Density-functional approximations for exchange and correlation

Viktor N. Staroverov
Department of Chemistry, The University of Western Ontario, London, Ontario N6A 5B7, Canada

July 28, 2012

Published in *A Matter of Density: Exploring the Electron Density Concept in the Chemical, Biological, and Materials Sciences*, edited by N. Sukumar (John Wiley & Sons, Hoboken, NJ, 2013), pp. 125-156

The challenge of density-functional theory

Density-functional theory (DFT) is based on two pivotal theorems due to Hohenberg and Kohn [1]. The first theorem states that the ground-state density $\rho(\mathbf{r})$ of a system of electrons uniquely determines the Hamiltonian and hence all properties that can be derived from it. Using mathematical language we can say that the total electronic energy of the system is a functional of the electron density,

$$E = E[\rho]. \tag{1}$$

The second Hohenberg–Kohn theorem demonstrates that the exact ground-state density and energy of the system can be found by minimizing the functional $E[\rho]$ over all admissible densities. The task of minimizing $E[\rho]$ amounts to solving the many-electron Schrödinger equation but, on the face of it, appears much simpler. Even a vague appreciation of the immense complexity of the Schrödinger equation makes one suspect that it cannot be tamed so easily, and that the almost miraculous solution of the electronic structure problem by DFT must come at a price. At least, there must be a catch.

There is actually not one catch, but two. First, the exact functional $E[\rho]$ is not known and, some believe, is so complicated that it is practically unknowable. The second catch is that $E[\rho]$, even if it *were* known, is not explicit, meaning that the exact mapping from ρ to E cannot in general be written down as a formula with E on the left-hand side and ρ on the right. Not all is lost, however, since we are free to approximate $E[\rho]$ with expressions involving standard functions and usual mathematical operations. Development of approximate density functionals that yield accurate electronic energies for the widest possible range of systems and properties is the chief preoccupation of DFT.

As of 2011, there are hundreds of density-functional approximations to choose from. Most of them perform remarkably well for certain types of problems and fail for others. For example, the B3LYP and PBE functionals are very good at predicting structural and thermodynamical properties, but not for charge transfer excitation energies, barriers of chemical reactions, polarizabilities, and noncovalent interactions. Sometimes, approximate functionals are designed to perform well for a particular property. However, this works like a see-saw: improvement for one target property often results in deterioration for others. The great proliferation of approximate density functionals and their uneven performance are in part responsible for certain skepticism toward DFT as a method.

In fairness to DFT, one should always keep in mind that practical computational chemistry never deals with the exact density functional, but only with density-functional *approximations*. If we knew the exact functional, then every DFT calculation would be exact. When we say "DFT fails", we mean that the density-functional approximation we chose to use fails to give the correct prediction. Such failures are not surprising and even should be expected, given how simple some approximations are. What *is* surprising is that compact closed-form density-to-energy expressions developed by theorists work as well as they do.

There are currently two views on the status of DFT. One view is that theorists have done everything they could, but the problem of approximating the exact functional is so hard that the hopes of making further progress may be fading. This sentiment is sometimes felt by users who have been growing impatient with the incremental progress of density functionals since the early 1990s when functionals such as B3LYP entered computational chemistry and completely transformed it. One should be reminded, however, that DFT has been in a similar position before: the significance of the Hohenberg–Kohn theorems was realized back in 1964, but it took two decades to understand the limitations of the early density-functional approximations and develop density functionals that were usefully accurate for chemical applications. Likewise, the limitations of present-day DFT only reflect the inadequacies of density-functional approximations that have been invented so far. Over the last decade, theorists have been busy trying to understand the reasons for successes and failures of currently available functionals, and made great strides in this regard. As a result, there is a basis for an optimistic view that DFT is ripe for a "paradigm shift" which will eventually lead to qualitatively better functionals.

