Regular Article

Density-functional energy gaps of solids demystified[*]

John P. Perdew[a] and Adrienn Ruzsinszky

Department of Physics, Temple University, Philadelphia, PA 19118, USA

Abstract. The fundamental energy gap of a solid is a ground-state second energy difference. Can one find the fundamental gap from the gap in the band structure of Kohn–Sham density functional theory? An argument of Williams and von Barth (WB), 1983, suggests that one can. In fact, self-consistent band-structure calculations within the local density approximation or the generalized gradient approximation (GGA) yield the fundamental gap *within the same approximation for the energy*. Such a calculation with the exact density functional would yield a band gap that also underestimates the fundamental gap, because the exact Kohn–Sham potential in a solid jumps up by an additive constant when one electron is added, and the WB argument does not take this effect into account. The WB argument has been extended recently to generalized Kohn–Sham theory, the simplest way to implement meta-GGAs and hybrid functionals self-consistently, with an exchange–correlation potential that is a non-multiplication operator. Since this operator is continuous, the band gap is again the fundamental gap within the same approximation, but, because the approximations are more realistic, so is the band gap. What approximations might be even more realistic?

1 Introduction

The fundamental energy gap G is arguably the most-important property of a solid, since it distinguishes insulators (non-zero gap) from metals (zero gap). It is a quasiparticle excitation energy, but it is also a ground-state second energy difference

$$G = I_N - A_N = [E_{N-1} - E_N] - [E_N - E_{N+1}], \quad (1)$$

the difference between the ionization energy or least energy cost to remove one electron from the neutral solid with N electrons and the electron affinity or least energy cost to remove one electron from the solid with one excess electron.

The Kohn–Sham [1] and generalized Kohn–Sham [2] density functional theories to be discussed here are only for ground-state properties of a system. The Runge–Gross time-dependent density functional theory [3] would be needed for excited-state properties [4].

2 Energy gaps in Kohn–Sham density functional theory

Kohn–Sham density functional theory [1] in principle reduces the calculation of the ground-state energy E_N and density $n(\boldsymbol{r})$ for interacting electrons in an external multiplicative potential $v(\boldsymbol{r})$ to the solution of self-consistent one-electron Schrödinger equations. The one-electron wavefunctions provide an exact treatment of the non-interacting kinetic energy, and the one-electron Schrödinger or Kohn–Sham equations minimize the total energy with respect to the density. The energy functional includes Hartree-like terms plus an exchange–correlation energy $E_{xc}[n]$ that is demonstrably a functional of the electron density, and whose functional derivative provides an exchange–correlation potential $v_{xc}(\boldsymbol{r})$. The exact exchange–correlation energy, which would make the theory exact, is not calculable in practice, and so must be approximated. Useful approximations of increasing accuracy define the rungs of a ladder: the local density approximation (LDA) [1], the generalized gradient approximation (GGA) [5], the meta-GGA [6,7], and the hybrid functionals [8–10] that mix a fraction of exact exchange with GGA or meta-GGA exchange.

The eigenvalues e of the one-electron Schrödinger equation define the band structure of a solid. The band gap

$$g = e_N^{LU} - e_N^{HO}, \quad (2)$$

is the difference between the lowest unoccupied (LU) and highest occupied (HO) one-electron energies. For non-interacting electrons, $I_N = -e_N^{HO}$ and $A_N = -e_N^{LU}$, so $G = g$. But does $G = g$ for interacting electrons?

Kohn–Sham density functional theory and its LDA were proposed in 1965 [1]. By the early 1970s, this theory was being applied to solids, with some success. It became clear that LDA band gaps were smaller, typically by a factor of two, than experimental gaps. But why?

In 1978, Janak proved a theorem [11] relating the ith one-electron energy e_i to the occupation number $f_i (0 < f_i < 1)$ and the energy E:

$$e_i = \frac{dE}{df_i}. \tag{3}$$

This theorem is true for finite or infinite systems, and for approximate or exact exchange–correlation energy functionals. Soon after, Williams and von Barth (WB) [12] combined Janak's theorem with the explicit assumption that the electron density changes only infinitesimally when an electron is added to or removed from an infinite periodic solid, and the implicit assumption that the exchange–correlation potential also changes only infinitesimally. From this, they concluded that $G = g$ within Kohn–Sham density functional theory.

