of the *GW* approximation to large systems. The extrapolar method does not change the scaling of the calculation with respect to the system size, but it divides the prefactor by a 5-10. The number of empty states that needs to be stored in the computer memory is reduced with the same ratio. This is an attractive feature, since most often the memory is the actual bottleneck for *GW* calculations.

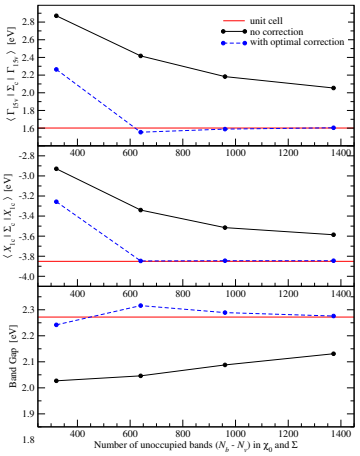

Figure 5.3: Convergence study of the correlation self-energy at top valence (upper panel), at bottom conduction (middle panel), and of the band gap (lower panel) of 3C-SiC in a 64-atom cubic supercell as a function of the number of unoccupied states explicitly included in the calculation of the polarizability and in the self-energy.

Chapter 6

DFT+GW construction for point defects

This section summarizes, further explains, and exemplifies the article by Bruneval (2012b) printed in Appendix D.

The necessity to have the correct band gap for the description of point defects has been stressed in the introduction. The most effective and reliable technique to obtain systematically the right band gap of semiconductors and insulators is nowadays the GW approximation. The GW approximation to the self-energy is the only truly *ab initio* technique which can predict band gaps with no empirical parameter. Furthermore, the GW approximation has been shown to be reliable for both delocalized states (the Bloch wavefunctions in solids) and localized states (wavefunctions in molecules). The description of point defect in crystals precisely involves these two types of electronic states: by breaking the translation invariance, the point defects induce a few rather localized states among a huge amount of crystal-like states. The GW approximation is hence a very attractive technique for the calculation of point defects.

Unfortunately, the GW approximation is an approximation for the self-energy, and therefore only provides quasiparticle eigenvalues. The quasiparticle eigenvalues are relevant for the properties of defects, since the HOMO and LUMO quasiparticle energies are, by definition, the total energy differences

$$\epsilon^{GW}_{\rm HOMO} = E_{\rm total}(0) - E_{\rm total}(+) \tag{6.1a}$$
$$\epsilon^{GW}_{\rm LUMO} = E_{\rm total}(-) - E_{\rm total}(0). \tag{6.1b}$$

In the context of a defect X^q with charge q, one sees that these definitions give access to the the vertical transitions of the defect

$$\begin{aligned} \epsilon_v(q \to q+1) &= E_{\rm total}(X^q, q) - E_{\rm total}(X^q, q+1) \\ &= \epsilon^{GW}_{\rm HOMO}(X^q, q) \qquad (6.2a) \\ \epsilon_v(q \to q-1) &= E_{\rm total}(X^q, q-1) - E_{\rm total}(X^q, q) \\ &= \epsilon^{GW}_{\rm LUMO}(X^q, q), \qquad (6.2b) \end{aligned}$$

where the atomic structures have been explicitly stated (X^q stands for the relaxed atomic structure of the defect X with charge state q). Note that these definitions are referred to the zero of the potential so far. In practical case, the results are usually presented with respect to the valence band maximum $\epsilon_{\rm VBM}$, as explained in Part I. Then the GW approximation straightforwardly gives access to the vertical transitions, which are the transitions measured with photoluminescence.

However for the thermodynamical stability, the important quantity is rather the thermodynamical transition

$$\epsilon_{\rm th}(q/q-1) = E_{\rm total}(X^{q-1}, q-1) - E_{\rm total}(X^q, q). \tag{6.3}$$

This quantity involves the total energy for two different atomic structures and therefore cannot be addressed by one single GW calculation. Since the GW method is an approximation for the self-energy, it does not give access to the total energies or to forces to relax the atomic positions. Of course, GW-like approximations do exist for the total energy but they lead to very different expressions for the total energy. One of them, the Random-Phase Approximation (RPA) energy, will be discussed here after in Chapter 8. Indeed in terms of Feynman diagrams, the energy involves other diagrams than the self-energy diagrams (the diagrams need to be closed by a line and prefactors are required). Furthermore, the RPA energy converges very slowly with the calculation parameters.

1 Introducing the DFT+GW method

In this context, Rinke et al. (2009) proposed an interesting combination of DFT and GW in order to calculate the thermodynamical quantities. It is readily achieved by inserting intermediate total energies in Eq. (6.3):

$$\begin{aligned}\epsilon_{\rm th}(q/q-1) = E_{\rm total}(X^{q-1}, q-1) - E_{\rm total}(X^{q-1}, q) \\ + E_{\rm total}(X^{q-1}, q) - E_{\rm total}(X^q, q).\end{aligned} \tag{6.4}$$

Then the vertical transition energy can be identified and evaluated within the GW approximation:

$$\epsilon_{\rm th}(q/q-1) = \epsilon^{GW}_{\rm HOMO}(X^{q-1}, q-1) + E_{\rm total}(X^{q-1}, q) - E_{\rm total}(X^q, q). \tag{6.5}$$

A total energy difference remains to be evaluated. However these total energies have both the same number of electrons. One expects that DFT will perform well in this case and will not be affected by the band gap problem. This method is the DFT+GW method.

One can immediately anticipate several problems for the presented method. First, the quality of the DFT+GW approach will depend on the choice of the exchange-correlation approximation of the DFT part. In the original paper of Rinke et al. (2009) and my subsequent articles (Bruneval and Roma, 2011;

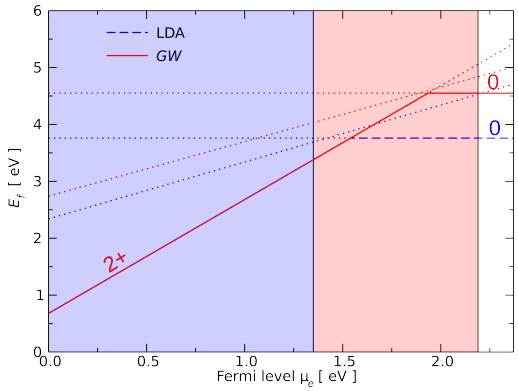

Figure 6.1: Formation energy of the carbon vacancy V_C in 3C-SiC as a function of the Fermi level μ_e in silicon-rich conditions. The dashed line stands for LDA and the solid line for DFT+GW. The dotted lines show the data used for the construction. The vertical lines delimit the conduction edge for LDA at 1.35 eV and for GW at 2.19 eV.

Bruneval, 2012b), the choice was of course LDA or PBE. Nowadays, it would be clever to use the higher quality hybrid functionals. Second, it is not granted that DFT will be accurate for the structures and for the total energy differences at fixed charge. Think of the Jahn-Teller distortions in silicon, which are poorly described within LDA or PBE (Probert and Payne, 2003). Third, the energy inserted in Eq. (6.4) was arbitrarily chosen. I could have inserted $E_{\rm total}(X^q, q-1)$ instead.

Then to obtain formation energies, a reference formation energy is still missing. Indeed, the DFT+GW method gives access only to the energetical transition $\epsilon_{\rm th}$. Rinke et al. (2009) proposed to use the DFT energy when the defect levels in the band gap are all empty and construct the other formation energies on this basis. If the reference formation energy is chosen for charge state q, the construction reads

$$E_f^{\rm DFT+\mathit{GW}}(q-1) = E_f^{\mathit{DFT}}(q) + \epsilon_{\rm th}^{\rm DFT+\mathit{GW}}(q/q-1) + \mu_e + \epsilon_{\rm VBM}, \qquad (6.6)$$

where μ_e is the Fermi level. Once again, the DFT+GW method is questionable for the choice of the reference formation energy.

I have insisted so far on the anticipated deficiencies of the DFT+GW method. Let me now show the performance of the DFT+GW for a problematic case for LDA. Figure 1 shows the formation energy of the carbon vacancy V_C in cubic SiC (3C-SiC) as obtained from 216 atom supercells. The DFT calculations have been performed with 2x2x2 **k**-point sampling, whereas the GW calculations have considered only the Γ point due to the numerical cost. Here and in the following, GW will always refer to G_0W_0 based LDA inputs (GW@LDA). It is difficult to conclude on the nature of the defect based on LDA calculations,

since all the observed transitions occur in the vicinity of the LDA conduction band edge. Note that the transitions are slightly above the LDA conduction band edge due to the use of shifted **k**-points (the conduction minimum is at Γ). Then legitimate questions arise: Is the defect level deep or shallow in reality? By fixing the band problem, would the defect transitions follow the conduction band edge or remain in the same position? These questions are answered by the DFT+*GW* technique as shown in Figure 1. The DFT+*GW* approach used the formation energy of the 2+ charge state as a reference, since then there is no occupied defect level in the band gap. The defect states indeed follow the conduction band edges. However the charge transitions do occur inside the *GW* band gap. As a consequence, the presence of numerous carbon vacancies in irradiated samples may push the Fermi level μ_e much upwards. As the defect levels are rather shallow, the use of charge corrections may not be fully justified in this case. That is why the results shown in Figure 1 have been obtained without such corrections. My DFT+*GW* results on V_C have been recently confirmed by the hybrid functional of Oda et al. (2013).

2 Assessing the DFT+*GW* method in a complex case: the carbon vacancy in SiC

The DFT+*GW* technique is a promising approach, however as written above, several technical approximations have been introduced. Here I would like to describe how the degree of freedom for the insertion of total energies in Eq. (6.4) could be rationalized for the carbon vacancy case in SiC.

This case is rather delicate, since the defect level move above the conduction band minimum for some atomic configurations. In Figure 6.2, the LDA band structure is provided for three charge states. When changing the panels, both the atomic structure and the electronic state occupation are modified. For charge state 2+, the empty defect state is far above the conduction band minimum at Γ point.

Figure 6.3 shows different paths that could be used to calculate the transition $\epsilon_{\rm th}(2+/1+)$. Two obvious paths are i) first change the charge, then the structure or ii) first change the structure and the charge. However, any other choice could be also legitimate. Imagine one uses a path through the atomic position of V_C^0:

$$\begin{aligned}
\epsilon_{\rm th}(2+/1+) &= E_{\rm total}(V_C^{1+},1+) - E_{\rm total}(V_C^{2+},2+) \\
&= E_{\rm total}(V_C^{1+},1+) - E_{\rm total}(V_C^{0},1+) \\
&\quad + E_{\rm total}(V_C^{0},1+) - E_{\rm total}(V_C^{0},2+) \\
&\quad + E_{\rm total}(V_C^{0},2+) - E_{\rm total}(V_C^{2+},2+). \qquad (6.7)
\end{aligned}$$

This path requires two structural changes (the first and third energy differences evaluated within LDA) and one charge change (the second energy difference evaluated within *GW*). A delicate manipulation comes from the fact that DFT total energy needs to be evaluated when the vacancy bears a +1 charge state

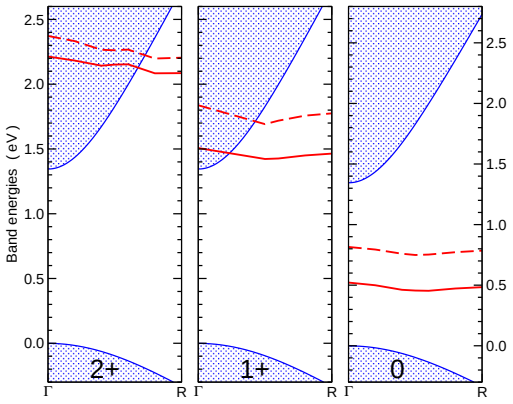

Figure 6.2: Defect levels of the carbon vacancy V_C^+ in 3C-SiC (red lines) as obtained from LDA for three different equilibrium geometries for charge states 2+, 1+, and 0, along the Γ-R line. The position of the defect state is compared to the valence and conduction bands of the pristine SiC in a 216 atom supercell drawn with the shaded areas. When the vacancy bears a 1+ charge, the defect level for spin up is occupied with one single electron (solid red line) and the defect level for spin down remains empty (dashed red line).

in the 2+ atomic configuration. As shown in the band structure plots in Figure 6.2, the defect state has to be filled, however it is above the conduction band. A standard self-consistent DFT calculation would obtain the minimum total energy with the extra electron placed in the lowest energy level available, that is the conduction band minimum. This would not reflect the charged defect situation I would have liked to described. However, with constrained occupation DFT, it is possible to enforce the occupation of the defect level, even though this is not the minimal total energy situation. Also when extracting the quasiparticle energy from the *GW* calculation with atomic configuration 2+, I had to carefully pick up the quasiparticle energy associated with the defect level and not the LUMO, which is in this case the conduction band minimum. Finally, if care is taken to populate or depopulate the defect level, the thermodynamic transition $\epsilon_{\rm th}(2+/1)$ can be alternatively obtained from any of the three paths described above with an accuracy of 0.1 eV.

This very encouraging result shows that the DFT+*GW* is a consistent method to evaluate the thermodynamical transitions of defects, even when a Jahn-Teller distortion occurs and induces a lowering of the point group symmetry of the defect, as shown in Figure 6.3).

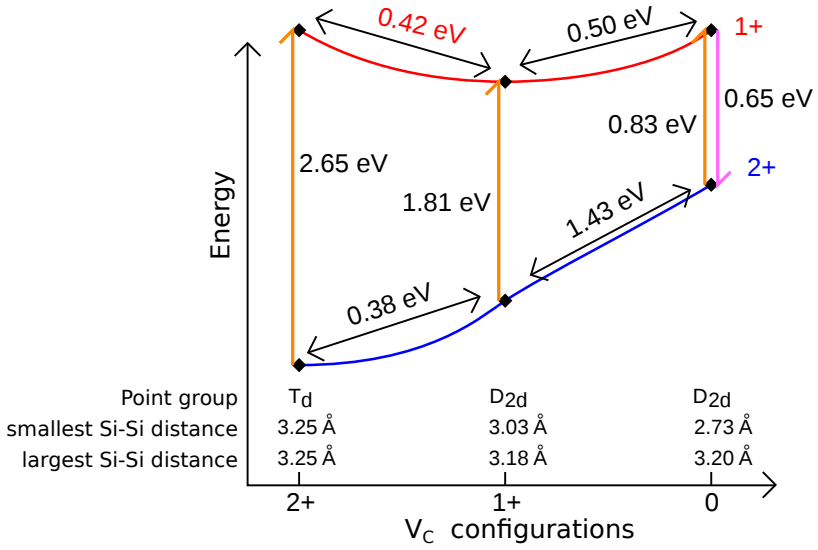

Figure 6.3: Schematic Born-Oppenheimer surfaces for charge states 2+ and 1+ of the carbon vacancy in 3C-SiC. The horizontal axis designates the three equilibrium structures for charge states 2+, 1+, and 0. The corresponding point group of the configurations as well as distances between the first neighbors of the vacancies are specified. The energy differences on the same surface are obtained from LDA, whereas the vertical transitions at constant geometry are obtained from the GW approximation. The energy values for the vertical transitions are referred to the bulk valence band maximum. The (orange) upward arrows designate the energy for adding an electron to V_C^{2+}. The (pink) downward arrow shows the energy for removing an electron to V_C^{+}. The red difference of energies has been obtained by quenching the occupations of the 2+ geometry, in order to populate precisely the defect states above the conduction edge.

3 Final remarks on DFT+GW

The present validation of the DFT+GW method is limited to one particular example. Other examples have been reported in the literature for which the combination seems to fail (Martin-Samos et al., 2010) or to be more problematic (Chen and Pasquarello, 2013).

It is common sense to understand that the smaller the structural change the more reliable the DFT+GW combination. In the limit of two charge states having the same atomic structure, DFT+GW approximation would be perfect. In the context of amorphous SiO_2, Martin-Samos et al. (2010) have pushed the DFT+GW to its limits by calculating different types of oxygen interstitials with DFT+GW. Indeed, in amorphous SiO_2, the prevailing structure of the oxygen interstitial depends on the charge state. This is a rather extreme case that DFT+GW technique does not seem to describe properly.

So far, here and in the literature, only the LDA or PBE approximations were employed in the DFT part of DFT+GW. As the hybrid functionals become wider and wider spread nowadays, I can foresee that the future of DFT+GW has to be based on these approximations. This combination will have numerous advantages: i) Better description of the relaxed structures (in particular for Jahn-Teller distortions); ii) Rarer issues with defect states outside the band gap, as in the specific case of V_C in SiC; iii) Improving the quality of the GW run, since in the G_0W_0 approach the results depends on the starting point; iv) Improved consistency between the DFT and the GW calculations, since hybrid functionals contains a part of exact-exchange, as GW does. For all the above reasons, I believe that the quality of the DFT+GW combination will improve in the next future by using hybrid functionals. However, if the hybrid functional become so much predictive, it might be that the GW step would not be crucial anymore...

Chapter 7

Concavity issue of the *GW* approximation, as exemplified for defects and atoms

This section summarizes, further explains, and exemplifies the article by Bruneval (2009) and Bruneval (2012a) printed in Appendix D.

1 A systematic inconsistency in defect levels

As previously noted, the description of the electronic structure of defects in crystal is a stringent test for the exchange-correlation approximations. Breaking the translational symmetry with a point defect induces the formation of a couple of localized electronic states among the numerous delocalized Bloch electronic states. The electronic structure method one employs has to be reliable for describing both types of electronic states co-existing in the same system. Whereas the two limiting cases are well characterized (molecules on the one hand and perfect crystals on the other hand), the defect situation is much less studied. Of course, many defect supercells have been performed since the advent of modern DFT, however there is a lack of reference values to compare with. Experimental photoluminescence values characterize very accurately the shallow donors or acceptors, which are precisely the one one cannot calculate in the supercell setup. Higher accuracy methods, such as Quantum Monte Carlo, have been so far limited to very small supercells, containing say 10-50 atoms (Leung et al., 1999; Batista et al., 2006). The situation is even more striking for *GW* calculations. Due to the numerical cost of a *GW* calculation even in a unit cell, the exploration of the performance of the *GW* approximation for defects is just beginning.

In the article Bruneval (2009), I pointed out a apparent problem of *GW* calculations when calculating the charge transition level with the DFT+*GW*

method described in the previous chapter. In this method, it is assumed that the GW approximation gives a reliable estimate of the total energy difference thanks to either of the quasiparticle energies:

$$E_{\text{total}}(N) - E_{\text{total}}(N-1) = \epsilon_{\text{HOMO}}(N) = \epsilon_{\text{LUMO}}(N-1), \tag{7.1}$$

where N is the number of electrons in the calculation. There is no theoretical reason to prefer the evaluation through $\epsilon_{\text{HOMO}}(N)$ or through $\epsilon_{\text{LUMO}}(N-1)$. Remember the Lehmann energies for the exact Green's function in Eq. (4.3). However there are some practical reason to use one rather the other, for instance to avoid spin-polarized calculations (Rinke et al., 2009). Unfortunately, when I performed the calculation in practice, I observed a discrepancy between the two values (Bruneval, 2009). This error was shown, but not commented, in the previous chapter in Figure 6.3 for the carbon vacancy in SiC. The transition at the structure V_C^0 presents two arrows: one from charge $+2$ to $+1$, corresponding to $\epsilon_{\text{LUMO}}(+2)$, and one from $+1$ to $+2$, corresponding to $\epsilon_{\text{HOMO}}(+1)$. The two energies differ by about 0.2 eV. This value is not large, however it is systematic as I observed it for several defects. This difference can be ascribed to the so-called "concavity error" of the GW approximation, that I am going to define in the next lines.

2 Concavity/convexity of the exchange-correlation approximations

As demonstrated by Perdew et al. (1982) in the beginning of the 80's, the exact exchange-correlation functional of DFT should present a piece-wise linear behavior in between in the integral numbers of electrons, as shown in Figure 7.1. This result comes from the extension of DFT to fractional numbers of electrons in an ensemble description. This demonstration has been considered for long as a very nice theoretical result for the exact DFT, but little influence on the "real world", or in other words, on quest for reliable exchange-correlation approximations. However, nowadays the practical consequences of the piece-wise linear behavior have been made obvious by the work of the group of Yang in multiple pedagogical articles (Yang et al., 2000; Mori-Sánchez et al., 2008; Cohen et al., 2008). Then the research works using the piece-wise linear idea to develop exchange-correlation approximation have flourished in the literature, to fit the U parameter in LDA+U (Cococcioni and de Gironcoli, 2005) or to fit the hybrid functional parameters (Refaely-Abramson et al., 2012; Atalla et al., 2013).

Indeed, the exact exchange-correlation functional should be linear in between the integral numbers of electrons. Discontinuities of the derivative are only allowed for integers. However in practice, all the usual exchange-correlation approximations fail badly with this property. All the semi-local approximations (LDA, PBE, etc.) and also the standard hybrid functionals are all much convex (see Figure 7.1). The Hartree-Fock approximation is concave instead. The

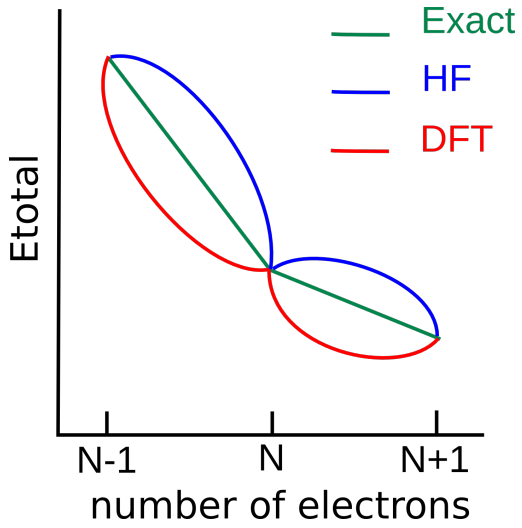

Figure 7.1: Schematic behavior of the total energy as a function of the number of electrons considered as a real number. The exact exchange-correlation functional should yield a piece-wise linear total energy. In practice, most of the DFT approximations are convex (positive curvature) and HF is concave (negative curvature).

deviation from the linear behavior for the fractional numbers of electrons is not only a benchmark for theoretical physicists: it has practical consequences on the accuracy of the exchange-correlation approximations. As first shown by Cohen et al. (2008), the fractional numbers of electrons can indeed be realized in practical cases: Imagine the example of the H_2^+ molecule for large separation of the protons. It is indeed equivalent to two protons with half an electron on each. According to the schematic energy from Figure 7.1, Hartree-Fock would minimize the total energy by having one electron on one side and no electron on the other side. This is a satisfying picture for the intuition. On the contrary, the DFT approximations would minimize the energy by having half an electron on each proton. This is a picture one intuitively does not like much. However the exact DFT functional should yield the same total energy for both situations.

With the H_2^+ example, one realizes that the convex approximations leads to rather nonphysical situations. So far, the concave approximations are not problematic and yield the same total energy as the exact functional having the linear behavior. The problems of concave approximations are less obvious to figure out. They occur when turning to infinite systems. Let us consider the case of a polymer made of n units of H_2 molecules and let the value of n increase as shown in Figure 7.2. A convex approximation, as well as the exact functional, will have the same limit for the ionization energy of an infinite chain, what ever this energy is evaluated through the total energy difference or through the

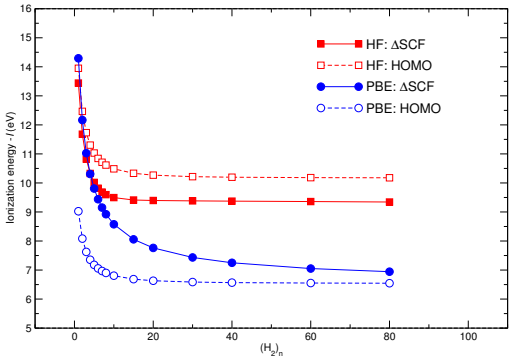

Figure 7.2: Opposite ionization potential $-I$ of a chain of aligned n H_2 molecules as a function of n, as evaluated from the total energy difference (ΔSCF procedure) $E_{\rm total}(0) - E_{\rm total}(+)$ (solid line) or from the HOMO energy $\epsilon_{\rm HOMO}(0)$ (dashed line). HF approximation is plotted with red squares, whereas PBE functional is plotted with blue circles.

HOMO eigenvalue. For an approximation suffering from a concave behavior, the two techniques will differ instead. As shown in Figure 7.2, two far apart evaluations for the HF ionization energy are obtained ($\sim$ 1 eV difference). This implies that for solid the evaluation of the band gap with the ΔSCF procedure is to differ from the eigenvalue difference. This would be quite severe problem, since some calculations, such as for charged defects, requires to compare total energy differences with eigenvalues (See e.g. Eq. (15)).

It is therefore important to look for exchange-correlation approximation that are nor convex nor concave. This is the path followed by several recent works (Refaely-Abramson et al., 2012; Atalla et al., 2013). However, I have followed the opposite path by asking: Are the known approximations convex or concave? In particular, is the GW approximation convex or concave?

3 Slight concavity of the GW approximation

3.1 How to measure the concavity/convexity for the GW approximation

Answering the question of the convexity/concavity of GW is not that straightforward, since the GW self-energy only gives access to quasiparticle energies and not to total energies. Therefore, the simple curvature test that could be performed with the total energy of PBE or HF cannot be considered here.

However, the properties of the GW quasiparticle energies can also be related to the convex/concave behavior. As introduced in the beginning of this section, in Ref. (Bruneval, 2009) I observed that the GW quasiparticle energies experience a shift when the levels are emptied or occupied. This systematic behavior

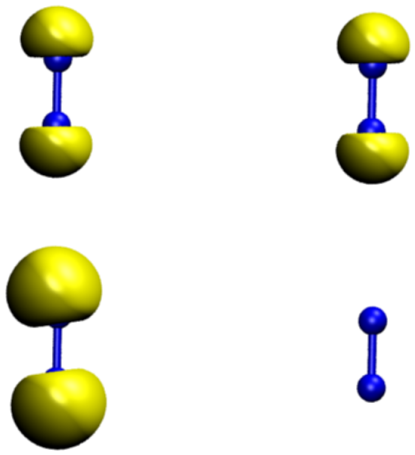

Figure 7.3: Isosurface of the HOMO of the $(2\ Na_2)^-$ system, i.e. localization of an extra electron added in the system of two distant Na_2 molecules, Upper panel represents the LDA results, whereas the lower panel shows the QSGW result.

can be linked to the concavity of the GW approximation.

Indeed, the Janak theorem (Janak, 1978) states that the eigenvalues in DFT and HF can be obtained as derivative of the total energy with respect to the occupation of the corresponding state. This translates for the particular case of the frontier orbital (HOMO or LUMO) into

$$\epsilon_f = \frac{\partial E_{\text{total}}(N)}{\partial N}, \tag{7.2}$$

where N is the number of electrons in the system. In other words, the frontier orbital energy is the tangent of the total energy curve represented in Figure 7.1.

First of all, the concavity of GW can be appreciated by performing self-consistency and looking at the frontier orbital when adding an electron in a system consisting of two well separated subsystems. This is the same numerical experiment as the one proposed by Cohen et al. (2008). In Figure 7.3, I reproduce the results of Bruneval (2009) obtained within the Quasiparticle-Self-consistent GW (QSGW) approximation to the fully self-consistent GW. The advantage of this self-consistent framework is the constrain of well-defined orthogonal wavefunctions, which helps the calculations and also the visualization. Whereas LDA had favored the situation with half an electron on each subsystem (upper panel), after self-consistency QSGW ends up with one additional electron on one subsystem and none on the other one. This shows that the GW approximation in its QSGW flavor in not convex. Then it is most likely concave, since the self-consistency was initialized with the LDA wavefunctions

with the HOMO spread on the two molecules and after iterations, the HOMO wavefunction becomes localized on one single molecule. When initializing the self-consistency with HF wavefunctions, for which the HOMO wavefunctions is already localized on one of the two molecules, QSGW HOMO wavefunction remains localized on one molecule. It is a clear sign that the GW approximation is indeed concave. If it would present the exact straight-line behavior, there would be no reason to favor the localization of the HOMO against the delocalization over the two sites.

Secondly, although I have no formal proof for that, I assumed that the Janak relation can be extended to the Green's function theory. With this assumption, the ordering of the eigenvalues should be a sign of the convexity/concavity. Looking at the schematic representation in Figure 7.1 and assuming that the quasiparticle energies can be obtained as the derivative of a corresponding total energy, a convex approximation would satisfy

$$\epsilon_{\rm LUMO}(N) < E_{\rm total}(N+1) - E_{\rm total}(N) < \epsilon_{\rm HOMO}(N+1) \tag{7.3}$$

whereas a concave approximation would have

$$\epsilon_{\rm LUMO}(N) > E_{\rm total}(N+1) - E_{\rm total}(N) > \epsilon_{\rm HOMO}(N+1). \tag{7.4}$$

Of course, the perfect piece-wise linearity would restore equalities.

In Ref. (Bruneval, 2009), I have advocated that the GW approximation is concave since its quasiparticle energies satisfy the inequality

$$\epsilon_{\rm LUMO}(N) > \epsilon_{\rm HOMO}(N+1), \tag{7.5}$$

which is a consequence of Eq. (7.4). This was demonstrated for some defect levels and for the ionization potential of isolated sodium clusters. This calculations were performed in the supercell geometry subjected to possible artifacts, as explained in Part II, and I was looking at small effects. That is why I decided to further assess the results with precise calculations on atoms (Bruneval, 2012a).

3.2 Developing an accurate GW code for isolated molecules

To that purpose, I have developed on my own a GW code for isolated atoms and molecules based on the favorite quantum chemistry basis set, namely the Gaussian basis set (Szabó and Ostlund, 1996). This choice was guided by the existing successful examples of Rohlfing (2000); Blase et al. (2011), which showed that the GW calculations are feasible and converge with small basis sets.

The GW code I developed, now named MOLGW, relies heavily on external libraries so that the most painful parts in the implementation have been avoided. In a few words, this code is capable of calculating the exact G_0W_0 result within a basis set. All the delicate technicalities of the usual GW calculations are avoided: there is no plasmon-pole model, no numerical frequency integration, no use of an auxiliary basis set to represent the dielectric matrix. Once the basis set is selected, then there is only one valid GW result. This statement

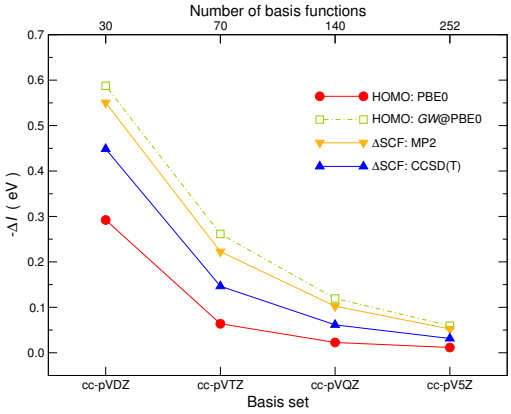

Figure 7.4: Convergence of the evaluation of the ionization potential of the CO molecule within several approximations, as the function of the basis set size in the Dunning sets. The zero has been set to the extrapolated ionization potential within each approximation.

can appear as mild, however when one knows the confused situation for crystal calculations, this is an appreciated feature. This feature is obtained at the expense of very cumbersome calculations. To be more precise, the scaling of the calculations that poses problem, both in terms of CPU time than in terms of memory. Currently, the calculations are limited to systems containing less than 300-400 basis functions, which corresponds to accurate calculations for 4-5 atom molecules or loosely converged calculations for 10-12 atoms.

Indeed, one of the first surprises I met was the rather slow convergence rate of the *GW* calculations in Gaussian basis, as demonstrated for the CO molecule in Figure 7.4. Using the correlation-consistent basis sets developed by Dunning (1989), which are well adapted to extrapolation to the complete basis set limit, For the CO molecule, the G_0W_0 energy of the HOMO is obtained with an error bar below 0.2 eV only by using the "quadruple zeta - triple polarized" basis, which consists of 140 basis functions with angular momentum up to g ($l = 4$). The convergence is noticeably slower that the convergence of the HOMO within DFT, for instance here within PBE0. However, the slow convergence is not completely unexpected, when one tries other correlated methods from quantum-chemistry such as MP2 or coupled-cluster (Szabó and Ostlund, 1996). This is code I am now able to provide *GW* results for atoms and small molecules with a 0.1 eV accuracy.

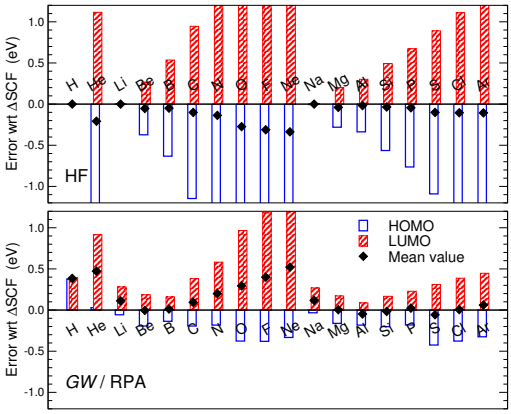

Figure 7.5: Deviation from ΔSCF reference for the atom HOMO energy ϵ^N_{HOMO} (open bars) and of the cation LUMO energy (striped bars). The mean value of the HOMO and the LUMO is displayed with the diamond symbol. The upper panel compares HF orbital energies to the HF total energy difference. The lower panel compares *GW*@HF orbital energies to the RPA@HF total energy difference.

3.3 Concavity of the *GW* approximation confirmed with atoms

In Bruneval (2012a), I have performed a systematic comparison of the HOMO or LUMO eigenvalues of the atoms and cations against the total energy difference for the first row atoms to check which of Eq. (7.3) or Eq. (7.4) is fulfilled. In Figure 7.5, the comparison is performed for both HF (upper panel) and *GW* (lower panel). Whereas for HF, which is calculated self-consistently, there is consistency between the eigenvalues and the total energy, for *GW* this is much more delicate. If one had results for the complete self-consistent *GW* framework based on the Green's function itself, everything would be consistent. However today, besides a few pioneering attempts (Stan et al., 2006; Caruso et al., 2012), *GW* results are usually obtained in the non-self-consistent manner G_0W_0 and therefore the different formulas for the total energy, Galitskii-Migdal (Galitskii and Migdal, 1958), Klein (Klein, 1961), or Luttinger-Ward (Luttinger and Ward, 1960) would give different results. For Figure 7.5, I choose to evaluate the self-energy and the total energy based on HF inputs, which are a rather good guess for atoms. Then I selected the RPA functional for the total energy (Furche, 2008), which is another name for the Klein functional. The RPA total energy will be further described in the next section.

Besides for H and He, the LUMO of the cation in red and the HOMO of the neutral atom always bracket the total energy difference for *GW*. The LUMO energy is systematically above and the HOMO energy is systematically below. This shows that *GW* indeed satisfies the inequality (7.4) and that *GW*

is a concave approximation. However, by measuring the difference between the LUMO and the HOMO, one realizes that GW is much less concave than HF. Furthermore, it appears that the concavity error is smaller for the larger atoms in the second row (from Na to Ar). Note that the underlying choice of HF for the G_0W_0 does not affect the concavity conclusions. I have also tested PBE based G_0W_0 with the same final conclusions.

4 Reconciling quasiparticle energies and total energy differences

In the usual GW calculations, there is no access to total energies. How do we reconcile the discrepancy between the LUMO of the positively charged ion and the HOMO of the neutral atom? Which value should be used? There is no theoretical reason to prefer one of them: none of this two quasiparticle energies is closer to the total energy difference.

Slater (Slater, 1974) recognized this long time ago and proposed to evaluate the total energy thanks to the eigenvalue of the frontier orbital, when it is half occupied. Indeed, if the total energy is not linear for fractional occupation numbers, then the first order correction is a second order term in N. In this second order approximation, it can easily been shown that the derivative of the total energy at a half integer is equal to the total energy difference (See Figure 7.1). This is the celebrated Slater one-half procedure. The extension of the GW approximation to fractional occupation number is not straightforward in my opinion. For example, the generalization of the RPA total energy to fractional occupation number is detailed in (Yang et al., 2013) and it is far from trivial.

I found much easier to invoke another combination inspired by Slater's trick. Under the same approximation as Slater (the total energy is a second order polynomial N in between integer), the total energy difference can alternatively be obtained as

$$E_{\rm total}(N+1) - E_{\rm total}(N) \approx \frac{1}{2}\left[\epsilon_{\rm LUMO}(N) + \epsilon_{\rm HOMO}(N+1)\right]. \qquad (7.6)$$

I propose this expression in Bruneval (2009) to reconcile the different evaluations of defect transition levels in GW. The same expression was proposed by Komsa et al. (2010) in the same period in the context of hybrid functionals.

Using this "mean value trick", I was able to improve much the agreement between the ΔSCF procedure and the quasiparticle energies in GW/RPA for atoms. In Figure 7.5, the mean value between the LUMO of the cation and the HOMO of the atom is plotted with the black diamond. Whereas the agreement with the total energy difference is still not perfect for the first row atoms, it becomes rather good for the second row. There are many possible origins for the discrepancy: a non purely second order polynomial behavior of the total energy as a function of N, the lack of self-consistency (and therefore the spread of all the possible total energy formulas)... The mean value trick improves much

the consistency between the total energy and the eigenvalues, which does not necessarily mean an improvement with respect to experiment.

Turning back to defect calculations, the previous considerations advocate for performing two *GW* calculations for each charge transition. For instance in the case of V_C in SiC studied in the previous chapter, Figure 6.3 shows the value for the LUMO in charge state +2 and the HOMO in charge +1, drawn with the upward (orange) and downward (pink) arrows. The average between the two evaluations will give the most consistent estimate for the transition $\epsilon_v(2+/1+)$. I used this approach in my subsequent *GW* study (Bruneval and Roma, 2011).

Chapter 8

RPA total energies applied to defects

This section summarizes the article by Bruneval (2012c), printed in Appendix D. I take this opportunity to provide technical details that did not fit the length requirement of the original paper.

I insisted in the previous chapters that it would be highly desirable for defects to have total energies from the GW framework. The main issue is not to have a formula for the total energy, it is rather to have only one formula! There are indeed several formulas for the total energy obtained out of a Green's function. However, these formulas give the same result only when the complete self-consistency on the Green's function has been reached. As in practice, there exists almost no study in solids with self-consistency on the Green's function. One can anticipate that this situation will persist for long for the defects, due to the supercell size required to properly account for defects.

1 Random Phase Approximation to the total energy

In the recent years, the Random Phase Approximation (RPA) formula for the total energy has started being used for real crystal calculations. This formula is a good compromise between the simplest formula, namely Galistkii-Migdal (Galitskii and Migdal, 1958), and the more intricate formulas, such as Luttinger-Ward (Luttinger and Ward, 1960) or ABL (Almbladh et al., 1999). The Galitskii-Migdal is not a stationary point with respect to the Green's function and therefore it is highly sensitive to the input in non-self-consistent calculations. On the contrary, the Luttinger-Ward and even better the ABL formulas are believed to be much stationary, which implies a weak sensitivity to the input Green's function, however at the expense of rather involved formulas. The RPA expression for the total energy (also known as the Klein formula (Klein, 1961) or as

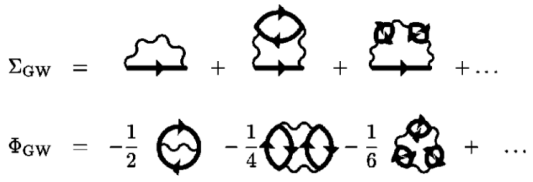

Figure 8.1: Feynman diagrams contained in the GW self-energy (upper panel) compared to the Feynman diagrams contained in the RPA correlation energy (lower panel).

the Pines-Nozières formula (Pines and Nozères, 1966)) is good compromise: it is stationary to some extent and it is not impossible to evaluate in realistic systems. Another feature of the RPA total energy has attracted a lot of attention: its ability to capture the van der Waals interactions (Dobson and Wang, 1999), which are beyond the standard DFT approximations. However, so far, all the reported applications of the RPA functional have been limited to crystal unit cells or at most to surfaces (Schimka et al., 2010).

The relation between the GW self-energy Σ_{GW} and the RPA correlation energy, label here Φ_{GW} is quite obvious in terms of Feynman diagrams, as shown in Figure 8.1. The GW self-energy is indeed Φ-derivable in the Baym-Kadanoff sense (Baym and Kadanoff, 1961):

$$\Sigma(\mathbf{r}, \mathbf{r}', \omega) = \frac{\delta \Phi_{GW}}{\delta G(\mathbf{r}', \mathbf{r}, -\omega)}, \tag{8.1}$$

which ensures that many conservation properties are indeed fulfilled by the GW self-energy. In terms of diagrams, the derivation with respect to G consists in removing one propagation line in Φ_{GW}. Owing to the numerical prefactors in Φ_{GW}, the derivative of the products gives precisely the GW expansion of the self-energy. Dropping the space and frequency indexes for clarity, the RPA correlation energy can be written as

$$\Phi_{GW} = -\frac{1}{2}\mathrm{Tr}\left[v\chi_0 + \frac{1}{2}(v\chi_0)^2 + \frac{1}{3}(v\chi_0)^3 + \dots\right], \tag{8.2}$$

where Tr stands for the trace and where the ring diagrams GG in Figure 8.1 are the independent particle polarizability χ_0 and the wiggly lines are the Coulomb interactions v. Inside the square brackets of Eq. (8.2), one can recognize the expansion series of a logarithm. This gives the final expression for Φ_{GW}:

$$\Phi_{GW} = -\frac{1}{2}\mathrm{Tr}\left[\ln(1 - v\chi_0)\right]. \tag{8.3}$$

This is this formula expressed in plane-waves that I implemented in the Abinit code. The formula for the correlation energy in Eq. (8.3) could have been

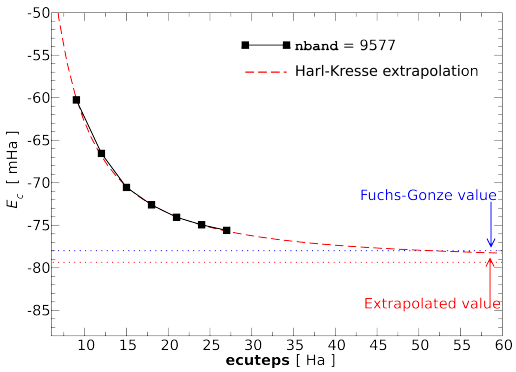

Figure 8.2: Convergence of the RPA correlation energy of a single helium atom in a box with respect to the cutoff energy for the plane-wave representation of $\chi_{0\mathbf{GG}'}$. The black square are actual calculations. The dashed line is the extrapolation fit proposed by Harl and Kresse (2008).

obtained from the integration of the coupling constant (adiabatic connection fluctuation dissipation theorem) as pedagogically explained in Refs. (Fuchs and Gonze, 2002; Niquet et al., 2003).

In practice, the calculations are performed non-self-consistently. For instance, if based on LDA, the total energy finally reads

$$E_{\text{total}}^{\text{RPA}} = E_{\text{total}}^{\text{LDA}} - E_{xc}^{\text{LDA}} + E_x + \Phi_{GW}, \tag{8.4}$$

where E_x stands for the exact-exchange energy

$$E_x = -\frac{1}{2}\sum_{ij} f_i f_j \int d\mathbf{r}d\mathbf{r}'\phi_i(\mathbf{r})\phi_i^*(\mathbf{r}')\frac{1}{|\mathbf{r}-\mathbf{r}'|}\phi_j(\mathbf{r}')\phi_j^*(\mathbf{r}). \tag{8.5}$$

Unfortunately, the RPA total energy presents a very slow convergence behavior with respect to the number of plane-waves for $\chi_{0\mathbf{GG}'}(q,\omega)$ and with respect to the number of empty states. The slow convergence is illustrated in Figure 8.2 for an isolated Helium atom. Without a fitting procedure, it would

Table 8.1: RPA correlation energy of bulk silicon at experimental lattice constant $a = 10.26$ bohr in eV/atom from different published works. QMC correlation energy is also provided for comparison.

	Marini et al. (2006)	Nguyen et al. (2009)	Harl et al. (2008)	This work	Hood et al. (1998)
		RPA			QMC
E_c	-5.44	-6.12	-6.13	-6.094	-4.08

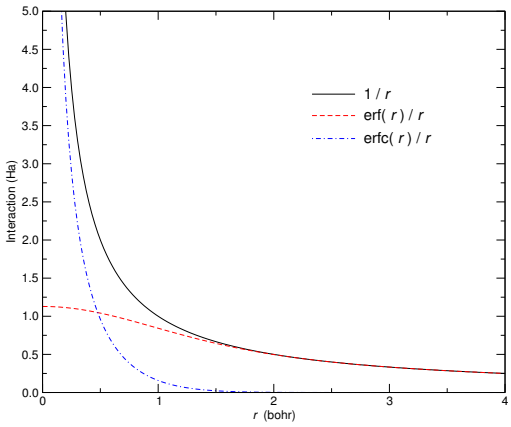

Figure 8.3: Full range Coulomb interaction $1/r$ compared to the long-range only interaction $\mathrm{erf}(r/r_c)/r$ and the short-range only interaction $\mathrm{erfc}(r/r_c)/r$. Here r_c has been set to 1 bohr.

be simply impossible to extract reasonable values: the largest calculations are still 8 % off from the extrapolated value. Note that the convergence of the absolute correlation energy is also an issue within Gaussian basis set. With the code MOLGW I described in the previous chapter the correlation energy changes by 1 mHa when moving from quintuple zeta (cc-pV5Z) to sextuple zeta basis set (cc-pV6Z)! The Gaussian basis set correlation energy (82.8 mHa) is slightly lower than the extrapolated plane-wave value.

The difficulty to converge the calculation explains the spread of the calculated correlation energies in solids. In Table 8.1, I summarized the various published data for the correlation energy of silicon within RPA based on LDA inputs, most of them are based on norm-conserving pseudopotentials, with the exception of Harl and Kresse (2008), which employs PAW. Note that the RPA correlation energy is noticeably lower than the QMC estimate. This is a known failure for the RPA correlation expression, which is well documented for the homogeneous electron gas (Vosko et al., 1980).

Owing to the convergence issues already for unit cells, it appears impossible to perform RPA correlation energy calculations for supercells containing a defect (50-100 atoms at least). The application of RPA for large cells would require further approximations. In Bruneval (2012c), I made use the range-separation idea to that particular aim.

2 Range-separation for RPA

The powerful concept of range separation has been introduced by Toulouse et al. (2004a,b). It allows one to treat each range with the most suitable theory. In

the original works, the idea was to treat the long-range with the best theory for long-range, wavefunctions methods, and the short-range with the best theory for short-range, DFT. The range separation is generally introduced with the same functional form as in Ewald summations(Ewald, 1921):

$$\frac{1}{r} = \frac{\mathrm{erf}(r/r_c)}{r} + \frac{\mathrm{erfc}(r/r_c)}{r}, \tag{8.6}$$

where the range-separation is governed by the screening length $r_c = 1/\mu$, using the notations of Toulouse et al. (2004b). The long-range/short-range splitting is shown in Figure 8.3.

The range separation was already employed in Toulouse et al. (2009) for small molecules. The idea of this work lies in the fact that RPA functional overestimates the correlation energy (see the correlation energy of silicon in Table 8.1 for instance), mainly due to errors in the short range part. In order to fix that deficiency, these authors have retained RPA for the long-range part, which is responsible for the nice account of the van der Waals interactions and selected a DFT functional based on the generalized gradient approximation for the short-range. Their idea is to cure the error of RPA in the short-range by introducing a new functional.