The purpose of this chapter is to explain the inner workings of density-functional approximations and to give a sense of where DFT is heading in the near future. For a more technical account of some of the older topics discussed in this chapter, the reader is referred to Ref. 2. To keep things simple, we will write many equations in this chapter in the form applicable only to closed-shell ("spin-unpolarized") systems. For spin-polarized systems, where the spin-up and spin-down densities are not equal, some modifications may be required (see Section 8.2 in Ref. 3). This convention eliminates the need to include spin subscripts and sums over spins. Spin-specific quantities will be discussed only when necessary.

Exchange and correlation functionals

The starting point for approximating the electronic energy functional is to think of $E[\rho]$ as a sum of several terms. The idea is to identify those terms that are known exactly, define others in some convenient way, and then focus on the only unknown term that remains. This is precisely what Kohn and Sham [1] did by writing the total energy functional as

$$E[\rho] = T_s[\rho] + V[\rho] + U[\rho] + E_{\rm xc}[\rho]. \tag{2}$$

In this equation,

$$T_s[\rho] = -\frac{1}{2}\sum_k^{\rm occ.} \int \phi_k^*(\mathbf{r})\nabla^2\phi_k(\mathbf{r})\,d\mathbf{r} \tag{3}$$

is the kinetic energy of a hypothetical system of noninteracting electrons whose total ground-state density is exactly equal to $\rho(\mathbf{r})$, and $\phi_k(\mathbf{r})$ are the so-called Kohn–Sham orbitals occupied by these electrons, such that

$$\rho(\mathbf{r}) = \sum_k^{\rm occ.} |\phi_k(\mathbf{r})|^2. \tag{4}$$

The symbol $\sum_k^{\rm occ.}$ means that each term in the sum must be included as many times as there are electrons occupying the orbital ϕ_k (one, two, or zero). The functional

$$V[\rho] = \int \rho(\mathbf{r})v(\mathbf{r})\,d\mathbf{r} \tag{5}$$

is the electrostatic energy of the electron density interacting with the external potential $v(\mathbf{r})$, whereas

$$U[\rho] = \frac{1}{2}\int d\mathbf{r}_1 \int d\mathbf{r}_2\, \frac{\rho(\mathbf{r}_1)\rho(\mathbf{r}_2)}{|\mathbf{r}_1-\mathbf{r}_2|} \tag{6}$$

is the electrostatic energy of $\rho(\mathbf{r})$ interacting with itself. The last term, $E_{\rm xc}[\rho]$, incorporates everything else and is called the *exchange-correlation* energy. It is the only term that is unknown. In a way, the Kohn–Sham method packs all the complexity of the total energy functional into the exchange-correlation functional. But this is a clever reshuffle because approximating a part ($E_{\rm xc}$) is safer than directly approximating the whole (E).

The first step in tackling the exchange-correlation functional involves the same trick as for $E[\rho]$: we divide $E_{\rm xc}[\rho]$ into two parts, one large and one small, in such a way that the large part can be defined and computed exactly. These two parts are called, respectively, *exchange* and *correlation functionals*:

$$E_{\rm xc}[\rho] = E_{\rm x}[\rho] + E_{\rm c}[\rho]. \tag{7}$$

For closed-shell systems, where each Kohn–Sham orbital is doubly occupied, the exchange part is *defined* exactly by the expression

$$E_{\rm x}^{\rm exact}[\rho] = -\sum_{k,l=1}^{N/2} \int d\mathbf{r}_1 \int d\mathbf{r}_2\, \frac{\phi_k(\mathbf{r}_1)\phi_k^*(\mathbf{r}_2)\phi_l^*(\mathbf{r}_1)\phi_l(\mathbf{r}_2)}{|\mathbf{r}_1-\mathbf{r}_2|}. \tag{8}$$

This definition is borrowed from the closed-shell Hartree–Fock theory, where an equation identical to Eq. (8) represents the Hartree–Fock exchange energy (see Section 2.3.5 in Ref. 4). The functional $E_{\rm x}^{\rm exact}[\rho]$ is an *implicit* functional of the density: it depends on ρ through the Kohn–Sham orbitals which are related to ρ by Eq. (4).