The WB conclusion was correct [13] within simple explicit density functional approximations such as the LDA, or the GGA that started to appear around the time of their work. In other words, if the (somewhat similar) band gaps in those two approximations underestimate the experimental fundamental gap, that must be because the ground state second energy differences in those approximations yield similar underestimations. But work on the density functional theory for open systems with non-integer average electron number [14–16] showed that the exact Kohn–Sham potential of a finite system jumps up by a finite positive constant C when the electron number crosses above an integer. In an infinite solid, this jump can occur when a band fills, making

$$G = g + C. \tag{4}$$

The discontinuity of the exact Kohn–Sham potential would be disturbing if this potential were physical, but in fact the exact Kohn–Sham potential is a useful mathematical fiction. By definition, it yields the exact interacting ground-state electron density when acting on fictitious non-interacting electrons, and it also makes the highest occupied or partly-occupied one-electron energy equal to minus the minimum electron removal energy. As the filling of the valence band is approached from below, the exact Kohn–Sham potential approaches 0 far outside the surface of the solid and makes $e_N^{HO} = -I_N$ [17]. But then it will make $e_N^{LU} = -A_N - C$, and there is no reason to expect that $C = 0$. A single effective potential $v_{eff}(\boldsymbol{r})$ has enough flexibility to reproduce the electron density $n(\boldsymbol{r})$ and either but not both of the relevant band edges.

To the extent that LDA or GGA predict accurate electron densities for solids, they will predict the same band-structure gap as will the exact Kohn–Sham potential for the N-electron system, but this prediction will underestimate the experimental gap. Roughly speaking, LDA and GGA average over the discontinuity of the exact Kohn–Sham potential, making $e_N^{HO} \sim -I_N + C/2$ and $e_N^{LU} \sim -A_N - C/2$. What the LDA and GGA one-electron eigenvalues best predict is not the fundamental gap but the center of the fundamental gap relative to the vacuum level [13]: $-(I_N + A_N)/2 \sim (e_N^{HO} + e_N^{LU})/2$. In a metal this quantity becomes minus the work function, which is well reproduced by standard functionals, even for transition metals [18].

3 Energy gaps in generalized Kohn–Sham theory

Generalized Kohn–Sham (GKS) theory was derived in reference [2]. Bartlett [19] has analogized the spatially nonlocal and frequency-independent GKS exchange–correlation potential to the self-energy in coupled-cluster theory.

To understand generalized Kohn–Sham theory [2], consider first the exchange energy, which can be expressed exactly in terms of the one-electron wavefunctions and their occupation numbers, and more precisely in terms of the non-interacting one-particle density matrix $n(\boldsymbol{r}, \boldsymbol{r}')$ whose diagonal is the electron density: $n(\boldsymbol{r}, \boldsymbol{r}) = n(\boldsymbol{r})$. Hartree–Fock theory, with an effective exchange potential that is an integral operator, is recovered if we minimize the exact energy (neglecting correlation) with respect to all possible orthonormal orbitals and fermion occupations. Hartree–Fock theory can be regarded as a generalization of exact exchange-only Kohn–Sham theory (in which the effective exchange potential is constrained to be a multiplication operator $v_x(\boldsymbol{r})$). Talman and Shadwick [20] discovered how to construct this "optimized effective potential", also known as the exact exchange-only Kohn–Sham potential, and found for atoms that the total energy and electron density are very nearly equal in Hartree–Fock theory and exact-exchange-only Kohn–Sham theory. This result suggests that generalized Kohn–Sham theory (in which the exchange or exchange–correlation energy is constructed from the non-interacting one-particle density matrix, and the effective exchange or exchange–correlation potential is not constrained to be a multiplication operator) is only a small step outside Kohn–Sham theory. The only appreciable difference between these two theories lies in their one-electron energies. In particular, the experimental fundamental energy gaps of solids are strongly *overestimated* in the Hartree–Fock band structure, while they are *underestimated* in GGA. For rare-gas solids, they are underestimated again in the band structure of the exact exchange-only optimized effective potential [21], resembling the optical gaps or exciton energies more than the fundamental gaps [21].