Incidentally, the range-splitting has some beneficial numerical consequences for the plane-wave approaches for periodic calculations. Indeed in reciprocal space, the long-range interaction becomes short-ranged in G. The Fourier transform of Eq. (8.6) reads

$$\frac{4\pi}{G^2} = \frac{4\pi}{G^2} e^{-r_c^2 G^2/4} + \frac{4\pi}{G^2}\left(1 - e^{-G^2/4r_c^2}\right) \tag{8.7}$$

In G space, the long-range Coulomb interaction has an additional exponential decaying factor. Using this mathematical fact, the convergence with respect to the plane-wave cutoff used to represent the non-interacting polarizability $\chi_{0\mathbf{GG}'}$ will be far superior when using a long-range interaction, since in Eq. (8.3), χ_0 is multiplied by the Coulomb interaction. The larger the cutoff radius r_c the faster the expected convergence.

In Bruneval (2012c), I make use of the numerical advantages of the range separation to accelerate the RPA convergence. My purpose is not to define a new functional as in Toulouse et al. (2009), but simply to reproduce the RPA correlation energy at smaller expense. In this approach, the cutoff radius r_c is to be considered as a convergence parameter: when r_c tends to zero, the original RPA results is fully recovered. The proposed expression for the total energy reads

$$E_{\mathrm{total}}^{\mathrm{RPA}} = E_{\mathrm{total}}^{\mathrm{LDA}} - E_x^{\mathrm{LDA}} - E_c^{\mathrm{LDA\ beyond\ RPA}} - E_c^{\mathrm{LDA\ RPA}}(r_c) + E_x + E_c^{\mathrm{RPA}}(r_c). \tag{8.8}$$

$E_c^{\mathrm{RPA}}(r_c)$ is obtained from the explicit calculation of Eq. (8.3) with the Coulomb interaction v replaced by the long-range only interaction. The LDA calculation for the homogeneous electron gas RPA long-range energy $E_c^{\mathrm{LDA\ RPA}}(r_c)$ is only missing piece in the energy.

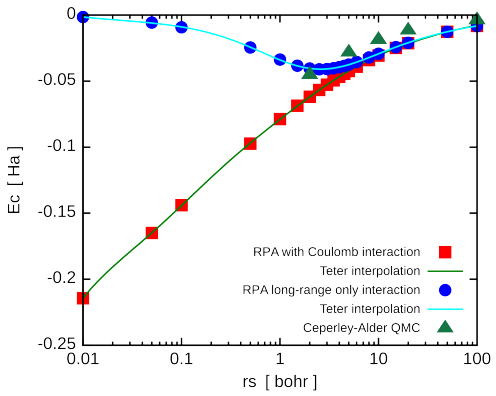

Figure 8.4: RPA correlation energy density in the homogeneous electron gas as a function of r_s, the Wigner radius. The RPA correlation energy density has been calculated for the Coulomb interaction $1/r$ and for the long-range only interaction $\mathrm{erf}(r/r_c)/r$ with $r_c = 2$ bohr.

The long-range RPA correlation energy density for the homogeneous electron gas can be obtained by straightforwardly modifying the Coulomb interaction in the RPA expression of von Barth and Hedin (1972). Then with a numerical integration, the correlation energy density can be obtained for selected values of the electronic density, or equivalently of the Wigner radius r_s (the radius of the sphere enclosing one electron). Finally, with the rational fraction proposed by Goedecker et al. (1996), the correlation energy density is interpolated to any value of r_s:

$$\epsilon_c^{r_c}(r_s) = -\frac{a_0 + a_1 r_s + a_2 r_s^2 + a_3 r_s^3}{r_s + b_2 r_s^2 + b_3 r_s^3 + b_4 r_s^4}. \tag{8.9}$$

A set of coefficients is obtained for each choice of the cutoff radius r_c. The performance of the interpolation is shown in Figure 8.4. In the low density limit (large r_s), the difference between the full RPA and the long-range only vanishes. In the high density limit (low r_s), the long-range only correlation energy vanishes, whereas the RPA correlation diverges. The turning point in the long-range only correlation density lies in the vicinity of $r_s = r_c$.

With this machinery, I am now able to address realistic calculations. Several solids have been studied in Bruneval (2012c) using much lower cutoff energies, but I would like to focus now on defects.

3 RPA results for the self-diffusion in pure silicon

The main interest of the previous technical development is the possibility to calculate larger supercells. In my study about the self-diffusion in silicon (Bruneval, 2012c), I had to calculate supercells as large as 216 atoms. This system size

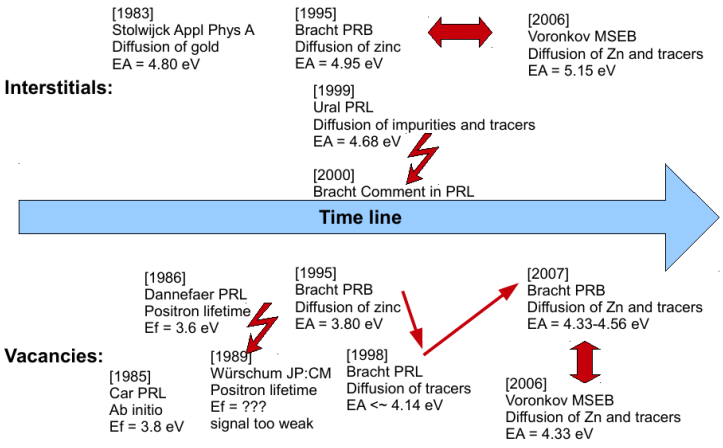

Figure 8.5: Partial review of the experimental literature concerning the silicon vacancy and self-interstitial diffusion activation energy represented as a function of time. The cited papers are (Stolwijk et al., 1983; Car et al., 1985; Dannefaer et al., 1986; Würschum et al., 1989; Bracht et al., 1995, 1998; Ural et al., 1999; Bracht and Haller, 2000; Voronkov and Falster, 2006; Bracht et al., 2007).

was absolutely required for the silicon vacancy. It is well documented that the silicon vacancy (even when they are charge neutral) converge very slowly with respect to supercell size. In particular, the too small supercell sizes prevents the formation of the physical Jahn-Teller distortion (Probert and Payne, 2003).

I first decided to work on the self-diffusion in silicon as a benchmark for the theory exposed in the previous section. However for several decades, the experimental values for the diffusion in silicon have been constantly revised, as illustrated in the literature time-line in Figure 8.5. It is difficult to extract secure values that would be the benchmark reference for my newest calculations. As a consequence, I also decided to perform hybrid functional calculations within the HSE functional (Heyd et al., 2006), as they were surprisingly missing also in the literature.

The RPA functional allows one to obtain total energies. However, to describe defects that breaks the translational ordering of crystal, having forces to relax the structures is also necessary. Within RPA, there is no expression for forces, at least nowadays. I therefore performed most of my calculations on LDA or PBE relaxed structures or saddle points. The only exception is the silicon vacancy which experiences the strong Jahn-Teller distortion that LDA or PBE functionals have difficulty to capture. I therefore preferred to optimize the energy by varying manually the two bond lengths involved the Jahn-Teller distortion and relaxing the other degrees of freedom with standard DFT. Of course, nowadays I would rather relax all the structures within HSE. But at the time when I performed the calculations, the HSE capable codes were still rare

and slow.

As explained in Part I, the self-diffusion activation energy is obtained as the sum of the barrier energy, which hinders the diffusion, and the formation energy, which governs the concentration of migrating defects [see Eq. (22)]. Furthermore, the self-diffusion can be mediated by self-interstitials or by vacancies. In a self-diffusion experiment, it is impossible to distinguish between the formation and the migration, or between self-interstitials and vacancies. To that respect, the calculations provide many useful intermediate energies.

The self-diffusion energies are summarized in Table 8.2. These data contain three original striking features. First of all, the HSE results are good agreement with the RPA results. However, they depart noticeably from the LDA or PBE results. Second, (not visible from the table) the self-interstitial migration mechanism is changed between LDA/PBE (hexagonal to hexagonal) and HSE/RPA (hexagonal to split), with potential consequences on the correlation factors (Posselt et al., 2008). Third, the activation energy of vacancies and interstitials are not so different. This is quite strong disagreement with the experimental guesses. However the experimental estimate of the vacancy contribution are only indirect (total diffusion minus self-interstitial diffusion) and are subjected to debate. Note that, as shown in Figure 8.5, the vacancy activation value has been constantly reevaluated (as higher) for the last two decades.

It is quite satisfactory to see that my first calculations have been later confirmed (within 0.2-0.3 eV) by several other works (Gao and Tkatchenko, 2013; Śpiewak and Kurzydłowski, 2013; Kresse, 2014).

Table 8.2: Summary of the calculated energies in Ref. (Bruneval, 2012c) for the description of the self-diffusion in silicon. The experimental values have extracted from (Voronkov and Falster, 2006; Bracht et al., 2007)

	LDA	PBE	HSE	RPA	Expt.
Vacancies					
E_f	3.58	3.72	–	4.33	
E_m	0.40	0.28	0.40	0.58	
E_A	3.98	5.00	–	4.91	4.33-4.56
Self-interstitials					
E_f	3.48	3.67	4.40	4.49	
E_m	0.12	0.21	0.47	0.77	
E_A	3.60	3.88	4.84	5.26	4.95-5.15

Part IV

Outlook

In this final part, I provide a few clues for what could be my research activities in the next years. Some leads are already been conducted right now, some other are left for the future. Some follow a straight-line extrapolation of my past activities, some would require deeper changes.

1 Technological applications

This memoir has focused on the methodological developments of my recent activity. However, this activity has always been accompanied with applications to technological materials. Most prominently, I have been constantly working on two particular materials SiC and ZnO having applications in the domain of the energy production/saving.

The cubic SiC is seriously considered as a coating materials to encapsulate the nuclear fuel (UO_2) for next generation nuclear plants owing to his resistance to irradiation and to his high thermal conductivity. Besides the nuclear applications, hexagonal SiC (4H-SiC) is also used for electronics and maybe tomorrow for quantum computing (Castelletto et al., 2014). I have co-authored a series of papers on defects in cubic SiC (3C-SiC) (Bruneval, 2009; Bruneval and Roma, 2011; Bruneval, 2012b). I plan to keep on working on that material, especially for its intricate spin-states I will further describe in Section 3.

ZnO is already been used in the thin film photovoltaic cells. The n-type ZnO is used as a transparent conductive oxide for the front electrode. However its usage would be even more important if one could also obtain the reverse doping, namely p-type ZnO. I have devoted quite some time so far to the study of p-type doping in ZnO, with increasing accuracy methods Cui and Bruneval (2010); Gabás et al. (2011); Petretto and Bruneval (2014). Unfortunately, the conclusions of the *ab initio* calculations have constantly pointed out that p-type doping cannot be achieved. I guess most of the obvious doping elements have been tried, both experimentally and theoretically. I have the feeling that the only chances of success rely now in heterodox solutions. To that aim, the systematic searches techniques developed in many different groups in the world, such as the "Materials project" just to name one, appear as the only viable route.

2 Improving the DFT+GW with hybrid functionals for the DFT level

In the direct continuation of the combination of GW with DFT for defects, one straightforward improvement of the technique would be the introduction of hybrid functionals in the DFT level. The DFT+GW (Rinke et al., 2009) has been so far always applied in combination with the old-fashioned DFT functionals, such as LDA or GGA Martin-Samos et al. (2010); Bruneval and Roma (2011).

Introducing much more accurate hybrid functionals would have many benefits:

- The perturbative G_0W_0 procedure performs better in general when starting from hybrid functionals, since G_0W_0 employs the starting wavefunctions and eigenvalues without any further self-consistency. The importance of a better starting point has been demonstrated for complex oxides (Isseroff and Carter, 2012) and for molecules (Bruneval and Marques, 2013).

- As the expression of the hybrid functionals is closer to GW (they both include some content of exact exchange), the problematic combination of GW and DFT (GW for charge changes and DFT for structural changes) can be believed to be more consistent.

- The relaxed structure are arguably better within hybrid functionals than within semi-local functionals. For instance, the Jahn-Teller distortion are difficult to obtain within PBE (Bruneval, 2012c)

- The GW formation energies are finally obtained by considering a DFT formation for a given charge state as the reference. The formation energy within hybrid functional will certainly be more accurate.

With the availability of hybrid functional codes, this step is now within reach with very little effort. Of course, careful comparison for a wide range of defects should be performed to evaluate the need to perform the final GW step on top of the already rather accurate hybrid functional calculations.

3 Spin states in open-shell defects

There are a few interesting defects, whose electronic structure cannot be described by standard DFT and GW. These two frameworks are based on a mean-field assumption that imposes the electronic wavefunctions to be a single Slater determinant. In isolated systems, there are many well documented situation in which the single Slater determinant is insufficient. For instance, the stretched H_2 molecule (with only two electrons) already shows the failure of the single Slater determinant assumption. The stretched H_2 would minimize its energy by having one electron on each proton. When performing spin-restricted calculations, the spin-up and spin-down wavefunctions are constrained to remain equal and as a consequence the energy does not tend correctly to twice the isolated hydrogen value when the distance between the protons goes to infinity. A solution would be to break the spin restriction constraint: then the correct limiting energy is obtained for large separations, however that the total spin (S^2 observable) is incorrect: it is not any more a pure spin singlet. This problem, named "static correlation" by quantum chemists, is also present in the context of point defects.

Several technologically relevant defects experience the static correlation problem. This generally occurs when several states with the same energy can be populated or not. In other words, one has to deal with an open electronic shell. Let me name a few defects that are affected by this puzzling situation. The color of gemstones is generally given by d element impurities in the matrix. For instance, in ruby, these are chromium impurities in alumina Al_2O_3. The optical transitions giving rise the red color originate from inner shell Cr3d electronic transitions (which are dipole forbidden transitions). An accurate description of these d states is required, however the single Slater determinant can only capture the high spin configuration. All the other spin arrangements are out of reach. The same open-shell situation however happens without the need of d electrons. A single vacancy in diamond or in SiC has a multi-determinental ground-state. In the quantum computing field, researchers plan to use the NV^- defects in diamond (a complex made of a vacancy and of a nitrogen impurity) to store the quantum bit. The quantum computing experiments play with the long-lived triplet spin-state that cannot spontaneously decay into spin singlet. All these situations can only be described with several Slater determinants.

As of today, these defects are not treated properly. I would like to devote some efforts to the incorporation of multi-determinental techniques in the defect context. Whereas the solution to static correlation exists for atoms and small molecules, namely with configuration interaction (CI) and related approaches, they cannot be used immediately for defects in supercells, which contains several hundreds of electrons. I think that defining effective interactions could make the accurate quantum-chemistry calculations on a smaller system feasible. There exists techniques to evaluate the effective interaction from *ab initio* calculations. Constrained Random Phase Approximation is one of them (Aryasetiawan et al., 2004). I have recently participated to a study on that particular topic (Amadon et al., 2014). With reliable effective interactions, one could in principle treat the correlation only for those defect orbitals. This proposed approach bears similarities with a simplified version of Dynamical Mean-Field Theory (DMFT) (Georges et al., 1996).

4 Shallow defects

Besides very rare studies(Zhang et al., 2013), the shallow defects are always disregarded in *ab initio* supercell studies. The accurate evaluation of the charge transition level of these defects is of high technological relevance, since the efficiency of doping is tightly related to this quantity. But as the defect wavefunction is extended, it is almost impossible to fit it in a tractable supercell. Furthermore, to compete with experimental measurements, the requested accuracy for the charge transition levels is much higher that for deep defects. The targeted accuracy should yield error bars of less than 50 meV.

In my opinion, it is very important to address this family of defects. However this cannot be done with brute force calculations. A change of computational framework should happen. Right now I can foresee two interesting possibilities.

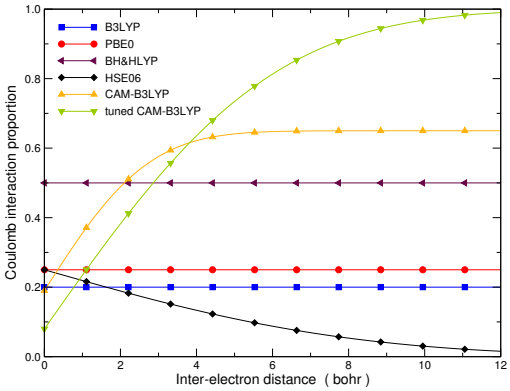

Figure 8.6: Coulomb interaction proportion $\alpha + \beta \mathrm{erf}(\omega r)$ effectively used by different hybrid functionals of DFT as a function of the inter-electron distance r. On this plot, HF would be an horizontal line at 1.0 and LDA, PBE would be an horizontal line at 0.0. Hybrid functionals offer a wide variety of choice in between these two extremes: full range, short-range, or long-range.

Either the order N electronic structure methods (Goedecker, 1999) will become competitive and reliable, or the Green's function methods, such as KKR (Ebert et al., 2011), will allow the clean treatment of dopants considered as impurities. Considering my own background in Green's function theory, I may consider devoting myself to the latter possibility in the future.

5 Confronting hybrid functionals with GW

Hybrid functionals have still a limited accuracy for solids. Many authors (including me) resort to a further fitting of the exact exchange mixing parameter α in order to improve the description of a chosen physical quantity (in general the band gap). On the other hand, the GW calculations are usually more reliable, but at a much higher computational cost.

Hybrid functionals and GW approximations share some common aspects. Both include exact exchange and then reduce it with a screening procedure. In GW calculations, the screening is calculated *ab initio* within RPA. The obtained screening contains "local-fields" (it is a function of two separated space indexes $\mathbf{r}$ and $\mathbf{r}'$) and it is dynamical (the screening is more efficient around the plasmon frequency for instance). The hybrid functional screening is much cruder. Each hybrid functional flavor (e.g. B3LYP, PBE0, HSE06, TD-CAM-B3LYP etc.) makes a choice for the screening amount and keeps it constant whatever the material or the molecule. As plotted in Figure 8.6, PBE0 chooses to consider 25 % of the exact exchange, which means a 75 % screening. B3LYP selects a 20 % amount of exact-exchange. HSE06 and CAM-B3LYP are more involved

since they introduce also a range-dependent screening. HSE06 retains only 25 % of short-range exact-exchange and 0 % of long-range. TD-CAM-B3LYP makes the reverse assumption: the full long-range exact-exchange is retained and a very strong screening is assumed at short-range (Okuno et al., 2012). In other words, all the hybrid functionals are not dynamic and do not include local-fields (they are a simple function of $|\mathbf{r} - \mathbf{r}'|$)

I think the careful comparison between GW and hybrid functionals could induce improvements and introduce novel ideas in two directions. As the hybrid functionals are already very successful, their approximations should bear some meaningful physical content. Obtaining information from the successes of hybrid functionals should help speeding up the GW calculations by designing approximations to GW that anyway conserve most of the physical content. The direction goes towards an improvement of hybrid functionals, making them more complicated but more accurate. By comparing GW and hybrid functionals, it could bring some of the missing material dependence in the framework of hybrid functionals.

Acknowledgments

It is a great pleasure for me to arrive at the acknowledgments part. Not only because, at this stage, the manuscript is at last finalized, but also mainly because research is a collaborative job. Although the writing of the "Habilitation" is an individual task, the underlying studies have involved a lot of different people, students, colleagues, collaborators, friends, and bosses. Note that the different categories have non-vanishing overlaps. I would seize one of these rare opportunities to express how much I enjoyed working with them.

Having a permanent position in a public research institution, such as CEA, is invaluable. For hiring me in his lab and trusting me during these 7 years, I would like to thank François Willaime, head of the Service de Recherches de Métallurgie Physique at CEA Saclay.

My colleagues at SRMP, especially the semiconductors guys, namely Jean-Paul Crocombette and Guido Roma, have been influential on my recent years activities. I am indebted to Cosmin Marinica, Laurent Proville, Lisa Ventelon and Emmanuel Clouet, for explaining me the details of quantum statistical physics and/or the life and death of dislocations in materials, during the coffee breaks.

I have to name a bunch of brave young people that, consciously or not, have taken the risk working under my supervision: Ying Cui, Samuel Taylor, Hichem Ben Hamed, Arnaud Bourrasseau, Guido Petretto, and Abdullah Shukri. Your courage shall not be forgotten.

Outside my lab, I would like to stress the influence on my research of two very active research communities: the ABINIT software community, with most prominently for me, Xavier Gonze, Gian-Marco Rignanese, Matteo Giantomassi, Bernard Amadon, François Jollet, Marc Torrent; and the ETSF research network, with most importantly for me Lucia Reining, Silvana Botti, Miguel Marques, Matteo Gatti, Francesco Sottile. I thank you all for your availability and your kindness.

I am honored to thank the members of the jury for the "Habilitation" defense that all kindly and immediately accepted to sign up: Alfredo Pasquarello, Silvana Botti, Xavier Blase, Jean-Paul Crocombette, Bruno Masenelli, Gian-Marco Rignanese, and Alfonso San Miguel.

No words are necessary for Agnès, Clément, and Simon. They already know.

Bibliography

Adamo, C. and Barone, V. (1999). Toward reliable density functional methods without adjustable parameters: The pbe0 model. *J. Chem. Phys.*, 110(13):6158–6170.

Adler, S. L. (1963). Theory of the range of hot electrons in real metals. *Phys. Rev.*, 130:1654–1666.

Almbladh, C.-O., Barth, U. V., and Leeuwen, R. V. (1999). Variational total energies from Φ- and Ψ- derivable theories. *International Journal of Modern Physics B*, 13:535–541.

Amadon, B., Applencourt, T., and Bruneval, F. (2014). Screened coulomb interaction calculations: cRPA implementation and applications to dynamical screening and self-consistency in uranium dioxide and cerium. *Phys. Rev. B*, 89:125110.

Anglade, P.-M. and Gonze, X. (2008). Preconditioning of self-consistent-field cycles in density-functional theory: The extrapolar method. *Phys. Rev. B*, 78:045126.

Aryasetiawan, F. and Gunnarsson, O. (1998). The *GW* method. *Rep. Prog. Phys.*, 61:237–312.

Aryasetiawan, F., Imada, M., Georges, A., Kotliar, G., Biermann, S., and Lichtenstein, A. I. (2004). Frequency-dependent local interactions and low-energy effective models from electronic structure calculations. *Phys. Rev. B*, 70:195104.

Atalla, V., Yoon, M., Caruso, F., Rinke, P., and Scheffler, M. (2013). Hybrid density functional theory meets quasiparticle calculations: A consistent electronic structure approach. *Phys. Rev. B*, 88:165122.

Aulbur, W. G., Jönsson, L., and Wilkins, J. W. (2000). Quasiparticle calculations in solids. *Solid State Phys.*, 54:1.

Batista, E. R., Heyd, J., Hennig, R. G., Uberuaga, B. P., Martin, R. L., Scuseria, G. E., Umrigar, C. J., and Wilkins, J. W. (2006). Comparison of screened hybrid density functional theory to diffusion monte carlo in calculations of total energies of silicon phases and defects. *Phys. Rev. B*, 74:121102.

Baym, G. and Kadanoff, L. P. (1961). Conservation laws and correlation functions. *Phys. Rev.*, 124:287–299.

Becke, A. D. (1993). Density-functional thermochemistry. iii. the role of exact exchange. *J. Chem. Phys.*, 98(7):5648–5652.

Berger, J. A., Reining, L., and Sottile, F. (2010). Ab initio calculations of electronic excitations: Collapsing spectral sums. *Phys. Rev. B*, 82:041103.

Bergersen, B., Kus, F. W., and Blomberg, C. (1973). Single particle green's function in the electron–plasmon approximation. *Canadian Journal of Physics*, 51(1):102–110.

Blase, X., Attaccalite, C., and Olevano, V. (2011). First-principles *GW* calculations for fullerenes, porphyrins, phtalocyanine, and other molecules of interest for organic photovoltaic applications. *Phys. Rev. B*, 83:115103.

Blöchl, P. E. (1994). Projector augmented-wave method. *Phys. Rev. B*, 50:17953–17979.

Blomberg, C. and Bergersen, B. (1972). Spurious structure from approximations to the dyson equation. *Canadian Journal of Physics*, 50(19):2286–2293.

Bourgoin, J. and Lannoo, M. (1983). *Point defects in semiconductors: Experimental aspects*. Springer series in solid-state sciences. Springer-Verlag.

Bracht, H. and Haller, E. E. (2000). Comment on "self-diffusion in silicon: Similarity between the properties of native point defects". *Phys. Rev. Lett.*, 85:4835–4835.

Bracht, H., Haller, E. E., and Clark-Phelps, R. (1998). Silicon self-diffusion in isotope heterostructures. *Phys. Rev. Lett.*, 81:393–396.

Bracht, H., Silvestri, H. H., Sharp, I. D., and Haller, E. E. (2007). Self- and foreign-atom diffusion in semiconductor isotope heterostructures. ii. experimental results for silicon. *Phys. Rev. B*, 75:035211.

Bracht, H., Stolwijk, N. A., and Mehrer, H. (1995). Properties of intrinsic point defects in silicon determined by zinc diffusion experiments under nonequilibrium conditions. *Phys. Rev. B*, 52:16542–16560.

Bruneval, F. (2009). *GW* approximation of the many-body problem and changes in the particle number. *Phys. Rev. Lett.*, 103:176403.

Bruneval, F. (2012a). Ionization energy of atoms obtained from *GW* self-energy or from random phase approximation total energies. *J. Chem. Phys.*, 136(19):–.

Bruneval, F. (2012b). Methodological aspects of the *GW* calculation of the carbon vacancy in 3C-SiC. *Nucl. Instrum. Meth. B*, 277:77.

Bruneval, F. (2012c). Range-separated approach to the RPA correlation applied to the van der waals bond and to diffusion of defects. *Phys. Rev. Lett.*, 108:256403.

Bruneval, F., Crocombette, J.-P., Gonze, X., Dorado, B., Torrent, M., and Jollet, F. (2014). Consistent treatment of charged systems within periodic boundary conditions: The projector augmented-wave and pseudopotential methods revisited. *Phys. Rev. B*, 89:045116.

Bruneval, F. and Gatti, M. (2014). Quasiparticle self-consistent *GW* method for the spectral properties of complex materials. In Di Valentin, C., Botti, S., and Cococcioni, M., editors, *First Principles Approaches to Spectroscopic Properties of Complex Materials*, Topics in Current Chemistry, pages 1–37. Springer Berlin Heidelberg.

Bruneval, F. and Gonze, X. (2008). Accurate *GW* self-energies in a plane-wave basis using only a few empty states: Towards large systems. *Phys. Rev. B*, 78:085125.

Bruneval, F. and Marques, M. A. L. (2013). Benchmarking the starting points of the gw approximation for molecules. *J. Chem. Theory Comput.*, 9(1):324–329.

Bruneval, F. and Roma, G. (2011). Energetics and metastability of the silicon vacancy in cubic sic. *Phys. Rev. B*, 83:144116.

Car, R., Kelly, P. J., Oshiyama, A., and Pantelides, S. T. (1985). Microscopic theory of impurity-defect reactions and impurity diffusion in silicon. *Phys. Rev. Lett.*, 54:360–363.

Caruso, F., Rinke, P., Ren, X., Scheffler, M., and Rubio, A. (2012). Unified description of ground and excited states of finite systems: The self-consistent *GW* approach. *Phys. Rev. B*, 86:081102.

Castelletto, S., Johnson, B. C., Iviády, V., Stavrias, N., Umeda, T., Gali, A., and Ohshima, T. (2014). A silicon carbide room-temperature single-photon source. *Nature Mater.*, 13:151–156.

Chen, W. and Pasquarello, A. (2013). Correspondence of defect energy levels in hybrid density functional theory and many-body perturbation theory. *Phys. Rev. B*, 88:115104.

Cococcioni, M. and de Gironcoli, S. (2005). Linear response approach to the calculation of the effective interaction parameters in the lda+u method. *Phys. Rev. B*, 71:035105.

Cohen, A. J., Mori-Sánchez, P., and Yang, W. (2008). Insights into current limitations of density functional theory. *Science*, 321:792.

Cui, Y. and Bruneval, F. (2010). p-type doping and codoping of zno based on nitrogen is ineffective: An ab initio clue. *Applied Physics Letters*, 97(4):042108.

Dannefaer, S., Mascher, P., and Kerr, D. (1986). Monovacancy formation enthalpy in silicon. *Phys. Rev. Lett.*, 56:2195–2198.

Dobson, J. F. and Wang, J. (1999). Successful test of a seamless van der waals density functional. *Phys. Rev. Lett.*, 82:2123–2126.

Dunning, T. H. (1989). Gaussian basis sets for use in correlated molecular calculations. i. the atoms boron through neon and hydrogen. *The Journal of Chemical Physics*, 90(2):1007–1023.

Ebert, H., Ködderitzsch, D., and Minár, J. (2011). Calculating condensed matter properties using the kkr-green's function method—recent developments and applications. *Rep. Prog. Phys.*, 74(9):096501.

Ewald, P. P. (1921). Die berechnung optischer und elektrostatischer gitterpotentiale. *Annalen der Physik*, 369(3):253–287.

Fetter, A. L. and Walecka, J. D. (1971). *Quantum theory of Many-Particle Systems*. MacGraw-Hill, New York.

Freysoldt, C., Neugebauer, J., and Van de Walle, C. G. (2009). Fully ab initio finite-size corrections for charged-defect supercell calculations. *Phys. Rev. Lett.*, 102:016402.

Fuchs, M. and Gonze, X. (2002). Accurate density functionals: Approaches using the adiabatic-connection fluctuation-dissipation theorem. *Phys. Rev. B*, 65:235109.

Furche, F. (2008). Developing the random phase approximation into a practical post-kohn–sham correlation model. *The Journal of Chemical Physics*, 129(11):114105.

Gabás, M., Torelli, P., Barrett, N. T., Sacchi, M., Bruneval, F., Cui, Y., Simonelli, L., Díaz-Carrasco, P., and Ramos Barrado, J. R. (2011). Direct observation of al-doping-induced electronic states in the valence band and band gap of zno films. *Phys. Rev. B*, 84:153303.

Galitskii, V. M. and Migdal, A. B. (1958). Application of quantum field theory methods to the many body problem. *Sov. Phys. JETP*, 139:96.

Gao, W. and Tkatchenko, A. (2013). Electronic structure and van der waals interactions in the stability and mobility of point defects in semiconductors. *Phys. Rev. Lett.*, 111:045501.

Georges, A., Kotliar, G., Krauth, W., and Rozenberg, M. J. (1996). Dynamical mean-field theory of strongly correlated fermion systems and the limit of infinite dimensions. *Rev. Mod. Phys.*, 68:13–125.

Giannozzi, P., Baroni, S., Bonini, N., Calandra, M., Car, R., Cavazzoni, C., Ceresoli, D., Chiarotti, G. L., Cococcioni, M., Dabo, I., Corso, A. D., de Gironcoli, S., Fabris, S., Fratesi, G., Gebauer, R., Gerstmann, U., Gougoussis, C., Kokalj, A., Lazzeri, M., Martin-Samos, L., Marzari, N., Mauri, F., Mazzarello, R., Paolini, S., Pasquarello, A., Paulatto, L., Sbraccia, C., Scandolo, S., Sclauzero, G., Seitsonen, A. P., Smogunov, A., Umari, P., and Wentzcovitch, R. M. (2009). Quantum espresso: a modular and open-source software project for quantum simulations of materials. *J. Phys. Condens. Matter*, 21(39):395502.

Godby, R. W. and Needs, R. J. (1989). Metal-insulator transition in kohn-sham theory and quasiparticle theory. *Phys. Rev. Lett.*, 62(10):1169–1172.

Goedecker, S. (1999). Linear scaling electronic structure methods. *Rev. Mod. Phys.*, 71:1085–1123.

Goedecker, S., Teter, M., and Hutter, J. (1996). Separable dual-space gaussian pseudopotentials. *Phys. Rev. B*, 54:1703–1710.

Gonze, X. (1997). First-principles responses of solids to atomic displacements and homogeneous electric fields: Implementation of a conjugate-gradient algorithm. *Phys. Rev. B*, 55:10337–10354.

Gonze, X., Amadon, B., Anglade, P. M., Beuken, J. M., Bottin, F., Boulanger, P., Bruneval, F., Caliste, D., Caracas, R., Cote, M., Deutsch, T., Genovese, L., Ghosez, P., Giantomassi, M., Goedecker, S., Hamann, D. R., Hermet, P., Jollet, F., Jomard, G., Leroux, S., Mancini, M., Mazevet, S., Oliveira, M. J. T., Onida, G., Pouillon, Y., Rangel, T., Rignanese, G. M., Sangalli, D., Shaltaf, R., Torrent, M., Verstraete, M. J., Zerah, G., and Zwanziger, J. W. (2009). Abinit: First-principles approach to material and nanosystem properties. *Comput. Phys. Commun.*, 180:2582.

Grosso, G. and Pastori Paravicini, G. (2000). *Solid State Physics*. Academic Press.

Guzzo, M., Lani, G., Sottile, F., Romaniello, P., Gatti, M., Kas, J. J., Rehr, J. J., Silly, M. G., Sirotti, F., and Reining, L. (2011). Valence electron photoemission spectrum of semiconductors: *Ab Initio* description of multiple satellites. *Phys. Rev. Lett.*, 107:166401.

Harl, J. and Kresse, G. (2008). Cohesive energy curves for noble gas solids calculated by adiabatic connection fluctuation-dissipation theory. *Phys. Rev. B*, 77:045136.

Hedin, L. (1965). New method for calculating the one-particle green's function with application to the electron-gas problem. *Phys. Rev.*, 139:A796–A823.

Hedin, L. (1999). On correlation effects in electron spectroscopies and the *GW* approximation. *J. Phys. Condens. Matter*, 11(42):R489.

Hedin, L., Lundqvist, B., and Lundqvist, S. (1967). New structure in the single-particle spectrum of an electron gas. *Solid State Commun.*, 5(4):237 – 239.

Hedin, L. and Lundqvist, S. (1970). Effects of electron-electron and electron-phonon interactions on the one-electron states of solids. In Frederick Seitz, D. T. and Ehrenreich, H., editors, , volume 23 of *Solid State Physics*, pages 1 – 181. Academic Press.

Heyd, J., Scuseria, G. E., and Ernzerhof, M. (2006). Erratum: "hybrid functionals based on a screened coulomb potential" [j. chem. phys.118, 8207 (2003)]. *The Journal of Chemical Physics*, 124(21):–.

Hohenberg, P. and Kohn, W. (1964). Inhomogeneous electron gas. *Phys. Rev.*, 136:B864–B871.

Hood, R. Q., Chou, M. Y., Williamson, A. J., Rajagopal, G., and Needs, R. J. (1998). Exchange and correlation in silicon. *Phys. Rev. B*, 57:8972–8982.

Hybertsen, M. S. and Louie, S. G. (1985). First-principles theory of quasiparticles: Calculation of band gaps in semiconductors and insulators. *Phys. Rev. Lett.*, 55:1418–1421.

Hybertsen, M. S. and Louie, S. G. (1986). Electron correlation in semiconductors and insulators: Band gaps and quasiparticle energies. *Phys. Rev. B*, 34:5390–5413.

Ihm, J., Zunger, A., and Cohen, M. L. (1979). Momentum-space formalism for the total energy of solids. *Journal of Physics C: Solid State Physics*, 12(21):4409.

Ishii, S., Iwata, S., and Ohno, K. (2010). All-electron GW calculations of silicon, diamond, and silicon carbide. *Mater. Trans.*, 51(12):2150–2156.

Isseroff, L. Y. and Carter, E. A. (2012). Importance of reference hamiltonians containing exact exchange for accurate one-shot GW calculations of Cu_2O. *Phys. Rev. B*, 85:235142.

Janak, J. F. (1978). Proof that $\partial E/\partial n_i = \epsilon_i$ in density-functional theory. *Phys. Rev. B*, 18:7165–7168.

Karch, K., Bechstedt, F., Pavone, P., and Strauch, D. (1996). Pressure-dependent dynamical and dielectric properties of cubic sic. *J. Phys. Condens. Matter*, 8(17):2945.

Klein, A. (1961). Perturbation theory for an infinite medium of fermions. ii. *Phys. Rev.*, 121:950–956.

Kohn, W. and Sham, L. J. (1965). Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140:A1133–A1138.

Komsa, H.-P., Broqvist, P., and Pasquarello, A. (2010). Alignment of defect levels and band edges through hybrid functionals: Effect of screening in the exchange term. *Phys. Rev. B*, 81:205118.

Komsa, H.-P., Rantala, T. T., and Pasquarello, A. (2012). Finite-size supercell correction schemes for charged defect calculations. *Phys. Rev. B*, 86:045112.

Kresse, G. (2014). private communication.

Kresse, G. and Furthmüller, J. (1996). Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set. *Phys. Rev. B*, 54:11169–11186.

Kresse, G. and Joubert, D. (1999). From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B*, 59:1758–1775.

Krieger, J. B., Li, Y., and Iafrate, G. J. (1990). Exact relations in the optimized effective potential method employing an arbitrary $E_{xc}[\psi_{i\sigma}]$. *Phys. Lett. A*, 146:256.

Kümmel, S. and Kronik, L. (2008). Orbital-dependent density functionals: Theory and applications. *Rev. Mod. Phys.*, 80:3–60.

Lannoo, M. and Bourgoin, J. (1981). *Point Defects in Semiconductors: Theoretical aspects*. Springer Series in Solid-State Sciences. Springer-Verlag.

Lany, S. and Zunger, A. (2008). Assessment of correction methods for the band-gap problem and for finite-size effects in supercell defect calculations: Case studies for zno and gaas. *Phys. Rev. B*, 78:235104.

Lebègue, S., Arnaud, B., Alouani, M., and Bloechl, P. E. (2003). Implementation of an all-electron *GW* approximation based on the projector augmented wave method without plasmon pole approximation: Application to Si, SiC, AlAs, InAs, NaH, and KH. *Phys. Rev. B*, 67(15):155208.

Leslie, M. and Gillan, M. J. (1985). The energy and elastic dipole tensor of defects in ionic crystals calculated by the supercell method. *J. Phys. C: Solid State Phys.*, 18:973.

Leung, W.-K., Needs, R. J., Rajagopal, G., Itoh, S., and Ihara, S. (1999). Calculations of silicon self-interstitial defects. *Phys. Rev. Lett.*, 83:2351–2354.

Lundqvist, B. (1967). Single-particle spectrum of the degenerate electron gas. *Physik der Kondensierten Materie*, 6(3):193–205.

Luttinger, J. M. and Ward, J. C. (1960). Ground-state energy of a many-fermion system. ii. *Phys. Rev.*, 118:1417–1427.

Mahan, G. D. (2000). *Many-Particle Physics*. Kluwer Academic/Plenum Publishers, 3rd edition edition.

Makov, G. and Payne, M. C. (1995). Periodic boundary conditions in ab initio calculations. *Phys. Rev. B*, 51:4014–4022.

Marini, A., García-González, P., and Rubio, A. (2006). First-principles description of correlation effects in layered materials. *Phys. Rev. Lett.*, 96:136404.

Martin, R. M. (2004). *Electronic Structure: Basic Theory and Practical Methods (Vol 1)*. Cambridge University Press.

Martin-Samos, L., Roma, G., Rinke, P., and Limoge, Y. (2010). Charged oxygen defects in sio_2: Going beyond local and semilocal approximations to density functional theory. *Phys. Rev. Lett.*, 104:075502.

Møller, C. and Plesset, M. S. (1934). Note on an approximation treatment for many-electron systems. *Phys. Rev.*, 46:618–622.

Mori-Sánchez, P., Cohen, A. J., and Yang, W. (2008). Localization and delocalization errors in density functional theory and implications for band-gap prediction. *Phys. Rev. Lett.*, 100:146401.

Nguyen, H.-V. and de Gironcoli, S. (2009). Efficient calculation of exact exchange and RPA correlation energies in the adiabatic-connection fluctuation-dissipation theory. *Phys. Rev. B*, 79:205114.

Niquet, Y. M., Fuchs, M., and Gonze, X. (2003). Exchange-correlation potentials in the adiabatic connection fluctuation-dissipation framework. *Phys. Rev. A*, 68:032507.

Oda, T., Zhang, Y., and Weber, W. J. (2013). Study of intrinsic defects in 3c-sic using first-principles calculation with a hybrid functional. *J. Chem. Phys.*, 139(12):–.

Okuno, K., Shigeta, Y., Kishi, R., Miyasaka, H., and Nakano, M. (2012). Tuned cam-b3lyp functional in the time-dependent density functional theory scheme for excitation energies and properties of diarylethene derivatives. *Journal of Photochemistry and Photobiology A: Chemistry*, 235(0):29 – 34.

Paier, J., Marsman, M., Hummer, K., Kresse, G., Gerber, I. C., and Ángyán, J. G. (2006). Screened hybrid density functionals applied to solids. *J. Chem. Phys.*, 124(15):154709.

Parr, R. G. and Yang, W. (1989). *Density-Functional Theory of Atoms and Molecules*. Oxford University Press, New York.

Payne, M. C., Teter, M. P., Allan, D. C., Arias, T. A., and Joannopoulos, J. D. (1992). Iterative minimization techniques for *ab initio* total-energy calculations: molecular dynamics and conjugate gradients. *Rev. Mod. Phys.*, 64:1045–1097.

Peelaers, H., Partoens, B., Giantomassi, M., Rangel, T., Goossens, E., Rignanese, G.-M., Gonze, X., and Peeters, F. M. (2011). Convergence of quasiparticle band structures of si and ge nanowires in the *GW* approximation and the validity of scissor shifts. *Phys. Rev. B*, 83:045306.

Perdew, J. P., Burke, K., and Ernzerhof, M. (1996). Generalized gradient approximation made simple. *Phys. Rev. Lett.*, 77:3865–3868.

Perdew, J. P., Parr, R. G., Levy, M., and Balduz, J. L. (1982). Density-functional theory for fractional particle number: Derivative discontinuities of the energy. *Phys. Rev. Lett.*, 49:1691–1694.

Petretto, G. and Bruneval, F. (2014). Comprehensive *Ab Initio* study of doping in bulk ZnO with group-V elements. *Phys. Rev. Applied*, 1:024005.

Pines, D. and Nozères, P. (1966). *Theory of Quantum Liquids*. Benjamin, New York.

Posselt, M., Gao, F., and Bracht, H. (2008). Correlation between self-diffusion in si and the migration mechanisms of vacancies and self-interstitials: An atomistic study. *Phys. Rev. B*, 78:035208.

Probert, M. I. J. and Payne, M. C. (2003). Improving the convergence of defect calculations in supercells: An ab initio study of the neutral silicon vacancy. *Phys. Rev. B*, 67:075204.

Refaely-Abramson, S., Sharifzadeh, S., Govind, N., Autschbach, J., Neaton, J. B., Baer, R., and Kronik, L. (2012). Quasiparticle spectra from a nonempirical optimally tuned range-separated hybrid density functional. *Phys. Rev. Lett.*, 109:226405.

Reining, L., Onida, G., and Godby, R. W. (1997). Elimination of unoccupied-state summations in ab initio self-energy calculations for large supercells. *Phys. Rev. B*, 56:R4301–R4304.

Rinke, P., Janotti, A., Scheffler, M., and Van de Walle, C. G. (2009). Defect formation energies without the band-gap problem: Combining density-functional theory and the *GW* approach for the silicon self-interstitial. *Phys. Rev. Lett.*, 102:026402.

Rohlfing, M. (2000). Excited states of molecules from green's function perturbation techniques. *Int. J. Quantum Chem.*, 80:807.

Schimka, L., Harl, J., Stroppa, A., Grueneis, A., Marsman, M., Mittendorfer, F., and Kresse, G. (2010). Accurate surface and adsorption energies from many-body perturbation theory. *Nat. Mater.*, 9:741–744.

Sharp, R. T. and Horton, G. K. (1953). A variational approach to the unipotential many-electron problem. *Phys. Rev.*, 90:317–317.

Shirley, E. L. and Martin, R. M. (1993). *GW* quasiparticle calculations in atoms. *Phys. Rev. B*, 47:15404–15412.

Slater, J. C. (1974). *The Self-Consistent Field for Molecules and Solids*, volume 4. McGraw-Hill, New York.

Śpiewak, P. and Kurzydłowski, K. J. (2013). Formation and migration energies of the vacancy in si calculated using the hse06 range-separated hybrid functional. *Phys. Rev. B*, 88:195204.

Stan, A., Dahlen, N. E., and van Leeuwen, R. (2006). Fully self-consistent *GW* calculations for atoms and molecules. *Europhys. Lett.*, 76:298.

Stolwijk, N., Schuster, B., Hölzl, J., Mehrer, H., and Frank, W. (1983). Diffusion and solubility of gold in silicon. *Physica B+C*, 116(1-3):335 – 342.

Strinati, G. (1988). Application of the greens-functions method to the study of the optical-properties of semiconductors. *Riv. Nuovo Cimento*, 11:1.

Strinati, G., Mattausch, H. J., and Hanke, W. (1980). Dynamical correlation effects on the quasiparticle bloch states of a covalent crystal. *Phys. Rev. Lett.*, 45:290–294.

Szabó, A. and Ostlund, N. S. (1996). *Modern quantum chemistry : introduction to advanced electronic structure theory*. Dover Publications, Mineola (N.Y.).

Taut, M. (1985). Frequency moments of the dielectric function for an inhomogeneous electron gas. *J. Phys. C: Solid State Phys.*, 18:2677.

Taylor, S. E. and Bruneval, F. (2011). Understanding and correcting the spurious interactions in charged supercells. *Phys. Rev. B*, 84:075155.

Tiago, M. L. and Chelikowsky, J. R. (2006). Optical excitations in organic molecules, clusters, and defects studied by first-principles green's function methods. *Phys. Rev. B*, 73:205334.

Tiago, M. L., Ismail-Beigi, S., and Louie, S. G. (2004). Effect of semicore orbitals on the electronic band gaps of si, ge, and gaas within the *GW* approximation. *Phys. Rev. B*, 69:125212.

Toulouse, J., Colonna, F. m. c., and Savin, A. (2004a). Long-range–short-range separation of the electron-electron interaction in density-functional theory. *Phys. Rev. A*, 70:062505.

Toulouse, J., Gerber, I. C., Jansen, G., Savin, A., and Ángyán, J. G. (2009). Adiabatic-connection fluctuation-dissipation density-functional theory based on range separation. *Phys. Rev. Lett.*, 102:096404.

Toulouse, J., Savin, A., and Flad, H. J. (2004b). Short-range exchange-correlation energy of a uniform electron gas with modified electron–electron interaction. *Int. J. Quantum Chem.*, 100:1047.

Ural, A., Griffin, P. B., and Plummer, J. D. (1999). Self-diffusion in silicon: Similarity between the properties of native point defects. *Phys. Rev. Lett.*, 83:3454–3457.

Van Noorden, R. (2014). The rechargeable revolution: A better battery. *Nature*, 507:26.

van Schilfgaarde, M., Kotani, T., and Faleev, S. (2006). Quasiparticle self-consistent *GW* theory. *Phys. Rev. Lett.*, 96:226402.

van Schilfgaarde, M., Kotani, T., and Faleev, S. V. (2006). Adequacy of approximations in *GW* theory. *Phys. Rev. B*, 74:245125.

von Barth, U. and Hedin, L. (1972). Local exchange-correlation potential for spin polarized case .1. *J. Phys. C: Solid State Phys.*, 5:1629.

Voronkov, V. and Falster, R. (2006). Properties of vacancies and self-interstitials in silicon deduced from crystal growth, wafer processing, self-diffusion and metal diffusion. *Mater. Sci. Eng. B-Solid State Mater. Adv. Technol.*, 134(2-3):227 – 232. {EMRS} 2006, Symposium V; Advanced Silicon for the 21st Century.