In atoms and molecules near their equilibrium geometries, the correlation energy $E_{\rm c}$ is roughly an order of magnitude smaller than the exchange energy $E_{\rm x}$. We seem to be making progress: instead of approximating the total energy we now need to approximate only a relatively small part, $E_{\rm c}$. Yet anyone who has ever run DFT calculations knows that the exchange energy is usually approximated. This brings up the question: Why would one want to use an approximate functional for exchange when an exact formula is readily available? The short answer is that the pairing of $E_{\rm x}^{\rm exact}$ with standard correlation functionals gives poor accuracy in calculations of most properties of interest. In order to understand how this comes about and why theorists work so hard to approximate something that is already known, we need to invoke the concept of exchange and correlation holes.

As explained in Section 1.3.5 of Ref. 5, the exchange-correlation energy can be written *exactly* as

$$E_{\rm xc}[\rho] = \frac{1}{2}\int d\mathbf{r}_1\, \rho(\mathbf{r}_1) \int d\mathbf{r}_2\, \frac{\rho_{\rm xc}(\mathbf{r}_1,\mathbf{r}_2)}{|\mathbf{r}_1-\mathbf{r}_2|}, \tag{9}$$

where $\rho_{\rm xc}(\mathbf{r}_1,\mathbf{r}_2)$ is a function called the (coupling-constant-averaged) exchange-correlation hole density. Equation (9) is physically revealing: it suggests that we think of the exchange-correlation energy as Coulombic interaction between an electron at $\mathbf{r}_1$ and the surrounding exchange-correlation hole charge $\rho_{\rm xc}(\mathbf{r}_1,\mathbf{r}_2)$. Note that hole charge at $\mathbf{r}_2$ is not static but depends on the current position of the electron $\mathbf{r}_1$—as if the hole were riding along with the electron.

The exchange-correlation hole may be subdivided into exchange and correlation holes,

$$\rho_{\rm xc}(\mathbf{r}_1,\mathbf{r}_2) = \rho_{\rm x}(\mathbf{r}_1,\mathbf{r}_2) + \rho_{\rm c}(\mathbf{r}_1,\mathbf{r}_2), \tag{10}$$

so we can write for the exchange functional

$$E_{\rm x}[\rho] = \frac{1}{2}\int d\mathbf{r}_1\, \rho(\mathbf{r}_1) \int d\mathbf{r}_2\, \frac{\rho_{\rm x}(\mathbf{r}_1,\mathbf{r}_2)}{|\mathbf{r}_1-\mathbf{r}_2|}. \tag{11}$$

The analogous expression for the correlation functional is

$$E_{\rm c}[\rho] = \frac{1}{2}\int d\mathbf{r}_1\, \rho(\mathbf{r}_1) \int d\mathbf{r}_2\, \frac{\rho_{\rm c}(\mathbf{r}_1,\mathbf{r}_2)}{|\mathbf{r}_1-\mathbf{r}_2|}. \tag{12}$$

By comparing Eq. (11) with Eq. (8) we see that the exact exchange hole for a closed-shell system is

$$\rho_{\rm x}^{\rm exact}(\mathbf{r}_1,\mathbf{r}_2) = -\frac{2}{\rho(\mathbf{r}_1)}\sum_{k,l=1}^{N/2} \phi_k(\mathbf{r}_1)\phi_k^*(\mathbf{r}_2)\phi_l^*(\mathbf{r}_1)\phi_l(\mathbf{r}_2), \tag{13}$$

where ϕ_k are the occupied Kohn-Sham orbitals. The exact correlation hole is, of course, not known.

It turns out [6] that the exact exchange hole in a molecule is delocalized, meaning that for a given position of the reference electron $\mathbf{r}_1$, the plot of $\rho_\mathrm{x}(\mathbf{r}_1, \mathbf{r}_2)$ as a function of $\mathbf{r}_2$ has deep minima at other nuclei, no matter how remote. By contrast, the *total* exchange-correlation hole, $\rho_\mathrm{xc}(\mathbf{r}_1, \mathbf{r}_2)$, is typically localized around the reference electron. This implies that the exact correlation hole, $\rho_\mathrm{c}(\mathbf{r}_1, \mathbf{r}_2)$, must also be highly delocalized in order to cancel out the nonlocality of the exact exchange hole. Thus, if we want to combine the exact exchange functional with a density-functional approximation for correlation, we need to devise a very sophisticated, highly nonlocal functional.