The last sentences of the previous paragraph suggest that realistic band gaps in generalized Kohn–Sham theory could be found by mixing a fraction x of exact exchange with a fraction $(1-x)$ of GGA exchange and full GGA correlation [8,9]. For semiconductors, the optimum fraction

of exact exchange for band gaps is around 0.25 [10,22], about the same as is needed to improve the overestimated GGA atomization energies of molecules [9,10]. Perhaps this is not a coincidence, but reflects a similarity between semiconductors and molecules. While the fraction of exact exchange in such a hybrid functional is somewhat empirical, it is clear that no exact constraint satisfied by the GGA exchange energy is lost in the corresponding global hybrid. Range-separated hybrid functionals [10,22] include about 0.25 of exact exchange at intermediate range, with other fractions at long range; 0.00 at very long range is the most computationally-efficient choice for solids [10,22].

While the hybrid functionals are fully nonlocal functionals of the non-interacting one-particle density matrix, the more computationally-efficient meta-GGAs are semilocal functionals of the density matrix that employ the non-interacting kinetic energy density (derivatives of the density matrix, evaluated at $r' = r$). Like hybrids, meta-GGAs are most easily implemented in a generalized Kohn–Sham theory, but their effective exchange–correlation potentials are differential [23] and not integral operators. The SCAN meta-GGA [6,7] was constructed recently to satisfy all 17 known exact constraints on the exchange–correlation energy that a meta-GGA can respect, for all possible electron densities. (Compare 11 exact constraints for the Perdew–Burke–Ernzerhof GGA [5].) SCAN was also fitted to appropriate norms, non-bonded systems such as the uniform electron gas or atoms in which a meta-GGA can be accurate for exchange alone and for correlation alone, and not just for their sum as in bonded systems. (The understood but uncontrolled error cancellation between approximate exchange and approximate correlation is important in many applications of density functional theory, including the applications to energy gaps.) Without being fitted to any bonded system, SCAN correctly predicts diverse bonds in molecules and solids [7], being sometimes more and sometimes less accurate than the hybrid functionals.

Hybrid band gaps calculated in generalized Kohn–Sham theory are highly accurate for semiconductors, and even for some "strongly-correlated" systems, providing about 100% of the needed correction to GGA [22]. For semiconductors, SCAN provides only about 40% [24] of the needed correction to GGA band gaps. But, if an optimized effective potential or Kohn–Sham potential is constructed for the SCAN meta-GGA, the semiconductor band gaps go back to GGA-like values [24]. These statements are illustrated for the semiconductor silicon in Table 1.

The authors of this work were thinking about these results in the summer of 2015, when they visited Hardy Gross in Halle. They suspected that they had proved Janak's theorem (Eq. (3)) [11] within generalized Kohn–Sham theory, and that they had applied the WB argument [12] to prove that the fundamental gap G equals the band gap g within the same approximation in generalized Kohn–Sham implementations of hybrid functionals and meta-GGAs (for which the potential operators change infinitesimally when an electron is added to an infinite solid). But they were unsure if they had done this correctly, or if it had been done earlier by others. Hardy encouraged them to persevere, and the upshot was the

Table 1. The band gap g (in eV) of equation (2) for the semiconductor silicon, calculated self-consistently in various approximations, compared to the experimental fundamental gap G. The KS implementation of SCAN employs an optimized effective potential for SCAN (from Ref. [24]).