Vosko, S., Wilk, L., and Nusair, M. (1980). Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Can. J. Phys.*, 58:1200.

Wiser, N. (1963). Dielectric constant with local field effects included. *Phys. Rev.*, 129:62–69.

Würschum, R., Bauer, W., Maier, K., Seeger, A., and Schaefer, H.-E. (1989). Defects in semiconductors after electron irradiation or in high-temperature thermal equilibrium, as studied by positron annihilation. *J. Phys. Condens. Matter*, 1(SA):SA33.

Yang, W., Mori-Sánchez, P., and Cohen, A. J. (2013). Extension of many-body theory and approximate density functionals to fractional charges and fractional spins. *J. Chem. Phys.*, 139(10):104114.

Yang, W., Zhang, Y., and Ayers, P. W. (2000). Degenerate ground states and a fractional number of electrons in density and reduced density matrix functional theory. *Phys. Rev. Lett.*, 84:5172–5175.

Zhang, G., Canning, A., Gronbech-Jensen, N., Derenzo, S., and Wang, L.-W. (2013). Shallow Impurity Level Calculations in Semiconductors Using Ab Initio Methods. *Phys. Rev. Lett.*, 110(16).

Part V

Appendix

Appendix A

Curriculum Vitae

Dr. Fabien Bruneval

Age:	34
Born:	01 May 1979
Marital status:	married, 2 children
Citizenship:	French
Address:	Service de Recherches de Métallurgie Physique CEA Saclay 91191 Gif-sur-Yvette, France
Phone:	+33.1.69.08.43.49
Email:	fabien.bruneval@cea.fr

Current position: (since Dec. 2007)

Permanent Researcher at CEA (Atomic Energy Commission)
Service de Recherches de Métallurgie Physique
CEA Saclay, France

Former positions:

Jan. 2006 – Nov. 2007: Post-doctoral researcher
at **ETH Zurich**, Department of Chemistry and Applied Biosciences, Switzerland
Supervisor: Prof. Michele Parrinello

Oct. 2005 – Dec. 2005: Post-doctoral researcher
at **Ecole Polytechnique**, Laboratoire des Solides Irradiés, France
Supervisors: Dr. Lucia Reining and Dr. Nathalie Vast

Education:

2002 – 2005: PhD in Physics at **Ecole Polytechnique**, France
"Exchange and correlation in solids, from silicon to cuprous oxyde"
supervisors: Dr. Nathalie Vast and Dr. Lucia Reining

2001 – 2002: Master of Science in Material Science
at **University Pierre et Marie Curie - Paris VI**

1999 – 2002: Engineering degree
at **Ecole Centrale Paris**, France
specialization in Applied Physics

Scientific publication highlights:

Number of peer-reviewed articles:	31
Book chapters:	1
H-index:	16

Most cited articles (> 50 citations in ISI-database as of April 1st 2014):

1. *A brief introduction to the ABINIT software package*, X. Gonze *et al.,* Z. Kristallogr. **220**, 558 (2005).
   531 citations
2. *ABINIT: First-principles approach to material and nanosystem properties*, X. Gonze *et al.*, Comput. Phys. Commun. **180**, 2582 (2009).
   504 citations
3. *Effect of self-consistency on quasiparticles in solids*, F. Bruneval, N. Vast, and L. Reining, Phys. Rev. B **74**, 045102 (2006).
   99 citations
4. *Understanding correlations in vanadium dioxide from first principles*, M. Gatti, F. Bruneval, V. Olevano, and L. Reining, Phys. Rev. Lett. **99**, 266402 (2007).
   77 citations
5. *Accurate GW self-energies in a plane-wave basis using only a few empty states: Towards large systems,* F. Bruneval and X. Gonze, Phys. Rev. B **78**, 085125 (2008).
   61 citations
6. *Many-body perturbation theory using the density-functional concept: Beyond the GW approximation*, F. Bruneval *et al.*, Phys. Rev. Lett. **94**, 186402 (2006).
   60 citations
7. *Exchange and correlation effects in electronic excitations of Cu2O*, F. Bruneval *et al.*, Phys. Rev. Lett. **97**, 267601 (2006).
   51 citations

Invited conference communications:

July 2014 *Recent advances in the electronic structure calculations of charged defects*, in ICMR workshop "Charged systems and solid/liquid interfaces from first principles", Santa Barbara (USA)

March 2013 *Hybrid functionals in Abinit à la GW*, in Abinit 6th developers' workshop, Dinard, (France)

June 2011 *The GW approximation when the number of particles changes for real*, in CECAM workshop "Challenges and Solutions in GW Calculations for Complex Systems", Lausanne (Switzerland)

April 2011 *The RPA total energy in Abinit*, in Abinit 5th developers' workshop, Han-sur-Lesse, (Belgium)

Nov. 2010 *The GW approximationin less than 60 minutes,* in "First Yarmouk School for Computational Condensed Matter and Nano Systems", Irbid (Jordan)

Sept. 2010 *GW approximation for electron number changes: application to point defects,* in Psik conference 2010, Berlin (Germany)

Sept. 2009 *Approche GW pour les défauts chargés dans les isolants in final workshop of the ANR project LN3M, Lyon (France)*

May 2009 *Introduction to the GW approximation* in CECAM tutorial "Theoretical Spectroscopy Lectures: theory and codes", Zurich (Switzerland)

Dec. 2007 *Introduction to the GW approximation* in CECAM tutorial "Theoretical Spectroscopy Lectures: theory and codes", Lyon (France)

Jan. 2007 *Self-consistent GW electronic structure of solids* in "13 th International Workshop on Computational Physics and Materials Science: Total Energy and Force Methods", Trieste (Italy)

Dec. 2006 *Introduction to the GW approximation* in CECAM tutorial "Electronic excitations and spectroscopies : Theory and Codes", Lyon (France)

Sept. 2005 *Electronic Structure of Cu2O within self-consistent GW* in "The 2005 Nanoquanta Workshop", Bad Honnef (Germany)

Teaching:

- Tutorial classes "Quantum and statistical physics", Ecole Centrale Paris (France), 3rd year of university
  2012: 55 hours
  2013: 55 hours
  2014: 55 hours
- Tutorial classes in "Solid-State Physics", Ecole Centrale Paris (France), 4th year of unversity
  2012: 18 hours

Jury:

- External reviewer for the PhD thesis of David Waroquiers, Université Catholique de Louvain-la-Neuve (Jan. 2013)

Organization:

- International organizing comittee member of the "5th ABINIT International developer workshop", April 11-14 2011, Han-sur-Lesse (Belgium).
- Local organizer of the "Nanoquanto Young Researchers' Meeting", May 6-8 2004, Palaiseau (France).

Grants/Fundings:

- 2009-2013: principal investigator for project "Materials for energy" granted by GENCI (high performance computational ressources obtained)
- 2011-2013: partner for MAD-FIZ project "Doping of ZnO nanowires" funded by Agence Nationale de la Recherche (2 years post-doctoral position obtained)
- 2009-2011: principal investigator for project "p-type doping in ZnO" funded by Advanced Material Program at CEA (2 years post-doctoral position obtained)

Mentoring:

- Post-docs:
  - Guido Petretto (2012-), "p-type doping in ZnO thick nanowires"
  - Ying Cui (2009-2011), "Doping capabilities of ZnO, a photovoltaic material"

- PhD students:
  - Abdullah Shukri (2012-), "Ab initio calculation of the electronic stopping power in materials"
  - Samuel E. Taylor (2010-2011), "Charged defects in supercells"

- Internship students:
  - 2012: Arnaud Bourasseau, Orsay University (France)
  - 2011: Hichem Ben Hamed, Tunis University (Tunisia)

Other activities in the scientific community:

- Member of the advisory board of the ab initio software **Abinit**
- Member of the committee for allocation of French high-performance computing resources (**GENCI** CT 9)
- Member of the board of the French research networks **GdR-DFT++** and then **GdR-coDFT**
- Reviewer for journals: *Physical Review B*, *Physical Review Letters*, *Journal of Physics: Condensed Matter*, *New Journal of Physics, Physica Status Solidi (b)*, *Physica Status Solidi (c)*, *Journal of Chemical Physics, Journal of Chemical Theory and Computation, Reports on Progress in Physics*, *European Physical Journal B*, *Nanotechnology*
- External reviewer for funding agencies: *National Science Foundation* (USA), *Agence Nationale de la Recherche* (France), *Swiss National Science Foundation* (Switzerland), *Platform of Advanced Scientific Computing* (Switzerland), *PRACE Prioritization Panel* (EU)

Appendix B

Complete publication list

My complete publication list with citation counts can be found on
`http://www.researcherid.com/rid/C-6923-2009`.

1 Book chapters

1. *Quasiparticle self-consistent GW method for the spectral properties of complex materials*,
F. Bruneval and M. Gatti,
Chapter in volume "First Principle Approaches to Spectroscopic Properties of Complex Materials",
Eds. C. Di Valentin, S. Botti, and M. Cococcioni,
Springer series "Current topic in Quantum Chemistry"(2014)
`http://dx.doi.org/10.1007/128_2013_460`

2 Peer-reviewed articles

33. *Comprehensive Ab Initio Study of Doping in Bulk ZnO with Group-V Elements*,
G. Petretto and F. Bruneval,
Phys. Rev. Applied **1**, 024005 (2014).

32. *Screened Coulomb interaction calculations: cRPA implementation and applications to dynamical screening and self-consistency in uranium dioxide and cerium*,
B. Amadon, Th. Applencourt, and F. Bruneval,
Phys. Rev. B **89**, 125110 (2014).

31. *Consistent treatment of charged systems within periodic boundary conditions: The projector augmented-wave and pseudopotential methods revisited*,

F. Bruneval, J.P. Crocombette, X. Gonze, B. Dorado, M. Torrent, and F. Jollet,
Phys. Rev. B **89**, 045116 (2014).

30. *Point defect modeling in materials: Coupling ab initio and elasticity approaches*,
C. Varvenne, F. Bruneval, M.C. Marinica, and E. Clouet,
Phys. Rev. B **88**, 134102 (2013).

29. *Benchmarking the Starting Points of the GW Approximation for Molecules*,
F. Bruneval and M.A.L. Marques,
J. Chem. Theory Comput. **9**, 324 (2013).

28. *Ab initio formation volume of charged defects*,
F. Bruneval and J.-P. Crocombette,
Phys. Rev. B **86**, 140103(R) (2012).

27. *Formation and migration energy of native defects in silicon carbide from first principles: an overview*,
G. Roma *et al.*,
Defect and Diffusion Forum **323-325**, 11 (2012).

26. *Range-Separated Approach to the RPA Correlation Applied to the van der Waals Bond and to Diffusion of Defects*,
F. Bruneval,
Phys. Rev. Lett. **108**, 256403 (2012).

25. *Ionization energy of atoms obtained from GW self-energy or from random phase approximation total energies*,
F. Bruneval,
J. Chem. Phys. **136**, 194107 (2012).

24. *Methodological aspects of the GW calculation of the carbon vacancy in 3C-SiC*,
F. Bruneval,
Nucl. Instrum. Meth. B **277**, 77 (2012).

23. *Direct observation of Al-doping-induced electronic states in the valence band and band gap of ZnO films*,
M. Gabás *et al.*,
Phys. Rev. B **84**, 153303 (2011).

22. *Understanding and correcting the spurious interactions in charged supercells*,
S.E. Taylor and F. Bruneval,
Phys. Rev. B **84**, 075155 (2011).

21. *Energetics and metastability of the silicon vacancy in cubic SiC*,
F. Bruneval and G. Roma,
Phys. Rev. B **83**, 144116 (2011).

20. *Electronic properties of interfaces and defects from many-body perturbation theory: Recent developments and applications*,
M. Giantomassi *et al.*,
Phys. Status Solidi B **248**, 275 (2011).

19. *p-type doping and codoping of ZnO based on nitrogen is ineffective: An ab initio clue*,
Y. Cui and F. Bruneval,
Appl. Phys. Lett. **97**, 042108 (2010).

18. *Effects of Electronic and Lattice Polarization on the Band Structure of Delafossite Transparent Conductive Oxides*,
J. Vidal *et al.*,
Phys. Rev. Lett. **104**, 136401 (2010).

17. *Dynamic structure factor and dielectric function of silicon for finite momentum transfer: Inelastic x-ray scattering experiments and ab initio calculations*,
H.-C. Weissker *et al.*,
Phys. Rev. B **81**, 085104 (2010).

16. *ABINIT: First-principles approach to material and nanosystem properties*,
X. Gonze *et al.*,
Comput. Phys. Commun. **180**, 2582 (2009).

15. *GW Approximation of the Many-Body Problem and Changes in the Particle Number*,
F. Bruneval,
Phys. Rev. Lett. **103**, 176403 (2009).

14. *A Molecular Dynamics Study of the Early Stages of Calcium Carbonate Growth*,
G.A. Tribello, F. Bruneval, C.C. Liew, and M. Parrinello,
J. Phys. Chem. B **113**, 11680 (2009).

13. *Accurate GW self-energies in a plane-wave basis using only a few empty states: Towards large systems*,
F. Bruneval and X. Gonze,
Phys. Rev. B **78**, 085125 (2008).

12. *New Lennard-Jones metastable phase*,
H. Eshet, F. Bruneval, and M. Parrinello,
J. Chem. Phys. **129**, 026101 (2008).

11. *Molecular dynamics study of the solvation of calcium carbonate in water*,
F. Bruneval, D. Donadio, and M. Parrinello,
J. Phys. Chem. B **111**, 12219 (2007).

10. *Understanding correlations in vanadium dioxide from first principles*,
M. Gatti, F. Bruneval, V. Olevano, and L. Reining,
Phys. Rev. Lett. **99**, 266402 (2007).

9. *Electronic excitations: Ab initio calculations of electronic spectra and application to zirconia ZrO_2, titania TiO_2 and cuprous oxide Cu_2O*,
L. K. Dash *et al.*,
Comput. Mater. Science **38**, 482 (2007).

8. *Exchange and correlation effects in electronic excitations of Cu_2O*,
F. Bruneval *et al.*,
Phys. Rev. Lett. **97**, 267601 (2006).

7. *Effect of self-consistency on quasiparticles in solids*,
F. Bruneval, N. Vast, and L. Reining,
Phys. Rev. B **74**, 045102 (2006).

6. *Beyond time-dependent exact exchange: The need for long-range correlation*,
F. Bruneval, F. Sottile, V. Olevano, and L. Reining,
J. Chem. Phys. **127**, 144113 (2006).

5. *Signatures of short-range many-body effects in the dielectric function of silicon for finite momentum transfer*,
H.C. Weissker *et al.*,
Phys. Rev. Lett. **97**, 237602 (2006).

4. *Many-body perturbation theory using the density-functional concept: Beyond the GW approximation*,
F. Bruneval *et al.*,
Phys. Rev. Lett. **94**, 186402 (2006).

3. *A brief introduction to the ABINIT software package*,
X. Gonze *et al.*,
Z. Kristallogr. **220**, 558 (2005).

2. *Comment on "Quantum confinement and electronic properties of silicon nanowires"*,
F. Bruneval, S. Botti, and L. Reining,
Phys. Rev. Lett. **94**, 219701 (2005).

1. *TDDFT from molecules to solids: The role of long-range interactions*,
F. Sottile *et al.*,
Int. J. Quant. Chem. **102**, 684 (2005).

Appendix C

Articles for Part II

PHYSICAL REVIEW B **89**, 045116 (2014)

Consistent treatment of charged systems within periodic boundary conditions: The projector augmented-wave and pseudopotential methods revisited

Fabien Bruneval and Jean-Paul Crocombette
CEA, DEN, Service de Recherches de Métallurgie Physique, F-91191 Gif-sur-Yvette, France

Xavier Gonze
European Theoretical Spectroscopy Facility, Institute of Condensed Matter and Nanosciences, Université catholique de Louvain, Chemin des étoiles 8, bte L07.03.01, B-1348 Louvain-la-neuve, Belgium

Boris Dorado, Marc Torrent, and François Jollet
CEA, DAM, DIF, F-91297 Arpajon, France

(Received 3 October 2013; revised manuscript received 20 December 2013; published 13 January 2014)

The *ab initio* calculation of charged defect properties in solids is not straightforward because of the delicate interplay between the long-range Coulomb interaction and the periodic boundary conditions. We derive the projector augmented-wave (PAW) energy and Hamiltonian with special care taken on the potentials from the Coulomb interaction. By explicitly treating the background compensation charge, we find additional terms in the total energy of the charged cells and in the potential. We show that these background terms are needed to accurately reproduce all-electron calculations of the formation energy of a charged defect. In particular, the previous PAW expressions were spuriously sensitive to the pseudization conditions and this artifact is removed by the background term. This PAW derivation also provides insights into the norm-conserving pseudopotential framework. We propose then an alternative definition for the total energy of charged cells and for the Kohn-Sham potential within this framework that better approximates the all-electron results.

DOI: 10.1103/PhysRevB.89.045116 PACS number(s): 71.15.Dx, 71.55.−i

I. INTRODUCTION

In order to calculate the properties of charged defects in semiconductors, of polarons in insulators, or of isolated ions, it is often required to consider charged systems in *ab initio* calculations. The combined use of charged simulation cells and of periodic boundary conditions leads to intricacies that require a lot of care [1–4]. First, a truly charged periodic system would have an infinite energy. This problem is circumvented by adding a compensating background charge to restore the global charge neutrality. Second, even with a compensating background, the electrostatic potential is still not uniquely defined. Indeed, the electrostatic potential induced by a lattice of point charges is a conditionally convergent series. This complicated mathematical behavior unfortunately leads to many delicate consequences in solid-state physics. One famous example is the dependence of the Madelung constant upon the shape of the truncation of the Coulomb series [1]. Another occurrence of this phenomenon is the well-known dependence of the work function upon the surface type [5].

In practical *ab initio* implementations, the subtleties related to the definition of the electrostatic potential are hidden deeply, owing to the choice of the Ewald summation technique together with the convention of zero average potentials [3,6]. This is of course an arbitrary choice. However, once this convention has been chosen, it has to be consistently propagated in the different electrostatic terms of the Hamiltonian: the ion-ion repulsion, the electron-ion attraction, and the electron-electron repulsion. For the all-electron (AE) methods that consider straightforwardly the physical nucleus attraction potential Z/r (in atomic units), this is not much of a problem. The situation is more complicated for the plane-wave methods using pseudopotentials. The valence electrons do not experience the bare ionic potential, but rather a smooth pseudopotential that induces an extra term in the total energy, namely, the difference between the average bare potential and the average pseudopotential [1,7]. This is the origin of the so-called "$Z\alpha$" term, first derived by Ihm and co-workers [8].

The situation becomes even more complex when turning to the projector augmented-wave (PAW) method. The PAW method, introduced by Blöchl [9], is an improvement over the pseudopotential approach. Owing to the PAW transformation, the pseudo-wave functions are mapping the true AE wave functions. The PAW bears many similarities with pseudopotentials, as demonstrated a few years later by Kresse and Joubert [10]. Most noticeably for our discussion, the pseudo-wave functions experience a pseudopotential, which requires a subtle treatment of the compensating background.

In this paper, we demonstrate that the current PAW total energy and Hamiltonian do not incorporate the compensating background contribution in a consistent manner. Extra terms have to be added to the potential of any system and to the energy of charged systems. It may appear counterintuitive that the origin of the potentials may have an effect on the physical properties. However, we show that the formation energy of a charged defect is indeed affected by an inconsistent treatment of the background. Although potential alignment techniques have been devised to circumvent the problem [11–14], a unanimous agreement in the literature about a unique definition that would work whatever the nature of the charged defect is still lacking. Only with these terms properly included could the PAW results be independent from the details of the PAW pseudopotential and could they adequately reproduce the reference AE calculations. As a by-product, we also propose

a modification of the total energy in the norm-conserving pseudopotential framework.

The paper is organized as follows: In Sec. II, we review the peculiarities of the Coulomb interaction in periodic systems. In Sec. III, we derive the PAW equations with a proper account of the compensating background density. Section IV provides the applications: the validation of the additional potential and energy terms by comparing to AE calculations and an application to a highly charged defect. Section V is devoted to the extension to the norm-conserving framework.

We will work in atomic units of length (1 bohr = 1), energy (1 hartree = 1), and action ($\hbar = 1$). However, two conventions for the atomic unit of charge are possible. While the common choice is to select a negative sign for the electronic charge, so that $e = -1$ (e.g., a charged Li vacancy in LiH is negatively charged), on the contrary, in the PAW literature an electron is given a positive charge, so that $e = +1$ [see, e.g., Ref. [10], shortly after Eq. (9)]. Such a choice does not affect the quantities in which two charges are multiplied by each other, namely, all contributions to the energy, as detailed below. However, it does have an influence on the sign of the electrostatic potential. Still, the potential felt by the electrons (e.g., the one present in the Schrödinger equation), obtained by multiplying the electrostatic potential by the electronic charge, is free of such a convention problem.

In Sec. IV, dealing with applications, we rely on the usual convention ($e = -1$). For the other sections, we avoid the problem of convention, either because the relevant quantities are invariant upon a charge sign change, or because we refer to the potential felt by the electrons (electronic potential) instead of the electrostatic potential.

II. COULOMB INTERACTIONS IN SOLIDS

To highlight the role of the different Coulomb interactions in a solid, the total energy E of a unit cell of solid can be grouped as different contributions to

$$E = T + E_{\text{Coul}} + E_{\text{xc}}, \tag{1}$$

where T is the kinetic energy, E_{Coul} the Coulomb energy, and E_{xc} the exchange-correlation energy. In the present paper, we focus on the Coulomb term; the details of the other two terms will not be discussed any further. With these notations, the Coulomb energy gathers all the electrostatic interactions in the solid: the electron-electron interaction (also named the Hartree energy), the nucleus-nucleus interaction, and the electron-nucleus interaction [also referred to as the external potential in the density-functional theory (DFT) language].

A. Coulomb interaction

In the following, an extensive use of Coulomb integrals, potential, and energy will be necessary. Let us introduce some useful notations here.

The Coulomb interaction between charge densities n_1 and n_2 is defined as

$$\langle n_1, n_2 \rangle = \iint d\mathbf{r}_1 d\mathbf{r}_2 n_1(\mathbf{r}_1) \frac{1}{|\mathbf{r}_1 - \mathbf{r}_2|} n_2(\mathbf{r}_2), \tag{2}$$

where the integrals run over the complete space. The Coulomb interaction is linear, symmetric, and positive definite. It is then a scalar product.

The potential created by a charge density $n(\mathbf{r})$ is obtained from the Poisson equation, which reads

$$v_H[n](\mathbf{r}) = \int d\mathbf{r}' \frac{n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}. \tag{3}$$

The potential is obviously linear with respect to its argument n.

The Coulomb self-energy $E[n]$ of charge distribution n reads

$$E[n] = \tfrac{1}{2}\langle n, n \rangle. \tag{4}$$

The factor $\frac{1}{2}$ comes from the double counting of the interactions. This is the energy of the entire system. Note that this Coulomb self-energy could also be obtained from the potential

$$E[n] = \frac{1}{2} \int d\mathbf{r} n(\mathbf{r}) v_H[n](\mathbf{r}). \tag{5}$$

When we turn to periodic system, it is more convenient to work with the energy per unit cell, $E[n]/N_{\mathbf{R}}$, where $N_{\mathbf{R}}$ stands for the number of unit cells contained in the full solid.

As explained in the Introduction, the *ab initio* implementations in periodic solids generally rely on the Ewald technique [4,6,15], which presents several subtle points due to the long-ranged Coulomb interaction $1/r$.

The Ewald technique proposes to evaluate the potential of a lattice of point charges with a compensating background:

$$\sum_{\mathbf{R}} \delta'_{\mathbf{R}}(\mathbf{r}) = \left[\sum_{\mathbf{R}} \delta(\mathbf{r} - \mathbf{R}) \right] - \frac{1}{\Omega}. \tag{6}$$

$\mathbf{R}$ stands for the lattice vectors and Ω for the unit cell volume. Here and below, we emphasize that a charge distribution is charge neutral by adding a prime. The direct solution of the Poisson equation for this charge density is impossible, since the lattice of point charges would induce a semiconvergent series, the value of which is undefined.

Using the following short-range/long-range decomposition,

$$\frac{1}{r} = \frac{\text{erfc}(\eta r)}{r} + \frac{\text{erf}(\eta r)}{r}, \tag{7}$$

the potential can be split into two contributions. The value of η does not influence the final result and can be tuned for numerical convenience. After some algebra, the potential created by the charge density of Eq. (6) can be written as the sum of two absolutely converging series up to a constant A:

$$v_H \left[\sum_{\mathbf{R}} \delta'_{\mathbf{R}} \right](\mathbf{r}) = \sum_{\mathbf{R}} \frac{\text{erfc}(\eta |\mathbf{r} - \mathbf{R}|)}{|\mathbf{r} - \mathbf{R}|} + \frac{4\pi}{\Omega} \sum_{\mathbf{G} \neq 0} \frac{e^{-G^2/4\eta}}{G^2} e^{i\mathbf{G}\cdot\mathbf{r}} + A. \tag{8}$$

The first term in real space arises from the short-range interaction, whereas the second term arises from the long-range part and is conveniently evaluated in reciprocal space. In Eq. (8), there should be an additional dipole term which

is disregarded in the Ewald sums. In other words, this corresponds to immersing the solid inside a metallic cavity that would perfectly compensate the global dipole of the considered solid. Furthermore, the choice for the value of A is purely conventional. In general, the average value of v_H is set to zero and consequently $A = -\frac{\pi}{\eta^2\Omega}$.

Within these conventions, the energies and potentials of neutral and charged solids are completely fixed. We insist that the choice of A is conventional. However, once the conventions are settled, the calculated energies and potentials should not depend on the *ab initio* calculation technique: In practice, PAW and AE should produce the same results. Furthermore, for charged systems it is customary to postprocess the results with so-called charge corrections [11–14,16,17]. It should be noted that the correction schemes also use the Ewald technique with a fixed convention for the value of A. This convention, however, rarely explicitly stated, is in all practical cases the zero average convention. It is then important that the same convention is used for both the electronic structure calculation and the postprocessing scheme.

B. The physical densities and the background compensated densities

In a solid, there are two charged particles: the electrons and the protons. The electrons will be treated quantum mechanically, whereas the protons will simply act as point charges. Furthermore, it is also convenient to distinguish the core electrons from the valence electrons.

In a solid, all the densities are periodic: They are unchanged by a translation of any lattice vector $\mathbf{R}$. Then, the valence electron density n integrates to N_v per unit cell,

$$\frac{1}{N_{\mathbf{R}}}\int d\mathbf{r} n(\mathbf{r}) = N_v. \tag{9}$$

All the integrals in this paper run over the whole solid. The quantities per unit cell are readily obtained by dividing by $N_{\mathbf{R}}$.

The core electron density n_c can be written as a sum over atomic sites:

$$n_c(\mathbf{r}) = \sum_{\mathbf{R}a} n_c^{\mathbf{R}a}(\mathbf{r}) \tag{10}$$

$$= \sum_{\mathbf{R}a} n_c^a(\mathbf{r} - \mathbf{R} - \tau_a), \tag{11}$$

where τ_a is the position of atom a in the unit cell. Here and consistently in the following, the densities with superscript $\mathbf{R}a$ are referred in the solid coordinates, whereas the densities with superscript a have the origin in the position of atom a. The core density is most commonly kept frozen in the atomic configuration within the PAW framework. The atomic core densities n_c^a are then considered as spherical, $n_c^a(r)$, and are obtained from the atomic data file in general. The core density integrates to N_c electrons per unit cell, so that the total electronic density $n + n_c$ integrates to $N_v + N_c = N$ in a unit cell.

The charge density of the nuclei is a sum of point charges:

$$q_Z(\mathbf{r}) = \sum_{\mathbf{R}a} q_Z^{\mathbf{R}a}(\mathbf{r}) \tag{12}$$

$$= \sum_{\mathbf{R}a} -eZ^a\delta(\mathbf{r} - \mathbf{R} - \tau_a). \tag{13}$$

Note that the latter definition is independent of the choice of sign for e, as discussed at the end of the Introduction. In order to link with the PAW literature, we introduce

$$n_Z(\mathbf{r}) = q_Z(\mathbf{r})/e. \tag{14}$$

n_Z integrates to $-Z$ in a unit cell, with

$$Z = \sum_a Z^a. \tag{15}$$

For convenience, the frozen densities that can be decomposed as a sum over atomic sites are often treated together, as an "ionic density" n_{Zc}:

$$n_{Zc}(\mathbf{r}) = n_Z(\mathbf{r}) + n_c(\mathbf{r}) \tag{16}$$

$$= \sum_{\mathbf{R}a} n_{Zc}^a(|\mathbf{r} - \mathbf{R} - \tau_a|). \tag{17}$$

The ionic density integrates to $-Z_{\text{ion}} = N_c - Z$.

Finally, the total charge density en_T which contains contributions from all charges (electrons and protons) is computed from

$$n_T = n_Z + n_c + n = n_{Zc} + n. \tag{18}$$

Some physical properties require one to calculate charged unit cells. Nonzero charges q are obtained whenever the number of protons is not balanced by the number of electrons in the cell:

$$q = Z - N = Z_{\text{ion}} - N_v. \tag{19}$$

However, the potential obtained from such an unbalanced density would diverge. In practice, a compensating background is added in order to ensure the global charge neutrality:

$$n_T'(\mathbf{r}) = n_T(\mathbf{r}) + \frac{q}{\Omega}. \tag{20}$$

We remind that we introduced the prime notation for charge compensated densities that average to zero.

With these definitions, the total Coulomb energy per unit cell in a solid can be explicitly written

$$E_{\text{Coul}} = \frac{1}{2N_{\mathbf{R}}}\langle n_T', n_T'\rangle - \frac{1}{2N_{\mathbf{R}}}\sum_{\mathbf{R}a}\langle n_Z^{\mathbf{R}a}, n_Z^{\mathbf{R}a}\rangle. \tag{21}$$

This is the Coulomb self-energy of n_T' with explicit removal of the nuclei self-interaction. The self-interaction energy of a point charge is infinite and therefore each term in the previous equations is infinite. Fortunately, the difference between the two terms remains finite. Although not mathematically correct, this way of writing the equations is extremely practical. The mathematical correctness would be recovered by considering Gaussian shaped nuclei instead of point nuclei and then performing the limit to vanishing Gaussian widths. This would unfortunately make the equations less readable. Equation (21)

is valid not only for neutral systems, but also for charged systems owing to the use of the background compensated n'_T.

III. PAW BACKGROUND TERMS

The PAW method allows one to reconstruct AE wave functions, AE densities, and AE expectation values of operators out of pseudoquantities. The technique has many conceptual and numerical advantages, but they are obtained at the expense of introducing intermediate densities and potentials. This makes the derivation of the equation less straightforward. Indeed, we show in this section that background terms in the potential and in the energy have been omitted so far.

A. PAW charge densities

The PAW transformation maps the physical valence densities n using smooth densities $\tilde{n}$ [9]. The physical densities have a full nodal structure in the vicinity of the atoms, whereas the smooth densities do not. The smooth density deviates from the physical density only inside spheres centered on atoms. In order to transform smooth densities into physical densities, just on-site corrections to the densities are necessary.

Then the physical valence density is written

$$n(\mathbf{r}) = \tilde{n}(\mathbf{r}) - \tilde{n}^1(\mathbf{r}) + n^1(\mathbf{r}), \tag{22}$$

where $\tilde{n}^1(\mathbf{r})$ is the spurious smooth density in the PAW spheres and $n^1(\mathbf{r})$ is the physical density in the spheres. We follow the standard notations [10,18,19]: The smooth quantities have a tilde and the on-site quantities have an exponent 1.

The PAW technique requires also the introduction of the smooth density $\tilde{n}_{Zc}(\mathbf{r})$ to mimic the ionic core density (core electrons plus protons) that we introduced in the previous section, $n_{Zc}(\mathbf{r})$. The potential created by $\tilde{n}_{Zc}$ plays the role of a pseudopotential for the smooth valence density $\tilde{n}$. $v_H[\tilde{n}_{Zc}]$ can be thought of as the local component in the pseudopotential scheme. It is the sum over atomic contribution

$$\tilde{n}_{Zc}(\mathbf{r}) = \sum_{\mathbf{R}a} \tilde{n}^{\mathbf{R}a}_{Zc}(\mathbf{r}) \tag{23}$$

$$= \sum_{\mathbf{R}a} \tilde{n}^{a}_{Zc}(|\mathbf{r} - \mathbf{R} - \tau_a|). \tag{24}$$

As consequence, the total density n_T including the core and valence electrons and the protons can be recast into three terms following in the same spirit as for Eq. (22):

$$n_T = \tilde{n}_T - \tilde{n}^1_T + n^1_T, \tag{25}$$

where

$$\tilde{n}_T = \tilde{n} + \hat{n} + \tilde{n}_{Zc}, \tag{26}$$

$$\tilde{n}^1_T = \tilde{n}^1 + \hat{n} + \tilde{n}_{Zc}, \tag{27}$$

$$n^1_T = n^1 + n_{Zc}. \tag{28}$$

The technical compensation charge $\hat{n}$ has been further added and subtracted in the total density. This last density is chosen so that the moments in the multipole expansion of $n^1_T - \tilde{n}^1_T$ are zero. This is necessary to eliminate electrostatic interactions between PAW spheres.

Since n_{Zc} and $\tilde{n}_{Zc}$ charge distributions are monopole, carrying the same charge $-Z_{\text{ion}}$, $\hat{n}$ makes the moments of $n^1 - \tilde{n}^1 - \hat{n}$ vanish. Although the smooth density does not necessarily conserve the number of electrons, the sum $\tilde{n} + \hat{n}$ does.

B. Coulomb energy within PAW

The electrostatic energy per unit cell in the PAW framework can be written as

$$E_{\text{Coul}} = \frac{1}{2N_{\mathbf{R}}}\langle n'_T, n'_T\rangle - \frac{1}{2N_{\mathbf{R}}}\sum_{\mathbf{R}a}\langle n^{\mathbf{R}a}_{Zc}, n^{\mathbf{R}a}_{Zc}\rangle. \tag{29}$$

This expression is similar to Eq. (21) except that the core-core and core-nucleus interactions have been removed, since they only account for a change of origin in the total energies. We stress that the expression for the Coulomb self-energy of the charge distribution departs from the usual expression [10], as we explicitly introduced the compensating background in the n'_T densities. This has no consequence for the energy of neutral systems. However, it has one for charged systems, as we will show in the following.

The compensating background is homogeneous in the solid. It is practical then to include it in the smooth density,

$$n'_T = \tilde{n}'_T - \tilde{n}^1_T + n^1_T. \tag{30}$$

Then, we transform

$$\langle n'_T, n'_T\rangle = \langle \tilde{n}'_T, \tilde{n}'_T\rangle + 2\langle n^1_T - \tilde{n}^1_T, \tilde{n}'_T\rangle + \langle n^1_T - \tilde{n}^1_T, n^1_T - \tilde{n}^1_T\rangle. \tag{31}$$

This follows the usual PAW derivation except for the account of the compensating background in the smooth density. The last two terms can be evaluated on-site, since the charge distribution $n^1_T - \tilde{n}^1_T$ has vanishing moments outside the PAW spheres. Compared to the standard derivation, only two terms need to be explicitly derived: $\langle \tilde{n}'_T, \tilde{n}'_T\rangle$ and $\langle n^1_T - \tilde{n}^1_T, \tilde{n}'_T\rangle$. The following focuses on these two modified terms. We refer the reader to Ref. [10] for the usual terms, which are not detailed here.

1. The smooth density Coulomb self-energy

In fact, the explicit introduction of the compensating background in the term $\langle \tilde{n}'_T, \tilde{n}'_T\rangle$ does not yield any change compared to the standard implementations, since the average values of the Hartree potential and of the ionic pseudopotential are usually set to zero manually. Let us demonstrate this equivalence here.

The total smooth density $\tilde{n}'_T$ can be recast into two charge-neutral terms:

$$\tilde{n}'_T = \left(\tilde{n} + \hat{n} - \frac{N_v}{\Omega}\right) + \left(\tilde{n}_{Zc} + \frac{Z_{\text{ion}}}{\Omega}\right) \tag{32}$$

$$= (\tilde{n} + \hat{n})' + \tilde{n}'_{Zc}. \tag{33}$$

When inserted in $\langle \tilde{n}'_T, \tilde{n}'_T \rangle$, this decomposition turns to the familiar sum of Hartree energy, local pseudopotential energy, and ion-ion repulsion energy:

$$\frac{1}{2N_{\mathbf{R}}}\langle \tilde{n}'_T, \tilde{n}'_T \rangle = \frac{1}{2N_{\mathbf{R}}}\langle (\tilde{n}+\hat{n})', (\tilde{n}+\hat{n})' \rangle + \frac{1}{N_{\mathbf{R}}}\langle (\tilde{n}+\hat{n})', \tilde{n}'_{Zc} \rangle + \frac{1}{2N_{\mathbf{R}}}\langle \tilde{n}'_{Zc}, \tilde{n}'_{Zc} \rangle. \quad (34)$$

Owing to the explicit introduction of the compensating background charge, we immediately recognize that the Hartree energy in the previous equation is half the integral of the background compensated valence electron density $(\tilde{n}+\hat{n})'$ times the Hartree potential induced by the same charge density $v_H[(\tilde{n}+\hat{n})']$. Note that the average value of the Hartree potential is explicitly set to zero due to the vanishing average of $(\tilde{n}+\hat{n})'$ in the argument of v_H. The same remarks hold for the local pseudopotential energy (the second term on the right-hand side of the previous equation), which arises from the zero-averaged pseudopotential $v_H[\tilde{n}'_{Zc}]$. We have then identified the Hartree energy and local pseudopotential energy as they are usually calculated in practical codes.

The last term, together with the removal of the self-interaction [last sum in Eq. (29)], is the pseudo-ion/pseudo-ion repulsion with the convention of a vanishing average potential. It reduces to the usual point-charge/point-charge repulsion, usually named E_{Ewald}, plus the so-called $Z\alpha$ term [8]

$$E_{Z\alpha} = \frac{Z_{\text{ion}}}{\Omega}\sum_a \alpha^a, \quad (35)$$

where the integral α^a,

$$\alpha^a = \int d\mathbf{r}\left\{ v_H\left[\tilde{n}^a_{Zc}\right](\mathbf{r}) + \frac{Z^a_{\text{ion}}}{|\mathbf{r}|}\right\}, \quad (36)$$

measures the deviation in average potential between the pseudodensity $v_H[\tilde{n}_{Zc}]$ and a point charge $-Z_{\text{ion}}\delta(\mathbf{r})$

In the Appendix, we provide the full derivation of these two terms, since several expressions exist in the literature. The energy $E_{Z\alpha}$ is sometimes written with a factor Z_{ion} [8,20] or with a factor N_v [1,7]. As long as neutral systems are considered, the choice does not matter. However, for charged systems $N_v \neq Z_{\text{ion}}$, the total energy depends on the particular expression implemented. The Appendix demonstrates that the consistent expression should employ the factor Z_{ion}.

2. *The background terms in the Coulomb energy*

In Eq. (31), there is another occurrence of background charge density from the term $\langle n^1_T - \tilde{n}^1_T, \tilde{n}'_T \rangle$. We show in the following that this term adds extra terms in the total energy of a charged system.

In the usual derivation of the PAW energies, the density $\tilde{n}_T$ is replaced by its on-site projection $\tilde{n}^1_T$, since the integral in $\langle n^1_T - \tilde{n}^1_T, \tilde{n}_T \rangle$ does not have any contribution from outside the sphere, due to the vanishing moments of $n^1_T - \tilde{n}^1_T$. This transformation is approximate but believed to be very accurate [9,10]. It would be exact in the completeness limit of the projectors inside the PAW sphere.

When the background is also included, the transformation reads

$$\tilde{n}'_T \approx \tilde{n}^1_T + \frac{q}{\Omega}. \quad (37)$$

Whereas the term $\langle n^1_T - \tilde{n}^1_T, \tilde{n}^1_T \rangle$ is treated in the existing PAW derivations and will not be discussed further here, the background term

$$E_{\text{PAW bg}} = \frac{1}{N_{\mathbf{R}}}\left\langle n^1_T - \tilde{n}^1_T, \frac{q}{\Omega}\right\rangle \quad (38)$$

has not been explored so far.

This correcting term can be split for numerical convenience into

$$E^{(1)}_{\text{PAW bg}} = \frac{1}{N_{\mathbf{R}}}\left\langle n_{Zc} - \tilde{n}_{Zc}, \frac{q}{\Omega}\right\rangle, \quad (39)$$

$$E^{(2)}_{\text{PAW bg}} = \frac{1}{N_{\mathbf{R}}}\left\langle n^1 - \tilde{n}^1 - \hat{n}, \frac{q}{\Omega}\right\rangle. \quad (40)$$

These expressions can be decomposed on the PAW spheres, due to the vanishing moments of the left-hand side arguments of the Coulomb integrals. The first term bears striking similarities to the $Z\alpha$ term in the pseudo-ion/pseudo-ion part:

$$E^{(1)}_{\text{PAW bg}} = \frac{q}{\Omega}\sum_a \beta^a, \quad (41)$$

where the integral

$$\beta^a = \int d\mathbf{r}\left\{ v_H\left[n^a_{Zc}\right](\mathbf{r}) - v_H\left[\tilde{n}^a_{Zc}\right](\mathbf{r})\right\} \quad (42)$$

can be precalculated from the PAW atomic data. The meaning of β^a is clear: It measures the difference between the physical potential due to nucleus and the core electrons $v_H[n^a_{Zc}]$ and the pseudopotential $v_H[\tilde{n}^a_{Zc}]$.

Figure 1 shows the potential used for the calculation of β^a for silicon and carbon. These two elements have the same number of valence electrons, however, the value of β^a depends much on the cutoff radius and on the pseudization scheme. In these examples, within these pseudization conditions, silicon has $\beta^a = -13.1$ hartree, whereas carbon has $\beta^a = -4.4$ hartree. This shows the wide range of possible values for β^a. Note that for carbon, the core electrons are few and therefore there is only a very small difference between the point-charge potential $-Z^a_{\text{ion}}/r$ and the physical potential $v_H[n^a_{Zc}]$. As a consequence, the integral β^a is close the opposite of integral $\alpha^a = 4.0$ hartree. For silicon, the core electrons are more widely spread, as can be appreciated in Fig. 1, and therefore the deviation of β^a from the opposite of $\alpha^a = 7.5$ hartree is noticeable.

Let us turn now to the second background term $E^{(2)}_{\text{PAW bg}}$. This term is slightly more complicated, since it explicitly depends on the valence density. It cannot be precalculated. However, it can be expanded on coefficients that can be precalculated and stored.

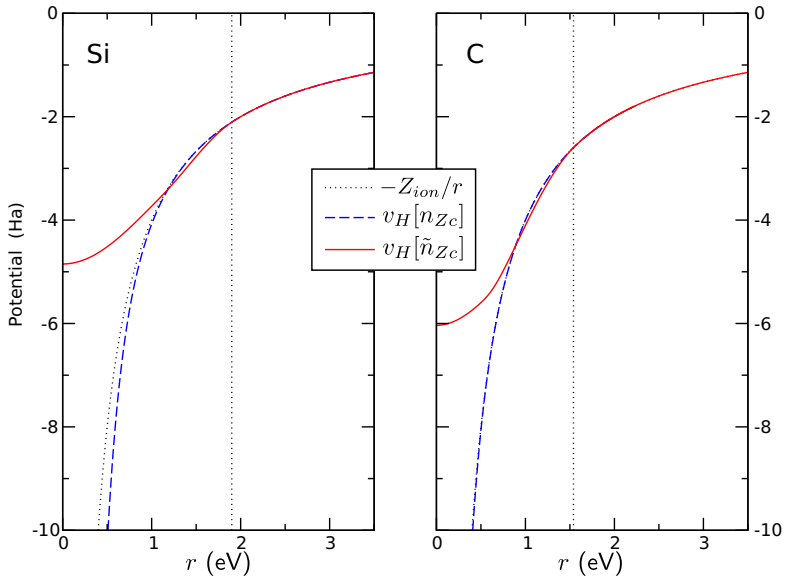

FIG. 1. (Color online) Spherical potentials involved in the integrals α^a and β^a exemplified for silicon in the left-hand panel and carbon in the right-hand panel. The point-charge potential is the dotted black line, the physical nucleus plus core electron potential $v_H[n_{Zc}]$ is the dashed blue line, and the pseudopotential $v_H[\tilde{n}_{Zc}]$ is the solid red line. Both pseudopotentials were generated with the Vanderbilt technique [21,22] using the cutoff radius symbolized with the vertical dotted line.

Indeed, the three densities in Eq. (40) can be expanded on the projectors inside each PAW sphere as

$$n^1(\mathbf{r}) = \sum_{\mathbf{R}aij} \rho^a_{ij} \phi^{a*}_i(\mathbf{r}-\mathbf{R}-\tau_a)\phi^a_j(\mathbf{r}-\mathbf{R}-\tau_a), \quad (43)$$

$$\tilde{n}^1(\mathbf{r}) = \sum_{\mathbf{R}aij} \rho^a_{ij} \tilde{\phi}^{a*}_i(\mathbf{r}-\mathbf{R}-\tau_a)\tilde{\phi}^a_j(\mathbf{r}-\mathbf{R}-\tau_a), \quad (44)$$

$$\hat{n}(\mathbf{r}) = \sum_{\mathbf{R}aij} \rho^a_{ij} \sum_{LM} \hat{Q}^{aLM}_{ij}(\mathbf{r}-\mathbf{R}-\tau_a). \quad (45)$$

The details of the algebra can be found, for instance, in Ref. [18]. Index a runs over atomic sites. Index i (and j) is a composite index for projector number n_i, and angular momenta l_i and m_i. ρ^a_{ij} is the density matrix in the basis of the projectors of site a. ϕ^a_i and $\tilde{\phi}^a_i$ are respectively the AE and the pseudo-wave functions for atom a and projector i. $\hat{Q}^{aLM}_{ij}$ are the coefficients of the multipole expansion of the compensation charge, with L and M being the angular momentum indexes.

Besides the density matrix ρ^a_{ij}, all these coefficients can be precalculated at the beginning of a PAW run. Let us gather these coefficients under the name γ^a_{ij}, so that the second background term can be written as

$$E^{(2)}_{\text{PAW bg}} = \frac{q}{\Omega} \sum_{aij} \rho^a_{ij} \gamma^a_{ij}. \quad (46)$$

The γ^a_{ij} are the average of the following potentials inside the sphere a,

$$\gamma^a_{ij} = \int d\mathbf{r} v_H \left[\phi^{a*}_i \phi^a_j - \tilde{\phi}^{a*}_i \tilde{\phi}^a_j - \sum_{LM} \hat{Q}^{aLM}_{ij} \right](\mathbf{r}). \quad (47)$$

Note that the multipole expansion of the density in the argument of v_H has zero moments, since the charge distribution $n^1 - \tilde{n}^1 + \hat{n}$ has a vanishing multipole expansion by construction of $\hat{n}$. However, this does not imply that the induced potential vanishes inside the sphere. It just vanishes outside the sphere.