For a long time, all attempts to marry the exact-exchange expression with an approximate correlation functional were defeated, although recently there has been some progress which we will discuss toward the end of the chapter. A simpler, pragmatic alternative is to abandon the exact exchange functional and use instead an approximation that is based on a localized hole and so is compatible with an approximate correlation functional. Of course, by giving up exact exchange in favor of approximations, one introduces an error into $E_\mathrm{x}[\rho]$. Fortunately, this error tends to be canceled out by a similar opposite-sign error in the approximation for $E_\mathrm{c}[\rho]$. This built-in cancellation of errors has proved to be a very fruitful idea and it was the principal reason for the tremendous success of exchange-correlation functionals developed in the 1980s and 1990s.

Ingredients and techniques for constructing density functional approximations

Development of density-functional approximations is a bold enterprise with relatively few strict guidelines. This means that one can be creative and try different routes. In fact, it is the absence of any mechanical prescriptions for systematic improvement of approximate functionals that makes DFT such an interesting subject.

The central objective of Kohn–Sham DFT is to come up with accurate approximations to the exact exchange-correlation functional. These approximations are usually cast in the form of integral expressions of the type

$$E_\mathrm{xc}[\rho] = \int e_\mathrm{xc}(\rho, \ldots)\, d\mathbf{r}, \tag{14}$$

where e_xc is some function of $\rho(\mathbf{r})$ and other density-dependent ingredients. Since the dimension of this quantity is $\frac{\text{energy}}{\text{volume}}$, e_xc is called the exchange-correlation energy density.

The most common ingredients of e_xc are: the modulus of the gradient of the density

$$g = |\nabla\rho|, \tag{15}$$

the Laplacian of the density

$$l = \nabla^2\rho, \tag{16}$$

the Kohn–Sham (non-interacting) kinetic energy density

$$\tau = \frac{1}{2}\sum_k^{\text{occ.}} |\nabla\phi_k|^2, \tag{17}$$

the (closed-shell) exact-exchange energy density

$$e_\mathrm{x}^\mathrm{exact}(\mathbf{r}_1) = -\sum_{k,l=1}^{N/2} \int \frac{\phi_k(\mathbf{r}_1)\phi_k^*(\mathbf{r}_2)\phi_l^*(\mathbf{r}_1)\phi_l(\mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|}\, d\mathbf{r}_2, \tag{18}$$

which is just the inner integral of Eq. (8), and the paramagnetic current density, defined in atomic units by

$$\mathbf{j} = \frac{1}{2i}\sum_k^{\text{occ.}} \left(\phi_k^*\nabla\phi_k - \phi_k\nabla\phi_k^*\right). \tag{19}$$

Observe that in the last equation, the expression in parentheses is purely imaginary, so that $\mathbf{j}$ itself is always real. Obviously, if the Kohn–Sham orbitals are real, the current density is zero.

Both g and l depend on ρ explicitly, whereas τ, $e_\mathrm{x}^\mathrm{exact}$, and $\mathbf{j}$ cannot be written entirely in terms of ρ, although they are uniquely determined by it. Accordingly, density-functional approximations of the type

$$E_\mathrm{xc}[\rho] = \int e_\mathrm{xc}(\rho, g, l)\, d\mathbf{r} \tag{20}$$

are called explicit, whereas functionals of the type

$$E_\mathrm{xc}[\rho] = \int e_\mathrm{xc}(\rho, g, \tau, e_\mathrm{x}^\mathrm{exact}, \ldots)\, d\mathbf{r} \tag{21}$$

are called implicit. Orbital-dependent functionals [7] are the most practically important type of implicit density functionals.