LDA (KS = GKS)	0.60
PBE GGA (KS = GKS)	0.71
SCAN meta-GGA (GKS)	0.97
SCAN meta-GGA (KS)	0.78
HSE hybrid (GKS)	1.11
Exper.	1.17

14-author reference [25], with all formal arguments presented in detail and with numerical demonstrations for finite linear chains of dimerized hydrogen atoms, extrapolated to infinite length, and for real periodic solids. The theorem applies not only to the band gap but also to its edges for a semi-infinite solid. It was also found that these conclusions for periodic solids follow directly from (but were not presented in) earlier and more general work by Yang et al. [26–29].

4 Energy gaps beyond global and range-separated hybrid functionals

The standard way to compute fundamental energy gaps within density functional theory is to find the band gap from a generalized Kohn–Sham implementation of a global or range-separated hybrid with about 1/4 of exact exchange [9,10,22]. Sometimes it is necessary to allow for symmetry breaking by introducing anti-ferromagnetism in an enlarged unit cell [30]. The rock-salt transition-metal monoxides MnO, FeO, CoO, and NiO find generalized Kohn–Sham band gaps [30] in this approach, and do so even above the temperature at which the spins disorder when spin correlations are taken into account in a very large supercell [31].

While these hybrid functionals predict accurate gaps in the semiconductor range, they overestimate the gaps of small-gap systems and underestimate those of large-gap systems. This effect has been explained by an exchange–correlation effect called screened exchange [32,33]: at long range, the exchange interaction in a solid is screened by or divided by the dielectric constant, which is roughly 4 for typical semiconductors but infinite for zero-gap metals and close to 1 for large-gap insulators. This idea leads to a method that works for gaps of any size in periodic solids [33], but is not general for all systems. It also explains why range-separated hybrids with full exact exchange at long-range are correct for isolated molecules in vacuum (with dielectric constant 1), but incorrect for typical solids.

For metals (zero gap), where the SCAN meta-GGA is often remarkably accurate, the hybrid functionals can be both difficult to converge and wrong. The valence bandwidth and exchange-splitting in metals are seriously overestimated by the hybrids [34]. The hybrids incorrectly favor ferromagnetic ground states for the metals Rh, Pd, and Pt [35]. The hybrids incorrectly predict a vanishing or

severely depressed density of states at the Fermi level of a metal, and overly-enhanced electron–phonon deformation potentials [36]. Hybrid functionals can also create spurious fundamental gaps in systems that are really metallic, as in the high-temperature superconducting material La_2CuO_4 doped with Sr, a system in which SCAN is remarkably accurate for the gap and other properties as functions of doping concentration [37].

A general-purpose method might be a local hybrid functional [38], in which the fraction of exact exchange energy density at a given point increases with the inhomogeneity of the electron density near that point. A complication is that the exact exchange energy density is non-unique, and must be taken in the same "gauge" [38–40] as the semi-local (e.g., SCAN meta-GGA) exchange energy density with which it is mixed. A local hybrid can also be range-separated.

Another possibility is suggested by the fact that improvements in the generalized Kohn–Sham band gap are obtained when the approximate functional becomes more nonlocal and satisfies more exact constraints. Thus realistic band gaps might well be found from a self-interaction correction [41,42] to SCAN. Note that SCAN is already self-correlation-free, but no meta-GGA can be exact for exchange for all one-electron densities.

Some exchange–correlation potential [43,44] or energy [45] functionals have been constructed to yield realistic band gaps of solids. Typically they achieve this by sacrificing the possibility or accuracy of other energy differences, but this sacrifice may not be necessary.

5 Conclusions

Standard approximations to the exchange–correlation energy can lead to band gaps g (Eq. (2)) in solids that equal the fundamental gaps G (Eq. (1)) within the same approximation for the total energy. This has long been known [13] for LDA and GGA implemented in the self-consistent Kohn–Sham (KS) scheme, which is equivalent to generalized Kohn–Sham (GKS) for these explicit density functionals. But it has also been shown [25] to be true for meta-GGAs and hybrids, when implemented self-consistently in GKS. However, when these functionals, or the exact functional, are implemented in KS via an optimized effective multiplicative potential, it is no longer true. Thus, within GKS, more realistic functionals typically lead to more realistic gaps, and there are several possibilities (Sect. 4) for still more realistic functionals, such as local hybrids and self-interaction-corrected meta-GGAs.