All the terms in the argument of v_H in Eq. (47) need not be calculated. Indeed, performing the average in a sphere only selects the monopole of the potential and, as the Coulomb interaction $1/|\mathbf{r}-\mathbf{r}'|$ is diagonal in a multipole expansion, only the monopole of the charge distribution yields a nonvanishing contribution. Using the definition of $\hat{Q}^{aLM}_{ij}$ (see, e.g., Ref. [18]) and after some algebra, the only contribution in the argument of v_H in Eq. (47) that needs to be calculated, R^a_{ij}, reads

$$R^a_{ij}(r) = \frac{\delta_{l_i,l_j}\delta_{m_i,m_j}}{4\pi} \left\{ \frac{\phi^a_{n_i l_i}(r)\phi^a_{n_j l_j}(r) - \tilde{\phi}^a_{n_i l_i}(r)\tilde{\phi}^a_{n_j l_j}(r)}{r^2} - g_0(r) \int dr' \left[\phi^a_{n_i l_i}(r')\phi^a_{n_j l_j}(r') - \tilde{\phi}^a_{n_i l_i}(r')\tilde{\phi}^a_{n_j l_j}(r') \right] \right\}, \quad (48)$$

with $\phi_{n_i l_i}(r)$ and $\tilde{\phi}_{n_i l_i}(r)$ the radial AE and pseudo-wave functions, and $g_0(r)$ a shape function for angular momentum $l = 0$ [18].

The final expression of γ_{ij}^a is simply

$$\gamma_{ij}^a = \int dr 4\pi r^2 v_H\big[R_{ij}^a\big](r). \tag{49}$$

As the first background term, the origin of the second background term is due to the introduction of a working quantity that modifies the smooth density compared to the AE density. This second background term is related to the existence of the compensation charge $\hat{n}$. In a norm-conserving (NC) framework, the density $\tilde{n}$ would integrate to N_v and no working density $\hat{n}$ would be required. The magnitude of this second term is not easily appreciated from its analytic expression. In the following section, we will show in a practical case that this term, though smaller than the first one, is indeed not negligible.

The total PAW background energy can be written as

$$E_{\text{PAW bg}} = \frac{q}{\Omega}\sum_a \left(\beta^a + \sum_{ij}\rho_{ij}^a\gamma_{ij}^a\right), \tag{50}$$

where the coefficients β^a and γ_{ij}^a can all be calculated from the PAW atomic data at the beginning of a solid-state calculation. This extra term has to be added to the usual PAW total energy. It is zero for charge-neutral cells, however, it will modify the charged cell total energy.

C. Extra contribution to the PAW stress tensor

Since Eq. (50) depends on the volume of the cell Ω, there will be an additional term in the diagonal of the stress tensor of charged cells. The stress tensor needs to be corrected with the addition of $\sigma_{\text{PAW bg}}^{xx'}$ (x and x' indexes over the Cartesian axis):

$$\sigma_{\text{PAW bg}}^{xx'} = -\delta_{xx'}\frac{q}{\Omega^2}\sum_a \left(\beta^a + \sum_{ij}\rho_{ij}^a\gamma_{ij}^a\right). \tag{51}$$

D. Extra contribution to the PAW potential

A less obvious consequence of the additional background energy is its influence on the PAW potential for both charged and neutral systems. Indeed, the Kohn-Sham potential is defined as a functional derivative with respect to the (physical) density [23]. In the PAW framework, this implies to differentiate with respect to the pseudodensity operator [9,10]. The extra term in the energy gives rise to a contribution to the potential named $v_{\text{PAW bg}}$.

The energy $E_{\text{PAW bg}}$ has an obvious dependence on the density matrix ρ_{ij}^a. However, it also has a dependence with respect to n through the factor $q = Z_{\text{ion}} - N_v$. Indeed, the number of valence electrons is a functional of the density

$$N_v = \int d\mathbf{r} n(\mathbf{r}) = \int d\mathbf{r}[\tilde{n}(\mathbf{r}) + \hat{n}(\mathbf{r})]. \tag{52}$$

Taking the derivative of N_v with respect to $\tilde{n}$ and to ρ_{ij}^a (contained in $\hat{n}$) introduces the overlap operator $\hat{S}$ [10].

Therefore, the new background contribution to the nonlocal PAW potential is

$$\hat{v}_{\text{PAW bg}} = -\frac{\hat{S}}{\Omega}\sum_a \left(\beta^a + \sum_{ij}\rho_{ij}^a\gamma_{ij}^a\right) + \frac{q}{\Omega}\sum_{aij}\big|\tilde{p}_i^a\big\rangle\gamma_{ij}^a\big\langle\tilde{p}_j^a\big|, \tag{53}$$

where $\tilde{p}_i^a$ are the PAW projectors.

The striking result is the existence of a correcting term in the potential even for neutral systems. Even though the energy correction of the neutral system is zero, its derivative with respect to the density is nonvanishing.

The changes introduced in the absolute value of the potential would affect all the eigenvalues with the same rigid shift. For instance, when referring the position of the band edges to the position of core states or to the average position of the electrostatic potential, the difference would remain unchanged and, consequently, the calculations of band offsets would remain unaffected [24,25]. However, when the composition of the solid is changed with the introduction of defects, the extra terms in the potential have a finite effect, as we will show in the following section.

IV. PAW APPLICATIONS TO CHARGED SYSTEMS

We have derived additional terms in the PAW energy, potential, and stress. This section provides practical examples for the influence of the extra contributions. The additional terms have been implemented in the PAW code ABINIT [26].

A. Lattice of protons

Our first example is a gedanken experiment that will not require any numerical calculation. Let us consider a lattice of protons, let us say, one proton per unit cell to fix the ideas, with no electrons. Of course, a neutralizing background is required to keep the total energy finite.

In this simplistic system, all the components of the total energy related to electrons are zero. In the conventional derivations of PAW, two terms remain: the Ewald point-charge/point-charge repulsion energy and the $Z\alpha$ energy. But one of these is actually spurious. Indeed, the Ewald repulsion energy is precisely the electrostatic self-energy of the charge distribution of the point-charge protons with their compensating background. However, the energy $E_{Z\alpha}$ should not be present, since it adds a contribution that depends on the local pseudopotential $v_H[\tilde{n}_{Zc}]$.

We demonstrate now that adding the background terms fixes the problem. The coefficients γ_{ij}^a have no effect since the density matrix ρ_{ij}^a vanishes. The physical core plus nucleus potential $v_H[n_{Zc}]$ reduces to $-Z_{\text{ion}}/r$, as there is no core electron either. As a consequence, comparing Eqs. (36) and (42), $\beta^a = -\alpha^a$ and thus $E_{\text{PAW bg}}^{(1)} = -E_{Z\alpha}$.

Owing to the background energy term $E_{\text{PAW bg}}^{(1)}$, the spurious $E_{Z\alpha}$ contribution is eliminated from the total energy. The total energy of the lattice of a proton with no electrons is then independent of the pseudization details, as we expect.

B. Charged vacancy in LiH: Benchmark against AE results

In order to check the validity of the PAW derivation and the magnitude of the additional terms in the energy and potential, it would be desirable to have a valid reference calculation for a charged system. We choose here to focus on the charged lithium vacancy in rocksalt LiH, V_{Li}^{-}. This particular system was selected for the small charge of the nuclei, so that AE calculations within a plane-wave basis set were tractable.

In practice, we employ model pseudopotentials with just a local component $-Z_a \operatorname{erf}(r/r_c)/r$ instead of the full potential $-Z_a/r$. We converge the calculation with respect to both the radius r_c and the plane-wave basis cutoff energy E_{cut}. Using this type of pseudopotential is equivalent to considering that the nuclei are Gaussian charge distributions with a spread r_c. Although the absolute energies are impossible to converge, the total energy differences and the potentials show a much smoother behavior with respect to r_c and E_{cut}. This procedure allows us to extract unambiguous AE data with parameters $r_c = 0.0025$ bohrs and $E_{\mathrm{cut}} = 2000$ hartree, which still keeps the calculation cost low enough even for the eight-atom supercells.

The formation energy of the lithium vacancy $E_f(V_{\mathrm{Li}}^{-})$ is evaluated through

$$E_f(V_{\mathrm{Li}}^{-}) = E(\mathrm{Li_3H_4}^{-}) - E(\mathrm{Li_4H_4}) + E(\mathrm{Li}) - \epsilon_{\mathrm{VBM}}(\mathrm{LiH}), \tag{54}$$

where $E(\mathrm{Li_3H_4}^{-})$ is the total energy of a supercell, $E(\mathrm{Li})$ is the total energy of the isolated Li atom, and $\epsilon_{\mathrm{VBM}}(\mathrm{LiH})$ is the valence band maximum of bulk LiH. This is the usual formation energy [27] with $q = -1$, with the Fermi level set to the top valence band, and with the chemical potential of Li fixed to the atom energy. Of course, supercells with only seven and eight atoms do not give a proper evaluation of the true formation energy, but our purpose is simply to compare the *ab initio* methods. For the same reason, we do not include any charge correction nor potential alignment [11–14,16,17]. The PAW and the AE quantities should in principle match on an absolute scale, since the same convention has been retained for the evaluation of the Coulomb potentials without the need for a correcting post-treatment.

In practice, the supercell consists of seven or eight unrelaxed atoms in a cubic supercell with an edge 7.60 bohrs, and the k-point sampling is a Γ-centered $4 \times 4 \times 4$ grid, within the local density approximation (LDA). The isolated lithium atom is placed in a supercell with the same dimensions, using a Γ-only sampling. The corresponding PAW calculations use an over converged cutoff energy of 30 hartree for the wave functions and 60 hartree for the dense grid. In Fig. 2, the PAW total energies and valence band maximum are compared to the AE results. The first series shows the total energy difference $E(\mathrm{Li_3H_4}^{-}) - E(\mathrm{Li_4H_4}) + E(\mathrm{Li})$ with or without inclusion of the background terms. The second series shows the valence band maximum of the bulk $\epsilon_{\mathrm{VBM}}(\mathrm{LiH})$. The last series is the formation energy, i.e., the difference between the two previous series. Including the additional background terms has a sizable effect on the total energies and on the top valence energy. Even if these changes cancel each other out to some extent, the final physical quantity E_f is modified by the inclusion of the background terms. The two additional terms arising from

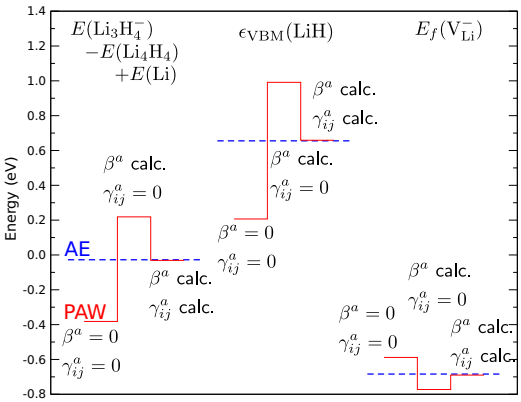

FIG. 2. (Color online) Contributions to the formation energy of a charged lithium vacancy V_{Li}^{-}, calculated with different PAW terms (solid red lines), compared to an AE reference (dashed blue lines). The first PAW calculations do not include β^a nor γ_{ij}^a, the second calculation does include β^a but not γ_{ij}^a, and the third calculation includes both terms. The formation energy is evaluated for the Fermi level fixed at the valence band maximum.

integrals β^a and γ_{ij}^a both have a visible effect. Only when the two are properly included could the PAW calculations reproduce the AE results.

As mentioned earlier, the PAW results without the background terms show a spurious dependence on the pseudization procedure. We evaluate this effect in Fig. 3 by varying the cutoff radius r_c^{loc} for the generation of the local pseudopotential $v_H[\tilde{n}_{Zc}]$ using the Vanderbilt procedure [21,22]. The corrected PAW results including the background terms are much more stable with respect to a change of pseudopotential than the uncorrected PAW data. The statement is not only true for the intermediate components such as the total energy difference or the top valence band energy [Figs. 3(a) and 3(b)], but also holds for the physical formation energy $E_f(V_{\mathrm{Li}}^{-})$ [Fig. 3(c)].

C. Highly charged interstitial SiC, $\mathrm{Si_{TC}^{4+}}$: Benchmarking different codes

The magnitude of the background terms is proportional to the charge q of the defect. We now turn to a well-documented [28,29] charged defect of cubic silicon carbide, the silicon interstitial tetrahedrically coordinated to carbon atoms, $\mathrm{Si_{TC}^{4+}}$.

In Fig. 4, we compare the Perdew-Burke-Ernzerhof (PBE) [30] formation energy of $\mathrm{Si_{TC}^{4+}}$ from three different PAW codes: VASP [31], QUANTUM ESPRESSO [32], and ABINIT [26]. In ABINIT, we have switched on and off the background correcting terms for the calculation of the formation energy. Note that for consistency QUANTUM ESPRESSO and ABINIT use the same PAW atomic data. This has not been possible for VASP, unfortunately, so that the VASP curve has been shifted up quite arbitrarily.

Carefully looking at the PAW implementation in the different codes, VASP and ABINIT without background have the same convention of setting the average smooth potential to zero, $\langle v_H[\tilde{n}_{Zc}] \rangle = 0$. Indeed, the slope of the corresponding

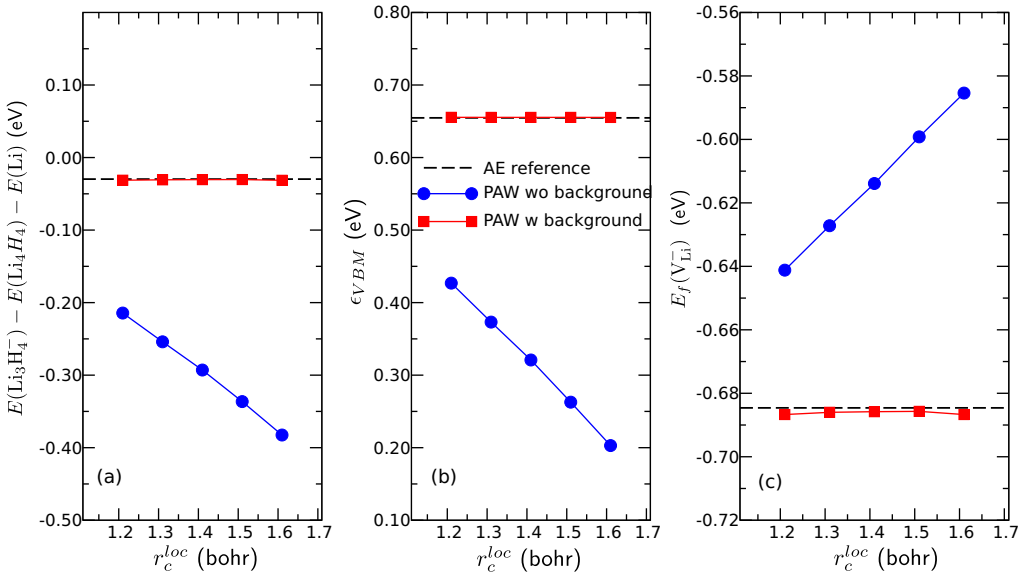

FIG. 3. (Color online) Dependence of uncorrected (blue circles) and corrected (red squares) PAW results with respect to the local potential cutoff radius r_c^{loc} in LiH. (a) represents the total energy difference in Eq. (54), i.e., the three first terms. (b) shows the valence band maximum of bulk LiH, i.e., the last term in Eq. (54). (c) shows the formation energy of the negatively charged Li vacancy with the Fermi level fixed at the valence band maximum.

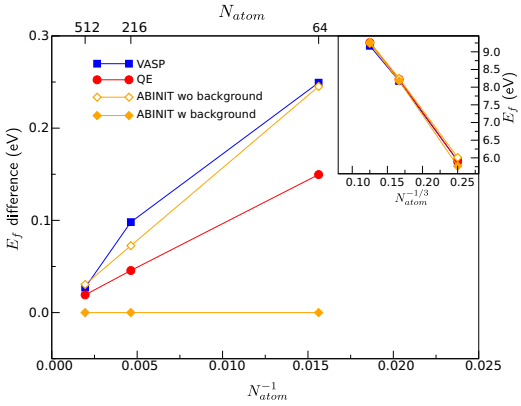

FIG. 4. (Color online) PAW formation energy E_f deviation for the silicon interstitial $\mathrm{Si}_{\mathrm{TC}}^{4+}$ in cubic SiC as a function of supercell size within PBE with the Fermi level fixed at the valence band maximum and in silicon-rich conditions. Results have been obtained from VASP (squares), QUANTUM ESPRESSO (circles), ABINIT without the background terms (open diamonds), and ABINIT including the background terms (solid diamonds). This last curve has been chosen as a zero, so as to highlight the difference between the implementations. The absolute formation energy E_f is given in the inset, where the spurious charge-charge interaction dominates [16] (convergence as $N_{\mathrm{atom}}^{-1/3}$). The QUANTUM ESPRESSO and ABINIT calculations use the same PAW atomic data, whereas it was necessary to shift up the VASP results by 100 meV.

curves matches. The QUANTUM ESPRESSO convention sets the average smooth potential to $\langle v_H[\tilde{n}_{Zc}]\rangle = \sum_a \beta^a/\Omega$. This produces a different slope. The meaning of this choice will be discussed in detail in the next section. Finally, the results from ABINIT with background terms are built to match the AE formalism using a total electrostatic potential that averages to zero $\langle v_H[n_{Zc}]\rangle = 0$.

For the interstitial $\mathrm{Si}_{\mathrm{TC}}^{4+}$, the difference between the conventions implemented in the codes quite significantly impacts the formation energy. For the 64-atom supercell, the difference can be as large as 0.25 eV and it is still 0.10 eV for the 216-atom supercell. Fortunately, the background terms are proportional to $1/\Omega$ and their effect should vanish with increasing supercell sizes.

V. CONSEQUENCES FOR THE PSEUDOPOTENTIAL METHOD

So far, we have stressed the importance of having a consistent convention for the potentials in the PAW method. The derivation for PAW in Sec. III highlights the role of different potentials: the pseudopotential, written $v_H[\tilde{n}_{Zc}]$ in the PAW language, and the true physical core electron plus nucleus potential, labeled $v_H[n_{Zc}]$. In an AE calculation, the situation is clear: The average value of the physical potential $v_H[n_{Zc}]$ is set to zero. However, within the pseudopotential framework, different choices can be found in the available implementations.

For instance, prior to version 7.5, the ABINIT code [26] uses the $Z\alpha$ energy term [8] and consequently sets the average

pseudopotential $v_H[\tilde{n}_{Zc}]$ to zero [33]. If the total energy expression uses instead a factor N_v in the $Z\alpha$ term [1,7], then the potential will be shifted accordingly. The potential is obtained as a functional derivative with respect to the electronic density and the number of valence electrons N_v is indeed a functional of the density. In QUANTUM ESPRESSO [32], for instance, this choice is made and the pseudopotential $v_H[\tilde{n}_{Zc}]$ averages to $\sum_a \alpha^a/\Omega$. This corresponds to a specific choice of the constant A introduced by the Ewald convention for the calculation of the potential $v_H[\tilde{n}_{Zc}]$. It could seem surprising to choose an inconsistent definition for the constant A.

Indeed, a consistent choice of the constant A that induces zero average for all the electrostatic potentials naturally leads to the absolute values as obtained with the $Z\alpha$ energy term, as derived in Ref. [8]. However, this choice induces a dependence of the energies and potentials on the pseudization details. It would be appreciated to devise a scheme which is independent of the pseudopotential and, even better, which reproduces as far as possible the absolute AE results. This could be achieved in practice by introducing the frozen core density, as we demonstrate in this section. This also gives an *a posteriori* justification for the convention using N_v in the $Z\alpha$ term.

A. Accounting for the physical core density in norm-conserving pseudopotentials

It is straightforward to adapt the PAW derivation of Sec. III to the simpler case of NC pseudopotentials. First of all, there is no equivalent to the integrals γ_{ij}^a in the pseudopotential framework, since there is no on-site representation of the charge density. But the integrals β^a, which measure the difference between the physical core electron+nucleus potential and the pseudopotential, still exist. Therefore, setting the physical potential to a zero average introduces an extra term in the total energy. The core electron and nucleus electrostatic energy reads

$$E_{\text{Coul}}^{Zc} = E_{\text{Ewald}} + E_{Z\alpha} + E_{\text{PAW bg}}^{(1)}, \tag{55}$$

where $E_{\text{PAW bg}}^{(1)}$ has been defined in Eq. (39), and detailed in Eq. (41). The origin of the usual terms E_{Ewald} and of $E_{Z\alpha}$ is recapitulated in the Appendix.

The addition of $E_{\text{PAW bg}}^{(1)}$ to the total energy does not modify the total energy of a charge-neutral cell. However, the potential as obtained from a functional derivative of the energy with respect to the density is affected, since $E_{\text{PAW bg}}^{(1)}$ has a dependence on N_v. The additional contribution to the potential $v_{\text{NC bg}}$ is a constant:

$$v_{\text{NC bg}} = -\frac{1}{\Omega}\sum_a \beta^a. \tag{56}$$

The calculation of the integrals β^a is straightforward: It just requires the knowledge of the physical frozen core electron density. This piece of information is available during the pseudopotential generation, but it is unfortunately generally not stored in the pseudopotential files. It would be direct to include it also.

Imagine now that the core density is much localized around the nucleus. In this case, the core+nucleus potential would only slightly deviate from the point-charge potential Z_{ion}/r, as we observed in the case of carbon in right-hand panel of Fig. 1. Then the β^a integrals would be very similar to α^a integrals:

$$\beta^a \approx -\alpha^a. \tag{57}$$

In this approximation, the following simplification occurs:

$$E_{Z\alpha} + E_{\text{PAW bg}}^{(1)} \approx \frac{N_v}{\Omega}\sum_a \alpha^a. \tag{58}$$

The present derivation gives an *a posteriori* justification for the total energy and potential formulas, which are written in some textbooks [1,7] and used in some codes, such as QUANTUM ESPRESSO [32].

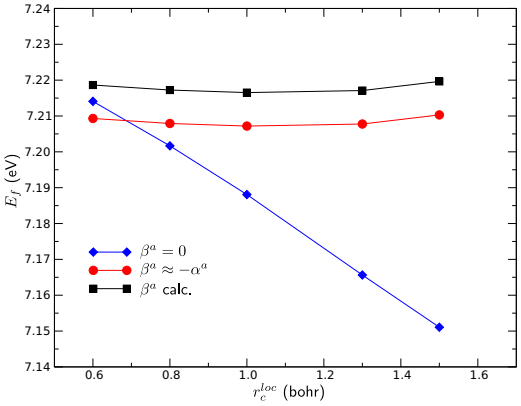

FIG. 5. (Color online) NC formation energy E_f within LDA of the unrelaxed doubly charged vacancy in diamond V_C^{2+} with the Fermi level fixed at the valence band maximum as a function of the cutoff radius of the local pseudopotential ($l = 2$), using $\beta^a = 0$ (blue diamonds), $\beta^a \approx -\alpha^a$ (red circles), or explicitly calculating β^a (black squares).

B. Charged defect examples

We now test the effect of modifying the total energy and potential expressions in NC calculations of the formation energy of charged defects. Within the NC framework, it would be perfection to hope to obtain the same results as with reference AE calculations. However, it would be more realistic to have a weak dependence of the physical properties upon the pseudopotential details.

We consider in Fig. 5 a doubly charged vacancy V_C^{2+} in a 64-atom cubic supercell of diamond. For simplicity, the atoms are not relaxed, the lattice constant is set to 6.75 bohrs, and the cutoff energy is set to a very large value of 150 hartree. The carbon pseudopotential that has been generated with the Troullier-Martins technique [34] with a local component is $l = 2$. As the carbon valence electrons have mainly a sp character, a change in the d component of the pseudopotential should only indirectly impact the physical properties. It is interesting to focus on this particular defect which has been recently shown to be out of reach of the usual correction schemes due to its delocalized nature [14]. In Fig. 5, we

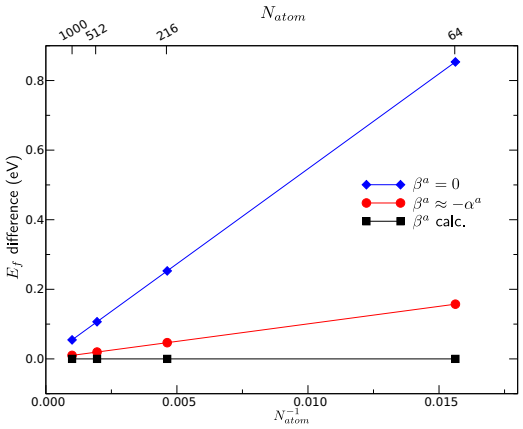

FIG. 6. (Color online) NC formation energy E_f deviation within LDA of the unrelaxed silicon interstitial Si_{TC}^{4+} in cubic SiC with the Fermi level fixed at the valence band maximum and in silicon-rich conditions as a function of supercell size, using $\beta^a = 0$ (blue diamonds), $\beta^a \approx -\alpha^a$ (red circles), or explicitly calculating β^a (black squares). This last curve has been chosen as a zero, so as to highlight the difference between the implementations. The absolute formation energy E_f shows the same behavior as in Fig. 4.

show the behavior of the formation energy of the charged vacancy as a function of the pseudopotential cutoff radius r_c^{loc}. The usual expression, which sets the average pseudopotential to zero ($\beta^a = 0$), presents a strong dependence on r_c^{loc}. If the zero of the potentials is defined with point-charge nuclei and core electrons ($\beta^a \approx -\alpha^a$), then the formation energy is remarkably stable. Furthermore, it deviates only slightly from the final results, which consistently set the origin of the potential through the true core electron plus nuclei density (actual calculation of β^a). It could be argued that all these changes have a small magnitude, however, the dependence on the details of the local component of the pseudopotential is clearly pathological.

Furthermore, when the defects involve adding or removing atoms with a wider core electron density, the effects can be significantly larger. Turning back to the silicon interstitial Si_{TC}^{4+} we used in the previous section, the added silicon atom has a larger core consisting of ten electrons. Figure 6 shows the difference in formation energy within LDA for the three choices of the potential origin. For clarity, the formation energy with an explicit calculation of β^a has been set to zero. The difference between the three schemes decreases as Ω^{-1}, however, for relatively large supercells (64–216 atoms), the difference can be as large as 0.25–0.80 eV. Except for the smallest supercell size, the approximation $\beta^a \approx -\alpha^a$ is a very decent approximation.

From these numerical applications, we conclude that it is important to include the background term in the total energy and in the potential. If the exact calculation of integrals β^a, though simple to perform, is not available, the approximation $\beta^a \approx -\alpha^a$ also yields reasonable results.

VI. CONCLUSIONS

In this paper, we derived two additional terms in the PAW energy of charged systems and in the PAW potential of all systems in order to reconcile AE calculations and PAW framework. These two terms [see Eq. (50)] arise from the proper treatment of the compensating background density, which is required by the use of periodic boundary calculations.

They are of a different nature. The first term measures the difference between the average smooth pseudopotential and the true physical nucleus plus the core electron potential. This contribution is usually the largest. Though smaller, the second term can also have a visible influence. The second term measures the difference in potential induced by the introduction of the compensation density $\hat{n}$.

The correct inclusion of these two terms has two positive consequences: It makes the PAW results directly comparable with AE calculations, and it makes the PAW results less sensitive to the PAW atomic data. We would like to stress that the proper treatment of the background terms not only affects the absolute energy values, but also impacts the physical quantities that are extracted, such as the formation energy or the relaxation volume of charged defects. The formation energy of the charged defects in small supercells or with a high charge state can be modified by several tenths of eV. Though these differences could also be reconciled by a potential alignment correction, a universal definition of the potential alignment is still missing.

For consistency with AE calculations, the first background term should also be introduced in the NC framework. This explains why different plane-wave codes could produce different physical results with the same pseudopotential data. The inclusion of the background term in Eq. (55) yields a total energy expression, which best approximates the AE results. However, its impact is even larger than experienced in the PAW framework. For the highly charged defect Si_{TC}^{4+} in SiC, the formation energy is still changed by 0.25 eV, even for the 216-atom supercell.

ACKNOWLEDGMENTS

This work was performed using HPC resources from GENCI-IDRIS and GENCI-TGCC (Grant No. 2013-gen6018).

APPENDIX: PSEUDO-ION/PSEUDO-ION INTERACTION

The expression for the pseudo-ion/pseudo-ion interaction varies in the literature. This Appendix is meant to fix this point.

The pseudo-ion/pseudo-ion repulsion energy is defined as

$$E_{\text{Coul}}^{Zc} = \frac{1}{2N_{\mathbf{R}}}\langle \tilde{n}'_{Zc}, \tilde{n}'_{Zc}\rangle - \frac{1}{2}\sum_a \langle \tilde{n}^a_{Zc}, \tilde{n}^a_{Zc}\rangle, \qquad \text{(A1)}$$

where $\tilde{n}_{Zc}$ and $\tilde{n}^a_{Zc}$ have been defined in Eqs. (23) and (24).

Indeed, it is sometimes written as the Ewald point-charge/point-charge repulsion E_{Ewald} plus the celebrated $Z\alpha$ term [8,20]. However, from some other sources, the $Z\alpha$ term is replaced by an $N\alpha$ term, with N_v replacing the Z_{ion} in Eq. (35) [1,7]. We have stressed in Sec. II that once the Ewald

convention has been chosen for the electrostatic interactions, there should only be one expression for the total energy (the unknown constant is fixed and the global dipole is assumed to be zero).

Let us demonstrate here that the correct expression is indeed the $Z\alpha$ with factor Z_{ion}. We have to bridge the difference between the original charge distribution n_{Zc}, which is a sum over atomic site contributions, and the point-charge distribution used in the Ewald energy. The point-charge distribution n_{pc}, which reads

$$n_{\text{pc}}(\mathbf{r}) = \sum_{\mathbf{R}a} n^a_{\text{pc}}(\mathbf{r}-\mathbf{R}-\tau_a) \tag{A2}$$

$$= \sum_{\mathbf{R}a} -Z^a_{\text{ion}}\delta(\mathbf{r}-\mathbf{R}-\tau_a), \tag{A3}$$

has the same multipole expansion as the original distribution n_{Zc}, under the mild assumption that the PAW spheres are nonoverlapping spheres. However, we have to then introduce the compensating background both in n'_{Zc} and n'_{pc}.

Then transforming the term $\langle \tilde{n}'_{Zc}, \tilde{n}'_{Zc}\rangle$, we write

$$\langle \tilde{n}'_{Zc}, \tilde{n}'_{Zc}\rangle = \langle n'_{\text{pc}}, n'_{\text{pc}}\rangle + \langle \tilde{n}'_{Zc} - n'_{\text{pc}}, \tilde{n}'_{Zc} + n'_{\text{pc}}\rangle. \tag{A4}$$

In the last term, the backgrounds in $\tilde{n}'_{Zc} - n'_{\text{pc}}$ compensate and therefore the primes can be dropped there. As $\tilde{n}_{Zc}$ and n'_{pc} are contained in the sphere and have the same multipole expansion, only the on-site terms remain:

$$\frac{1}{2N_{\mathbf{R}}}\langle \tilde{n}_{Zc} - n_{\text{pc}}, \tilde{n}'_{Zc} + n'_{\text{pc}}\rangle = \frac{1}{2}\sum_a \langle \tilde{n}^a_{Zc} - n^a_{\text{pc}}, \tilde{n}^a_{Zc} + n^a_{\text{pc}}\rangle + \sum_a \left\langle \tilde{n}^a_{Zc} - n^a_{\text{pc}}, \frac{Z_{\text{ion}}}{\Omega}\right\rangle, \tag{A5}$$

where the backgrounds have been written explicitly. The last term is precisely the $Z\alpha$ energy of Eq. (35).

Inserting the last equation in Eq. (A4), we obtain

$$\frac{1}{2N_{\mathbf{R}}}\langle \tilde{n}'_{Zc}, \tilde{n}'_{Zc}\rangle = \frac{1}{2N_{\mathbf{R}}}\langle n'_{\text{pc}}, n'_{\text{pc}}\rangle - \frac{1}{2}\sum_a \langle n^a_{\text{pc}}, n^a_{\text{pc}}\rangle + \frac{1}{2}\sum_a \langle n^a_{Zc}, n^a_{Zc}\rangle + E_{Z\alpha}. \tag{A6}$$

Reordering the terms in the last equation finally proves the announced result:

$$E^{Zc}_{\text{Coul}} = E_{\text{Ewald}} + E_{Z\alpha}. \tag{A7}$$

The result that the factor in energy $E_{Z\alpha}$ is not N_v could have been anticipated, since there is no reason to introduce the number of electrons in the energy E^{Zc}_{Coul} that only depends on the pseudopotential quantities.

[1] R. M. Martin, *Electronic Structure: Basic Theory and Practical Methods* (Cambridge University Press, Cambridge, U.K., 2004), Vol. 1.

[2] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids* (Oxford Science, New York, 1987).

[3] R. A. Coldwell-Horsfall and A. A. Maradudin, J. Math. Phys. **1**, 395 (1960).

[4] S. W. de Leeuw, J. W. Perram, and E. R. Smith, Proc. R. Soc. London, Ser. A **373**, 27 (1980).

[5] A. W. Dweydari and C. H. B. Mee, Phys. Status Solidi A **27**, 223 (1975).

[6] P. P. Ewald, Ann. Phys. **369**, 253 (1921).

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

[8] J. Ihm, A. Zunger, and M. L. Cohen, J. Phys. C **12**, 4409 (1979).

[9] P. E. Blöchl, Phys. Rev. B **50**, 17953 (1994).

[10] G. Kresse and D. Joubert, Phys. Rev. B **59**, 1758 (1999).

[11] S. Lany and A. Zunger, Phys. Rev. B **78**, 235104 (2008).

[12] C. Freysoldt, J. Neugebauer, and C. G. Van de Walle, Phys. Rev. Lett. **102**, 016402 (2009).

[13] S. E. Taylor and F. Bruneval, Phys. Rev. B **84**, 075155 (2011).

[14] H.-P. Komsa, T. T. Rantala, and A. Pasquarello, Phys. Rev. B **86**, 045112 (2012).

[15] L. M. Fraser, W. M. C. Foulkes, G. Rajagopal, R. J. Needs, S. D. Kenny, and A. J. Williamson, Phys. Rev. B **53**, 1814 (1996).

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

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

[18] M. Torrent, F. Jollet, F. Bottin, G. Zérah, and X. Gonze, Comput. Mater. Sci. **42**, 337 (2008).

[19] M. Torrent, N. Holzwarth, F. Jollet, D. Harris, N. Lepley, and X. Xu, Comput. Phys. Commun. **181**, 1862 (2010).

[20] X. Gonze, Phys. Rev. B **55**, 10337 (1997).

[21] D. Vanderbilt, Phys. Rev. B **41**, 7892 (1990).

[22] N. Holzwarth, A. Tackett, and G. Matthews, Comput. Phys. Commun. **135**, 329 (2001).

[23] R. G. Parr and W. Yang, *Density-Functional Theory of Atoms and Molecules* (Oxford University Press, New York, 1989).

[24] A. Baldereschi, S. Baroni, and R. Resta, Phys. Rev. Lett. **61**, 734 (1988).

[25] H. P. Komsa, E. Arola, E. Larkins, and T. T. Rantala, J. Phys.: Condens. Matter **20**, 315004 (2008).

[26] X. Gonze, B. Amadon, P. M. Anglade, J. M. Beuken, F. Bottin, P. Boulanger, F. Bruneval, D. Caliste, R. Caracas, M. Cote, T. Deutsch, L. Genovese, P. Ghosez, M. Giantomassi, S. Goedecker, D. R. Hamann, P. Hermet, F. Jollet, G. Jomard, S. Leroux, M. Mancini, S. Mazevet, M. J. T. Oliveira, G. Onida, Y. Pouillon, T. Rangel, G. M. Rignanese, D. Sangalli, R. Shaltaf, M. Torrent, M. J. Verstraete, G. Zerah, and J. W. Zwanziger, Comput. Phys. Commun. **180**, 2582 (2009).

[27] S. B. Zhang and J. E. Northrup, Phys. Rev. Lett. **67**, 2339 (1991).

[28] M. Bockstedte, A. Mattausch, and O. Pankratov, Phys. Rev. B **68**, 205201 (2003).

[29] G. Roma, F. Bruneval, T. Liao, O. N. B. Martinez, and J.-P. Crocombette, Defect Diffus. Forum **323–325**, 11 (2012).

[30] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).

[31] G. Kresse and J. Furthmüller, Phys. Rev. B **54**, 11169 (1996).
[32] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, A. D. Corso, S. de Gironcoli, S. Fabris, G. Fratesi, R. Gebauer, U. Gerstmann, C. Gougoussis, A. Kokalj, M. Lazzeri, L. Martin-Samos, N. Marzari, F. Mauri, R. Mazzarello, S. Paolini, A. Pasquarello, L. Paulatto, C. Sbraccia, S. Scandolo, G. Sclauzero, A. P. Seitsonen, A. Smogunov, P. Umari, and R. M. Wentzcovitch, J. Phys.: Condens. Matter **21**, 395502 (2009).
[33] Starting with ABINIT v7.5, the user has the possibility to change the definition of the average value of the Coulomb potential.
[34] N. Troullier and J. L. Martins, Phys. Rev. B **43**, 1993 (1991).

PHYSICAL REVIEW B **84**, 075155 (2011)

Understanding and correcting the spurious interactions in charged supercells

Samuel E. Taylor and Fabien Bruneval
CEA, DEN, Service de Recherches de Métallurgie Physique, F-91191 Gif-sur-Yvette, France
(Received 7 April 2011; published 15 August 2011)

The supercell technique is widely spread for the simulation of charged point defects. Charged defects in a supercell are unfortunately subjected to spurious image interactions, which are usually handled by introducing two correcting terms: a Madelung-type correction that accounts for the electrostatic interactions of repeated charges in a compensating background and a potential alignment term that refers the charged supercell to the electron reservoir. We demonstrate that the Madelung correction already brings a large potential shift that slowly converges as $1/L$ with increasing supercell sizes. We hence define a potential alignment devoid of any double counting. We finally propose a simple evaluation for the nearest-neighbor interaction that removes the remaining spurious hybridization of the defect wave functions between images. The application of these three corrections together drastically speeds up the convergence with respect to supercell size for all defects that are not too shallow.

DOI: 10.1103/PhysRevB.84.075155 PACS number(s): 71.15.Mb, 61.72.Bb

I. INTRODUCTION

The accurate prediction of the properties of point defects is a key target of computer simulations in condensed matter since defects govern many aspects of the physics of materials. For instance, applications in electronics, optoelectronics, and photovoltaics all rely on the fine control of charged defects in semiconductors.[1] With the advent of large supercomputers, it has been possible to address the *ab initio* calculation of point defects for over two decades now, thanks to density functional theory (DFT).[2]

The *ab initio* calculation of defects in condensed matter usually relies on the supercell approach.[3] In this framework, the *isolated* defect one intends to study is placed in a large cell, which is periodically replicated. The advantages of this approach are numerous, in particular the use of standard plane-wave codes. The supercell approach is so practical that it prevailed over competing frameworks, such as Mott-Littleton[4] or Korringa-Kohn-Rostoker Green's function.[5]

Nonetheless, the supercell approach suffers from one main drawback: the spurious interaction between the defect and its periodic images. This problem becomes particularly prominent for charged systems that are subjected to the long-range Coulomb interaction between images. No supercell size accessible to modern (or future) computers would be sufficient to render this interaction negligible. Indeed, the magnitude of this spurious contribution to the total energy scales as $N^{-1/3}$, with N being the number of atoms in the supercell.

This fact has given rise to the design of correction schemes that would accelerate the slow convergence of charged supercells. Correction schemes are numerous,[3,6–11] but they generally rely on the evaluation of two contributions: a correction of the energy and/or a correction of the potential. The correction for the energy ΔE_{el} is intended to remove the spurious long-ranged electrostatic interaction between the charged defect, its images, and the compensating background. The potential shift ΔV should account for the change of the reference energy for the electrons in the charged supercell compared to the electrons in the pristine bulk. Then the formation energy $E_f(D,q)$ of defect D with charge q in the Zhang and Northrup formalism[12] finally reads[13]

$$E_f(D,q) = E_{D,q} - E_{\text{Host}} - \sum_i n_i \mu_i + q(\epsilon_{\text{VBM}} + \epsilon_F + \Delta V) + \Delta E_{el}(q), \quad (1)$$

where $E_{D,q}$ is the raw energy of a supercell containing the defect D and an extra charge q and E_{Host} is the energy of the perfect supercell with no defect. The energy of the added or removed atoms n_i is referred to the chemical potential of reservoirs for the different elements μ_i. For electrons, the chemical potential is governed by the Fermi energy E_F, the zero of which is conventionally set at the valence band maximum of the bulk material ϵ_{VBM}.

Much effort has been devoted to the design of intelligent electrostatic corrections ΔE_{el}; comprehensive discussion on this point can be found elsewhere.[9,14–16] A multitude of conflicting ways have also been suggested to calculate the potential alignment, but no convincing arguments have yet been put forward for which is the most suitable. Some authors suggest taking an average of the total Kohn-Sham potential,[15,17,18] and others suggest an average of the electrostatic potential only.[13,19,20] This average is then taken either over the entire supercell[17,19] or in some localized region, usually as far as possible from the defect.[8,9,14] Some authors even refrain from including potential alignment at all, due to a (not entirely unfounded) fear of double counting some terms when employing an electrostatic correction and potential alignment together.[21] Even more worryingly, there seems to be a discontinuity in the community in the sign convention used when defining potential alignment. It appears that many authors take the potential shift defined in Eq. (1) as the average potential in the defect cell minus the average potential in the host cell, whatever their definition of these averages is,[20,22,23] while other authors do completely the opposite.[17,24]

Hence, the best way to proceed when attempting to improve the convergence for the supercell technique for charged defects is rather unclear. As a first illustrative example, we provide in Fig. 1 the convergence of the formation energy of two charged defects in silicon: the tetrahedrically coordinated

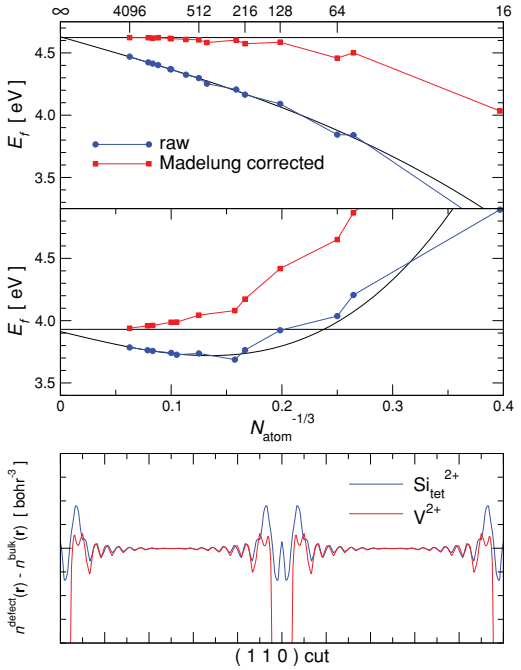

FIG. 1. (Color online) Convergence as a function of the supercell size of the formation energy (top) of a silicon interstitial Si_{Tet}^{2+} and (middle) of a silicon vacancy V_{Si}^{2+}. The raw energies are represented with circles. The Madelung corrected energies are represented with squares. The horizontal lines represent the converged values, and the thin dashed lines are tentative extrapolations with the usual function $\gamma_1 N^{-1/3} + \gamma_2 N^{-1}$. (bottom) A cut of the difference in electronic densities between the defective and host cells, $n^{\text{defect}}(\mathbf{r}) - n^{\text{bulk}}(\mathbf{r})$, along the (110) direction, passing through the bond centers for a 1000-atom cubic supercell.

self-interstitial Si_{Tet} (top panel) and the silicon vacancy V_{Si} (middle panel). The two defects have been considered in their 2+ charge state. For this charge state, they both have no occupied state in the band gap. They are both embedded in the same silicon host. In principle, one could have expected the same behavior as a function of the supercell size. Figure 1 obviously contradicts this prediction. The uncorrected data monotonously converge with a quite fair $N^{-1/3}$ behavior in the case of Si_{Tet}^{2+}. In the case of V_{Si}^{2+}, the convergence experiences a turning point. The inclusion of the simple Madelung electrostatic correction performs very well for the former and very poorly for the latter. This different behavior could not easily be anticipated from the electronic structure. The bottom panel of Fig. 1 shows a cut of the difference of electronic density between the defective and pristine supercells. Except in the vicinity of the defect, the electronic density differences at middle range simply show some Friedel's type oscillations with similar amplitudes. This clearly shows that the solution to the problem does not lie in an improved definition of electrostatic corrections.

In this article, we carefully address the different errors affecting the energy obtained in supercell calculations. We leave aside the elastic relaxations that produce much weaker effects, and we concentrate on the electronic structure problems. We summarize the computational aspects in Sec. II. In Sec. III, we demonstrate that the electrostatic interactions induce a position-dependent shift in the potential. As a consequence, the definition of the potential alignment should be revised to ensure the electrostatic contribution is not erroneously double counted. Furthermore, our proposed potential alignment is opposite in sign to some definitions (Sec. IV). We finally identify a prominent contribution to the error in the supercell technique: the spurious hybridization of defect wave functions onto several images. This contribution is usually completely disregarded. We provide in Sec. V a simple and practical way to evaluate this involved term. The performance of our three corrections is then demonstrated using various typical examples.

II. COMPUTATIONAL DETAILS

All the DFT calculations presented here utilized the local density approximation (LDA) for the exchange-correlation functional, as implemented within the plane-wave code ABINIT.[25] Norm-conserving Troulliers-Martins[26] pseudopotentials were used for sodium and chloride, with only the 1*s* electrons treated as core for sodium. For silicon, we developed an extremely smooth pseudopotential using the FHI-98PP program.[27] This pseudopotential has a very large cutoff radius of 4.0 bohr for both the *s* and the *p* channels. This somewhat crude pseudopotential yields a surprisingly good lattice parameter (5.408 Å) and defect formation energies. The very low plane-wave cutoff of 2.0 Ha enabled us to study phenomenally large supercells (up to 4096 atoms). For NaCl and for Si, $4 \times 4 \times 4$ and $2 \times 2 \times 2$ shifted Monkhorst-Pack[28] *k*-point grids were used for primitive cells and supercells, respectively. A lattice constant of 5.646 Å was used for NaCl, and the cells were left unrelaxed throughout the calculations. For silicon, following Ref. 9, the four neighbors nearest to the defect have been relaxed in the 64-atom cell, and these positions of the nearest neighbors are used for all further calculations.

III. EVALUATING THE POTENTIAL SHIFT INDUCED BY THE ELECTROSTATIC CORRECTION

In this section, we demonstrate that the electrostatic correction ΔE_{el} and the potential alignment ΔV are, indeed, connected quantities. Having this connection in mind will allow us to propose an evaluation of the potential alignment that does not double count the spurious electrostatic potential of the supercell approach.

In order to keep the discussion simple we consider here the simplest electrostatic correction, the monopole Madelung term, as first proposed by Leslie and Gillan:[3]

$$\Delta E_{el} = E_{el}^{\text{isolated}} - E_{el}^{\text{periodic}} \approx \frac{\alpha q^2}{2\epsilon L}, \tag{2}$$

where α is the Madelung constant of the lattice, q is the unbalanced charge, and L is the edge of the periodic box. The

monopole correction is designed to transform the electrostatic energy of a lattice of point charges in a neutralizing background into the electrostatic energy of a single point charge. In polarizable medium such as a solid, the Coulomb interaction is further screened by the electrons, and the electrostatic energy should be divided by the electronic dielectric constant ϵ. Here we use ϵ_∞ since the atoms are not allowed to relax in the present study.

Some authors attempt to improve convergence by including the third-order quadrupole in the electrostatic correction. We have avoided doing this for three reasons. First, one of our primary aims in this work was to produce an effective, useful, and, crucially, *simple* correction scheme. Hence, we utilize the simplest possible electrostatic correction. Second, as mentioned in Sec. I, the similarity in the electronic density difference between two silicon defects that converge at vastly different rates proves that improving our definition of the electrostatic correction will not solve the problem. In fact, this electronic density difference is the key quantity in the quadrupole term, lending further weight to this assumption. Finally, a relatively recent study[8] showed that the quadrupole correction does not always improve results, leaving its utility somewhat in question. In fact, since the quadrupole term always acts in the opposite direction to the Madelung monopole, it will always worsen results for defects that are converging from below after the monopole correction has been applied (e.g., the silicon interstitial in Fig. 1).

Let us now prove rigorously that the monopole term in Eq. (2) already introduces a shift in the potentials. The Kohn-Sham (KS) potential v_{KS} is obtained by the functional derivative of the total energy minus the kinetic energy with respect to the electronic density $n(\mathbf{r})$:[2]

$$v_{\mathrm{KS}}(\mathbf{r}) = \frac{\delta(E[n] - T[n])}{\delta n(\mathbf{r})}. \tag{3}$$

If the energy $E[n]$ requires an electrostatic correction ΔE_{el}, so will the obtained potential.

The functional derivative of the KS potential with the electrostatic correction can easily be tracted if the expression of the charge q as a function of the density is introduced:

$$q = \sum_i Z_i - \int d\mathbf{r} n(\mathbf{r}), \tag{4}$$

where $\sum_i Z_i$ is the total of the ionic charges in the cell.

Hence, the periodic KS potential also contains a spurious contribution when compared to the isolated KS potential, if one assumes a monopole correction:

$$v_{\mathrm{KS}}^{\mathrm{periodic}}(\mathbf{r}) = v_{\mathrm{KS}}^{\mathrm{isolated}}(\mathbf{r}) - \frac{d}{dq}(-\Delta E_{el}) \tag{5}$$

$$= v_{\mathrm{KS}}^{\mathrm{isolated}}(\mathbf{r}) + \frac{\alpha q}{\epsilon_\infty L}, \tag{6}$$

where the minus sign in the first line comes from the differentiation of Eq. (4) with respect to the electronic density. Finally, we see the KS potential in a periodic supercell is shifted with respect to the KS potential that an isolated charge would have, by the Madelung potential constant v_M, which reads[11]

$$v_M = -\frac{\alpha q}{\epsilon_\infty L}. \tag{7}$$

We have thus demonstrated that the charged-supercell approach introduces a significant shift in the KS potentials, which slowly decays as $1/L$. Therefore, one cannot consider independently correcting the electrostatic energy and correcting via potential alignment.

Keeping this in mind, what should be the practical procedure to perform a consistent, reliable potential alignment? In order to approach this problem, we have implemented a simple Poisson solver for periodic systems, based on fast Fourier transforms, completely analogous to the technique used in periodic DFT codes. This code allowed us to produce the data for Fig. 2 that present the electrostatic potential of a positive point charge (in reality, a Gaussian with a very small width) as it would be calculated in any periodic code. The parameters were chosen to represent a positive charge, located at zero in a cubic 512-atom supercell of sodium chloride. The interactions were scaled down with the calculated dielectric constant ϵ_∞. The choice of NaCl is governed by the desire to have localized defects that ease understanding.

A truly isolated point charge q in a medium should create a long-range Coulomb potential $q/\epsilon_\infty r$, as represented by a solid line in Fig. 2. The potential created by the truly isolated point charge goes asymptotically to zero. The calculated electrostatic potential of a point charge in a supercell with a compensating background, represented by the dashed line, deviates significantly from the single isolated charge. In

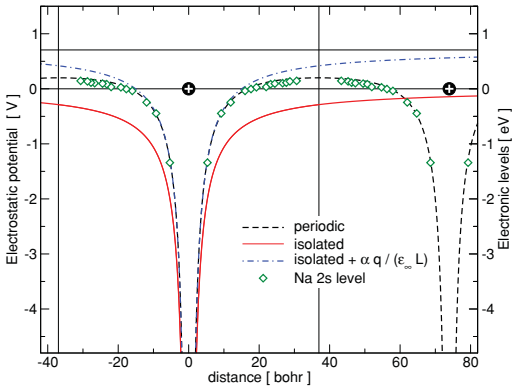

FIG. 2. (Color online) Electrostatic potentials created by a single point charge in an infinite sample (solid red line), an array of point charges with compensating background (dashed black line), and a single point charge in an infinite sample shifted by the Madelung potential $-v_M$ (dot-dashed blue line). The parameters (lattice constant, dielectric constant) have been chosen to mimic a cubic 512-atom supercell of NaCl. The green diamonds represent the deviation of the Na 2*s* core levels with respect to the bulk Na 2*s* levels, as obtained from a real calculation of a 512-atom supercell containing a vacancy V_{Cl}^{+}. The horizontal lines show the asymptotic values of the single point charge potentials.

the vicinity of the charge, the periodic potential appears as shifted with respect to the isolated potential. At the box boundary, the periodic potential experiences the two neighboring point charges equally and therefore shows a spurious plateau shape. We also introduced, with a dot-dashed line, the isolated potential shifted by the Madelung potential v_M following Eq. (5). We observe that this shifted potential closely reproduces the periodic potential in the vicinity of the point charge but asymptotically converges to $-v_M$. The divergence of the shifted isolated point-charge potential from the KS potential further away from the point charge would be reduced if higher-order terms in the Makov-Payne expansion were considered in Eq. (2). It is now obvious that the difference between the periodic potential and the shifted isolated potential is worst at the box boundary. Finally, in order to demonstrate that our modeling bears some connection to reality, we added the deviation in the 2*s* level of sodium with respect to bulk in an actual 512-atom supercell calculation of a chlorine vacancy V_{Cl}^{+}. As shown by the diamonds in Fig. 2, the calculated points and the periodic potential agree impressively well. The positions of the Na2*s* levels are simply governed by the screened electrostatic potential of the periodically replicated charges in a compensating background.

Many authors have prescribed performing the potential alignment by considering the electrostatic potential far from the charged defect as the zero of the potential.[8,9,13] In our opinion, this approach presents several problems. First of all, applying an electrostatic energy correction already brings about a shift in the potential, proportional to $1/L$. There is no need, therefore, to introduce another electrostatic potential alignment term that also goes as $1/L$, as this leads to double counting the same contribution. Second, when considering an energy correction brought about by an electrostatic potential shift, one needs to divide by a factor of 2, as shown when going from Eq. (7) to Eq. (2). This is not always clear in other potential alignment methods. Third, no matter how far from the defect one measures the potential alignment and no matter how large one makes the supercell, one can never recover the infinite-limit correct potential, with its long-range $1/r$ behavior. Fourth, considering the potential far from the defect is precisely the position where the deviation of the periodic potential from the isolated charge is the most striking: at the box boundary, the potential is equally generated by charges from different cells.

These conclusions show the crucial need to redefine the potential alignment. This is the topic of the next section.

IV. DEFINING THE PROPER POTENTIAL ALIGNMENT

Our goal is now to find a proper definition for the potential alignment ΔV introduced in Eq. (1). Potential alignment is needed for charged defects since the formation energy of a charged defect is a function of the Fermi level ϵ_F, i.e., the energy of the electrons from a reservoir. The energy zero is conventionally set to the top valence band of the bulk material, and the Fermi level is usually varied within the range of the band gap.

It was recognized very early on that the band structures of defective supercells are shifted with respect to their pristine counterparts and that, therefore, a potential alignment correction was needed.[29] Unfortunately, the potential alignment correction was mainly thought to correct for the spurious electrostatic potential, even though this contribution is usually already corrected through the electrostatic correction. Our definition for the potential alignment is, therefore, deliberately set up to ensure the electrostatic correction is not double counted. We suggest a correction that provides a naïve, extremely simple measurement of the potential shift yet performs surprisingly well, as we will show in the following.

We propose a scheme similar to that suggested in Ref. 17, whereby the average of the *total* potential over the *entire* supercell $\langle v_{KS} \rangle$ is considered:

$$\langle v_{KS} \rangle = \frac{1}{\Omega} \int_{\Omega} d\mathbf{r} v_{KS}(\mathbf{r}), \tag{8}$$

where Ω is the volume of the supercell. Why do we focus on this particular quantity?

First, the total average potential is completely free of any electrostatic contribution. Indeed, the average value of the electrostatic potential in a periodic cell is conventionally set to zero; otherwise, it would give rise to divergent terms. By considering the average potential we ensure that the electrostatic potential shift does not enter into the correction again, having already taken care of it via the previously defined ΔE_{el} term.

Second, a reference electron from the reservoir is one delocalized in a region infinitely far from the defect. In Fig. 3, this ideal situation is represented in the top schematic. The delocalized electron experiences the KS potential of the perfect bulk averaged over a large region. In practice, however, we perform a supercell calculation (schematic in the bottom panel of Fig. 3) where there is no region of space unaffected by the defect. An infinitely distant delocalized electron would

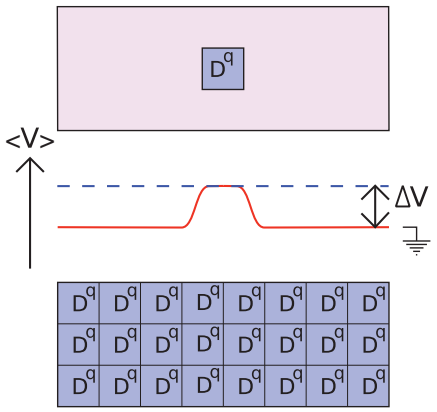

FIG. 3. (Color online) Schematics illustrating the role of the potential alignment ΔV. (top) The system we intend to simulate: a single charged defect D^q in a single supercell (dark blue), embedded in the infinite bulk (light pink). (bottom) The system we actually calculate with the supercell approach: an array of replicated defects D^q. (middle) The corresponding running average potentials, with a solid red line for the truly isolated defect and a dashed blue line for the supercell approach.

experience the KS potential of the defective supercell averaged over a large region. The potential alignment ΔV represented in the middle panel should, therefore, bring the average potential of the defective cell onto the reference average potential of the bulk cell:

$$\Delta V = \langle v_{\mathrm{KS}}^{\mathrm{bulk}} \rangle - \langle v_{\mathrm{KS}}^{\mathrm{defect}} \rangle. \tag{9}$$

Note that this definition of the potential alignment differs in sign with respect to the definition of some authors.[9,23] This potential alignment clearly states that the average potential obtained from supercell calculations is erroneous and should be corrected to fit the average potential of the pristine bulk. Finally, it should also be noted that the value defined in Eq. (8) is part of the standard output of the electronic structure code used in this study,[25] making evaluation of the potential alignment defined in Eq. (9) extremely quick and simple.

Let us demonstrate for a selected case the quality of the potential alignment we proposed in Eq. (9). As we intend to isolate the effect of potential alignment without the other corrections, we need it to be sizable. In Fig. 4 we considered the negatively charged sodium vacancy in NaCl. This particular case was chosen because one could expect a good performance of the Madelung correction in this defect. Indeed, the charge associated with the defect is very well localized; it is almost a point charge even for the smallest supercells. An informative sample case is a defect that is well converged after applying an electrostatic correction and potential alignment. We need this to hold even for small supercells, for which the potential alignment is large and its effect can be seen most clearly. The highly localized, nonshallow nature of the sodium vacancy allows it to agree with this demonstrative requirement.

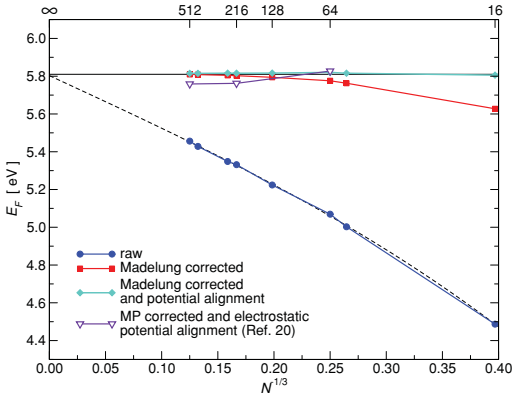

FIG. 4. (Color online) Convergence as a function of the supercell size of the formation energy of a sodium vacancy V_{Na}^{-}. The raw energies are represented with circles. The Madelung corrected energies are represented with squares. The data with potential alignment following Eq. (9) together with the Madelung correction are represented by diamonds. The data with potential alignment following Ref. 9 together with the Makov-Payne monopole and quadrupole corrections are represented by open triangles. The horizontal line represents our converged value, and the thin dashed line is a tentative extrapolation with the usual function $\gamma_1 N^{-1/3} + \gamma_2 N^{-1}$.

After the usual Madelung correction, the potential aligned data in Fig. 4 using Eq. (9) converge to the asymptotic value extremely quickly. Note that with a supercell as small as 16 atoms, the potential alignment $q\Delta V$ is as large as 0.18 eV, and applying it (along with the Madelung correction) leads to a corrected formation energy less than 10 meV from its converged value. This is somewhat compelling evidence that the sign convention we introduced in Eq. (9) is correct. For comparison, we also show in Fig. 4 results obtained with a quite popular alternative correction scheme, which combines the Makov-Payne correction (including terms up to the quadrupole) and an electrostatic potential alignment, as detailed in Ref. 9. As shown clearly in Fig. 4, our scheme appears to be converging to a slightly different value and at a much faster rate. Another correction scheme, detailed in Ref. 10, has already been shown to yield similar results to ours in the case of defects in NaCl, although it is rather more complicated to implement.

Note also that the potential alignment goes to zero very fast for larger supercells, as predicted. This may explain why, to date, it has proved difficult for the defect community to reach an agreement on the definition of potential alignment.

V. CORRECTING THE REMAINING NEIGHBOR'S INTERACTION

After correcting the electrostatic energy and the Fermi level with potential alignment, we are still left with some unexplained, slowly converging terms. For instance, neutral defects, which are unaffected by the two aforementioned corrections, may sometimes also experience a very slow convergence.[30,31] This behavior can be attributed, at least in part, to the quantum interaction between the defect and its images. Instead of being localized around one single defect, the defect-related wave functions can be delocalized over several images. This hybridization may lead to a change in the defect energy.

A similar behavior is observed and well documented in the context of adatoms on surfaces, where effective lattice gas models have been introduced.[32] We will now follow the same philosophy but simplify the situation by considering only a single kind of neighbors. The effective Hamiltonian H_n for a defect in a supercell interacting with n neighbors of the same kind reads

$$H_n = H_0 + nV, \tag{10}$$

where H_0 is the effective Hamiltonian with no neighbor interactions and V is the magnitude of the neighbor-neighbor interaction. The Hamiltonian H_0 is the target quantity, and H_n is the quantity obtained from a supercell calculation. In the modeling of Eq. (10) we assumed two-body interactions only.

We then propose to fit the two parameters H_0 and V of the model in Eq. (10) with two *ab initio* calculations. The first calculation is of a regular supercell, and the second calculation uses a nonregular supercell, for which one direction has been doubled. In doing so and assuming the next-nearest-neighbor interactions are small, we vary the number of interacting neighbors n and hence can extract the two parameters of the model.

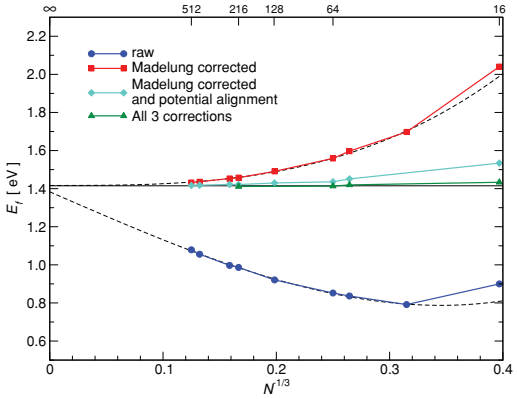

FIG. 5. (Color online) Convergence as a function of the supercell size of the formation energy of a chlorine vacancy V_{Cl}^{+}. The raw energies are represented with circles. The Madelung corrected energies are represented with squares. The data with potential alignment following Eq. (9) together with the Madelung correction are diamonds. The triangles represent the final data including the removal of the neighbor interaction according to Eq. (10). The horizontal line represents the converged value, and the thin dashed lines are tentative extrapolations with the usual function $\gamma_1 N^{-1/3} + \gamma_2 N^{-1}$.

The method is better explained with a practical example. We consider the chlorine vacancy V_{Cl}^{+} in NaCl in Fig. 5. In this case, again, the Madelung correction together with the potential alignment already yields a significantly improved result: the 16-atom supercell is converged to within 0.12 eV. NaCl is a textbook example for an ionic compound. The binding of the crystal is mediated through the isotropic Coulomb interaction. It is hence most probable that the defect states are isotropic too. As a consequence, we will assume that doubling the supercell in one direction will cut the magnitude of the interaction with neighbors by half. In practice, a calculation for a 16-atom face-centered-cubic supercell ($2 \times 2 \times 2$ unit cells) provided the value for $H_n = 1.50$ eV (after applying the Madelung correction and potential alignment), and a calculation for a 32-atom elongated face-centered-cubic supercell ($4 \times 2 \times 2$ unit cells) set the value for $H_{n/2} = 1.45$ eV. The extrapolated value for no defect-defect interactions is then easily obtained: $H_0 = 1.40$ eV, which lies within 0.02 eV of the converged value. The same procedure was also performed for larger supercells with a very good accuracy, as shown in Fig. 5. Our approach appears to be computationally relevant as well since the calculations of two small supercells (16 atoms and 32 atoms) offer an accuracy superior to the calculation with 64 atoms.

The model we propose considerably speeds up the convergence with respect to supercell size, at the expense of two calculations instead of one and some knowledge of the system under study. The approach crucially relies on the identification of the important directions of the crystal, with respect to the defect-defect interactions. In the case of NaCl, we assumed that all directions are equally important. However, returning to the case of silicon that we used as an introduction, we assume that the defect-defect interactions are preferentially mediated along the (110) zigzag chains of the diamond structure.[20] We thus considered the neighbors in these directions as the most relevant for the hybridization of defect states and set the values of n in Eq. (10) accordingly. When using a face-centered-cubic supercell, the (110) directions are, indeed, the first-nearest neighbors, and moving from a regular supercell to one doubled in one direction drops the number of nearest neighbors from 12 to 6.

When using a simple cubic supercell, it is the *second*-nearest neighbors that lie in the (110) directions. For extremely small simple cubic supercells, the assumption that the hybridization occurs only along the (110) directions is doubtful since the (100) neighbors are much closer. However, in order to assess the simplicity and the robustness of the present scheme, we will stick to our convention. For cubic supercells, the number of (110) neighbors drops from 12 to 4 when the length of one side of the cell is doubled. We used this framework to produce Fig. 6. First, note once again the performance of our potential alignment represented

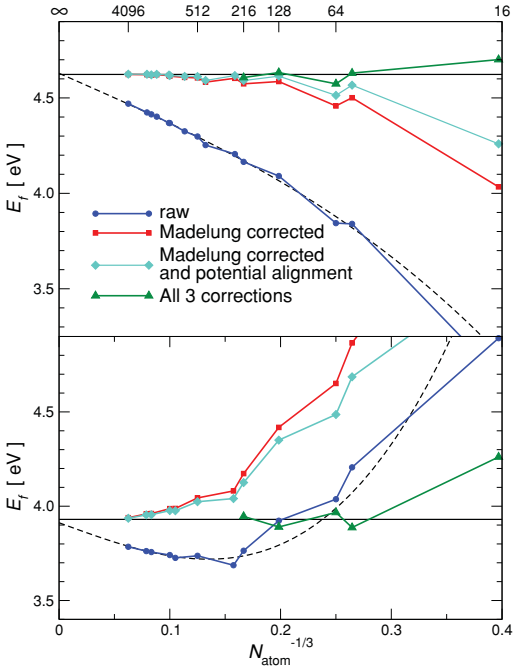

FIG. 6. (Color online) Convergence as a function of the supercell size of the formation energy of (top) a silicon interstitial Si_{Tet}^{2+} and (bottom) a silicon vacancy V_{Si}^{2+}. The raw energies are represented with circles. The Madelung corrected energies are represented with squares. The data with potential alignment following Eq. (9) together with the Madelung correction are diamonds. The triangles represent the final data, which include the removal of the neighbor interaction according to Eq. (10). The horizontal lines represent the converged value, and the thin dashed lines are tentative extrapolations with the usual function $\gamma_1 N^{-1/3} + \gamma_2 N^{-1}$.

with the diamonds. The triangles then show the final result of the present study including our three corrections. The agreement with the converged value is impressively good even for supercells as small as 54 atoms. The effect of the shape of the supercells becomes obvious: all the face-centered-cubic supercells converge to the asymptotic value from one side, and all the simple cubic supercells converge from the other. Finally, we should stress that this hybridization correction can also be utilized in the case of troublesome neutral defects. Indeed, the correction should prove particularly useful in these cases since the electrostatic and potential alignment terms do not apply.

VI. CONCLUSIONS

Calculations of charged point defects within the supercell approach are impossible to converge with a brute-force approach. Even our calculated 4096-atom supercells for defects in silicon still deviate largely from the asymptotic values. This makes it clear that more subtle approaches need to be designed and implemented. Many previous works addressed this issue, utilizing many different approaches, but the situation remains rather unsatisfactory.

The present contribution is twofold: a theoretical derivation that demonstrates that the spurious electrostatic energy introduced by the nonbalanced charge in the supercell induces a shift in the potential and a practical scheme using three simple corrections that significantly improve the convergence of the supercell approach.

The practical scheme we propose is extremely robust and simple and does not require additional coding. The only unconventional data needed here are the Madelung constant for nonregular cells. Our scheme reads (i) electrostatic correction, (ii) potential alignment, and (iii) hybridization correction. We showed that the simplest electrostatic correction of all, namely, the Leslie-Gillan Madelung correction,[3] is sufficient. We then showed that, if this electrostatic correction is applied, the potential alignment should be based on the *total average* Kohn-Sham potential to avoid double counting of the slowly converging $1/L$ term. Note that our definition uses a sign convention opposite to the definition of many authors. Finally, we could reduce the error due to the hybridization of defect states onto several images by using a simplistic model Hamiltonian and fitting it with two *ab initio* calculations. This hybridization correction could also be applied just as well, in principle, to the case of a slowly converging neutral defect.

Even though all the calculations presented here were based on local density approximation (LDA), the scheme could also be used in combination with hybrid functionals or the GW approximation.[33–35] The only cases that cannot be corrected within our scheme are those of shallow defect states, which are delocalized over regions that are impossible to fit into a tractable supercell. Besides this limitation, the efficiency and accuracy of our scheme has been impressive for all the cases tested so far.

ACKNOWLEDGMENTS

We acknowledge exciting discussions with J.-P. Crocombette and G. Roma and thank M.-C. Marinica for pointing out to us the literature on lattice gas models. This work was performed using HPC resources from GENCI-CINES and GENCI-CCRT (Grant No. 2011-gen6018).

[1] G. Grosso and G. Pastori Paravicini, *Solid State Physics* (Academic, San Diego, 2000).
[2] R. G. Parr and W. Yang, *Density-Functional Theory of Atoms and Molecules* (Oxford University Press, New York, 1989).
[3] M. Leslie and M. J. Gillan, J. Phys. C **18**, 973 (1985).
[4] A. B. Lidiard, J. Chem. Soc. Faraday Trans. 2 **85**, 341 (1989).
[5] A. R. Williams, J. F. Janak, and V. L. Moruzzi, Phys. Rev. B **6**, 4509 (1972).
[6] G. Makov and M. C. Payne, Phys. Rev. B **51**, 4014 (1995).
[7] P. A. Schultz, Phys. Rev. Lett. **84**, 1942 (2000).
[8] C. W. M. Castleton, A. Höglund, and S. Mirbt, Phys. Rev. B **73**, 035215 (2006).
[9] S. Lany and A. Zunger, Phys. Rev. B **78**, 235104 (2008).
[10] C. Freysoldt, J. Neugebauer, and C. G. Van de Walle, Phys. Rev. Lett. **102**, 016402 (2009).
[11] N. D. M. Hine, K. Frensch, W. M. C. Foulkes, and M. W. Finnis, Phys. Rev. B **79**, 024112 (2009).
[12] S. B. Zhang and J. E. Northrup, Phys. Rev. Lett. **67**, 2339 (1991).
[13] C. G. van de Walle and J. Neugebauer, J. Appl. Phys. **95**, 3851 (2004).
[14] A. Janotti and C. G. Van de Walle, Phys. Rev. B **76**, 165202 (2007).
[15] S. B. Zhang, J. Phys. Condens. Matter **14**, R881 (2002).
[16] P. Erhart, K. Albe, and A. Klein, Phys. Rev. B **73**, 205203 (2006).
[17] J. Shim, E.-K. Lee, Y. J. Lee, and R. M. Nieminen, Phys. Rev. B **71**, 035206 (2005).
[18] T. Mattila and A. Zunger, Phys. Rev. B **58**, 1367 (1998).
[19] D. B. Laks, C. G. Van de Walle, G. F. Neumark, P. E. Blöchl, and S. T. Pantelides, Phys. Rev. B **45**, 10965 (1992).
[20] S. Lany and A. Zunger, Modell. Simul. Mater. Sci. Eng. **17**, 084002 (2009).
[21] W. Chen, C. Tegenkamp, H. Pfnur, and T. Bredow, Phys. Rev. B **82**, 104106 (2010).
[22] C. Persson, Y.-J. Zhao, S. Lany, and A. Zunger, Phys. Rev. B **72**, 035211 (2005).
[23] S. Pöykkö, M. J. Puska, and R. M. Nieminen, Phys. Rev. B **53**, 3813 (1996).
[24] Y. Cui and F. Bruneval, Appl. Phys. Lett. **97**, 042108 (2010).
[25] X. Gonze, B. Amadon, P. M. Anglade, J. M. Beuken, F. Bottin, P. Boulanger, F. Bruneval, D. Caliste, R. Caracas, M. Cote, T. Deutsch, L. Genovese, P. Ghosez, M. Giantomassi, S. Goedecker, D. R. Hamann, P. Hermet, F. Jollet, G. Jomard, S. Leroux, M. Mancini, S. Mazevet, M. J. T. Oliveira, G. Onida, Y. Pouillon, T. Rangel, G. M. Rignanese, D. Sangalli, R. Shaltaf, M. Torrent, M. J. Verstraete, G. Zerah, and J. W. Zwanziger, Comput. Phys. Commun. **180**, 2582 (2009).
[26] N. Troullier and J. L. Martins, Phys. Rev. B **43**, 1993 (1991).

[27] M. Fuchs and M. Scheffler, Comput. Phys. Commun. **119**, 67 (1999).
[28] H. J. Monkhorst and J. D. Pack, Phys. Rev. B **13**, 5188 (1976).
[29] A. Garcia and J. E. Northrup, Phys. Rev. Lett. **74**, 1131 (1995).
[30] M. I. J. Probert and M. C. Payne, Phys. Rev. B **67**, 075204 (2003).
[31] A. F. Wright, Phys. Rev. B **74**, 165116 (2006).
[32] C. Stampfl, H. J. Kreuzer, S. H. Payne, H. Pfnür, and M. Scheffler, Phys. Rev. Lett. **83**, 2993 (1999).
[33] F. Oba, A. Togo, I. Tanaka, J. Paier, and G. Kresse, Phys. Rev. B **77**, 245202 (2008).
[34] F. Bruneval, Phys. Rev. Lett. **103**, 176403 (2009).
[35] M. Giantomassi, M. Stankovski, R. S. M. Grüning, F. Bruneval, P. Rinke, and G.-M. Rignanese, Phys. Status Solidi B **248**, 275 (2011).

Appendix D

Articles for Part III

PHYSICAL REVIEW B **78**, 085125 (2008)

Accurate *GW* self-energies in a plane-wave basis using only a few empty states: Towards large systems

Fabien Bruneval
CEA, DEN, Service de Recherches de Métallurgie Physique, F-91191 Gif-sur-Yvette, France

Xavier Gonze
European Theoretical Spectroscopy Facility, Université Catholique de Louvain, Place Croix du Sud 1, B-1348 Louvain-la-Neuve, Belgium

(Received 25 April 2008; revised manuscript received 4 July 2008; published 22 August 2008)

The *GW* approximation to the electronic self-energy yields band structures in excellent agreement with experimental data. Unfortunately, this type of calculation is extremely cumbersome even for present-day computers. The huge number of empty states required both in the calculation of the polarizability and of the self-energy is a major bottleneck in *GW* calculations. We propose an almost costless scheme, which allows us to divide the number of empty states by about a factor of 5 to reach the same accuracy. The computational cost and the memory requirements are decreased by the same amount, accelerating all calculations from small primitive cells to large supercells.

DOI: 10.1103/PhysRevB.78.085125 PACS number(s): 71.15.Qe, 71.20.Nr, 71.45.Gm

I. INTRODUCTION

Calculating the correct electronic band structure of a solid, especially in the band-gap region, is not a trivial task for *ab initio* methods. The commonly used density-functional theory (DFT)[1,2] is notoriously insufficient in that respect. To get realistic band structure from the computer, one has to resort to more accurate but also more cumbersome methods. In this context, Hedin's *GW* approximation[3–5] to the electronic self-energy has encountered a wide success for systems where correlation effects are not strong. Unfortunately, the cost of such calculations is generally two orders of magnitude higher that their DFT counterpart. Furthermore, the need to study nanowires, interfaces, or defects drives the interest of the scientific community towards larger and larger systems. It is urgent to find reliable techniques to speed up the *GW* approach and make it tractable for a wider range of applications.

The *GW* self-energy is a convolution of the Green's function G and the screened Coulomb potential $v\varepsilon^{-1}$

$$\Sigma(\omega) = \int d\omega' G(\omega+\omega') v\varepsilon^{-1}(\omega'), \tag{1}$$

where v is the Coulomb interaction and ε^{-1} is the inverse dielectric matrix. The dielectric matrix in turn is obtained from the random-phase approximation (RPA)

$$\varepsilon(\omega) = 1 - v\chi_0(\omega), \tag{2}$$

with $\chi_0(\omega)$ being the Kohn-Sham polarizability. A practical *GW* calculations consists of evaluating the polarizability $\chi_0(\omega)$ and then of performing the convolution in Eq. (1).

The main bottleneck in the efficiency of a *GW* calculation is the dependence with respect to the empty states. In contrast with Kohn-Sham DFT, the two ingredients in a *GW* run, i.e., the polarizability and the *GW* self-energy itself, both involve explicitly the unoccupied states. The evaluation of the *GW* band structure requires, first, calculating a huge quantity of empty Kohn-Sham eigenvectors and eigenvalues and, second, using them in sums running over all the states of the system. The poor convergence of the *GW* approximation with respect to the empty states has been recognized long ago.[5–7] In order to be exact, the number of states should be the same as the dimension of the Hilbert space that is equal to the number of basis functions. In a plane-wave basis or in a real-space representation, the dimension of the space is huge (typically from thousand to millions). As a consequence, a speedup of the *GW* approach should address the elimination or the reduction in the number of empty states in the calculations. This direction has already been identified by several other groups.[8,9] However, the previously proposed techniques are not widely used nowadays because of either low efficiency or because the cost exceeds the benefits for the available system sizes.

In the context of the optimized effective potentials, the same problem arises—one needs to invert the empty state-dependent Kohn-Sham polarizability in order to obtain the local Kohn-Sham potential that represents best a nonlocal exchange-correlation operator.[10] For this framework, several schemes have been developed to get rid of the empty states dependence. In the 50's, Sharp and Horton[11] have already proposed a rough approximation, which was then reused in the celebrated Krieger-Li-Iafrate approximation of the exchange-only potential.[12] More recently, Gritsenko and Baerends[13] improved much on this approach with their common energy denominator approximation (CEDA).

In this paper, we propose a technique that allows us to reduce the number of unoccupied states required in the two steps of a *GW* calculation (for the polarizability and the self-energy) in a plane-wave implementation. In order to achieve this goal, we transpose the CEDA trick of Gritsenko and Baerends[13] into the framework of the *GW* approximation. For the polarizability step, this corresponds to a simple extension of the extrapolar method of Anglade and Gonze.[14] By replacing the eigenenergies of the states that are not treated explicitly by a common energy, determined with respect to the highest computed eigenstate through a single adjustable parameter, we will be able to take into account all the states, which are not explicitly included in the calculation through the closure relation

$$\sum_{i>N_b} |i\rangle\langle i| = \mathbb{1} - \sum_{i\leq N_b} |i\rangle\langle i|, \tag{3}$$

where N_b is the number of states explicitly included in the calculation. The principle can be extended to other formulations which have a large Hilbert-space dimension such as real-space approach. This permits us to provide a correction to the polarizability and to the self-energy that approximates the effect of the states not explicitly taken into account. The only drawback of the extrapolar approximation is the introduction of a parameter that can be thought as *ad hoc*. In this paper, we also present a formula that gives an *ab initio* evaluation of this parameter.

In Sec. II, we focus on the computation of the polarizability with a limited number of empty states. After developing the corresponding equations, we examine the effect of the number of empty states treated in the polarizability on the *GW* corrections for the band gap of SiC. In Sec. III, we propose to use the sum rule for the first moment of the dielectric function to analyze the effect of the eigenenergy approximation as a function of the transferred momentum. Thanks to a proper weight factor; the best value for the adjustable parameter might be determined. In Sec. IV, we perform the same approximation in the self-energy expression assuming the dynamically screened Coulomb interaction to be well represented by a generalized plasmon-pole model for the large-energy transfers. The adequacy of this approximation increases with the number of states explicitly computed. Unlike for the polarizability, no sum rule exists for the self-energy. However, we argue that the value of the adjustable parameter, optimized for the polarizability, is likely close to the optimal value for the self-energy. We apply the methodology to the case of bulk SiC, to a 64-atom supercell of SiC, to the insulator of argon, and also to an isolated benzene molecule.

II. POLARIZABILITY WITH A LIMITED NUMBER OF EMPTY STATES

In this section, we recapitulate the extrapolar approximation of Ref. 14 for the empty states that are not included explicitly in the calculation and derive the corresponding correction to the independent-particle polarizability χ_0.

The formulas are written here for spin-unpolarized and nonmetallic systems, but they can be straightforwardly extended to spin-polarized systems and to metals by introducing fractional occupations. Using the time-reversal symmetry, the independent-particle polarizability in reciprocal space and frequency reads

$$\chi_{0\mathbf{G}\mathbf{G}'}(\mathbf{q},\omega) = \frac{2}{N_\mathbf{k}\Omega} \sum_{\substack{\mathbf{k} \\ N_v<i\leq N_b \\ j\leq N_v}} M_{\mathbf{k}ij}(\mathbf{q}+\mathbf{G}) M^*_{\mathbf{k}ij}(\mathbf{q}+\mathbf{G}') \times \left[\frac{1}{\omega - (\epsilon_{\mathbf{k}j} - \epsilon_{\mathbf{k}-\mathbf{q}i}) - i\eta} - \frac{1}{\omega - (\epsilon_{\mathbf{k}-\mathbf{q}i} - \epsilon_{\mathbf{k}j}) + i\eta} \right], \tag{4}$$

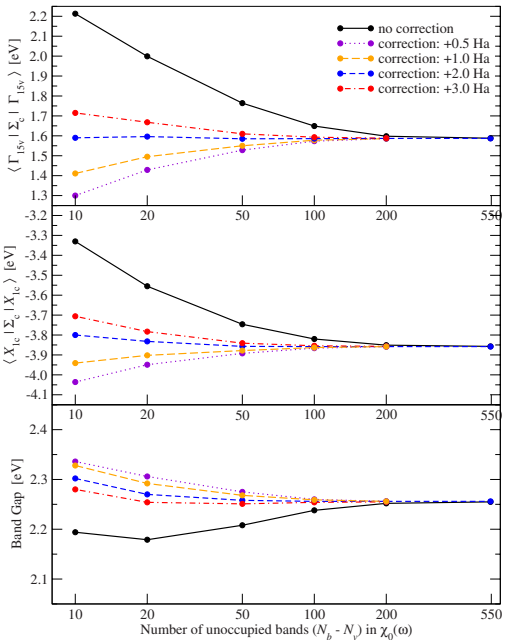

FIG. 1. (Color online) Convergence study of the correlation part of the self-energy at top valence (upper panel) and at bottom conduction (middle panel) and of the band gap (lower panel) of β-SiC as a function of the number of unoccupied states explicitly included in the calculation of the polarizability. The solid curve shows the usual *GW* result with no correction. The other curves include the correction of Eq. (8) with different values for the energy parameter $\bar{\epsilon}_{\chi_0}$: 0.5 Ha , 1.0 Ha , 2.0 Ha, and 3.0 Ha above the last explicitly calculated band.

where Ω is the volume of the unit cell, N_v is the number of valence states, $N_\mathbf{k}$ is the number of $\mathbf{k}$ points in the Brillouin zone (the index $\mathbf{k}$ runs over the $\mathbf{k}$ points of the Brillouin zone), and where the matrix elements

$$M_{\mathbf{k}ij}(\mathbf{q}+\mathbf{G}) = \langle \mathbf{k}-\mathbf{q}i | e^{-i(\mathbf{q}+\mathbf{G})\cdot\mathbf{r}} | \mathbf{k}j \rangle \tag{5}$$

are the so-called oscillator strengths.

In practice, the number of unoccupied states needed in Eq. (4) can be very large in order to converge the value of the self-energy calculated using this polarizability. In this paper, the numerical applications are performed first on bulk β-SiC, which is slightly more sensitive than the prototypical bulk silicon to the number of states. The solid curve in Fig. 1 illustrates the convergence of the correlation part of the self-energy at the top of the valence band and at the bottom of the conduction band (and the resulting band gap) as a function of the number of unoccupied bands considered in the calculation of χ_0. More than 100 empty states are required to obtain the self-energy at the top valence Γ_{15v} with the typical accuracy of *GW* calculations, i.e., 50 meV. Furthermore, Fig. 1 shows that the convergence rate of the bottom conduction

band X_{1c} is not the same as of Γ_{15v}. Therefore, it would be interesting to accelerate this poor convergence; thanks to a properly defined correction.

The extrapolar approximation proposes to attribute to all the states above N_b the same "average" high energy $\bar{\varepsilon}_{\chi_0}$. Obviously, this energy should lie higher than the energy of the last actually calculated band. But so far, this energy is considered as a parameter.

Let us define the correction $\Delta_{\mathbf{GG}'}(\mathbf{q},\omega)$, which is the quantity neglected in Eq. (4) due to the truncation of the unoccupied state sum. Introducing this average energy in this correction allows one to change the order of the sums as follows:

$$\Delta_{\mathbf{GG}'}(\mathbf{q},\omega) = \frac{2}{N_{\mathbf{k}}\Omega}\sum_{\substack{\mathbf{k}\\ j\le N_v}}\left[\frac{1}{\omega-(\epsilon_{\mathbf{k}j}-\bar{\epsilon}_{\chi_0})-i\eta} - \frac{1}{\omega-(\bar{\epsilon}_{\chi_0}-\epsilon_{\mathbf{k}j})+i\eta}\right] \times \sum_{i>N_b} M_{\mathbf{k}ij}(\mathbf{q}+\mathbf{G})M^*_{\mathbf{k}ij}(\mathbf{q}+\mathbf{G}'), \quad (6)$$

in which the only quantity depending on the empty states i are the oscillator strengths $M_{\mathbf{k}ij}(\mathbf{q}+\mathbf{G})$.

The closure relation can be straightforwardly applied to

$$\sum_{i>N_b} M_{\mathbf{k}ij}(\mathbf{q}+\mathbf{G})M^*_{\mathbf{k}ij}(\mathbf{q}+\mathbf{G}') = \langle \mathbf{k}j\ e^{i(\mathbf{G}'-\mathbf{G}).\mathbf{r}}\ \mathbf{k}j\rangle - \sum_{i\le N_b} M_{\mathbf{k}ij}(\mathbf{q}+\mathbf{G})M^*_{\mathbf{k}ij}(\mathbf{q}+\mathbf{G}'), \quad (7)$$

in order to get rid of the states above N_b. The final expression for the correction to the independent-particle polarizability within the extrapolar approximation is

$$\Delta_{\mathbf{GG}'}(\mathbf{q},\omega) = \frac{2}{N_{\mathbf{k}}\Omega}\sum_{\substack{\mathbf{k}\\ j\le N_v}}\langle \mathbf{k}j|e^{i(\mathbf{G}'-\mathbf{G}).\mathbf{r}}|\mathbf{k}j\rangle\left[\frac{1}{\omega-(\epsilon_j-\bar{\epsilon}_{\chi_0})-i\eta} - \frac{1}{\omega-(\bar{\epsilon}_{\chi_0}-\epsilon_{\mathbf{k}j})+i\eta}\right] - \frac{2}{N_{\mathbf{k}}\Omega}\sum_{\substack{\mathbf{k}\\ i\le N_b\\ j\le N_v}} M_{\mathbf{k}ij}(\mathbf{q}+\mathbf{G})M^*_{\mathbf{k}ij}(\mathbf{q}+\mathbf{G}') \times\left[\frac{1}{\omega-(\epsilon_{\mathbf{k}j}-\bar{\epsilon}_{\chi_0})-i\eta} - \frac{1}{\omega-(\bar{\epsilon}_{\chi_0}-\epsilon_{\mathbf{k}j})+i\eta}\right]. \quad (8)$$

The calculation of this correction does not require much coding when the polarizability is already available. Furthermore, it produces very little overhead in the calculation time. Indeed, the first term in Eq. (8) does not have any sum over empty states and basically requires one fast Fourier transform per $\mathbf{k}$ point and per occupied state. The second term can be merged with the corresponding part in the calculation of χ_0 for each triplet index $(\mathbf{k}ij)$. So it does not add any complex operation. The calculation of the correction is really for free.

The extrapolar approximation is not designed to yield the right frequency-dependent polarizability. As this approximation replaces the many neglected high-energy transitions by a single transition with a large weight, the imaginary part of the polarizability would look like a single δ peak at high energy instead of a continuous spectrum. Nevertheless, one can reasonably hope that this approximated polarizability, when integrated, retains some physics. The *GW* self-energy is precisely an integrated quantity as seen in Eq. (1). Let us check before this assumption on the static dielectric matrix. The static dielectric matrix can be considered as an integration of the frequency-dependent dielectric function through the Kramers-Kronig relation

$$\mathrm{Re}\{\varepsilon(\omega=0)\} = 1+\frac{1}{\pi}\mathcal{P}\int_{-\infty}^{+\infty} d\omega' \frac{\mathrm{Im}\{\varepsilon(\omega')\}}{\omega'}. \quad (9)$$

Table I shows the convergence with or without correction of the static inverse dielectric matrix. The value of the average extrapolar energy $\bar{\epsilon}_{\chi_0}$ is referenced with respect to the highest calculated energy ϵ_{N_b}. It can be observed from the data that whatever the choice of the extrapolar energy (in a reasonable range), the convergence of diagonal and off-diagonal elements of the dielectric matrix is accelerated.

Now let us describe the quality of the extrapolar correction for the polarizability when it is used to evaluate the *GW* self-energy. The performance for β-SiC is shown in Fig. 1. Whatever the value of $\bar{\epsilon}_{\chi_0}$, the convergence of the self-energy and of the band gap is improved significantly; thanks to the correction. When $\bar{\epsilon}_{\chi_0}$ is chosen too high with respect to the last calculated band, the correction vanishes and the results tend to the uncorrected one. When $\bar{\epsilon}_{\chi_0}$ is chosen too close to the last calculated band, the correction is then slightly overestimated. The best fit is obtained for an average energy of around 2.0 Ha above ϵ_{N_b}. If a reasonable value for $\bar{\epsilon}_{\chi_0}$ is used, the number of empty states can be lowered to around 20 to achieve the 50-meV accuracy. This corresponds to five times fewer states as without correction. As a conclusion, the numerical application strongly supports the use of the correction to the polarizability.

III. USING A SUM RULE TO DETERMINE THE ENERGY PARAMETER

In Sec. II, we have shown that the precise determination of the average energy $\bar{\epsilon}_{\chi_0}$ is not crucial as it provides an accurate correction for a wide range of $\bar{\epsilon}_{\chi_0}$. However, it would be desirable to have a tool, which measures the quality of a choice of an average energy $\bar{\epsilon}_{\chi_0}$ without knowing before the exact target result.

A common procedure to assign the value of parameters is to enforce the fulfillment of exact relations. For response functions, there exists a class of integrals, of which the value is known exactly. For instance, the first moment of the in-

TABLE I. Convergence study of some selected elements of the inverse dielectric matrix of β-SiC as a function of the number of empty bands explicitly included in the calculation of the polarizability. The first element is the macroscopic static dielectric constant. The second element is a diagonal element and the last one is off diagonal.

		Number of empty states				(N_b-N_v)
	Extrapolar energy (Ha)	4	10	20	50	200
$1/\varepsilon^{-1}_{(000),(000)}(\mathbf{q}\to 0,\omega=0)$	No correction	6.617	6.722	6.737	6.748	6.755
	0.5	6.728	6.790	6.762	6.753	6.754
	1.0	6.700	6.776	6.758	6.753	6.754
	2.0	6.673	6.761	6.753	6.752	6.754
	3.0	6.659	6.752	6.750	6.751	6.754
$\varepsilon^{-1}_{(100),(100)}(\mathbf{q}\to 0,\omega=0)$	No correction	0.792	0.708	0.662	0.645	0.645
	0.5	0.637	0.645	0.646	0.644	0.645
	1.0	0.676	0.658	0.648	0.644	0.645
	2.0	0.715	0.673	0.652	0.645	0.645
	3.0	0.735	0.681	0.654	0.645	0.645
$\varepsilon^{-1}_{(100),(0\bar{1}0)}(\mathbf{q}\to 0,\omega=0)$	No correction	0.037	0.030	0.027	0.026	0.026
	0.5	0.026	0.026	0.026	0.026	0.026
	1.0	0.029	0.027	0.026	0.026	0.026
	2.0	0.031	0.028	0.027	0.026	0.026
	3.0	0.032	0.028	0.027	0.026	0.026

verse dielectric matrix $\int d\omega\omega\ \mathrm{Im}\ \varepsilon^{-1}[\mathbf{q},\omega]$ is fixed by the so-called *f -sum rule*. This relation allowed Hybertsen and Louie[15] to calculate the free parameters of their model to represent the inverse dielectric matrix.

In the present case, the evaluation of the *f*-sum rule would not be adequate because it would require either modeling the inverse dielectric function or performing a numerical frequency integration subjected to discretization error. Instead, another sum rule exists for the first moment of the dielectric function itself;[16,17]

$$\int_0^{+\infty} d\omega\omega\ \mathrm{Im}[\varepsilon_{\mathbf{GG}}(\mathbf{q},\omega)] = \frac{\pi}{2}\omega_p^2, \tag{10}$$

where $\omega_p=\sqrt{4\pi n}$ is the classical plasma frequency (n being the average electronic density). The sum rule in Eq. (10), although not valid in general, has been shown to be true for the RPA dielectric matrix.[17] The RPA is precisely the approximation used for the *GW* self-energy. In the present discussion, we will concentrate only on the diagonal elements of the dielectric matrix since these elements yield by far the largest contribution to the *GW* self-energy.