A particular implication is that the Hartree–Fock band gap of a solid is also the Hartree–Fock fundamental gap. But that is not a new result, because the Hartree–Fock exchange potential is known to be the exchange part of the Hartree–Fock quasiparticle self-energy operator. The fundamental gaps of many real solids are radically reduced by electron correlation.

For a numerical example of band gaps from standard functionals, consider the iconic semiconductor, silicon in the diamond structure. Table 1 shows band gaps for silicon

calculated in various ways [24], for comparison with the experimental fundamental gap. LDA and PBE [5] GGA underestimate, with only a slight improvement from LDA to GGA. The SCAN [6,7] meta-GGA improves the gap significantly in GKS, but most of the improvement is lost for SCAN implemented in KS. The HSE [10,22] hybrid, implemented in GKS, yields close agreement with experiment. But, as pointed out before, the local and range-separated hybrid functionals represent a solution for semiconductors, and not for metals or large-gap insulators. For the gaps and other properties of the cuprates [37] under doping, and perhaps more generally for trends, the SCAN meta-GGA is correct where these hybrids are not.

The hybrid functionals are also much more accurate than GGA for the interstitial formation energy [46] in silicon and for the critical pressure [46] of its phase transition to the metallic beta-tin structure. It has been suggested [46] that the improvement in band gap from GGA to hybrid is responsible for this difference, but that cannot be the whole story since SCAN is just as accurate as HSE for the interstitial formation energy [7] and critical pressure [7,47].

Author contribution statement

JPP wrote the first draft. AR and JPP revised the manuscript.

References

1. W. Kohn, L.J. Sham, Phys. Rev. **140**, A1133 (1965)
2. A. Seidl, A. Görling, P. Vogl, J.A. Majewski, M. Levy, Phys. Rev. B **53**, 3764 (1996)
3. E. Runge, E.K.U. Gross, Phys. Rev. Lett. **52**, 997 (1984)
4. M. Petersilka, U.J. Gossmann, E.K.U. Gross, Phys. Rev. Lett. **76**, 1212 (1996)
5. J.P. Perdew, K. Burke, M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996)
6. J. Sun, A. Ruzsinszky, J.P. Perdew, Phys. Rev. Lett. **115**, 036402 (2015)
7. J. Sun, R.C. Remsing, Y. Zhang, Z. Sun, A. Ruzsinszky, H. Peng, Z. Ying, A. Paul, U. Waghmare, X. Wu, M.L. Klein, J.P. Perdew, Nat. Chem. **8**, 831 (2016)
8. A.D. Becke, J. Chem. Phys. **98**, 5648 (1993)
9. J.P. Perdew, M. Ernzerhof, K. Burke, J. Chem. Phys. **105**, 9982 (1996)
10. J. Heyd, G.E. Scuseria, M. Ernzerhof, J. Chem. Phys. **118**, 8207 (2003)
11. J.F. Janak, Phys. Rev. B **18**, 7165 (1978)
12. A.R. Williams, U. von Barth, in *Theory of the inhomogeneous electron gas*, edited by S. Lundqvist, N.H. March (Plenum, NY, 1983)
13. J.P. Perdew, Int. J. Quant. Chem. **S19**, 497 (1985)
14. J.P. Perdew, R.G. Parr, M. Levy, J.L. Balduz, Phys. Rev. Lett. **49**, 1691 (1982)