In the RPA approximation, the dielectric matrix is related to the independent-particle polarizability through $\epsilon(\omega)=1-v\chi_0(\omega)$, where v is the Coulomb potential. Hence, the sum rule of Eq. (10) reads

$$\int_0^{+\infty} d\omega\omega \frac{4\pi}{|\mathbf{q}+\mathbf{G}|^2}\mathrm{Im}[\chi_{0\mathbf{GG}}(\mathbf{q},\omega)] = -\frac{\pi}{2}\omega_p^2, \tag{11}$$

in which $4\pi/|\mathbf{q}+\mathbf{G}|^2$ is the Fourier transform of the Coulomb potential v.

The check of the validity of Eq. (11) provides a stringent test on the completeness in the calculation of χ_0. If the sum over states in Eq. (4) has been truncated, the integral in the left-hand side of Eq. (11) will be too small. The advantage of Eq. (11) with respect to the *f*-sum rule is that the evaluation of the integral can be performed analytically, as the frequency dependence of $\mathrm{Im}[\chi_0(\omega)]$ consists only of a series of δ peaks; thanks to the classical identity

$$\lim_{\eta\to 0}\frac{1}{\omega+i\eta} = \mathcal{P}\frac{1}{\omega} - i\pi\delta(\omega). \tag{12}$$

Practically, introducing the expression of $\mathrm{Im}[\chi_0]$ in Eq. (11), it reduces to

$$\frac{4\pi^2}{N_{\mathbf{k}}\Omega|\mathbf{q}+\mathbf{G}|^2}\sum_{\substack{\mathbf{k}\\ N_v<i\le N_b\\ j\le N_v}}|M_{\mathbf{k}ij}(\mathbf{q}+\mathbf{G})|^2(\epsilon_{\mathbf{k}-\mathbf{q}i}-\epsilon_{\mathbf{k}j}) = \frac{\pi}{2}\omega_p^2. \tag{13}$$

In the upper panel of Fig. 2, we report for β-SiC the evaluation of the left-hand side of Eq. (13) as a function of the transferred momentum $|\mathbf{q}+\mathbf{G}|$ of different number of unoccupied bands included in the calculation. When almost all the states available are included in the calculation (550 unoccupied bands), the sum rule is verified for any value of $|\mathbf{q}+\mathbf{G}|$. Contrastingly, when only a few empty states are taken into account (e.g., ten unoccupied bands), the sum rule is only approximately satisfied for low transferred momenta. This expresses the fact that well-separated occupied and unoccupied states can couple through electronic transition with

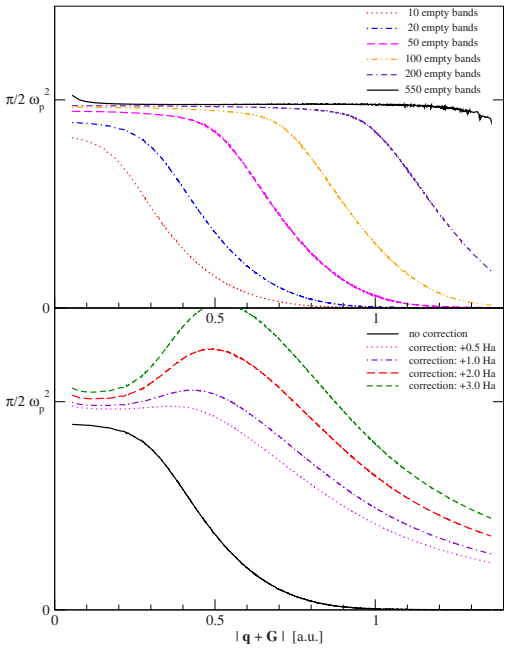

FIG. 2. (Color online) Upper panel: Value of the integral in Eq. (10) as a function of the transferred momentum $|\mathbf{q}+\mathbf{G}|$ without any correction using 10, 20, 50, 100, 200, and 550 empty bands in β-SiC. Lower panel: Value of the integral in Eq. (10) as a function of the transferred momentum $|\mathbf{q}+\mathbf{G}|$ with 20 empty bands using no correction or a correction with an average energy $\bar{\epsilon}_{\chi_0}$ of 0.5, 1.0, 2.0, and 3.0 Ha above the last explicitly calculated band in β-SiC.

a large momentum. For large transferred momenta, the coupling between far apart states cannot be neglected without damaging the polarizability.

The bottom panel of Fig. 2 shows how the extrapolar correction to χ_0 proposed in Eq. (8) affects the sum rule for a fixed number of 20 unoccupied bands with different values of the extrapolar energy parameter. All calculations that include the correction fulfill the sum rule with a much higher accuracy than the reference curve without correction. The completeness correction is a large step towards a fulfillment of the sum rule, especially for large transferred momenta. By using the present correction, it is possible that for some values of $|\mathbf{q}+\mathbf{G}|$, the sum rule gets overestimated. This can allow one to compensate for the underestimation of the highest values of $|\mathbf{q}+\mathbf{G}|$.

As a consequence, a sensible approach is to seek for the value of the extrapolar energy $\bar{\epsilon}_{\chi_0}$, which in average allows for the best compliance to the sum rule. The significance of each transferred momentum has to be weighted by its importance in the subsequent *GW* self-energy calculation since the ultimate goal is merely to evaluate *GW* band structures. The contribution of the polarizability χ_0 in the *GW* correlation is proportional to $(\varepsilon^{-1}-1)v$. In order to have a rough estimate of this weight, we can assume that all the matrices χ_0, ε, and ε^{-1} are diagonal and are considered in the static limit $\omega \to 0$ since this is the dominating contribution. Under these assumptions, the weight w assigned to the sum rule for the momentum $|\mathbf{q}+\mathbf{G}|$ is

$$w(\mathbf{q}+\mathbf{G}) \propto \frac{1}{|\mathbf{q}+\mathbf{G}|^2}[\varepsilon^{-1}_{\mathbf{GG}}(\mathbf{q},\omega=0)-1]. \tag{14}$$

When applying this procedure to β-SiC with 20 unoccupied states (the same conditions as the bottom panel of Fig. 2), the best choice of the average energy appears to be ~1.6 Ha above the last explicitly calculated band. This is in good agreement with the quality of the curve with the average energy $\bar{\epsilon}_{\chi_0}$ chosen at 2.0 Ha above the last band in Fig. 1.

IV. SELF-ENERGY WITH A LIMITED NUMBER OF EMPTY STATES

In the expression of the *GW* self-energy as in the formula of the polarizability, a sum over all the states is present. In the present section, using the same procedure as the one shown previously for the polarizability, we propose a correction to the self-energy that allows us to reduce drastically the number of empty states required.

For the sake of simplicity, we show the derivation of diagonal matrix elements, but the extension to the off-diagonal terms needed in self-consistent *GW* (Refs. 18 and 19) is straightforward. The exchange part of the *GW* self-energy is the Fock exchange operator and therefore does not involve empty states. The diagonal matrix element of the correlation part of the *GW* self-energy expressed in a plane-wave basis reads[15]

$$\langle \mathbf{k}j|\Sigma_c(\varepsilon_{\mathbf{k}j})|\mathbf{k}j\rangle = \frac{i}{2\pi N_{\mathbf{k}}\Omega}\int d\omega' \sum_{i\le N_b}\sum_{\mathbf{qGG}'}[W_{\mathbf{GG}'}(\mathbf{q},\omega') - \delta_{\mathbf{GG}'}v(\mathbf{q}+\mathbf{G})] \times \frac{M_{ji}(\mathbf{q}+\mathbf{G})M^*_{ji}(\mathbf{q}+\mathbf{G}')}{\omega' - \epsilon_{\mathbf{k}-\mathbf{q}i} + \epsilon_{\mathbf{k}j} \pm i\eta}, \tag{15}$$

where η is a vanishing positive real. The sign in front of η is plus when the state i is empty and minus otherwise.

The correction we propose is again to account for the states $i > N_b$ through an extrapolar energy $\bar{\epsilon}_\Sigma$. This permits us to interexchange the order of the sum over bands and all the rest except the oscillator strengths in Eq. (15). Then, we can apply the closure relation written in Eq. (7). Hence, the extrapolar correction $\Delta_{\mathbf{k}j}$ to the self-energy reads

$$\Delta_{\mathbf{k}j} = \frac{i}{2\pi N_{\mathbf{k}}\Omega}\int d\omega' \sum_{\mathbf{qGG}'} \frac{W_{\mathbf{GG}'}(\mathbf{q},\omega') - \delta_{\mathbf{GG}'}v(\mathbf{q}+\mathbf{G})}{\omega' - \bar{\epsilon}_\Sigma + \epsilon_{\mathbf{k}j} + i\eta} \times \left[\langle \mathbf{k}j|e^{i(\mathbf{G}'-\mathbf{G})\cdot\mathbf{r}}|\mathbf{k}j\rangle - \sum_{i\le N_b} M_{ji}(\mathbf{q}+\mathbf{G})M^*_{ji}(\mathbf{q}+\mathbf{G}')\right]. \tag{16}$$

We now have to evaluate the frequency integral in the expression of the correction $\Delta_{\mathbf{k}j}$. According to the polar structure of $W(\omega)$, which is a time-ordered quantity in the Green's

function theory, only the pole located in $\omega' = \bar{\epsilon}_\Sigma - \epsilon_{\mathbf{k}j}$ contributes to the integral by virtue of the residue theorem. The corresponding residue invokes just the antiresonant part of $W(\bar{\epsilon}_\Sigma - \epsilon_{\mathbf{k}j})$, i.e., the part of W having poles in the upper part of the complex plane. The energy difference $\bar{\epsilon}_\Sigma - \epsilon_{\mathbf{k}j}$ is large in practice because $\epsilon_{\mathbf{k}j}$ is typically located in the vicinity of the Fermi level and $\bar{\epsilon}_\Sigma$ will be a "high" energy (above the last one explicitly treated in the calculation). As a consequence, the function $W(\omega)$ in the residue is evaluated only for large values of ω. It is a general result[16] that the dielectric function of an electron gas tends to a single plasmon pole in the limit $\omega \to \infty$.

Hence, in the calculation of the correction $\Delta_{\mathbf{k}j}$, we will assume that the dynamically screened Coulomb interaction is well represented by a generalized plasmon-pole model[15] even though the plasmon-pole model is not used to calculate the self-energy itself. In other words, the plasmon-pole model is much better justified for the correction $\Delta_{\mathbf{k}j}$ to the self-energy than for the self-energy itself. The plasmon-pole approximation models the dynamically screened Coulomb interaction as

$$W_{\mathbf{GG}'}(\mathbf{q},\omega') = \delta_{\mathbf{GG}'} v(\mathbf{q}+\mathbf{G}) + \frac{\Omega^2_{\mathbf{GG}'}(\mathbf{q})}{2\tilde{\omega}_{\mathbf{GG}'}(\mathbf{q})}\left[\frac{1}{\omega - \tilde{\omega}_{\mathbf{GG}'}(\mathbf{q}) + i\eta} \times \left[-\frac{1}{\omega + \tilde{\omega}_{\mathbf{GG}'}(\mathbf{q}) - i\eta}\right] v(\mathbf{q}+\mathbf{G}'), \quad (17)$$

where $\tilde{\omega}_{\mathbf{GG}'}(\mathbf{q})$ and $\Omega^2_{\mathbf{GG}'}(\mathbf{q})$ are parameters determined by simple fits on the *ab initio* calculated dielectric matrices. With this model for $W(\omega)$, the frequency integration in Eq. (16) is performed analytically and yields the final expression for the correction

$$\Delta_{\mathbf{k}j} = \frac{1}{N_{\mathbf{k}}\Omega}\sum_{\mathbf{qGG}'} \frac{\Omega^2_{\mathbf{GG}'}(\mathbf{q}) v(\mathbf{q}+\mathbf{G})}{2\tilde{\omega}_{\mathbf{GG}'}(\mathbf{q})[\tilde{\omega}_{\mathbf{GG}'}(\mathbf{q}) + \bar{\epsilon}_\Sigma - \epsilon_{\mathbf{k}j} - i\eta]} \times \left[\langle \mathbf{k}j | e^{i(\mathbf{G}'-\mathbf{G})\cdot\mathbf{r}} | \mathbf{k}j\rangle - \sum_{i=1}^{N_b} M_{ji}(\mathbf{q}+\mathbf{G}) M^*_{ji}(\mathbf{q}+\mathbf{G}')\right]. \quad (18)$$

The first term of the correction $\Delta_{\mathbf{k}j}$ is almost costless since it does not involve the sum over states. The second term can be grouped with the usual evaluation of the *GW* self-energy where all the ingredients to build it are freely available. Note that usually the *GW* self-energy is calculated for several frequencies since the self-energy is a dynamical operator. Yet the correction can safely be assumed static because the energy $\epsilon_{\mathbf{k}j}$ is merely present inside the differences $\bar{\epsilon}_\Sigma - \epsilon_{\mathbf{k}j}$, which are large in all cases.

Here, we can compare our extrapolar-approximation-based correction to the coulomb hole plus screened exchange (COHSEX)-based correction of Tiago and Chelikowsky.[9] The basic idea is similar—make the denominator independent of the unoccupied state energy with index i—as in Eq. (18) so that one can factorize the denominator out of the sum and apply the closure relation. In Ref. 9, the authors chose to neglect energy differences $\epsilon_{\mathbf{k}-\mathbf{q}i} - \epsilon_{\mathbf{k}j}$ with respect to the plasmon frequencies $\tilde{\omega}_{\mathbf{GG}'}(\mathbf{q})$. In doing so, the static screening appears in the expression as in the COHSEX approximation to the self-energy. This choice may not be optimal since, when one wants to account for the high-energy bands, the energy differences are typically large compared to the plasma frequency of the system. Instead, the approximation proposed here is to disregard the energy dispersion of the high-energy bands and keep it fixed to an average value $\bar{\epsilon}_\Sigma$. In other words, our approximation assumes that the energy dispersion of the empty states $\epsilon_{\mathbf{k}-\mathbf{q}i} - \bar{\epsilon}_\Sigma$ is small compared to the difference $\tilde{\omega}_{\mathbf{GG}'}(\mathbf{q}) + \bar{\epsilon}_\Sigma - \epsilon_{\mathbf{k}j}$. This is much more realistic in the typical applications and becomes even better when the number of empty bands is increased.

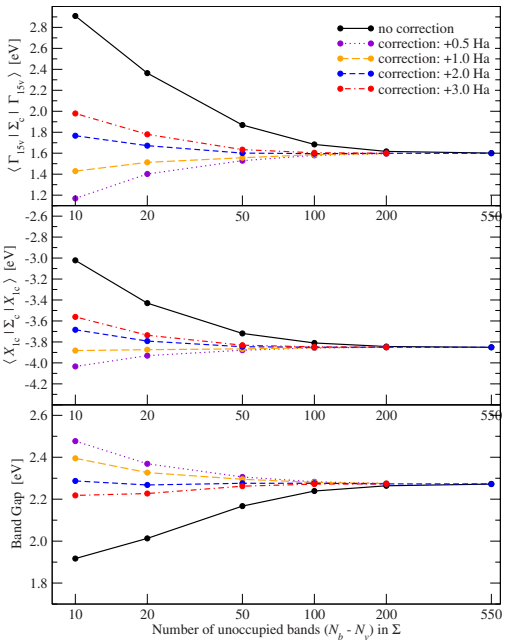

FIG. 3. (Color online) Convergence study of the correlation self-energy at top valence (upper panel) and at bottom conduction (middle panel) and of the band gap (lower panel) of β-SiC as a function of the number of unoccupied states explicitly included in the calculation of the self-energy.

In Fig. 3, we test the performance of the proposed correction in a convergence study of the matrix elements of the *GW* correlation self-energy and of the band gap of β-SiC with respect to the number of unoccupied states explicitly included in the calculation. Consistently with the convergence study on the number of empty states used in χ_0, the convergence of the results without correction is very slow. In order to achieve the typical accuracy of a *GW*-band structure ($\sim$50 meV), 100 to 200 unoccupied bands are required. Note that, as usual, the band gap converges faster than the absolute positions of the *GW* energies. When the correction is turned on, the convergence becomes much smoother and

the value of the correction does not depend strongly on the chosen average energy $\bar{\epsilon}_\Sigma$. Only 20 bands are necessary to converge the band gap within 50-meV, whereas 50 bands are needed to converge the absolute position of the top valence band named Γ_{15v}.

The evaluation of the optimum average energy $\bar{\epsilon}_\Sigma$ for the extrapolar approximation cannot be based on an exact scheme since no sum rule exists for the self-energy. Fortunately, a direct analogy between χ_0 and Σ can be underlined—they both contain a truncated sum over empty states and the summand is for both the squared modulus of the oscillator strengths divided by an energy difference. Yet, the denominator in Σ differs from the one of χ_0 by the presence of the plasmon-pole frequencies $\tilde{\omega}_{\mathbf{GG}'}(\mathbf{q})$. In average, these frequencies $\tilde{\omega}_{\mathbf{GG}'}(\mathbf{q})$ lie closely to the classical plasma frequency of the system. Considering that in the practical case the plasma frequency is small compared to the energy differences, we can expand the denominator as a function of the small quantity $\tilde{\omega}_{\mathbf{GG}'}(\mathbf{q})/(\epsilon_{\mathbf{k}-\mathbf{q}i}-\epsilon_{\mathbf{k}j})$. The leading term in this expansion of the denominator does not involve the quantity $\tilde{\omega}_{\mathbf{GG}'}(\mathbf{q})$;

$$\frac{1}{\epsilon_{\mathbf{k}-\mathbf{q}i}-\epsilon_{\mathbf{k}j}+\tilde{\omega}_{\mathbf{GG}'}(\mathbf{q})} \approx \frac{1}{\epsilon_{\mathbf{k}-\mathbf{q}i}-\epsilon_{\mathbf{k}j}}\left[1-\frac{\tilde{\omega}_{\mathbf{GG}'}(\mathbf{q})}{\epsilon_{\mathbf{k}-\mathbf{q}i}-\epsilon_{\mathbf{k}j}}\right]. \tag{19}$$

This shows that the determination of the extrapolar energy for the self-energy $\bar{\epsilon}_\Sigma$ is to the zeroth order—the same as the determination of the extrapolar energy for the polarizability

$$\bar{\epsilon}_\Sigma \approx \bar{\epsilon}_{\chi_0}. \tag{20}$$

In the case of β-SiC with 20 unoccupied states, the evaluation of $\bar{\epsilon}_{\chi_0}$ gave 1.6 Ha above the last band energy. This value would be also suitable for $\bar{\epsilon}_\Sigma$ as can be judged from Fig. 3.

V. APPLICABILITY

In order to show that the present scheme possesses a wide range of applications, we further carried out calculations for a supercell of bulk β-SiC and for two other very different systems: a wide band-gap insulator, solid argon and a benzene (C_6H_6) molecule in gas phase.

In Fig. 4, we examine the convergence with the number of unoccupied states explicitly included in the calculation for a 64-atom β-SiC supercell. The tendencies observed for the small unit cell here are even more pronounced. We could not achieve convergence within 50 meV of the uncorrected *GW* band gap even using almost 1400 empty states. In contrast, applying the proposed correction with the optimal extrapolar energy as evaluated according to the sum rule Eq. (10) allows us to have an accurate evaluation of the *GW* band gap with as few as 320 unoccupied states. By extrapolation, we can evaluate that the uncorrected *GW* would require about 3000 empty bands to achieve the same convergence.

As previously noticed,[19] the *GW* band gap of argon is much smaller than the experimental value (14.2 eV). However, this is not the point here. Figure 5 shows the convergence of the top valence, bottom conduction correlation self-energy, and band gap of this insulator. Once again, the extrapolar approximation performs extremely well even better than in the case of silicon carbide; the number of empty states can be reduced from 200 to 20.

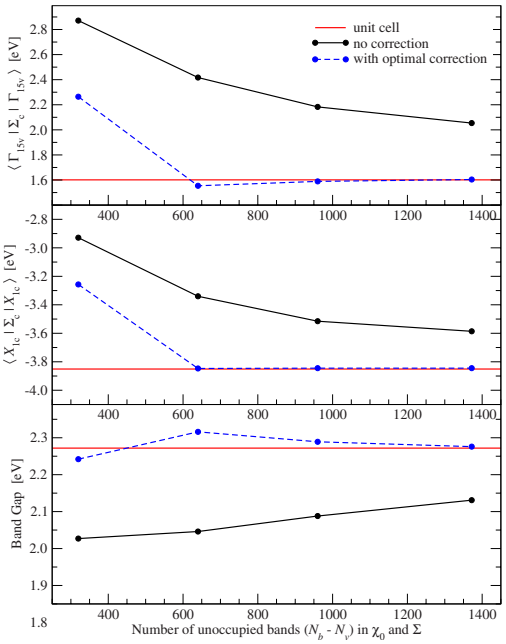

FIG. 4. (Color online) Convergence study of the correlation self-energy at top valence (upper panel) and at bottom conduction (middle panel) and of the band gap (lower panel) of β-SiC in a 64-atom cubic supercell as a function of the number of unoccupied states explicitly included in the calculation of the polarizability and in the self-energy.

The calculation of finite systems with periodic boundary-condition code should be considered with care, especially for the *GW* framework, which has long-range interactions. To mimic an isolated benzene molecule, we use a 40-bohr-long box in face-centered cubic geometry. In addition, the Coulomb interaction has been suppressed beyond 20 bohr.[20,21] As the stress is placed on the convergence behavior and not on the system properties, we applied the usual perturbative procedure for the *GW* evaluation. But we know that this is not sufficient as shown by Tiago and Cheliskowsky.[9] The convergence is displayed in Fig. 6. Without the extrapolar approximation, it would not have been possible to produce a reliable result even considering 1000 empty states. By contrast, we are able to obtain an evaluation of the highest occupied molecular orbital (HOMO)-lowest unoccupied molecular orbital (LUMO) gap with only 200 bands and an absolute energy position of the HOMO and the LUMO with 500 bands.

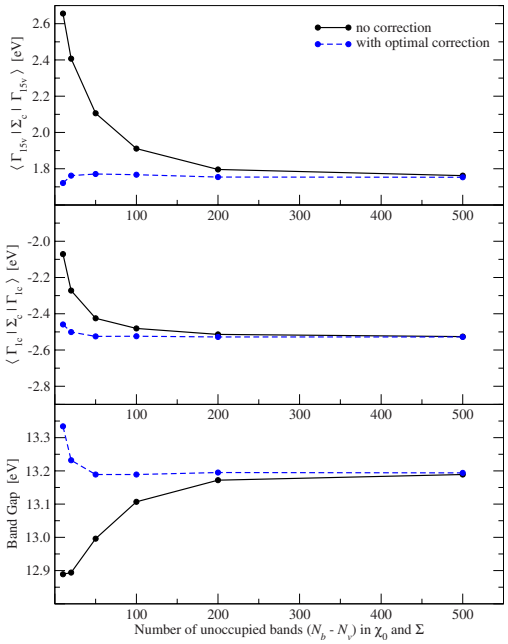

FIG. 5. (Color online) Convergence study of the correlation self-energy at top valence (upper panel) and bottom conduction (middle panel) and of the band gap (lower panel) of solid argon as a function of the number of unoccupied states explicitly included in the calculation of the polarizability and in the self-energy.

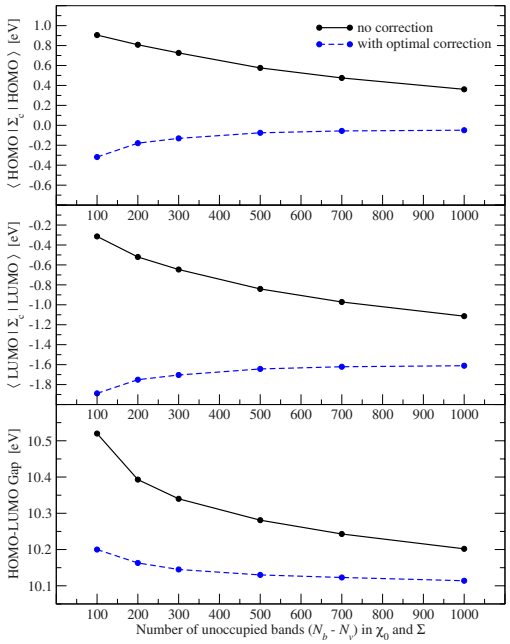

FIG. 6. (Color online) Convergence study of the correlation self-energy at HOMO (upper panel) and at LUMO (middle panel) and of the band gap (lower panel) of the benzene molecule (C_6H_6) as a function of the number of unoccupied states explicitly included in the calculation of the polarizability and in the self-energy.

VI. CONCLUSION

The number of empty states to be used in present implementations of the *GW* approximation with a plane-wave basis hinders the use of the method for large-scale applications. We have provided here a technique based on the closure relation and the adequate approximation for the eigenenergies of states not treated explicitly largely reduces the prefactor in CPU time and in memory needs. This technique is a generalization of the extrapolar approximation of Ref. 14 and is similar to CEDA developed in the framework of optimized effective potential method by Gritsenko and Baerends.[13] The gain is already large for bulk cells and it will allow one to consider applications to systems that were previously out of reach of the *GW* method.

We have emphasized that the completeness in the Hilbert space is critical in order to have full convergence of the *GW* band structure. A critical tool to measure this completeness is the fulfillment of the sum rule in Eq. (10). Using this relation, we have been able to estimate the correct range for the energy parameter to be introduced in the extrapolar approximation. With this determination, the proposed scheme can be considered as *ab initio*.

One immediate application of the presented acceleration technique is the *GW* evaluation of band alignment in junctions, which requires the *absolute* positions of the *GW* energy levels.[22]

ACKNOWLEDGMENTS

The present calculations and developments were based on the ABINIT code.[23,24] The new features will be made available in the forthcoming public version ABINITV5.6. One of the authors (X.G.) would like to acknowledge the support from the European Union through the 6th Framework Programme Network of Excellence under Contract No. NMP4-CT-2004-500198 (NANOQUANTA) and the 7th Framework Programme through the ETSF I3 e-Infrastructure project (Grant Agreement No. 211956).

[1] P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964).
[2] W. Kohn and L. J. Sham, Phys. Rev. **140**, A1133 (1965).
[3] L. Hedin, Phys. Rev. **139**, A796 (1965).
[4] G. Strinati, Riv. Nuovo Cimento **11**, 1 (1988).
[5] W. G. Aulbur, L. Jönsson, and J. W. Wilkins, Solid State Commun. **54**, 1 (2000).
[6] M. L. Tiago, S. Ismail-Beigi, and S. G. Louie, Phys. Rev. B **69**, 125212 (2004).
[7] M. van Schilfgaarde, T. Kotani, and S. V. Faleev, Phys. Rev. B **74**, 245125 (2006).
[8] L. Reining, G. Onida, and R. W. Godby, Phys. Rev. B **56**, R4301 (1997).
[9] M. L. Tiago and J. R. Chelikowsky, Phys. Rev. B **73**, 205334 (2006).
[10] S. Kümmel and L. Kronik, Rev. Mod. Phys. **80**, 3 (2008).
[11] R. T. Sharp and G. K. Horton, Phys. Rev. **90**, 317 (1953).
[12] J. B. Krieger, Y. Li, and G. J. Iafrate, Phys. Lett. A **146**, 256 (1990).
[13] O. V. Gritsenko and E. J. Baerends, Phys. Rev. A **64**, 042506 (2001).
[14] P.-M. Anglade and X. Gonze, Phys. Rev. B **78**, 045126 (2008).
[15] M. S. Hybertsen and S. G. Louie, Phys. Rev. B **34**, 5390 (1986).
[16] G. D. Mahan, *Many-Particle Physics*, 3rd ed. (Kluwer, Dordrecht/Plenum, New York, 2000).
[17] M. Taut, J. Phys. C **18**, 2677 (1985).
[18] M. van Schilfgaarde, T. Kotani, and S. Faleev, Phys. Rev. Lett. **96**, 226402 (2006).
[19] F. Bruneval, N. Vast, and L. Reining, Phys. Rev. B **74**, 045102 (2006).
[20] G. Onida, L. Reining, R. W. Godby, R. Del Sole, and W. Andreoni, Phys. Rev. Lett. **75**, 818 (1995).
[21] S. Ismail-Beigi, Phys. Rev. B **73**, 233103 (2006).
[22] R. Shaltaf, G.-M. Rignanese, X. Gonze, F. Giustino, and A. Pasquarello, Phys. Rev. Lett. **100**, 186401 (2008).
[23] http://www.abinit.org
[24] X. Gonze *et al.*, Z. Kristallogr. **220**, 558 (2005).

Nuclear Instruments and Methods in Physics Research B 277 (2012) 77–79

Contents lists available at SciVerse ScienceDirect

Nuclear Instruments and Methods in Physics Research B

journal homepage: www.elsevier.com/locate/nimb

Methodological aspects of the *GW* calculation of the carbon vacancy in 3C-SiC

Fabien Bruneval

CEA, DEN, Service de Recherches de Métallurgie Physique, F-91191 Gif-sur-Yvette, France

ARTICLE INFO

Article history:
Received 25 October 2011
Available online 30 December 2011

Keywords:
Point defects
Ab initio calculations
Silicon carbide

ABSTRACT

We employ the *GW* approximation to calculate the properties of the carbon vacancy, a prominent defect in irradiated 3C-SiC. The *GW* method has been recently proposed for point defects in order to cure the band gap problem of the usual approximations. However, its application relies on stringent approximations, such as the calculation of the relaxation energies of the atomic structures from another simpler approximation, namely the local density approximation. We assess here the validity of this approach in the complex case of the carbon vacancy. Finally, the calculated properties of the carbon vacancy are greatly affected by the use of the *GW* approximation with respect to earlier studies. The carbon vacancy is a rather shallow donor with a negative U behavior.

Cubic silicon carbide (3C-SIC) is a material with great potential for nuclear applications, both for next generation fission reactors and fusion reactors [1]. Materials in nuclear environments are subjected to high energy irradiations that create numerous intrinsic defects. In order to characterize and understand the elementary physical processes, it is highly relevant to gather accurate data for the important point defects created upon irradiation. A previous study of ours was devoted to the properties of the silicon vacancy in 3C-SiC [2]. The present article is meant to complement that data for another important defect of silicon carbide, the carbon vacancy V_C.

The calculation of the properties of point defects by a first-principles approach needs care for semiconductors and insulators. Indeed, defects in semiconductors and insulators can carry a charge due to the presence of defect states in the band gap. The defect states can be occupied or empty according to the relative position of the Fermi level μ_e with respect to the defect level. It is therefore crucial to have a proper description of the band gap region. Unfortunately, it is well known that the standard *ab initio* approaches based on the local density approximation (LDA) or generalized gradient approximation (GGA) completely fail in that respect. This shortcoming is named the *band gap problem* [3]. For many physical properties, this band gap problem is not particularly significant. However, for point defects in semiconductors and insulators, this drawback greatly affects the validity of the calculated data. This situation was clarified only recently.

In the present study, we follow the approach of Rinke and coauthors [4] that made the calculation of point defects accessible to the *GW* approximation [5,6]. The *GW* approximation is a well known approximation of many-body perturbation theory that yields very good band gaps at the expense of cumbersome calculations [7]. It is therefore very tempting to apply this approach devoid of any band gap problem to the study of point defects. However, the difficulties are of two kinds at least: firstly the calculations are so heavy that the calculation of a point defect in a large periodic supercell is not tractable, and secondly the *GW* approximation does not easily yield total energies nor forces (it just gives one-electron level energies).

The first bottleneck, the limitation due to the system size, has been pushed away due to the progresses in modern computers and also in algorithms [8], so that the calculations of this study dealing with supercells of as many as 215 atoms do not represent a real computational challenge. The second bottleneck, namely the missing expression for the total energy, requires combination of the *GW* approximation with the usual LDA, as proposed in Ref. [4]. The combination of the *GW* approximation for charge changes together with LDA for structural changes is a simple and attractive scheme. However, the validity of such a partition has always been taken for granted and no analysis about its limitations has been performed to our knowledge. We hence would like to seize the opportunity to clarify the situation in the present article.

A crucial quantity for charged defects is the thermodynamic transition energy, ε_{th}. This is the Fermi level at which the most stable charge changes. This quantity is defined in the specific case of the carbon vacancy as

$$\varepsilon_{th}(2+/1+) = E_0(V_C^{1+}, 1+) - E_0(V_C^{2+}, 2+) \quad (1)$$

where $E_0(V_C^{1+}, 1+)$ stands for the energy of the system in charge state 1+ having the relaxed geometry of V_C^{1+}. Note that these energies are usually referred to the valence band maximum of the pristine solid. The *GW* approximation allows one to obtain the transition energy

E-mail address: fabien.bruneval@cea.fr

0168-583X/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.nimb.2011.12.052

of a defect for a fixed structure only. This is the so-called *vertical transition energy*, ε_v,

$$\varepsilon_v(2+/1+) = E_0(V_C^{2+}, 1+) - E_0(V_C^{2+}, 2+) \quad (2)$$

The only difference between Eq. (1) and Eq. (2) is the structure in the first total energy. One can easily insert the definition of ε_v into Eq. (1) by adding and subtracting the same total energy:

$$\varepsilon_{th}(2+/1+) = E_0(V_C^{1+}, 1+) - E_0(V_C^{2+}, 1+) + E_0(V_C^{2+}, 1+) - E_0(V_C^{2+}, 2+) \quad (3)$$

The first two energies together account for a structural change at the constant charge 1+ and the last two energies are precisely the vertical transition energy that can be calculated within the *GW* approximation [2,4,9].

The point we want to address here is the arbitrariness in the inserted total energy in Eq. (3): any quantity that is added and subtracted could have done a similar job. For instance, one could have inserted the energy $E_0(V_C^{1+}, 2+)$, or even other energies. We are now about to investigate the quality of this assumption in the specific case of the carbon vacancy of 3C-SiC.

The case of V_C is indeed interesting because, due to the band gap problem of LDA, the ideal construction explained above is constrained by the position of the defect state. The band gap problem is indeed quite large in the case of 3C-SiC: the LDA band gap is only 1.35 eV to be compared to 2.19 eV within *GW*, and 2.37 eV in the experiment. The position of the important defect state that is to be filled when going from charge state 2+ to charge state 0 should lie in the band gap.

In Fig. 1 we provide the position of the defect state within LDA, superimposed onto the LDA band structure of bulk 3C-SiC along the Γ-R high symmetry line of the cubic supercell. The Brillouin zone of the bulk has been folded in order to mimic a cubic supercell of 216 atoms, similar to the one employed for the vacancy calculations. The position of the defect state is provided for the three relevant equilibrium geometries, V_C^{2+}, V_C^{1+}, and V_C^0. The defect levels do not require any sizeable shifting procedure to get aligned onto the pristine bulk bands. Furthermore, charge corrections are avoided here, since we experienced in our previous works that they rather deteriorates the convergence rate for vacancies [2,10].

One immediately notices that in the case of the 2+ and 1+ geometries, the defect level can lie above the conduction edge. For charge state 1+, this defect state should be occupied with one electron. This situation will cause problems since the standard minimization procedure would place the electron not in the defect state but rather in the bottom of the conduction band, which is lower in energy. However, this does not properly represent a carbon vacancy V_C^{1+}. Note that due the use of shifted ***k***-points following the Monkhorst–Pack scheme [11] the defect state remains below the conduction band in the center of the Γ-*R* line, even in the 1+ case. Of course, if we could employ a very large supercell, the Brillouin Zone would eventually be so small that the conduction band would be completely folded into its minimum at Γ and the defect state would again be higher than the conduction band.

The most simple workaround to deal with the situation shown in Fig. 1 would be to insert in Eq. (3) the intermediate energy of the neutral geometry, $E_0(V_C^0, 2+)$. This would be our advice in general, but we would like to show that this intuitive workaround is indeed justified.

In Fig. 2, we provide three different possible paths that could be implemented in Eq. (3) to calculate the thermodynamic transition $\varepsilon_{th}(2+/1+)$. We propose the use of either the intermediate structures V_C^{2+}, V_C^{1+}, or V_C^0. The combination of the different vertical transitions obtained with the *GW* approximation together with the different structural changes obtained within LDA should, in principle, give the same result. Let us start with the intermediate structure V_C^0: the system in charge 2+ is first distorted to the structure of charge 0 giving an increase in energy of 1.81 eV, then the charge is changed from 2+ to 1+ with an energetic cost of 0.83 eV obtained from *GW*, and finally the structure on the 1+ Born-Oppenheimer surface is relaxed to the equilibrium final position V_C^{1+}, with an energetic gain of 0.50 eV. Altogether this path gives a thermodynamic transition $\varepsilon_{th}(2+/1+)$ of 2.14 eV. If we use the same reasoning with the intermediate geometry V_C^{1+} we obtain a quite similar value of 2.19 eV. This error of only 0.05 eV is very encouraging: the usual uncertainty of any *GW* calculation is of the same order of magnitude.

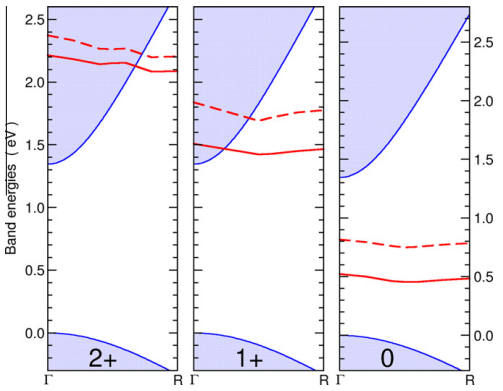

Fig. 1. Defect levels of the carbon vacancy V_C^+ (red lines) as obtained from LDA for three different equilibrium geometries for charge states 2+, 1+, and 0, along the Γ–*R* line. The position of the defect state is compared to the valence and conduction bands of the pristine SiC in a 216 atom supercell drawn with the shaded areas. When the vacancy bears a 1+ charge, the defect level for spin up is occupied with one single electron (solid red line) and the defect level for spin down remains empty (dashed red line). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

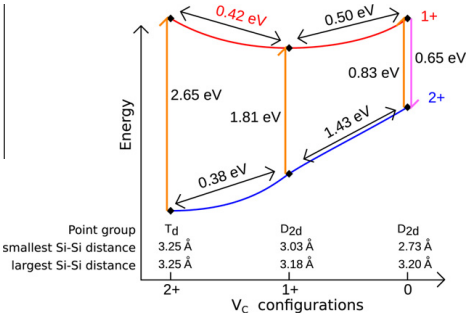

Fig. 2. Schematic Born-Oppenheimer surfaces for charge states 2+ and 1+ of the carbon vacancy. The horizontal axis designates the three equilibrium structures for charge states 2+, 1+, and 0. The corresponding point group of the configurations as well as distances between the first neighbors of the vacancies are specified. The energy differences on the same surface are obtained from LDA, whereas the vertical transitions at constant geometry are obtained from the *GW* approximation. The energy values for the vertical transitions are referred to the bulk valence band maximum. The (orange) upward arrows designate the energy for adding an electron to V_C^{2+}. The (pink) downward arrow shows the energy for removing an electron to V_C^+. The red difference of energies has been obtained by quenching the occupations of the 2+ geometry, in order to populate precisely the defect states above the conduction edge. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

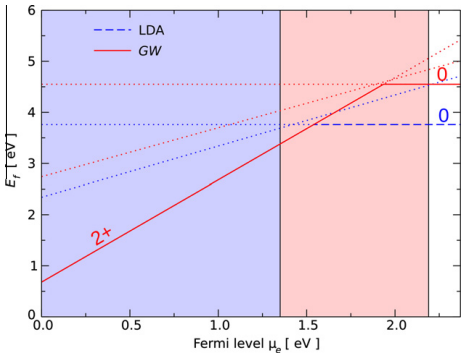

Fig. 3. Formation energy of the carbon vacancy V_C, E_f, as a function of the Fermi level μ_e in the silicon-rich conditions. The dashed line stands for LDA and the solid line for *GW*. The dotted lines show the data used for the construction. The vertical lines symbolize the conduction edge for LDA at 1.35 eV and for GW at 2.19 eV. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

If we would like to investigate the path through the geometry V_C^{2+} as well, we should be extremely cautious. As shown in Fig. 1, the defect state that we want to place an electron in lies above the conduction edge. Therefore, in order to obtain a truly occupied defect state, one should first identify which electronic state corresponds to the defect, and then one should enforce the occupation of this precise electronic state (one should not rely on the usual minimization scheme). In Fig. 2, the vertical transition reported for geometry V_C^{2+} is not the energy of the first (degenerate) empty states that correspond to the conduction states, but the transition to the threefold degenerate defect states. We then calculate the LDA energy of this geometry with a constrained occupation of the threefold degenerate defect states. The standard minimization procedure would have transferred this electron into the lowest available empty state, namely the bottom of the conduction band. Using this non-standard procedure, we have been able to evaluate the value in Eq. (3) properly, using the intermediate geometry V_C^{2+}. We finally obtained 2.23 eV for the transition $\varepsilon_{th}(2+/1+)$. This last value agrees impressively well with the other two paths ($\sim$0.1 eV) if we consider all the non-standard procedures we had to use in order to force the system through this peculiar path.

The combination of LDA for structural changes and *GW* for charge changes appears to be a reliable technique to obtain thermodynamic transitions of the carbon vacancy. Even though the different paths we have tested have very different structures, different point groups (T_d or D_{2d}), and different energies (up to 1.43 eV difference), the error bar of the combination LDA + *GW* is evaluated as less than 0.1 eV.

We finally use the previously described scheme to produce the formation energy of the carbon vacancy for the different relevant charge states. The formation energy is a central value for the comparison with respect to experiment. The formation energy of the carbon vacancy in 3C-SiC has been studied extensively in the past [12–15]. However, the previous studies all suffered from the ubiquitous band gap problem.

In Fig. 3, we provide the formation energy as a function of the Fermi level for the carbon vacancy for the LDA and *GW* approximations. The formation energy of the carbon vacancy with charge q, $E_f(V_C^q)$, is defined, in the silicon-rich conditions, as

$$E_f(V_C^q) = E_0(V_C^q) - E_0(\text{perfect}) + (\mu_{SiC}^0 - \mu_{Si}^0) + q\mu_e \quad (4)$$

where μ_{SiC}^0 and μ_{Si}^0 stand for the reference chemical potentials of bulk SiC and of bulk Si. The slope of the lines in Fig. 3 is hence given by the charge state q. The difference between the formation energies of different charge states is closely related to the thermodynamic transition energies, ε_{th}, as defined in Eq. (1).

The *GW* results in Fig. 3 were obtained by introducing the intermediate geometry V_C^0 in Eq. (3). The use of this geometry made the calculations more straightforward, as explained above. There is also another subtlety: the electron addition energies and electron removal energies are slightly different in the *GW* approximation, as shown by the upward and downward arrows in Fig. 2. This point was discussed extensively in Refs. [16,9] and is cured by considering the average value of the addition and removal energies.

Fig. 3 shows the differences and similarities between the LDA and *GW* formation energies. Both describe the carbon vacancy as a negative-U center: the charge state 1+ is never stable for any Fermi level. The charge states jumps from 2+ directly to 0 with increasing Fermi levels. The LDA charge transitions are bound to the too small LDA band gap. In the present evaluation they are not strictly below the LDA conduction edge because we employ a shifted $\boldsymbol{k}$-point grid that does not sample the conduction edge at Γ. In the limit of large supercells, it is doubtless that the LDA transitions would all decrease to the conduction edge. The *GW* transition energies span almost the full range of the experimental band gap. The *GW* results show that the carbon vacancy is a rather shallow donor type defect with $\varepsilon_{th}(2+/0) = E_c = 0.26$ eV. This value gives hints for the search of the photoluminescence peak related to the carbon vacancy [17].

As a conclusion, we have presented here some technical assessments about the promising technique LDA + *GW* that allows one to get rid of the famous band gap problem. Indeed, this framework relies on assumptions that had not been extensively tested previously. The carbon vacancy has defect states that are placed above the conduction edge for some configurations and therefore it is very intricate for the LDA + *GW* approach. Even in this complicated test case, the combination LDA + *GW* has a low error bar evaluated as less than 0.1 eV. With this reliable method at hand, we were able to provide the formation energy of the carbon vacancy, which showed that the carbon vacancy is a quite shallow double donor. Our results deviate noticeably from the earlier studies that predicted the carbon vacancy to be a deep double donor [13,15].

We are indebted to Guido Roma for discussions and to Samuel E. Taylor for carefully proofreading the manuscript. The present calculations were based on the ABINIT code [18]. This work was performed using HPC resources from GENCI-CINES and GENCI-CCRT (Grant No. 2011-gen6018).

References

[1] P. Yvon, F. Carré, J. Nucl. Mater. 385 (2009) 217.
[2] F. Bruneval, G. Roma, Phys. Rev. B 83 (2011) 144116.
[3] R.G. Parr, W. Yang, Density-Functional Theory of Atoms and Molecules, Oxford University Press, North Carolina, 1994.
[4] P. Rinke, A. Janotti, M. Scheffler, C.G. Van de Walle, Phys. Rev. Lett. 102 (2009) 026402.
[5] L. Hedin, Phys. Rev. 139 (1965) A796.
[6] W.G. Aulbur, L. Jonsson, J.W. Wilkins, Solid State Physics Vol. 54, Academic, San Diego, 2000. p. 1.
[7] M.S. Hybertsen, S.G. Louie, Phys. Rev. Lett. 55 (1985) 1418.
[8] F. Bruneval, X. Gonze, Phys. Rev. B 78 (2008) 085125.
[9] M. Giantomassi, M. Stankovski, R. Shaltaf, M. Grüning, F. Bruneval, P. Rinke, G.-M. Rignanese, Phys. Status Solidi B 248 (2011) 275.
[10] S.E. Taylor, F. Bruneval, Phys. Rev. B 84 (2011) 075155.
[11] H.J. Monkhorst, J.D. Pack, Phys. Rev. B 13 (1976) 5188.
[12] F. Bechstedt, A. Zywietz, J. Furthmüller, Europhys. Lett. 44 (1998) 309.
[13] A. Zywietz, J. Furthmüller, F. Bechstedt, Phys. Rev. B 57 (1999) 15166.
[14] B. Aradi, A. Gali, P. Deak, J.E. Lowther, N.T. Son, E. Janzen, W.J. Choyke, Phys. Rev. B 63 (2001) 245202.
[15] M. Bockstedte, A. Mattausch, O. Pankratov, Phys. Rev. B 69 (2004) 235202.
[16] F. Bruneval, Phys. Rev. Lett. 103 (2009) 176403.
[17] W.J. Choyke, L. Patrick, Phys. Rev. B 4 (1971) 1843.
[18] X. Gonze et al., Comput. Phys. Commun. 180 (2009) 2582.

GW Approximation of the Many-Body Problem and Changes in the Particle Number

Fabien Bruneval
CEA, DEN, Service de Recherches de Métallurgie Physique, F-91191 Gif-sur-Yvette, France
(Received 25 August 2009; published 23 October 2009)

A stringent test for an exchange-correlation approximation in electronic structure calculations is the equality between the ionization energy of the neutral system and the affinity of the singly positively charged system. All of the commonly used approximations (local, semilocal, hybrid) for the exchange correlation within density functional theory fail badly with this test. They consequently present a localization or delocalization error, resulting in a highest occupied molecular orbital or lowest unoccupied molecular orbital gap over- or underestimation. The *GW* approximation appears as the best available framework to describe particle number changes. The small remaining error can be further reduced by devising a ΔSCF-like method within the *GW* approximation. The proposed approach is necessary as soon as localized states are involved, e.g., in finite systems or defect states in crystals.

DOI: 10.1103/PhysRevLett.103.176403 PACS numbers: 71.15.−m, 32.10.Hq, 71.10.−w, 71.55.−i

For many years, density functional theory (DFT) has been seeking for the most correct approximation for the exchange-correlation functional [1]. The exact exchange-correlation term is of course unknown, since it has to account for all the quantum effects contained in the Schrödinger equation for interacting electrons. Several shortcomings of its existing approximations have been identified. A crucial issue for the available exchange-correlation functionals is the behavior of the energy along with the change of the number of electrons. Practical difficulties are tightly related to that problem: for instance, the infamous band gap problem arises from the poor description of electron addition and removal for most of the existing functionals.

Generalizing the idea of particle number changes to fractional number of electrons, it has been shown [2,3] from ensemble arguments that the total energy should be linear in between integral numbers of electrons. Unfortunately, all the usual approximations fail with this crucial property: local density approximation (LDA) and generalized gradient approximation (GGA) are convex, whereas Hartree-Fock (HF) is concave [4]. The straight line behavior arising from the exact exchange correlation is not just a playground for theoreticians. Indeed, the first ionization energy I can be obtained as the derivative on the left-hand side of the total energy with respect to the number of particles and the electron affinity A can be reached by the derivative on the right-hand side. If the slope of the total energy between $N-1$ and N electrons is not a constant, the electron affinity of the $N-1$ electron system is not to be equal to the ionization energy of the N electron system, even though they should represent the same total energy difference:

$$A(N-1) = -\frac{\partial E_0}{\partial n}\bigg|_{n=(N-1)^+} = E_0(N-1) - E_0(N) \quad (1)$$

$$I(N) = -\frac{\partial E_0}{\partial n}\bigg|_{n=N^-} = E_0(N-1) - E_0(N), \quad (2)$$

where $E_0(N)$ is the ground-state energy of the system with N electrons. The $+$ and $-$ signs indicate the side of the derivative. In most cases, the derivative reduces to the eigenvalue thanks to the Janak's theorem [5].

In a recent work, Cohen, Mori-Sánchez, and Yang [4,6] have clarified the relation between convexity or concavity, localization error and band gap error. The convex approximations, like LDA or GGA, lower the energy in spreading electrons as much as possible so that a fractional number of electrons is preferred. The concave approximations, like HF, instead find energetically favorable to localize electrons so that they integrate to an integer. HF is indeed a quasiparticle theory, which relies on integral number of particles. The true exchange correlation should be insensitive to these two situations. Only in this case, the total energy difference, called ΔSCF method, is to match the eigenvalue estimate of the ionization or affinity.

Originating from an other framework, many-body perturbation theory, the *GW* approximation to the exchange correlation [7] has been extremely powerful in describing the band structure of solids [8,9]. The *GW* approximation is an improvement over the HF approximation. It is based on the concept of screened Coulomb interaction. In practice, the *GW* approximation is usually evaluated as a first-order perturbation with respect to LDA, in the so-called G_0W_0 approach. This assumes, in particular, that the LDA and the G_0W_0 wave functions are identical. The currently best implementation of the *GW* approximation, the quasi particle self-consistent *GW* (sc*GW*) [10–12], proposes a static approximation to the complete *GW* case, which allows one to recalculate self-consistently the *GW* wave functions and eigenvalues. This approach is compulsory in the case of atoms and molecules, as shown in the following.

In the present Letter, we evaluate the quality of local, hybrid, and *GW* approximations to the exchange correlation in terms of localization error, band gap error, and ionization or affinity consistency. By testing them on small sodium clusters, we show that the *GW* approximation prevails over all the usual exchange-correlation approximations. The small remaining error in the ionization or affinity determination can be integrated in the framework in devising an extension of the ΔSCF procedure to the *GW* approach. We finally demonstrate the effectiveness of the procedure for localized defect states in a crystal.

It is unfortunately difficult in practice to obtain the total energy within the *GW* framework. Furthermore, the generalization of the *GW* equations to fractional numbers of electrons would require some care. As a consequence, we propose two alternative routes to evaluate the behavior of the *GW* approximation as a function of a fractional number of electrons. (i) Analyze the *GW* wave functions of an extra electron (or hole) in a doubled system: consider the system of two distant molecules with one additional electron. Will the extra electron be spread onto the two molecules or will it localize on one of them? (ii) Compare the ionization energy of a neutral system with the affinity of the positively charged one. If $-A(+) < -I(0)$ $[-A(+) > -I(0)]$, the approximation is to be convex (concave). Furthermore, the difference between $I(0)$ and $A(+)$ evaluates the magnitude of the localization error.

Following our route (i), we considered a system consisting of two distant molecules of Na_2 in a supercell with one extra electron. The technical details are provided below. If the chosen scheme were exact, the extra electron spread onto the two molecules or the electron localized on one single molecule should be two stationary points of the functional giving the same total energy. LDA is subjected to delocalization error so that the extra electron has an equal weight on the two molecules. The panel (a) of Fig. 1 represents the LDA highest occupied molecular orbital (HOMO) that places half an electron on the two molecules. Note that the lowest unoccupied molecular orbital (LUMO) is degenerate with the HOMO within LDA. The HF framework localizes easily the extra electron on one of the two molecules and breaks the HOMO/LUMO degeneracy. We performed then sc*GW* calculations starting either from HF and from LDA. Initiating the sc*GW* evaluation from the HF wave functions leads to a rapidly converging result, which maintains the extra electron on one Na_2 molecule [panel (b) of Fig. 1]. Starting the sc*GW* calculation from the LDA wave functions leads to a slowly converging result: after a dozen iterations with the extra electron spread over the two molecules, the HOMO wave function finally turns into the HF one. Whatever the starting point, the *GW* calculation ends up in the same stationary point, localizing the electron on one single Na_2 molecule. As a consequence, the *GW* approach yields a concave total energy and suffers from a localization error.

In the following, we quantify the nonlinearity of the sc*GW* approximation with calculations for small sodium

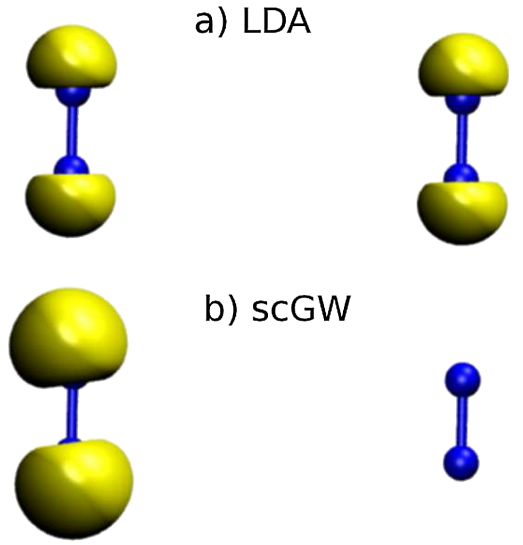

FIG. 1 (color online). Isosurface of an extra electron added in the system of two distant Na_2 molecules, or in other words, isosurface of the highest occupied molecular orbital of the $(2Na_2)^-$ system. Panel (a) represents the LDA results, whereas panel (b) provides the sc*GW* result.

clusters according to our route (ii): we compare the consistency of ionizations and affinities. The sodium clusters are a system of choice for such a study, since accurate experiments [13] and configuration-interaction calculations [14] are available. Furthermore, these systems are practical enough so that we still can use a plane-wave code, which is customary in the *GW* framework. We performed Γ-point calculations in large face-centered cubic supercells with a 60 Bohr cubic edge length. We use a plane-wave cutoff of 14 Ha for wave functions and of 2 Ha for dielectric matrices. We use a norm-conserving pseudopotential, where the semicore states ($2s^2 2p^6$) of sodium are treated as valence. These states are indeed very important for the exchange operator and are noticeably polarizable. We employ a plasmon-pole approximation, and we have verified that this is not an issue. The number of states is 512, which is reasonably low thanks to the acceleration scheme of Ref. [15]. The Coulomb interaction has been cut off in order to remove spurious image interactions [16,17]. As the geometries are not the issue here, we performed all the calculations in the neutral configurations obtained from a reference quantum chemistry work [14].

Some care is required due to the periodic approach. Periodic charged systems in presence a neutralizing background slowly converge [18]. However, the eigenvalues of neutral molecule experience also slow convergence in the supercell approach. The eigenvalues are shifted with respect to an isolated system calculation since the potential does not vanish at infinity in the periodic calculation. The

difference between a finite system calculation within Gaussian formalism using a $6{-}311{+}{+}G{**}$ basis and our periodic supercell approach appears to be a mere shift of the eigenvalues, whatever the charge, we simulate. Both neutral and charged systems can be corrected by shifting the supercell eigenvalues onto the isolated ones within LDA, for instance. Using the same shift, we were able to superimpose the HF eigenvalues from the periodic calculations onto the isolated results with a 0.1 eV accuracy. In the following, this shifting procedure is systematically applied. Note that we do not present results for the affinity of the clusters, as they show strong dependence with respect to the supercell size.

Table I compares the ionization energy of small sodium clusters, as obtained from removing an electron from Na_n or from adding an electron to Na_n^+, within LDA, HF, B3LYP, G_0W_0, and scGW. For all the approximations considered here, there is no discontinuity in the exchange-correlation functional so that the ionizations and affinities reduce to the HOMO and LUMO eigenvalues. Within LDA, the LUMO energy for the positively charged clusters is much lower than the HOMO energy of the neutral ones. LDA (and GGA, not presented here) is a convex approximation, which is consistent with the band gap underestimation problem [4]. HF gives the exact answer for a one electron system, since it is devoid of any self-interaction. The sodium atom, which has a single $3s$ valence electron, is well described within HF. The agreement between ionization and affinities then deteriorates up to $\sim$0.8 eV for the largest clusters. The HF approximation is clearly concave, which is consistent with the observed band gap overestimation. The hybrid functional family that mixes LDA, GGA, and exact-exchange could be a potential answer. The B3LYP functional [19] that includes 20% of exact-exchange is still convex: B3LYP predicts systematically the LUMO energies of the Na_n^+ clusters much lower than the HOMO of the Na_n.

Turning to GW calculations, we first provide for completeness the standard G_0W_0 results. Our results agree well with the published data for neutral species from Ref. [20]. Though reasonable compared to experiment, the G_0W_0 data are difficult to interpret and do not show clear trends. This is mainly due to the inadequacy of the perturbative approach in the case of the unoccupied states in a finite system. The scGW approach, which recalculates the wave functions, provides the most sensible results. The LUMO of Na_n^+ is systematically slightly higher than the HOMO of Na_n, but the difference is always lower than 0.45 eV. This shows a small, but noticeable, localization error in agreement with the result from route (i).

Because of the inconsistency between the eigenvalues, the ionizations and affinities are generally obtained from either total energy differences, the ΔSCF method, or from Slater's transition state theory [21]. Both approaches generally agree very well. The ΔSCF results provided in Table I for LDA, HF, and B3LYP supersede the eigenvalue estimate within the corresponding approximations. Following the argument of Slater, if the total energy within our approximation is not linear for fractional number of electrons as it should be, we may expand it as a second order polynomial. Under this assumption, the ionization energy can be approximated by the eigenvalue at the half charge $N - 1/2$. The Slater's transition state approach gives a very good estimate for the total energy difference. Following the same arguments, we observe that the total energy difference can be also evaluated as the mean value,

$$I(N) \approx -\frac{1}{2}\left[\frac{\partial E_0}{\partial n}\bigg|_{n=(N-1)^+} + \frac{\partial E_0}{\partial n}\bigg|_{n=N^-}\right]. \quad (3)$$

The evaluation of Eq. (3) does not require total energy nor fractional charge calculations, but only the derivatives with respect to the particle number, which reduce in most cases to the eigenvalues of the neutral and the charged system. It can be readily evaluated from the data provided in Table I. The final result, labeled ΔSCF within scGW approximation, gives the best estimate of all approximations for the ionization energy of the sodium clusters. Furthermore, this ΔSCF procedure allows for a reconciliation between total energy and eigenvalue approaches. A direct evaluation (beyond reach by now) of the scGW total energy differences would be consistent with the proposed procedure.

As a final illustration of the inconsistency between eigenvalues between charged systems, we consider the localized state in the band gap of crystal created by a point defect. We exemplify with the carbon split $\langle 100\rangle$ interstitial in cubic silicon carbide [22]. The calculations have been performed in a 65 atom cubic supercell with a $2 \times 2 \times 2$ k point sampling. The structure of the defect (inset of Fig. 2)

TABLE I. Ionization energy in eV of the small sodium clusters evaluated from the HOMO of the neutral species $-I(0)$, from the LUMO of the cationic species $-A(+)$, or from the difference in total energies (ΔSCF), within different approximations to the exchange correlation: LDA, HF, B3LYP, G_0W_0, and scGW.

	LDA			HF			B3LYP			G_0W_0		scGW			Expt. [13]
	$-A(+)$	$-I(0)$	ΔSCF	$-A(+)$	$-I(0)$	ΔSCF	$-A(+)$	$-I(0)$	ΔSCF	$-A(+)$	$-I(0)$	$-A(+)$	$-I(0)$	ΔSCF	
Na_1	-6.96	-3.08	-5.36	-4.94	-4.95	-4.94	-7.10	-3.48	-5.42	-4.88	-5.40	-5.05	-5.49	-5.27	-5.139
Na_2	-7.12	-3.20	-5.25	-3.90	-4.48	-4.08	-6.78	-3.52	-5.19	-5.10	-5.05	-4.66	-5.10	-4.88	-4.934
Na_3	-5.99	-2.62	-4.30	-3.47	-4.17	-3.80	-5.59	-2.91	-4.24	-4.42	-4.24	-4.15	-4.42	-4.29	-3.97
Na_4	-6.01	-2.77	-4.38	-3.00	-3.78	-3.34	-5.55	-2.98	-4.25	-4.71	-4.29	-4.18	-4.38	-4.28	-4.27
Na_5	-5.77	-2.78	-4.27	-3.34	-4.43	-3.94	-5.32	-3.03	-4.17	-4.54	-4.17	-4.18	-4.39	-4.28	-4.05

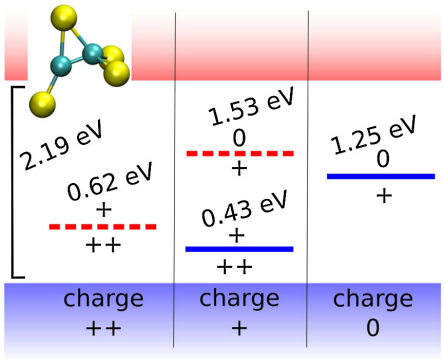

FIG. 2 (color online). Vertical charge transition levels of the carbon split ⟨100⟩ interstitial of 3C-SiC, evaluated within G_0W_0 using different charges for the 65 atom supercell. The occupied levels are plain lines, whereas the empty ones are dashed. The structure of the defect has been kept frozen in the neutral geometry, as shown in the inset in the upper left corner. The energy of the defect levels is referred to the top valence band.

has been frozen in the neutral optimized configuration in order to isolate the purely electronic behavior we want to address. We performed LDA and G_0W_0 evaluation of the position of the defect level inside the band gap for different charged supercells (charges 0, + and ++). In this case, scGW is not required since we have verified that the LDA wave functions constitute a good approximation for the scGW wave functions. Comparing the density of states, we carefully checked that the very tiny band shifts were not an issue. The trends are consistent with the sodium clusters. The eigenvalues of the levels within LDA deviate strongly when changing their occupation, showing a strong convex behavior (the HOMO of the neutral system is higher than the LUMO of the positive system). As seen from Fig. 2, the discrepancy between ionizations and affinities is small within G_0W_0 (∼0.2–0.3 eV) and confirms the slight localization error. For the defect calculations, we again propose the ΔSCF procedure within GW, which simulates total energy differences without the need to perform such calculations. The final ΔSCF value for the charge transitions are $\epsilon_{GW}(+/0) = 1.39$ eV and $\epsilon_{GW}(++/+) = 0.53$ eV.

In conclusion, we proposed to judge the quality of the exchange-correlation approximations on the discrepancy between ionization of the neutral system and affinity of the positively charged one. The exact exchange-correlation functional should not have any. From all the approximations tested here (LDA, HF, B3LYP, GW), the GW approach offers the lowest discrepancy. The small remaining error within GW is consistent with a systematic localization error and the slight band gap overestimation observed in practice [11,23]. In order to provide the most meaningful results, we support the use of a ΔSCF-like procedure to conciliate total energy and quasiparticle energy evaluations of the ionization and affinity energies. Finally, the ionization or affinity consistency can give insights concerning the properties of the vertex function that should fix the GW errors.

The present calculations were performed using the ABINIT code [24] and GAUSSIAN03 [25]. We are grateful to Silvana Botti for her helpful comments. This work was performed using HPC resources from GENCI-CINES (Grant No. 2009-GEN6018).

[1] W. Kohn, Rev. Mod. Phys. **71**, 1253 (1999).
[2] J. P. Perdew, R. G. Parr, M. Levy, and J. L. Balduz, Jr., Phys. Rev. Lett. **49**, 1691 (1982).
[3] W. T. Yang, Y. Zhang, and P. W. Ayers, Phys. Rev. Lett. **84**, 5172 (2000).
[4] P. Mori-Sánchez, A. J. Cohen, and W. T. Yang, Phys. Rev. Lett. **100**, 146401 (2008).
[5] J. F. Janak, Phys. Rev. B **18**, 7165 (1978).
[6] A. J. Cohen, P. Mori-Sánchez, and W. T. Yang, Science **321**, 792 (2008).
[7] L. Hedin, Phys. Rev. **139**, A796 (1965).
[8] F. Aryasetiawan and O. Gunnarsson, Rep. Prog. Phys. **61**, 237 (1998).
[9] M. S. Hybertsen and S. G. Louie, Phys. Rev. Lett. **55**, 1418 (1985).
[10] S. V. Faleev, M. van Schilfgaarde, and T. Kotani, Phys. Rev. Lett. **93**, 126406 (2004).
[11] M. van Schilfgaarde, T. Kotani, and S. Faleev, Phys. Rev. Lett. **96**, 226402 (2006).
[12] F. Bruneval, N. Vast, and L. Reining, Phys. Rev. B **74**, 045102 (2006).
[13] A. Herrmann, E. Schumacher, and L. Wöste, J. Chem. Phys. **68**, 2327 (1978).
[14] V. Bonačić-Koutecký, P. Fantucci, and J. Koutecký, J. Chem. Phys. **93**, 3802 (1990).
[15] F. Bruneval and X. Gonze, Phys. Rev. B **78**, 085125 (2008).
[16] G. Onida, L. Reining, R. W. Godby, R. Del Sole, and W. Andreoni, Phys. Rev. Lett. **75**, 818 (1995).
[17] J. Spencer and A. Alavi, Phys. Rev. B **77**, 193110 (2008).
[18] G. Makov and M. C. Payne, Phys. Rev. B **51**, 4014 (1995).
[19] A. D. Becke, J. Chem. Phys. **98**, 5648 (1993).
[20] Y. Noguchi, S. Ishii, K. Ohno, and T. Sasaki, J. Chem. Phys. **129**, 104104 (2008).
[21] J. C. Slater, *The Self-Consistent Field for Molecules and Solids* (McGraw-Hill, New York, 1974), Vol. 4.
[22] M. Bockstedte, A. Mattausch, and O. Pankratov, Phys. Rev. B **68**, 205201 (2003).
[23] M. Shishkin, M. Marsman, and G. Kresse, Phys. Rev. Lett. **99**, 246403 (2007).
[24] X. Gonze *et al.*, Z. Kristallogr. **220**, 558 (2005).
[25] M. J. Frisch *et al.*, *Gaussian 03, Revision C.02* (Gaussian, Inc., Wallingford, CT, 2004).

Ionization energy of atoms obtained from *GW* self-energy or from random phase approximation total energies

Fabien Bruneval
CEA, DEN, Service de Recherches de Métallurgie Physique, F-91191 Gif-sur-Yvette, France

(Received 1 March 2012; accepted 1 May 2012; published online 17 May 2012)

A systematic evaluation of the ionization energy within the *GW* approximation is carried out for the first row atoms, from H to Ar. We describe a Gaussian basis implementation of the *GW* approximation, which does not resort to any further technical approximation, besides the choice of the basis set for the electronic wavefunctions. Different approaches to the *GW* approximation have been implemented and tested, for example, the standard perturbative approach based on a prior mean-field calculation (Hartree-Fock *GW*@HF or density-functional theory *GW*@DFT) or the recently developed quasiparticle self-consistent method (QS*GW*). The highest occupied molecular orbital energies of atoms obtained from both *GW*@HF and QS*GW* are in excellent agreement with the experimental ionization energy. The lowest unoccupied molecular orbital energies of the singly charged cation yield a noticeably worse estimate of the ionization energy. The best agreement with respect to experiment is obtained from the total energy differences within the random phase approximation functional, which is the total energy corresponding to the *GW* self-energy. We conclude with a discussion about the slight concave behavior upon number electron change of the *GW* approximation and its consequences upon the quality of the orbital energies. © *2012 American Institute of Physics*. [http://dx.doi.org/10.1063/1.4718428]

I. INTRODUCTION

The Green's function approach to many-body problem has been extremely successful for the electronic structure in condensed matter physics. Most noticeably, the *GW* approximation[1] is known to outperform the local and semi-local approximations of density functional theory (DFT) for the description of band gaps and band structures.[2–4] For atoms and molecules, the performance of the *GW* approximation has been studied very little. The seminal results of Shirley and Martin[5] were rather promising, but have had follow-ups only recently.[6–10] However, no systematic study is available yet. Especially the *GW* calculations for open-shell and spin polarized atoms do not exist to the best of our knowledge.

There is no stringent need for accurate orbital energies for atoms, since the ionization energy I for instance can be obtained thanks to a total energy difference

$$I = E_0^{N-1} - E_0^N, \tag{1}$$

where E_0^N stands for the total energy for the N electron system. This is the so-called ΔSCF procedure that produces in general good quality ionization energies at the expense of two separate self-consistent calculations.

In DFT or in many-body perturbation theory, the ionization energy can also be evaluated from the eigenvalues. Within DFT, they are named Kohn-Sham eigenvalues, whereas within many-body perturbation theory, they are called quasiparticle energies. Hence, the ionization energy could alternatively be obtained from the Kohn-Sham or quasiparticle energy corresponding to the highest occupied molecular orbital (HOMO) $\epsilon_{\rm HOMO}^N$ or from the Kohn-Sham or quasiparticle energy corresponding to the lowest unoccupied molecular orbital (LUMO) energy of the $N-1$ electron system $\epsilon_{\rm LUMO}^{N-1}$,

$$I = -\epsilon_{\rm HOMO}^N = -\epsilon_{\rm LUMO}^{N-1}. \tag{2}$$

Generally speaking, the calculation of the ionization energy through the orbital energies yields rather poor results. This problem is not much acute for atoms, as the ΔSCF technique can be used. Nevertheless, it has been understood recently that the poor quality of the potentials (or orbital energies) has also deep consequences on the total energies.[11] The orbital energies are related to the fractional electron behavior, which in turn is related to a localization or delocalization of the wavefunctions. There is therefore a strong need to investigate higher levels of approximation for the potentials.

The *GW* approximation is a successful approximation for self-energies, which, with the Hartree potential, is the effective one-electron potential of a many-electron system. Unfortunately, the *GW* implementation is not unique in practice. Owing to the complexity of the calculations, several types of *GW* calculations have been designed in the last 50 years. The standard approach is not self-consistent and makes use of a prior mean-field calculation as a starting point: this is the so-called G_0W_0 procedure. This situation introduces a dependence of the *GW* result onto the underlying mean-field choice. For solids, the chosen mean-field is very often the local density approximation (LDA) or the generalized gradient approximation (GGA). For atoms and molecules, the starting mean-field happens to be Hartree-Fock (HF) or any other approximation of DFT. The choice of the starting point is unfortunately crucial, since the final *GW* result can be affected by deficiencies in the starting point.[12–15]

In order to get rid of the starting point dependence, self-consistent *GW* would appear appealing at first sight. However, according to the few studies available, the performance of such an approach for the quasiparticle energies is unclear.[7,8,10,16–18] For spectral properties, such as ionization energy, the full self-consistency has been shown to yield incorrect results for the homogeneous electron gas.[16] As far as finite systems are considered, the comprehensive study of 30 molecules by Rostgaard and co-workers[8] shows little or no improvement due to the fully self-consistent *GW* approach. Approximate static self-consistent schemes are then an interesting option;[19,20] they allow one to completely forget about the starting point and they are not affected by the dynamical caveats of the full self-consistency. The quasiparticle self-consistent *GW* (QS*GW*) approach of Faleev and co-workers[19] has been extremely successful for band gaps of solids.[21,22] Besides one single study on molecules,[10] its performance for atoms is however still to be determined.

The purpose of the paper is to evaluate the performance of the *GW* approximation for the ionization energy of the first row atoms. In order to calculate unambiguously converged results, we first present a novel implementation of the *GW* approximation for atoms that are free of the usual drawbacks of standard implementations. Our implementation uses Gaussian basis and does not rely on any further approximation besides the initial choice of the basis set for the wavefunctions. Second, we assess the so far unknown performance of QS*GW* approach for atoms and conclude it yields a small but noticeable improvement over *GW*@HF. Third, we compare three different methods to evaluate the ionization energy of atoms within *GW*: HOMO energy of the atom $-\epsilon^{N}_{\mathrm{HOMO}}$, LUMO energy of the cation $-\epsilon^{N-1}_{\mathrm{LUMO}}$, or total energy difference of the atom and the cation (ΔSCF procedure). The most accurate results for the ionization energy are obtained from ΔSCF and from the HOMO energy. The LUMO energies of the cations yield noticeably worse estimates. We conclude our study with a discussion about the slightly concave upon electron number changes behavior of the *GW* approximation that rationalizes the discrepancy between the three different paths towards the ionization energy.

II. A SHORT REVIEW OF THE *GW* APPROXIMATION

A. General theory

The *GW* self-energy arises from the many-body perturbation theory, when the considered perturbation is not in the bare Coulomb interaction $v(\mathbf{r},\mathbf{r}') = 1/|\mathbf{r}-\mathbf{r}'|$ (in atomic units), but is in the screened Coulomb interaction W. The effective interaction W accounts for the screening of the interactions by the electrons of the system. W is anticipated to be smaller and better behaved than v. Most importantly, the long-ranged part is damped out for metallic systems or reduced for the other systems. The *GW* self-energy may be thought of as a dynamically screened generalization of the Fock exchange.

In practice, the *GW* self-energy is built from the frequency convolution of the Green's function G with the screened Coulomb interaction W,

$$\Sigma^{GW}(\mathbf{r},\mathbf{r}',\omega) = \frac{i}{2\pi}\int d\omega' e^{i\eta\omega'} G(\mathbf{r},\mathbf{r}',\omega+\omega') \times W(\mathbf{r}',\mathbf{r},\omega'), \quad (3)$$

where η is a vanishing positive real number.

Introducing the polarizable part of the screened Coulomb interaction $W_p = W - v$, the self-energy can be conveniently split in the usual Fock exchange operator

$$\Sigma_x(\mathbf{r},\mathbf{r}') = \frac{i}{2\pi} v(\mathbf{r},\mathbf{r}') \int d\omega' e^{i\eta\omega'} G(\mathbf{r},\mathbf{r}',\omega') \quad (4)$$

and a remainder Σ_c^{GW}. By definition, the remainder accounts for the correlation effects. The term $e^{i\eta\omega'}$ in Eq. (4) retains only the contribution from the occupied states in the Green's function.

The polarizable part of the screened Coulomb interaction W_p is in turn a function of the random phase approximation (RPA) polarizability χ of the electronic system,

$$W_p(\mathbf{r},\mathbf{r}',\omega) = \int d\mathbf{r}_1 d\mathbf{r}_2 v(\mathbf{r},\mathbf{r}_1)\chi(\mathbf{r}_1,\mathbf{r}_2,\omega)v(\mathbf{r}_2,\mathbf{r}'). \quad (5)$$

Then the RPA polarizability χ can be related to the independent particle polarizability χ_0 through a Dyson-like equation,

$$\chi^{-1}(\mathbf{r},\mathbf{r}',\omega) = \chi_0^{-1}(\mathbf{r},\mathbf{r}',\omega) - v(\mathbf{r},\mathbf{r}') \quad (6)$$

that connects the non-interacting system to the interacting system. Finally, χ_0 has a simple expression in terms of two Green's functions,

$$\chi_0(\mathbf{r},\mathbf{r}',\omega) = -i\int d\omega' G(\mathbf{r},\mathbf{r}',\omega+\omega')G(\mathbf{r}',\mathbf{r},\omega'). \quad (7)$$

Using the diagrammatic language, χ_0 is a ring diagram and the RPA polarizability χ is the infinite sum over the ring diagrams. Symbolically, it reads

$$\chi = \chi_0 + \chi_0 v \chi_0 + \chi_0 v \chi_0 v \chi_0 + \dots . \quad (8)$$

The *GW* diagrams for the correlation self-energy are displayed in panel (b) of Fig. 1. It contains an infinite summation over all the ring diagrams. The second-order approximation (or MP2 when the Green's functions are HF ones[23]), on the other hand, not only contains the first of the ring diagrams, but also includes the second-order exchange diagram (panel (a)).

B. The perturbative *GW* approach

In principle, the Green's function appearing in Eqs. (3) and (7) should be obtained self-consistently from the iteration of the Dyson equation. In reality, this is hardly feasible and may be not desirable[7,8,16–18] as discussed in the Introduction. It is then common practice[1–3] to consider the Green's function from another simpler approximation: LDA, GGA, HF, etc. We denote these approaches as *GW*@LDA, *GW*@GGA, *GW*@HF, respectively.

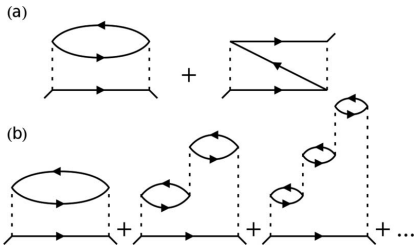

FIG. 1. Correlation self-energy diagrams included in the second-order approximation [panel (a)] and in the *GW* approximation [panel (b)]. The solid lines with arrows represent the one-particle Green's function G and the dashed lines represent the Coulomb interaction v. The second-order self-energy consists of the one-ring diagram and of the second-order exchange, whereas the *GW* self-energy contains the infinite sum over the ring diagrams. If the Green's functions are HF Green's functions, the second-order self-energy is named MP2.

Within a mean-field approach with eigenvalues $\epsilon_{i\sigma}$ and eigenvectors $\varphi_{i\sigma}(\mathbf{r})$, the Green's function simply reads

$$G(\mathbf{r},\mathbf{r}',\omega) = \sum_{i\sigma} \varphi_{i\sigma}(\mathbf{r})\varphi_{i\sigma}(\mathbf{r}') \times \left[\frac{f_{i\sigma}}{\omega - \epsilon_{i\sigma} - i\eta} + \frac{1-f_{i\sigma}}{\omega - \epsilon_{i\sigma} + i\eta} \right], \quad (9)$$

where the wavefunctions have been assumed to be real and $f_{i\sigma}$ is the occupation number of state i with spin σ. The Green's function depends on all the orbitals: occupied and virtual. Its poles are the eigenvalues, slightly shifted above or below the real axis.

With this definition for the Green's function, the equations presented in Sec. II A can be tracted numerically and finally, the *GW* quasiparticle energy reads

$$\epsilon_{i\sigma}^{GW} = \epsilon_{i\sigma}^{\mathrm{HF}} + \left\langle i\sigma \middle| \Sigma_c^{GW}(\epsilon_{i\sigma}^{GW}) \middle| i\sigma \right\rangle. \quad (10)$$

Please note that the self-energy is a dynamical operator and needs to be evaluated precisely at the unknown quasiparticle energy. This is not an issue since the equation can be solved for instance graphically as exemplified in Fig. 2.

C. Quasiparticle self-consistent *GW*

The dependence of the *GW* result onto the starting mean-field Green's function is not elegant in theory and can introduce additional issues in practice. It would be desirable to perform the calculations self-consistently, so that the starting point is forgotten.

Recently, the quasiparticle self-consistent *GW* (QS*GW*) was introduced by Faleev and co-workers[19] in order to get a simplified version of self-consistent *GW* calculations. They proposed a static and hermitian approximation to the *GW* self-energy,

$$\left\langle i\sigma \middle| \Sigma_c^{\mathrm{QS}GW} \middle| j\sigma \right\rangle = \frac{1}{2}\left[\left\langle i\sigma \middle| \Sigma_c^{GW}(\epsilon_{j\sigma}) \middle| j\sigma \right\rangle + \left\langle j\sigma \middle| \Sigma_c^{GW}(\epsilon_{i\sigma}) \middle| i\sigma \right\rangle \right] \quad (11)$$

that conserves the orthogonality of the underlying wavefunctions and the real-valued eigenvalues. The advantage of this

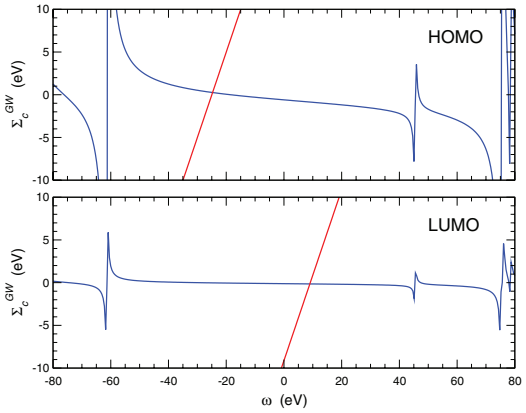

FIG. 2. HOMO (upper panel) and LUMO (lower panel) expectation values of the correlation part of the dynamical *GW* self-energy based on HF inputs (*GW*@HF) for He using a cc-pV5Z basis. The small real number η has been set to 0.25 eV. The crossing point between the straight line and the self-energy is the solution of the quasiparticle equation (10).

particular expression is that only the off-diagonal terms ($i \neq j$) are approximated. Once self-consistency has been reached, the diagonal expectation values of Σ_c are evaluated precisely at the quasiparticle energy, as they should be according to Eq. (10).

D. RPA total energy

Finally, we close the theoretical review by introducing the RPA expression for the total energy. This approximation is tightly bound the *GW* self-energy: the *GW* self-energy operator is obtained from the functional derivative of the RPA functional Φ_c^{RPA} with respect to the Green's function,[24]

$$\Sigma_c^{GW}(\mathbf{r},\mathbf{r}',\omega) = \frac{\delta\Phi_c^{\mathrm{RPA}}}{\delta G(\mathbf{r}',\mathbf{r},-\omega)}, \quad (12)$$

where the RPA functional symbolically reads

$$\Phi_c^{\mathrm{RPA}} = -\frac{1}{2}\mathrm{Tr}\left[\sum_{n=2}^{+\infty} \frac{(v\chi_0)^n}{n} \right]. \quad (13)$$

The symbol Tr is short for the triple integral over $\mathbf{r}$, $\mathbf{r}'$, and ω. More details can be found for instance in Ref. 25. The RPA functional is the infinite sum over the ring diagrams. In summary, the functional Φ_c^{RPA} yields the correlation energy corresponding to the *GW* approximation to the self-energy. The connection between the two frameworks will be numerically investigated in the following.

III. PRACTICAL IMPLEMENTATION

The main target of the present work is to obtain unambiguous values. Therefore, we resort to as few approximations as possible. Basically, all the calculations are exact, once the basis for the wavefunctions has been set. The computational efficiency is clearly not the issue here.

We adopt an all-electron formalism to solve the non-relativistic Schrödinger equation. The self-consistent field equations are solved in the unrestricted manner, for which, the spin-up and spin-down wavefunctions are allowed to differ. The wavefunctions are expanded in a Gaussian basis. Unlike previous implementations in a Gaussian basis,[6,9] we do not resort to any auxiliary basis set to expand the polarizabilities. Furthermore, the RPA polarizability is obtained in the product basis set $|ij\sigma\rangle$, so that its frequency dependence is exactly known and can be integrated analytically.[26] As a consequence, no plasmon-pole model,[3,27] nor analytic continuation[10,28] is needed. In summary, once the basis set has been chosen, there is no other convergence parameter of any kind.

A. The Gaussian basis set

For convenience, we adopt the Cartesian Gaussian basis functions

$$\phi_\alpha(\mathbf{r}) = N x^{n_x} y^{n_y} z^{n_z} e^{-\zeta r^2}, \tag{14}$$

where ζ is the decay rate, $l = n_x + n_y + n_z$ defines the angular momentum of the Gaussian basis function, and N is the normalization factor. The decay rates are obtained from the Dünning's correlation consistent sets[29] as reported in a web-available database.[30,31] Even though the original basis sets from Dünning were defined for pure Gaussian functions with spherical harmonics describing the angular part, the use of Cartesian Gaussian just adds a few basis functions and affects the final result very little. For instance, with Cartesian Gaussians, there are six d orbitals instead of the usual five; there are 10 f orbitals instead of the usual seven; etc. The use of the Dünning sets cc-pVXZ (with X = D, T, Q, 5, or 6) allows us to reduce the basis set error in a systematic manner.

The overlap, the kinetic, the nucleus attraction integrals are readily obtained from basic formulas. The numerical value of the four Gaussian electron repulsion integrals

$$(\alpha\beta|\gamma\delta) = \int d\mathbf{r}_1 d\mathbf{r}_2 \phi_\alpha(\mathbf{r}_1)\phi_\beta(\mathbf{r}_1)\frac{1}{|\mathbf{r}_1 - \mathbf{r}_2|}\phi_\gamma(\mathbf{r}_2)\phi_\delta(\mathbf{r}_2) \tag{15}$$

can be obtained from a web-available library.[32]

The most cumbersome part of a *GW* calculation is the transformation of the electron repulsion integrals into the eigenvector basis,

$$(ij\sigma|kl\sigma') = \sum_{\alpha\beta\gamma\delta} C_{\alpha i\sigma} C_{\beta j\sigma} C_{\gamma k\sigma'} C_{\delta l\sigma'} (\alpha\beta|\gamma\delta), \tag{16}$$

where $C_{\alpha i\sigma}$ are the expansion coefficients of the eigenvectors into the Gaussian basis set. This operation scales as N^5 with N being the number basis functions. The same bottleneck is also encountered in MP2 calculations.

B. RPA equation in the product basis

Once these electron repulsion integrals are available, we are able to evaluate the *GW* approximation for atoms. We first solve the RPA equation in the product basis set $|ab\sigma\rangle$, where a and b are indexes over the mean-field eigenstates. This equation requires the diagonalization of the RPA two-particle Hamiltonian H_{RPA},

$$H_{\mathrm{RPA}\,ab\sigma}^{\;\;cd\sigma'} = (\epsilon_{b\sigma} - \epsilon_{a\sigma})\delta_{ac}\delta_{bd}\delta_{\sigma\sigma'} + (f_{a\sigma} - f_{b\sigma})(ab\sigma|cd\sigma'). \tag{17}$$

The product basis is limited to occupied-virtual or virtual-occupied pairs. This operation is then a matrix diagonalization of dimension $2N_{\mathrm{occupied}}N_{\mathrm{virtual}}N_{\mathrm{spin}}$. The diagonalization problem is non-symmetric. Let us consider the matrix R containing the right-eigenvectors,

$$H_{\mathrm{RPA}} R = R D. \tag{18}$$

The matrix D stands for the diagonal matrix containing the eigenvalues E_λ. The eigenvalues E_λ represents the neutral excitations with positive energy (resonant part of the spectrum) and negative energy (antiresonant part of the spectrum). The right eigenvectors R_λ are then expanded in the product basis $|ab\sigma\rangle$. The problem could be recast in a symmetric manner using the so-called Casida equations.[33]

Using the eigenvectors and eigenvalues of the RPA Hamiltonian, the polarizability χ can be written in the product basis,

$$\chi_{ab\sigma}^{cd\sigma}(\omega) = \sum_\lambda R_{\lambda ab\sigma}(\tilde{R}^{-1})_{\lambda cd\sigma} \times \left[\frac{\Theta(E_\lambda)}{\omega - E_\lambda + i\eta} + \frac{\Theta(-E_\lambda)}{\omega - E_\lambda - i\eta}\right], \tag{19}$$

where the index λ runs over the solutions of the RPA equation and $\tilde{R}^{-1}$ is a short notation for the matrix inverse of R the columns of which were multiplied by the occupation number difference. The Heavyside function $\Theta(E_\lambda)$ ensures the correct polar structure for a time-ordered response function: the negative energies E_λ are located just above the real axis of the complex plane and the positive energies just below.

C. *GW* self-energy with exact frequency dependence

Hence, introducing the Green's function from Eq. (9) and $W_p = v\chi v$ from Eq. (19) into the correlation self-energy Σ_c^{GW}, the residue theorem allows one to perform exactly the frequency integral. The final expression for *GW* self-energy reads

$$\left\langle i\sigma\left|\Sigma_c^{GW}(\omega)\right|j\sigma\right\rangle = \sum_{k\lambda} M_{\lambda ik\sigma}\tilde{M}_{\lambda kj\sigma}\left[\frac{f_{k\sigma}\Theta(E_\lambda)}{\omega - \epsilon_{k\sigma} + E_\lambda - i\eta} - \frac{(1 - f_{k\sigma})\Theta(-E_\lambda)}{\omega - \epsilon_{k\sigma} + E_\lambda + i\eta}\right], \tag{20}$$

where the intermediate matrix products

$$M_{\lambda ik\sigma} = \sum_{ab} R_{\lambda ab\sigma}(ik\sigma|ab\sigma) \tag{21}$$

and

$$\tilde{M}_{\lambda kj\sigma} = \sum_{cd}(\tilde{R}^{-1})_{\lambda cd\sigma}(cd\sigma|kj\sigma) \tag{22}$$

have been introduced. Note that the *GW* self-energy is diagonal in spin.

The polar structure of Σ_c can be observed in Fig. 2. The poles are located $\epsilon_{k\sigma} + E_\lambda$. The self-energy is very weakly

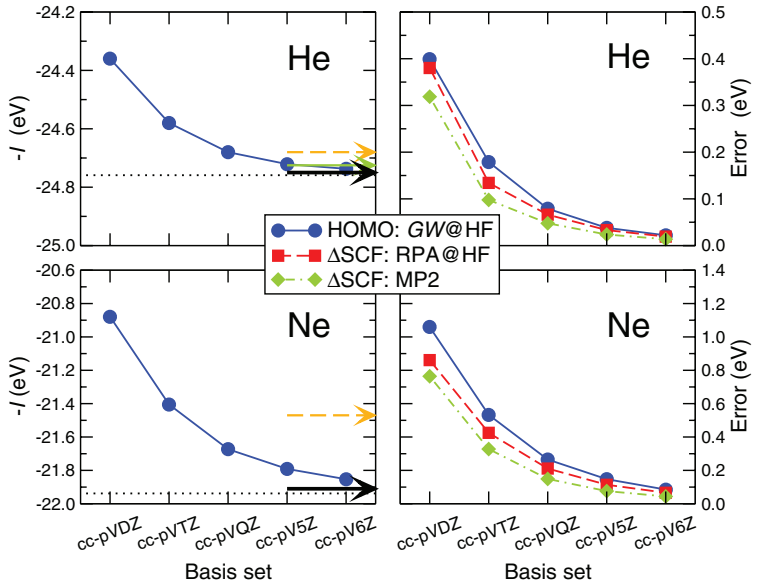

FIG. 3. Basis set convergence of the ionization energy of helium (upper panels) and neon (lower panels). The ionization energy is obtained from the HOMO energy of the atom within *GW*@HF in the left-hand panels. The horizontal arrows show the HOMO energy from literature's calculations in Ref. 6 (dashed arrow), in Ref. 7 (thick arrow), and in Ref. 10 (thin arrow). The horizontal doted line shows the complete-basis set limit using a simple extrapolation scheme.[35] The right-hand panels compare the convergence rates of the HOMO energy within *GW*@HF (circles) to the ΔSCF procedure using RPA@HF (squares) or using MP2 (diamonds).

frequency-dependent in the range of interest, i.e., in the region where the two curves intersect. In other words, the renormalization factor of the quasiparticle peaks is close to 1. The closest poles are approximately at HOMO energy minus the HOMO-LUMO gap and at the LUMO energy plus the HOMO-LUMO gap.

The implementation of QS*GW* is then straightforward. The full matrix $\langle i\sigma|\Sigma_c^{QSGW}|j\sigma\rangle$ is calculated and then transformed back into the basis representation $\langle\alpha|\Sigma_c^{QSGW}|\beta\rangle$ for spin up and spin down. The only additional difficulty arises from the self-consistency loop stabilization. We employ here a simple mixing scheme. We mix not only the density matrix, as it is customary for HF calculations, but also the self-energy matrix itself, since the self-energy depends directly onto the energies and the wavefunctions. With a mixing parameter of 0.5, we were able to achieve good convergence even with the largest basis sets of the present work using a maximum of 60 cycles.

D. RPA correlation energy

The RPA correlation energy is obtained as a by-product of the calculation of the polarizability χ. Indeed, Furche showed[34] that the RPA correlation energy can be obtained from the formula,

$$E_c^{\rm RPA} = \frac{1}{2}\sum_{\substack{\lambda \\ E_\lambda>0}} \left(E_\lambda - E_\lambda^{TDA}\right), \qquad (23)$$

where the sum has been limited to positive excitation energies and the excitation energies E_λ^{TDA} are obtained within the Tamm-Dancoff approximation that ignores the coupling between occupied to virtual excitations and virtual to occupied excitations. In practice, we perform a self-consistent HF (resp. QS*GW*) calculation and calculate the RPA correlation energy out of the HF (resp. QS*GW*) eigenvectors and eigenvalues. We label this procedure RPA@HF (resp. RPA@QS*GW*).

IV. CONVERGENCE AND ACCURACY

Before starting the systematic calculations, let us first check the reliability of the method. We first test the basis set convergence for some selected elements and then check our results against the very few published data for the first row atoms.

The basis set convergence is shown in Fig. 3 for the ionization energy of He and Ne. We observe the slow convergence of the *GW*@HF HOMO energy as a function of the basis set size. For Ne, an accuracy of 0.1 eV is obtained at the expense of a cc-pV5Z basis, which corresponds to 126 Cartesian Gaussian functions and a maximum angular momentum of $l = 5$. The rate of convergence is somewhat slower than reported for molecules in the previous Gaussian *GW* implementations.[6,9,10,36] However, it seems to be consistent with the convergence rate of RPA energies. RPA energies have been observed to require extremely complete basis sets in order to achieve chemical accuracy.[37] Figure 3 also shows the convergence rate of the ionization energy using the difference

TABLE I. Review of the previously published *GW* ionization energies and electron affinities of the first row atoms, and comparison with our results within the cc-pV5Z basis.

	HF+*GW*		LDA+*GW*		QS*GW*		
	This work	Earlier studies	This work	Earlier studies	This work	Earlier studies	Expt.[a]
Ionizations							
H			− 12.85	−12.66[b]			− 13.61
He	− 24.72	−24.68,[c] −24.73,[d] −24.75[e]	− 23.92	−23.65,[e] −24.20[f]			− 24.59
Be	− 9.16	−9.17,[d] −9.19[e]	− 9.02	−8.88,[e] −9.24[f]			− 9.32
B^+	− 24.88	−24.9[g]					− 24.15
Ne	− 21.79	−21.47,[c] −21.91[e]	− 20.97	−21.06,[e] −20.55[f]			− 21.56
Na			− 5.32	−5.40[h]	−5.43	−5.49[h]	− 5.15
Mg	− 7.62	−7.69[e]	− 7.53	−7.52[e]			− 7.65
Al^+	− 18.76	−18.9[g]					− 18.83
Ar	− 16.07	−15.94[c]					− 15.76
Electron affinities							
B^+	− 8.46	−8.5[g]					− 8.30
Na^+			− 4.71	−4.88[h]	−5.06	−5.05[h]	− 5.15
Al^+	− 6.01	−6.0[g]					− 5.99

[a]Reference 39.
[b]Reference 40.
[c]Reference 6.
[d]Reference 10.
[e]Reference 7.
[f]Reference 41.
[g]Reference 5.
[h]Reference 42.

of RPA total energies of the atom and of the positive ion. The convergence rate of the ΔSCF procedure nicely follows the convergence of the *GW* HOMO energy. Such a slow convergence is not surprising for perturbation theory. The MP2 calculations also shown in Fig. 3 are known to slowly converge to the complete basis set limit.[35,38]

From now on, all the calculations will be performed using the Dunning's cc-pV5Z basis set. This kind of basis appears as sufficient to ensure a 0.1 eV accuracy. The number of basis functions ranges from 70 for hydrogen to 130 for argon.