15. J.P. Perdew, M. Levy, Phys. Rev. Lett. **51**, 1884 (1983)
16. L.J. Sham, M. Schlueter, Phys. Rev. Lett. **51**, 1888 (1983)
17. M. Levy, J.P. Perdew, V. Sahni, Phys. Rev. A **30**, 2745 (1984)
18. A. Patra, J.E. Bates, J. Sun, J.P. Perdew, Proc. Natl. Acad. Sci. U.S.A. **114**, E9188 (2017)
19. R.J. Bartlett, Chem. Phys. Lett. **484**, 1 (2009)
20. J.D. Talman, W.F. Shadwick, Phys. Rev. A **14**, 36 (1976)
21. R.J. Magyar, A. Fleszar, E.K.U. Gross, Phys. Rev. B **69**, 04511 (2004)
22. A.J. Garza, G.E. Scuseria, J. Phys. Chem. Lett. **7**, 4165 (2016)
23. R. Neumann, R. Nobes, N.C. Handy, Mol. Phys. **87**, 1 (1996)
24. Z. Yang, H. Peng, J. Sun, J.P. Perdew, Phys. Rev. B **93**, 205205 (2016)
25. J.P. Perdew, W. Yang, K. Burke, Z. Yang, E.K.U. Gross, M. Scheffler, G.E. Scuseria, T.M. Henderson, I.Y. Zhang, A. Ruzsinszky, H. Peng, J. Sun, E. Trushin, A. Görling, Proc. Natl. Acad. Sci. U.S.A. **114**, 2801 (2017)
26. A.J. Cohen, P. Mori-Sanchez, W. Yang, Phys. Rev. B **77**, 115123 (2008)
27. P. Mori-Sanchez, A.J. Cohen, W. Yang, Phys. Rev. Lett. **100**, 146401 (2008)
28. P. Mori-Sanchez, A.J. Cohen, W. Yang, Phys. Rev. Lett. **102**, 066403 (2009)
29. W. Yang, A.J. Cohen, P. Mori-Sanchez, J. Chem. Phys. **136**, 204111 (2012)
30. H. Peng, J.P. Perdew, Phys. Rev. B **96**, 100101(R) (2017)
31. G. Trimarchi, Z. Wang, A. Zunger, Phys. Rev. B **97**, 035107 (2018)
32. S. Refaely-Abramson, S. Sharifzadeh, M. Jain, R. Baer, J.B. Neaton, L. Kronik, Phys. Rev. B **88**, 081204 (2013)
33. J.H. Skone, M. Govoni, G. Galli, Phys. Rev. B **89**, 1 (2014)
34. J. Paier, M. Marsman, K. Hummer, G. Kresse, I.C. Gerber, J.G. Angyan, J. Chem. Phys. **124**, 154709 (2006)
35. F. Tran, D. Koller, P. Blaha, Phys. Rev. B **86**, 134406 (2012)
36. W. Gao, T.A. Abtew, T. Cai, Y.-Y. Sun, S. Zhang, P. Zhang, Solid State Commun. **234–235**, 10 (2016)
37. J.W. Furness, Y. Zhang, C. Lane, I.G. Buda, B. Barbiellini, R.S. Markiewicz, A. Bansil, J. Sun, Nat. Commun. Phys. **1**, 11 (2018)
38. J.P. Perdew, V.N. Staroverov, J. Tao, G.E. Scuseria, Phys. Rev. A **78**, 052513 (2008)
39. J.P. Perdew, A. Ruzsinszky, J. Sun, K. Burke, J. Chem. Phys. **140**, 18A533 (2014)
40. T.M. Maier, M. Haasler, A.V. Arbuznikov, M. Kaupp, Phys. Chem. Chem. Phys. **18**, 21133 (2016)
41. J.P. Perdew, A. Zunger, Phys. Rev. B **23**, 5048 (1981)
42. M.R. Pederson, A. Ruzsinszky, J.P. Perdew, J. Chem. Phys. **140**, 121103 (2014)
43. F. Tran, P. Blaha, Phys. Rev. Lett. **102**, 226401 (2009)
44. M. Kuisma, J. Ojanen, J. Enkovaara, T.T. Rantala, Phys. Rev. B **82**, 115106 (2010)
45. R. Armiento, S. Kümmel, Phys. Rev. Lett. **111**, 036402 (2013)
46. E.R. Batista, J. Heyd, R.G. Hennig, B.P. Uberuaga, R.L. Martin, G.E. Scuseria, J.W. Wilkins, Phys. Rev. B **74**, 121102 (2006)
47. C. Shahi, J. Sun, J.P. Perdew, Phys. Rev. B **97**, 094111 (2018)