Table I compares our evaluation of *GW*@LDA, at *GW*@HF, and QS*GW* ionization energies and electron affinities with all the available results in the literature we are aware of. Results published to date use different basis sets: Gaussian basis sets for Refs. 6 and 10, numerical radial grid for Refs. 5, 7, 40, and 41 and plane-waves for Ref. 42. The overall agreement of our values with the published values is rather good, especially for *GW*@HF. The somewhat larger discrepancies for *GW*@LDA may possibly be attributed to the generalized Koopmans' theorem employed in Ref. 7. Note that agreement with the oldest results of Shirley and Martin is good.[5] The most similar implementation to ours[10] yields impressively similar results (within 0.01 eV). The only QS*GW* result for an atom from the literature also agrees well with our implementation.[42]

In Secs. V A–V C, we provide accurate evaluation of the ionization energy for all the first row atoms. These atoms include open-shell atoms, which are delicate to treat in a mean-field approach. Some approximations, such as local and semi-local approximations of DFT, minimize the total energy with fractional occupation numbers. Other approximations, such as HF and QS*GW*, do favor integral occupation numbers. In the present study, only approximations of the latter kind have been considered for the open-shell atoms and therefore, the occupation numbers have been safely set to integers.

V. IONIZATION OF ATOMS

A. Magnitude of the screening for atoms

Here, we evaluate the importance of the screening of the interactions for atoms. The perturbation theory in solids is often based on the screened Coulomb interaction W, whereas for atoms, it is rather based on the bare Coulomb interaction v. The rational behind this choice is the weak screening attributed to atoms. Indeed, the electrons in isolated atoms are localized and weakly polarizable. Therefore, it would be pointless using the complex W instead of the simple v for atoms. This explains why the *GW* approximation is prominent for the condensed matter, whereas the Møller-Plesset approximations are in use for gas phase calculations.

We would like to check explicitly the influence of using v or W for atoms. The comparison is exemplified with the HOMO expectation value for different approximations to the correlation self-energy Σ_c. The complete *GW* self-energy is compared to the first term in the ring diagram expansion (see Fig. 1). The one-ring self-energy is contained both in the MP2 approach and in the *GW* approach. The self-energy truncated to one-ring only is easily derived from Eq. (5), where χ is replaced by χ_0.

Figure 4 demonstrates that the difference between the infinite sum of ring diagrams and the truncation to the first

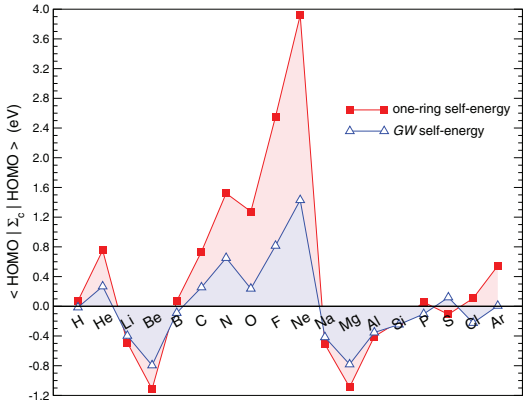

FIG. 4. HOMO expectation value of the correlation self-energy for the one-ring self-energy (squares) and for the *GW* self-energy (open triangles) for the light atoms. The calculations are based on HF inputs in a cc-pV5Z basis set.

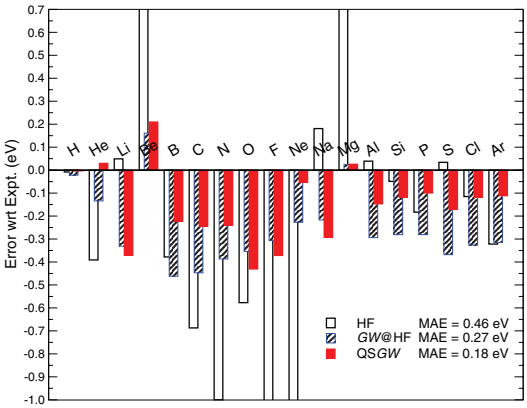

FIG. 5. Deviation from experiment[39] in the HOMO energy $\epsilon^N_{\rm HOMO} - (-I)$ of the light atoms within cc-pV5Z basis set for HF (open bars), *GW*@HF (striped bars), or QS*GW* (solid bars). The mean-absolute error (MAE) is also provided.

diagram is sizable. As expected, the *GW* self-energy is in general smaller than the one-ring counterpart. The statement is clear for the first row atoms and more contrasted for the second row. The closed-shell atoms (He, Be, Ne, Mg, Ne) are especially sensitive to the truncation of the sum over ring diagrams. We conclude that *W* is a better choice for a perturbation expansion, even for the light atoms.

B. HOMO of atoms and LUMO of cations

As explained in the introduction, there are several ways to evaluate the ionization energy of an atom. Most commonly, the total energy difference in the atom and the singly positively charged ion is taken. Alternative choices are the atom HOMO energy or the cation LUMO energy [Eq. (2)]. Within an exact theory, these three quantities are identical. Within DFT or HF, these three evaluations strongly deviate. There are some early indications *GW* should be much better.[42] We now consider these alternative forms for HF and two kinds of *GW*: *GW*@HF and QS*GW*.

Figure 5 shows the error with respect to experimental negative ionization energy: $\epsilon^N_{HOMO} - (-I)$. It is well known that the HF HOMO energy is not catastrophic in predicting the ionization energy. As might be expected because screening is weak, *GW*@HF and QS*GW* are quite similar. All the *GW* based approaches underestimate the position of the HOMO by a small amount. The perturbative *GW* approach seems to be justified for atoms: even when the HF starting point is noticeably wrong (e.g., for atoms of the end of the first row), the *GW*@HF performs almost as well as QS*GW*. Note that the *GW* correlation contains a small self-interaction: the HOMO of the hydrogen atom is not exact.

In Fig. 6, we provide the error with respect to experiment for the LUMO energy of singly positively charged ions within same three approximations. It is well known from text books that the LUMO in HF gives a very poor estimate to the ionization energy. We calculated a huge mean-absolute error (MAE) of 1.74 eV. The different *GW* flavors, *GW*@HF and QS*GW*, once again perform rather well in predicting the correct position of the cation LUMO energy. The MAE error is twice larger than for the positioning of the HOMO of atoms. Generally speaking, the self-consistency improves over the perturbative *GW*@HF for cations, except for carbon and silicon. When the HF starting point is completely off, the self-consistency can help much sometimes, as can be observed for the atoms of the end of the first row series and sometimes does not do much, as for the end of the second row series. In general, the *GW* based methods slightly overestimate the position of the LUMO of singly positively charged ions.

In order to reach the best agreement with experiment, it appears safer so far to evaluate the ionization energy with the *GW* approximation from the HOMO of the atom rather than from the LUMO of cations.

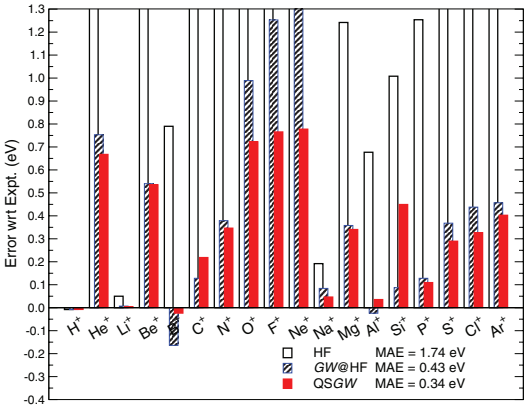

FIG. 6. Deviation from experiment[39] in the LUMO energy $\epsilon^{N-1}_{\rm LUMO} - (-I)$ of the light singly positively charged ions within cc-pV5Z basis set for HF (open bars), *GW*@HF (striped bars), or QS*GW* (solid bars). The mean-absolute error (MAE) is also provided.

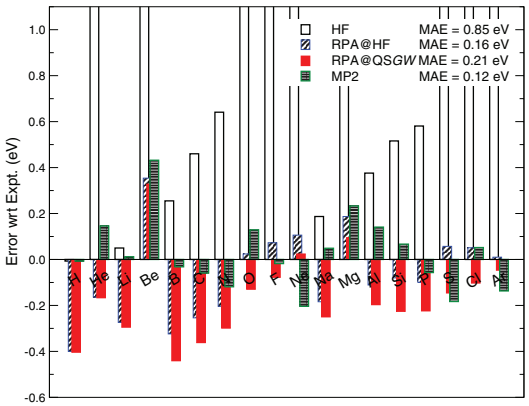

FIG. 7. Deviation from experiment[39] using the ΔSCF procedure $E_0^N - E_0^{N-1} - (-I)$ within cc-pV5Z basis set. Different levels of theory are shown: HF (open bars), RPA@HF (oblique striped bars), RPA@QS*GW* (full bars), MP2 (horizontally striped bars). The mean-absolute error (MAE) is also provided.

C. ΔSCF evaluation of the ionization energy

We now turn to the classical method to evaluate the ionization energy of atoms, namely the ΔSCF procedure. Figure 7 shows the error in calculating the ionization energy from the total energy difference described in Eq. (1). The RPA expression for the correlation energy corresponds to the *GW* approximation for the correlation self-energy. RPA energy and *GW* self-energy are, in principle, closely related. We hence employed four different approximations for the total energies: HF, RPA based on HF inputs, RPA based on QS*GW* inputs, and the standard MP2 approximation. For the ΔSCF procedure, MP2 clearly prevails over the other approximations. The second-order exchange diagram, as drawn in Fig. 1, which is present in MP2 and absent in RPA, is undoubtedly important for atoms. Thanks to this diagram, MP2 is devoid of self-interaction, whereas RPA suffers from self-interaction to some extent. This is clearly seen in the case of the hydrogen atom.

In the present work, we do not evaluate the RPA energy self-consistently with the corresponding RPA potential. However, the RPA functional is a stationary expression for the total energy. And even though the stationarity is believed to be limited,[25] the results should be weakly sensitive to the input Green's function. This is indeed what we observe in Fig. 7: RPA@HF and RPA@QS*GW* are in overall agreement. Surprisingly, RPA@HF appears slightly better than RPA@QS*GW* even when the HF starting point is clearly wrong. This statement calls for further investigations.

VI. CONCAVITY OF THE *GW* APPROXIMATION

How a particular approximation varies with fractional occupation number offers insight into its qualities and limitations. In the exact theory, the total energy should be a straight line in between the integral number of electrons.[43,44] Thus, the derivative of the energy with respect to electron number should be constant in between two consecutive integers and equal to the total energy difference. This last quantity is nothing else but the orbital energy (including a possible exchange-correlation discontinuity in the case of local Kohn-Sham potentials).[45,46]

In practice, the exchange-correlation approximations never induce the perfect straight line behavior. The deviation from the straight line is a sign of a localization or delocalization error.[11,47] In general, the approximations to DFT yield a convex total energy and therefore suffer from a delocalization error. An electron added to a system made of two identical well-separated subsystems minimizes its energy by splitting: half an electron goes on each subsystem. On the other hand, the HF approximation induces a concave total energy and is therefore affected by a localization error. The aforementioned extra electron lowers its energy by localizing on one single subsystem. In the exact theory, spreading or localizing the electron should not affect the total energy.

In Ref. 42, we established that for clusters in Na and certain defects in SiC, localized energy levels were slightly concave with respect to occupation. The concavity can be accessed from the ordering between the LUMO energy of a positive ion and the HOMO energy of the corresponding atom, if one assumes a monotonic behavior of the orbital energy as a function of the fractional number of electrons. This is generally the case, except for MP2 to a small extent.[48]

In Fig. 8, we recast the previously calculated orbital energies now considering ΔSCF as the reference. From the upper panel, one can observe that the HF approximation is clearly concave. Indeed, the HOMO energy of the atom is always much lower than the total energy difference, whereas the LUMO energy of the positive ion is always much higher. The atom HOMO energy, nor the cation LUMO energy, are a good estimate for the total energy difference. The deviation is very large in both cases. Following Slater's argument,[49]

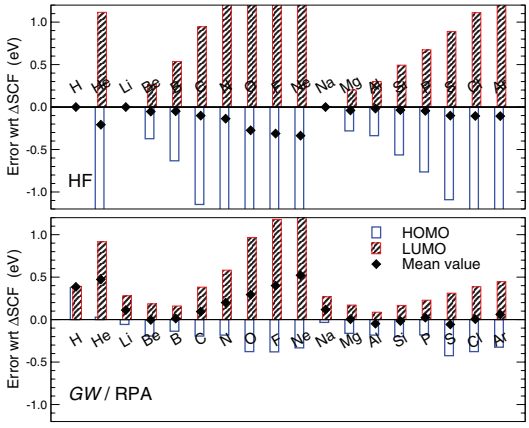

FIG. 8. Deviation from ΔSCF reference for the atom HOMO energy $\epsilon^N_{\mathrm{HOMO}}$ (open bars) and of the cation LUMO energy (striped bars). The mean value of the HOMO and the LUMO is displayed with the diamond symbol. The upper panel compares HF orbital energies to the HF total energy difference. The lower panel compares *GW*@HF orbital energies to the RPA@HF total energy difference.

let us assume that the deviation from the straight line can be approximated by a second order polynomial. Under this mild assumption, Slater proposed to use the orbital energy at half charge to best approximate the total energy difference. Alternatively, under the same assumption, we proposed in Ref. 42 to approximate the total energy difference with the mean-value of the HOMO of the atom and the LUMO of the positive ion,

$$I \approx -\frac{\epsilon^{N}_{\rm HOMO} + \epsilon^{N-1}_{\rm LUMO}}{2}. \qquad (24)$$

With this alternative evaluation of the ionization energy, there is no need to perform *GW* calculations for half charges. It requires nevertheless to perform two separate calculations. The outcome of the mean-value technique is given in Fig. 8 with the diamond symbols. For HF energies, the agreement between the mean-value and the ΔSCF energy difference is striking.

In the lower panel of Fig. 8, we compare the *GW* orbital energies to the RPA total energies. In this case, the HOMO energy of the atom is always slightly lower than the total energy difference, whereas the LUMO energy of the positive ion is always moderately higher. This proves the weak concavity of the *GW* approximations. Compared to HF, the *GW* orbital energies are much better estimates to the total energy difference. The associated localization error is then much weaker than the one of HF. The mean-value technique within *GW* yields a nice estimate of the RPA ΔSCF procedure. Only the end of the first row atom series deviates noticeably. This conclusion confirms our previous works on sodium clusters[42] on defects in solids[50,51] that first identified the slight concavity of the *GW* approximation. We confirm here that the mean-value technique is a more correct estimate to the total energy difference than the mere atom HOMO or cation LUMO.

VII. CONCLUSIONS

In this article, we described an implementation of the *GW* approximation to the electronic self-energy for atoms. This Gaussian basis set implementation does not need auxiliary functions and is based on an exact convolution in the frequency domain, so that no extra technical approximation is made besides the choice of the basis set. In addition to the usual perturbative approach to *GW* such as *GW*@LDA or *GW*@HF, we introduced the recently proposed self-consistent scheme named QS*GW*. The RPA correlation energies were obtained as a mere by-product of the code.

We considered different flavors of the *GW* approximation (*GW*@HF or QS*GW*) for the light atoms, from H to Ar. Noticeably, we calculated non-spherical atoms and spin-polarized systems, which have never been treated within *GW* to the best of our knowledge. An important technical conclusion of the present work is the slow convergence of the *GW* calculations with respect to the basis set size. This is however not completely surprising, when compared to RPA energy or MP2 energy convergence rates. The targeted error bar of 0.1 eV for HOMO/LUMO orbital energies could be reached only at the expense of a large cc-pV5Z basis set.

We then demonstrated the reliability of the *GW* approximation for the HOMO energy of atoms and for the LUMO energy of the cations compared to the experimental data. Since the HF approximation performs reasonably well for atoms and ions, the difference between perturbative *GW* based on HF (*GW*@HF) and self-consistent *GW* (QS*GW*) is not large, even though QS*GW* is slightly better on average. When turning to total energies, one could infer that the main ingredient missing in the RPA correlation is the second-order exchange diagram, which is contained in MP2. Comparing the total energy difference to the HOMO/LUMO orbital energy, we could confirm the weak concavity of the *GW* approximation for fractional electron numbers. The mean-value between the HOMO energy of the atom and the LUMO energy of the positive ion appears as a correct way to evaluate the total energy difference.

ACKNOWLEDGMENTS

We thank M. van Schilfgaarde and X. Blase for fruitful discussions and their comments on the manuscript.

[1] L. Hedin, Phys. Rev. **139**, A796 (1965).
[2] G. Strinati, H. J. Mattausch, and W. Hanke, Phys. Rev. B **25**, 2867 (1982).
[3] M. S. Hybertsen and S. G. Louie, Phys. Rev. B **34**, 5390 (1986).
[4] F. Aryasetiawan and O. Gunnarsson, Rep. Prog. Phys. **61**, 237 (1998).
[5] E. L. Shirley and R. M. Martin, Phys. Rev. B **47**, 15404 (1993).
[6] M. Rohlfing, Int. J. Quantum Chem. **80**, 807 (2000).
[7] A. Stan, N. E. Dahlen, and R. van Leeuwen, Europhys. Lett. **76**, 298 (2006).
[8] C. Rostgaard, K. W. Jacobsen, and K. S. Thygesen, Phys. Rev. B **81**, 085103 (2010).
[9] X. Blase, C. Attaccalite, and V. Olevano, Phys. Rev. B **83**, 115103 (2011).
[10] S.-H. Ke, Phys. Rev. B **84**, 205415 (2011).
[11] A. J. Cohen, P. Mori-Sánchez, and W. Yang, Science **321**, 792 (2008).
[12] J. C. Grossman, M. Rohlfing, L. Mitas, S. G. Louie, and M. L. Cohen, Phys. Rev. Lett. **86**, 472 (2001).
[13] F. Bruneval, N. Vast, L. Reining, M. Izquierdo, F. Sirotti, and N. Barrett, Phys. Rev. Lett. **97**, 267601 (2006).
[14] M. Gatti, F. Bruneval, V. Olevano, and L. Reining, Phys. Rev. Lett. **99**, 266402 (2007).
[15] H. Jiang, R. I. Gomez-Abal, P. Rinke, and M. Scheffler, Phys. Rev. B **82**, 045108 (2010).
[16] B. Holm and U. von Barth, Phys. Rev. B **57**, 2108 (1998).
[17] W. Ku and A. G. Eguiluz, Phys. Rev. Lett. **89**, 126401 (2002).
[18] A. Stan, N. E. Dahlen, and R. van Leeuwen, J. Chem. Phys. **130**, 114105 (2009).
[19] S. V. Faleev, M. van Schilfgaarde, and T. Kotani, Phys. Rev. Lett. **93**, 126406 (2004).
[20] F. Bruneval, N. Vast, and L. Reining, Phys. Rev. B **74**, 045102 (2006).
[21] M. van Schilfgaarde, T. Kotani, and S. Faleev, Phys. Rev. Lett. **96**, 226402 (2006).
[22] M. Shishkin, M. Marsman, and G. Kresse, Phys. Rev. Lett. **99**, 246403 (2007).
[23] C. Møller and M. S. Plesset, Phys. Rev. **46**, 618 (1934).
[24] G. Baym, Phys. Rev. **127**, 1391 (1962).
[25] N. E. Dahlen and U. v. Barth, Phys. Rev. B **69**, 195102 (2004).
[26] M. L. Tiago and J. R. Chelikowsky, Phys. Rev. B **73**, 205334 (2006).
[27] B. I. Lundqvist, Phys. Kondens. Mater. **6**, 206 (1967).
[28] H. N. Rojas, R. W. Godby, and R. J. Needs, Phys. Rev. Lett. **74**, 1827 (1995).
[29] T. H. Dunning Jr., J. Chem. Phys. **90**, 1007 (1989).
[30] D. Feller, J. Comp. Chem. **17**, 1571 (1996).
[31] K. Schuchardt, B. Didier, T. Elsethagen, L. Sun, V. Gurumoorthi, J. Chase, J. Li, and T. Windus, J. Chem. Inf. Model. **47**, 1045 (2007).
[32] See http://sourceforge.net/p/libint to obtain the library LIBINT, an efficient library for calculating the Coulomb integrals.
[33] M. E. Casida, *Recent Advances in Density Functional Methods, Part I* (World Scientific, Singapore, 1995), p. 155.

[34]F. Furche, J. Chem. Phys. **129**, 114105 (2008).
[35]A. Halkier, T. Helgaker, P. Jorgensen, W. Klopper, H. Koch, J. Olsen, and A. Wilson, Chem. Phys. Lett. **286**, 243 (1998).
[36]D. Foerster, P. Koval, and D. Sanchez-Portal, J. Chem. Phys. **135**, 074105 (2011).
[37]F. Furche, Phys. Rev. B **64**, 195120 (2001).
[38]E. C. Barnes and G. A. Petersson, J. Chem. Phys. **132**, 114111 (2010).
[39]*CRC Handbook of Chemistry and Physics*, 84th ed., edited by D. P. Lide (CRC, Boca Raton, Florida, 2003).
[40]W. Nelson, P. Bokes, P. Rinke, and R. W. Godby, Phys. Rev. A **75**, 032505 (2007).
[41]A. J. Morris, M. Stankovski, K. T. Delaney, P. Rinke, P. García-González, and R. W. Godby, Phys. Rev. B **76**, 155106 (2007).
[42]F. Bruneval, Phys. Rev. Lett. **103**, 176403 (2009).
[43]J. P. Perdew, R. G. Parr, M. Levy, and J. L. Balduz Jr., Phys. Rev. Lett. **49**, 1691 (1982).
[44]W. T. Yang, Y. Zhang, and P. W. Ayers, Phys. Rev. Lett. **84**, 5172 (2000).
[45]M. Grüning, A. Marini, and A. Rubio, J. Chem. Phys. **124**, 154108 (2006).
[46]F. Bruneval, F. Sottile, V. Olevano, and L. Reining, J. Chem. Phys. **124**, 144113 (2006).
[47]P. Mori-Sánchez, A. J. Cohen, and W. T. Yang, Phys. Rev. Lett. **100**, 146401 (2008).
[48]A. J. Cohen, P. Mori-Sánchez, and W. Yang, J. Chem. Theory Comput. **5**, 786 (2009).
[49]J. C. Slater, *The Self-Consistent Field for Molecules and Solids* (McGraw-Hill, New York, 1974), Vol. 4.
[50]M. Giantomassi, M. Stankovski, R. Shaltaf, M. Grüning, F. Bruneval, P. Rinke, and G.-M. Rignanese, Phys. Status Solidi B **248**, 275 (2011).
[51]F. Bruneval, Nucl. Instrum. Methods Phys. Res. B **277**, 77 (2012).

Range-Separated Approach to the RPA Correlation Applied to the van der Waals Bond and to Diffusion of Defects

Fabien Bruneval
CEA, DEN, Service de Recherches de Métallurgie Physique, F-91191 Gif-sur-Yvette, France
(Received 1 February 2012; published 19 June 2012)

The random-phase approximation (RPA) is a promising approximation to the exchange-correlation energy of density functional theory, since it contains the van der Waals (vdW) interaction and yields a potential with the correct band gap. However, its calculation is computationally very demanding. We apply a range-separation concept to RPA and demonstrate how it drastically speeds up the calculations without loss of accuracy. The scheme is then successfully applied to a layered system subjected to weak vdW attraction and is used to address the controversy of the self-diffusion in silicon. We calculate the formation and migration energies of self-interstitials and vacancies taking into account atomic relaxations. The obtained activation energies deviate significantly from the earlier calculations and challenge some of the experimental interpretations: the diffusion of vacancies and interstitials has almost the same activation energy.

DOI: 10.1103/PhysRevLett.108.256403 PACS numbers: 71.15.−m, 61.72.Bb, 61.72.uf

The quest for the exact exchange-correlation energy of density functional theory (DFT) is endless of course. However, the random-phase approximation (RPA) [1,2] is now believed to be a huge step forward. It is now commonplace to cite the RPA as *the* first-principles method that correctly describes the weak van der Waals (vdW) interaction [3], which is prominent for many important problems: physisorption [4–6] and layered system binding [7,8], for instance. A much less known feature of RPA is the correct prediction of band gaps, as opposed to the large underestimation of the local and semilocal approximation to DFT. It has been demonstrated recently [9] that the exchange-correlation potential obtained from RPA closely resembles the *GW* approximation [10], which is nowadays the most robust method to predict band gap of solids.

The correctness of the band gaps would make RPA a method of choice for the properties of defects in semiconductors and insulators. The underestimation of the band gap in the calculations is known to be the reason why the usual local and semilocal approximations fail for defects in semiconductors [11,12]. Whereas the calculation of the energetic of point defects relies on total energies only, a poor description of the band structure will still affect the final formation energies.

An approach with no band gap problem should be able to clarify the experimental controversy about the self-diffusion in silicon. Although silicon can be regarded as the best characterized material ever and although the self-diffusion is a key parameter for industrial processes, there is still no unanimous interpretation for this phenomenon for silicon [13–20]. Self-diffusion in solids is governed by the formation and the migration of point defects, namely, vacancies and self-interstitials.

For the above mentioned reasons, RPA is a very appealing framework. However, its application has been limited so far to simple systems cases, because of its numerical intricacies. Its convergence behavior is so bad that most groups had to employ extrapolation techniques [7,21,22] to infer the converged properties out of a few underconverged calculations. Furthermore, the scaling with system size is dramatically high and the application to point defects in supercells would be out of reach.

In this Letter, we introduce a range-separated framework for the calculation of the RPA correlation energy. The short-range (SR) part is to be approximated with a local density approximation (LDA), whereas the long-range (LR) part is to be calculated exactly. This approach speeds up the calculations with a controlled loss of accuracy. We demonstrate its robustness calculating a wide variety of covalent crystals and a vdW bonded system, namely, hexagonal boron nitride. Within this approach, the system size required for an accurate description of point defects is made accessible. We apply the method to the calculation of the self-diffusion in silicon and show that the commonly used *ab initio* values need to be drastically revised.

A RPA calculation relies on the calculation of the electronic polarizability. The convergence of the RPA energy is surprisingly slow against both the basis representation of the polarizability and the number of empty states that should be included in the sum-over-state formula [21,23]. We propose to overcome this situation thanks to the range-separation idea. Following Toulouse and co-workers [24,25], the Coulomb interaction v can be split into SR and LR components:

$$v(r) = \frac{1 - \mathrm{erf}(r/r_c)}{r} + \frac{\mathrm{erf}(r/r_c)}{r}, \qquad (1)$$

where r_c is a cutoff radius (Hartree atomic units are employed throughout the Letter).

At variance with Toulouse and co-workers, the purpose of the splitting is not to fix some SR deficiencies of the RPA, but simply to accelerate the convergence of the RPA energies. We intend to benefit from the fast decay of the LR part of the Coulomb interaction in Fourier space $4\pi/q^2 \exp[-(r_c q)^2/4]$. In comparison, the bare Coulomb interaction does not contain the exponential term.

A LDA evaluation of the RPA energy [26] noticeably overestimates the computed RPA energy. It is sensible to anticipate that the LDA is a reliable approximation for the SR part of RPA but not for the LR. Hence, we propose to evaluate the total RPA correlation energy as follows:

$$E_c^{\mathrm{RPA}} = \int d\mathbf{r}\epsilon_c^{\mathrm{RPA,jellium}}[n(\mathbf{r})]n(\mathbf{r}) - \int d\mathbf{r}\epsilon_c^{\mathrm{LR\text{-}RPA,jellium}}[n(\mathbf{r}), r_c]n(\mathbf{r}) + E_c^{\mathrm{LR\text{-}RPA,calc}}(r_c), \quad (2)$$

where $n(\mathbf{r})$ is the electronic density, $\epsilon_c^{\mathrm{RPA,jellium}}$ is the RPA correlation energy density of the jellium subjected to the bare Coulomb interaction, $\epsilon_c^{\mathrm{LR\text{-}RPA,jellium}}$ is the RPA correlation energy density of the jellium with the LR-only interaction, and finally $E_c^{\mathrm{LR\text{-}RPA,calc}}$ is the calculated RPA correlation energy with the same LR-only interaction. The expression for the LR-RPA correlation energy can be easily derived from the usual expression of the RPA energy (see, e.g., Ref. [22]). The modified interaction is governed by the cutoff radius r_c, which plays the role of a convergence parameter in our scheme. Indeed, if r_c is set to 0, the LR interaction turns out to be the full interaction and we recover the usual expression for the RPA correlation energy. If r_c is set to ∞, the LR interaction vanishes and the scheme turns out to be equal to the usual LDA evaluation to the RPA correlation energy.

In order to implement Eq. (2) in solid state calculations, an explicit expression for the LR-RPA correlation energy density of jellium had to be establish. We numerically evaluated the RPA integrals in jellium [27] using either the Coulomb interaction or the LR interaction for different r_c values. The calculated energies were then interpolated with a Padé approximant [28]. Figure 1 shows the behavior of the computed RPA energies based on LDA wave functions and energies as a function of the cutoff of the modified interaction. As the correlation energy is not linear with respect to the interaction, the SR contribution is defined as the difference between the total correlation and the LR-only correlation. First of all, our RPA correlation energy for the full interaction is in very good agreement with previously published values: 6.12 eV/atom to be compared to 6.11 eV/atom from Ref. [29]. The discrepancy between the LDA evaluation of the RPA correlation energy and the computed one arises mainly from the LR part: the LDA evaluation of the SR contribution nicely reproduces the explicit calculation for radius as large as $r_c = 4$ bohr.

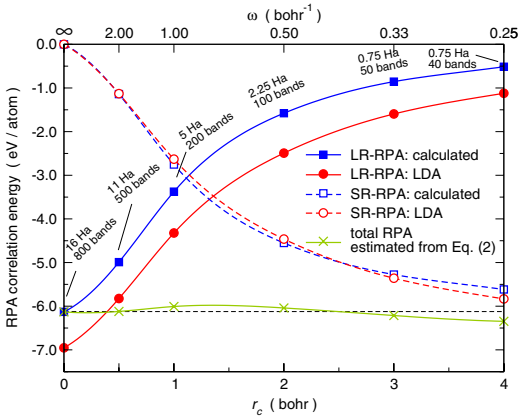

FIG. 1 (color online). LR-RPA (filled symbols) and SR-RPA (open symbols) correlation energies of bulk silicon for lattice constant $a = 10.26$ bohr, as a function of the cutoff radius $r_c = 1/\omega$. The SR-RPA correlation energy is obtained as the RPA correlation energy minus the LR-RPA correlation energy. The explicit calculation is displayed with squares and LDA with circles. The total correlation is displayed with crosses. The horizontal line emphasizes the calculated RPA correlation energies with full Coulomb interaction. Each point is associated with the corresponding convergence parameters necessary to achieve a 2 meV/atom accuracy: first, the cutoff energy for the polarizability (Ha), and second, the number of states to be included in the expression of the polarizability.

This observation confirms the assumption we made in Eq. (2) when approximating the total RPA correlation energy. The approximated total correlation energy (cross symbols) closely follows the full calculation shown with the horizontal line.

Figure 1 also shows the convergence parameters for the different values of r_c. The use of a LR-only interaction is very convenient for a plane-wave expansion. Firstly, the polarizabilities, which are required for a RPA calculation, are two point functions and therefore their calculation is massively accelerated when lowering in the plane-wave cutoff energy. Secondly, the number of empty states required to achieve convergence is largely reduced, since the exponential decay in the LR-only interaction in Fourier space drastically decreases the coupling between the occupied states and the high energy empty states.

For practical applications, one has to determine the largest radius r_c that still captures the desired physical effects. The cutoff radius has to be considered as a convergence parameter. In order to appreciate the relevant range for r_c, Table I shows the atomization energy of selected crystals. The list includes a metal, narrow and wide band gap semiconductors, zinc blende and wurzite semiconductors. The atomization energy can be considered as a difficult test for the range-separation technique, since it compares solids to atoms, which have noticeably

TABLE I. Atomization energy or binding energy of a selection of crystals in eV per atom. RPA evaluation is given with our range-separated scheme using different values of r_c and with the standard expression as a reference.

		RPA with r_c				
Crystal	PBE	2.0	1.0	0.5	RPA	Expt.
Al	3.56	3.45	3.53	3.50	3.44	3.39
Si	4.63	4.56	4.60	4.64	4.63	4.62
β-SiC	6.49	6.01	6.08	6.11	6.12	6.34
Diamond	7.82	7.11	7.27	7.34	7.27	7.37
w-AlN	6.54	4.96	4.92	5.52	5.65	5.83
c-BN	7.80	5.90	5.84	6.29	6.28	6.68

different spatial extension. If we compare the range-separated RPA to the standard RPA, we conclude that $r_c = 0.5$ bohr yield converged results. When dealing with larger atoms, not in the first row of the periodic table, a larger value for r_c can be safely retained. The overall agreement of RPA with respect to experiment is very good.

The proposed range-separated technique is highly relevant for vdW bonded system. Indeed, our scheme should automatically describe the covalent bond with the LDA quality and the distant vdW bonds with the RPA precision. This statement is exemplified in Fig. 2 with the interlayer spacing of hexagonal boron nitride. In this system, LDA is correct thanks to a fortunate compensation of errors and PBE [30] largely overestimates the interlayer spacing. Genuine RPA is known to be excellent for h-BN [7] and clearly superior to the modelized vdW-DF approach [31]. Whereas $r_c = 4$ bohr is definitely too large; the range separation using $r_c = 1$ or 2 bohr is sufficient to yield the correct interlayer spacing and the correct elastic constant C_{33}.

We now turn to the large supercells necessary to predict self-diffusion of silicon. The RPA potential yields a good evaluation of the band gap (1.30 eV to be compared to the experimental value of 1.17 eV) and therefore the energetics of point defects in silicon should be strongly corrected with respect to the LDA or PBE values. For completeness, we also performed hybrid functional calculations within HSE06 [32] and PBE0 [33]. For instance, HSE06 yields a nice band gap of 1.20 eV for silicon.

The defect calculations are performed within 16, 64, and 216 atom cubic supercells. Care was taken about the **k**-point convergence for RPA ($2 \times 2 \times 2$ grid) and for the exact exchange ($4 \times 4 \times 4$ grid) [29]. The validity of a cutoff radius $r_c = 1.0$ bohr was checked against a smaller radius of $r_c = 0.5$ eV. The number of empty states was then further reduced using an acceleration scheme [34].

The RPA scheme does not provide the forces easily. Thanks to the similarity between the elastic constant of LDA and RPA, we manually relaxed the few degrees of freedom directly involved in the defect structure, and the other ones were relaxed within LDA. This procedure yields basically negligible energetical changes except for the vacancy that experiences a large Jahn-Teller distortion as exemplified in Fig. 3. In contrast with LDA, RPA massively favors the Jahn-Teller configuration against the tetrahedric environment (0.7 eV gain within RPA, almost 0 eV within LDA).

The formation and migration energies relevant for the self-diffusion through neutral defects are summarized in Table II. HSE06 and RPA show very similar trends, even

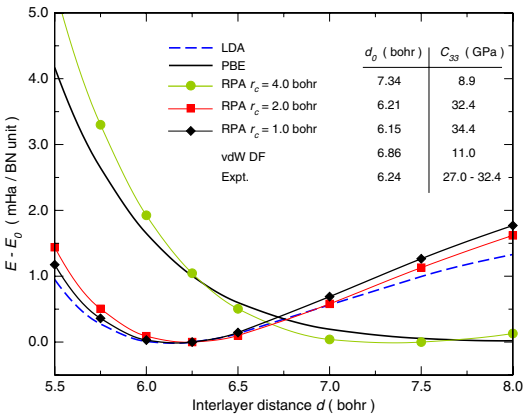

FIG. 2 (color online). Energy as a function of the interlayer spacing $d = c/2$ of hexagonal BN within LDA (dashed line), PBE (solid line), RPA with $r_c = 4$, 2, or 1 (respectively, circles, squares, or diamonds). The equilibrium spacing d_0 and elastic constant C_{33} are shown and compared to the modelized method vdW DF [31] and to experiment [41,42].

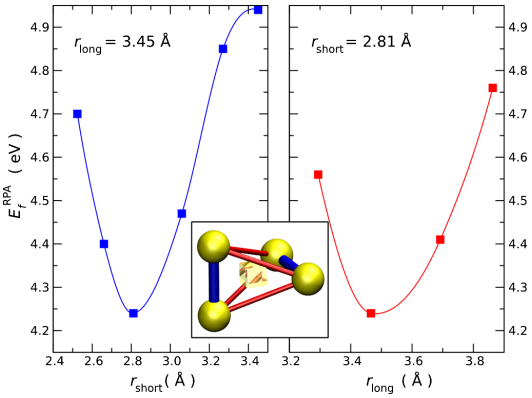

FIG. 3 (color online). Silicon vacancy V_{Si} formation energy in a 63-atom supercell as a function of the nearest neighbor atom distances. In the left-hand panel, the longest distance r_{long} is fixed and the shortest one r_{short} is varied. In the right-hand panel, r_{long} is fixed and r_{short} is varied. The absence of Jahn-Teller distortion corresponds to $r_{long} = r_{short}$. The atomic configuration is displayed in the inset: The cube stands in the empty lattice site.

TABLE II. Formation energies and migration barriers in eV of the self-interstitial and vacancy in silicon within different *ab initio* schemes for different supercell sizes. The results for HSE06 [32] and for PBE0 [33] are given for comparison. The QMC value is taken from Ref. [35].

	LDA	PBE	HSE06	PBE0	RPA	QMC
16-atom supercell						
$Si_{split\langle 110\rangle}$	3.45	3.65	4.50	4.61	5.06	4.94
64-atom supercell						
$Si_{split\langle 110\rangle}$	3.45	3.62	4.40	4.50	4.49	
Si_{hex}	3.48	3.67	4.52	4.63	4.74	
$Si_{split\langle 110\rangle} \rightarrow Si_{hex}$	0.37	0.40	0.47	0.49	0.77	
$Si_{hex} \rightarrow Si_{hex}$	0.12	0.21	0.49	0.69	1.01	
V_{Si}	3.66	3.71	4.56	4.64	4.24	
$V_{Si} \rightarrow V_{Si}$	0.40	0.28	0.40	0.58	0.83	
216-atom supercell						
V_{Si}	3.58	3.72			4.33	

though the energetics can differ in the details. We present the small 16-atom supercell in order to allow comparison with earlier quantum Monte Carlo (QMC) calculations [35,36]. RPA seems to nicely approximate the high level QMC method. However, our results show that the 16 atom supercell is too small to achieve convergence, mainly because of the long-ranged exchange interaction.

Generally speaking, the energy of all the defects is underestimated by 0.7–1.0 eV by LDA and PBE compared to HSE06 or RPA. The RPA formation energy compares favorably with earlier *GW* calculations [37]. The migration barriers are also underestimated with LDA and PBE. Noticeably, the migration of self-interstitial in the hexagonal sites Si_{hex} had a very low barrier (0.12 eV for LDA) and, as a consequence, was the preferred mechanism for self-interstitial migration for both LDA and PBE. When turning to RPA (and HSE06), this diffusion path is completely ruled out against the following two-step path: $Si_{split\langle 110\rangle} \rightarrow Si_{hex} \rightarrow Si_{split\langle 110\rangle}$. The corresponding diffusion activation energy (formation + barrier) is 4.87 eV within HSE06 and 5.26 eV within RPA. These values lie in the range of the experimental values 4.95 eV [38] and 5.15 eV [39], obtained from the self-interstitial assisted diffusion of zinc in silicon.

Concerning the vacancy diffusion, the situation is even more debated. The positron annihilation spectroscopy is not conclusive [13,14] and the diffusion measurements have difficulties isolating the vacancy contribution [15–20]. In calculations, the vacancy is known to converge slowly with system size [40]. We therefore performed a 216-atom supercell calculation to ensure a 0.1 eV convergence as shown in Table II. The vacancy diffusion activation energy within RPA 5.16 eV and HSE06 4.96 eV are much higher than the corresponding LDA estimate 4.06 eV. We confirm the warnings raised recently by some authors [18]: the agreement between theory at the LDA level and experiment seems to be completely fictitious. The diffusion activation energy of interstitials and of vacancies is almost the same: this piece of information is much of a surprise.

In conclusion, we demonstrated in this Letter the practical advantages of range separation when applied to RPA. The SR is approximated within LDA and the LR part is calculated exactly. This procedure is perfectly suited to Fourier space approaches. The efficiency gain without accuracy loss is so substantial that the application to the properties of point defects becomes accessible using supercells as large as 216 atoms. RPA is a relevant trade-off between the fast but not reliable LDA and the slow but accurate QMC calculations. The described scheme allowed us to produce an estimate for the activation energies of self-diffusion in silicon. Our energies, which significantly deviate from the corresponding LDA or PBE values, are in good agreement with respect to experiment for interstitials. The diffusion path of interstitials is identified as a transformation between two different configurations $Si_{split\langle 110\rangle}$ and Si_{hex}. Surprisingly, the vacancy and the interstitial activation energies are calculated to be very close.

We acknowledge insightful discussions with J.-P. Crocombette, G. Roma, and E. Clouet. The calculations presented here are performed with the plane-wave codes ABINIT [43] and QUANTUM-ESPRESSO [44]. This work was performed using HPC resources from GENCI-CINES and GENCI-CCRT (Grant No. 2012-gen6018).

[1] D. Bohm and D. Pines, Phys. Rev. **92**, 609 (1953).
[2] D. C. Langreth and J. P. Perdew, Phys. Rev. B **15**, 2884 (1977).
[3] J. F. Dobson and J. Wang, Phys. Rev. Lett. **82**, 2123 (1999).
[4] X. Ren, P. Rinke, and M. Scheffler, Phys. Rev. B **80**, 045402 (2009).
[5] L. Schimka, J. Harl, A. Stroppa, A. Grueneis, M. Marsman, F. Mittendorfer, and G. Kresse, Nature Mater. **9**, 741 (2010).
[6] J. Ma, A. Michaelides, D. Alfe, L. Schimka, G. Kresse, and E. Wang, Phys. Rev. B **84**, 033402 (2011).
[7] A. Marini, P. García-González, and A. Rubio, Phys. Rev. Lett. **96**, 136404 (2006).
[8] S. Lebègue, J. Harl, T. Gould, J. G. Ángyán, G. Kresse, and J. F. Dobson, Phys. Rev. Lett. **105**, 196401 (2010).
[9] Y.-M. Niquet and X. Gonze, Phys. Rev. B **70**, 245115 (2004).
[10] L. Hedin, Phys. Rev. **139**, A796 (1965).
[11] W. R. L. Lambrecht, Phys. Status Solidi B **248**, 1547 (2011).
[12] M. Giantomassi, M. Stankovski, R. Shaltaf, M. Grning, F. Bruneval, P. Rinke, and G.-M. Rignanese, Phys. Status Solidi B **248**, 275 (2011).
[13] S. Dannefaer, P. Mascher, and D. Kerr, Phys. Rev. Lett. **56**, 2195 (1986).
[14] R. Würschum, W. Bauer, K. Maier, A. Seeger, and H.-E. Schaefer, J. Phys. Condens. Matter **1**, SA33 (1989).

[15] H. Bracht, E. E. Haller, and R. Clark-Phelps, Phys. Rev. Lett. **81**, 393 (1998).
[16] A. Ural, P. B. Griffin, and J. D. Plummer, Phys. Rev. Lett. **83**, 3454 (1999).
[17] Y. Shimizu, M. Uematsu, and K. M. Itoh, Phys. Rev. Lett. **98**, 095901 (2007).
[18] H. Bracht and A. Chroneos, J. Appl. Phys. **104**, 076108 (2008).
[19] H. Bracht and E. E. Haller, Phys. Rev. Lett. **85**, 4835 (2000).
[20] A. Ural, P. B. Griffin, and J. D. Plummer, Phys. Rev. Lett. **85**, 4836 (2000).
[21] F. Furche, Phys. Rev. B **64**, 195120 (2001).
[22] J. Harl and G. Kresse, Phys. Rev. B **77**, 045136 (2008).
[23] M. Fuchs and X. Gonze, Phys. Rev. B **65**, 235109 (2002).
[24] J. Toulouse, F. Colonna, and A. Savin, Phys. Rev. A **70**, 062505 (2004).
[25] J. Toulouse, I. C. Gerber, G. Jansen, A. Savin, and J. G. Ángyán, Phys. Rev. Lett. **102**, 096404 (2009).
[26] S. Vosko, L. Wilk, and M. Nusair, Can. J. Phys. **58**, 1200 (1980).
[27] U. von Barth and L. Hedin, J. Phys. C **5**, 1629 (1972).
[28] S. Goedecker, M. Teter, and J. Hutter, Phys. Rev. B **54**, 1703 (1996).
[29] H.-V. Nguyen and S. de Gironcoli, Phys. Rev. B **79**, 205114 (2009).
[30] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).
[31] H. Rydberg, M. Dion, N. Jacobson, E. Schröder, P. Hyldgaard, S. I. Simak, D. C. Langreth, and B. I. Lundqvist, Phys. Rev. Lett. **91**, 126402 (2003).
[32] J. Heyd, G. E. Scuseria, and M. Ernzerhof, J. Chem. Phys. **124**, 219906 (2006).
[33] C. Adamo and V. Barone, J. Chem. Phys. **110**, 6158 (1999).
[34] F. Bruneval and X. Gonze, Phys. Rev. B **78**, 085125 (2008).
[35] E. R. Batista, J. Heyd, R. G. Hennig, B. P. Uberuaga, R. L. Martin, G. E. Scuseria, C. J. Umrigar, and J. W. Wilkins, Phys. Rev. B **74**, 121102 (2006).
[36] W.-K. Leung, R. J. Needs, G. Rajagopal, S. Itoh, and S. Ihara, Phys. Rev. Lett. **83**, 2351 (1999).
[37] P. Rinke, A. Janotti, M. Scheer, and C. G. Van de Walle, Phys. Rev. Lett. **102**, 026402 (2009).
[38] H. Bracht, N. A. Stolwijk, and H. Mehrer, Phys. Rev. B **52**, 16542 (1995).
[39] V. Voronkov and R. Falster, Mater. Sci. Eng. B **134**, 227 (2006).
[40] F. Corsetti and A. A. Mosto, Phys. Rev. B **84**, 035209 (2011).
[41] A. Bosak, J. Serrano, M. Krisch, K. Watanabe, T. Taniguchi, and H. Kanda, Phys. Rev. B **73**, 041402 (2006).
[42] J. Green, T. Bolland, and J. Bolland, J. Chem. Phys. **64**, 656 (1976).
[43] X. Gonze *et al.*, Comput. Phys. Commun. **180**, 2582 (2009).
[44] P. Giannozzi *et al.*, J. Phys. Condens. Matter **21**, 395502 (2009).