The ingredients g, l, τ, and $\mathbf{j}$ are called *semilocal* because they depend on the value of ρ or ϕ_k at $\mathbf{r}$ and/or in an infinitesimal neighborhood of $\mathbf{r}$. The exact-exchange energy density $e_\mathrm{x}^\mathrm{exact}$ is different in this respect because it depends on values of all ϕ_k everywhere, as reflected in the integration over $\mathbf{r}_2$. Such ingredients are said to be *nonlocal*. Semilocal density-functional approximations are those that involve one or more semilocal ingredients.

A significant portion of the vocabulary of modern DFT was developed by John Perdew in reference to a systematic approach called Jacob's ladder of density-functional approximations [8]. In this classification, density-functional approximations that are constructed using the electron density ρ and no other ingredients represent rung 1 of the ladder and are termed local density approximations (LDA),

$$E_\mathrm{xc}^\mathrm{LDA}[\rho] = \int e_\mathrm{xc}(\rho)\, d\mathbf{r}. \tag{22}$$

Approximations where $e_{\rm xc}$ depends on ρ and g represent rung 2 and are called generalized-gradient approximations (GGA),

$$E_{\rm xc}^{\rm GGA}[\rho] = \int e_{\rm xc}(\rho, g)\, d\mathbf{r}. \tag{23}$$

Rung 3 approximations depend, in addition to ρ and g, on l and/or τ, and are called meta-GGAs (MGGA),

$$E_{\rm xc}^{\rm MGGA}[\rho] = \int e_{\rm xc}(\rho, g, l, \tau)\, d\mathbf{r}. \tag{24}$$

The functionals of rung 4 involve dependence on a nonlocal ingredient, the exact-exchange energy density, and are termed hyper-GGAs (HGGA),

$$E_{\rm xc}^{\rm HGGA}[\rho] = \int e_{\rm xc}(\rho, g, l, \tau, e_{\rm x}^{\rm exact})\, d\mathbf{r}. \tag{25}$$

Approximations of rungs 1 through 4 involve only occupied Kohn–Sham orbitals. There is also a fifth rung where one finds approximations that involve occupied *and* virtual Kohn–Sham orbitals.

The historical development of density-functional approximations for exchange-correlation may be regarded as the process of climbing Jacob's ladder or as a story of passing the following milestones:

1. Analysis of exactly solvable models and introduction of various local density approximations.
2. Development of GGAs and meta-GGAs by bringing into play semilocal ingredients and by grafting selected properties of the exact functional.
3. Introduction of exact exchange into semilocal functionals (hybrid DFT).
4. Empirical construction (fitting).
5. Development of nonlocal correlation functionals compatible with exact exchange.

Most density functionals that are currently in use fall into groups 1 through 4, while functionals of group 5 are still at experimental stage. The rest of this chapter offers a close look at various strategies of devising density-functional approximations.

Nonempirical derivation and local density models

In an ideal world, we might be able to derive the exact exchange-correlation functional from first principles. In reality, we have to settle for less. One possible strategy is to obtain the exact functional for a solvable model system and hope that the same expression will work well in general. To illustrate this approach, let us consider a trivial example of one electron in an external potential $v(\mathbf{r})$. The Schrödinger equation for this system is identical with the Kohn–Sham equation,

$$\left[-\frac{1}{2}\nabla^2 + v(\mathbf{r})\right]\phi(\mathbf{r}) = E\phi(\mathbf{r}), \tag{26}$$

where $\phi(\mathbf{r})$ is the exact wavefunction and simultaneously the exact Kohn–Sham orbital. Suppose that $\phi(\mathbf{r})$ is normalized and real. (If ϕ is complex, it can always be made real as explained in Section 2.2 of Ref. 9.) Since there is only one electron in this system, the density is just $\rho = \phi^2$, so $\phi = \rho^{1/2}$. Let us multiply Eq. (26) from the left by $\phi(\mathbf{r})$, integrate over $\mathbf{r}$, and write the result as

$$E[\rho] = -\frac{1}{2}\int \rho^{1/2}(\mathbf{r})\nabla^2\rho^{1/2}(\mathbf{r})\, d\mathbf{r} + \int \rho(\mathbf{r})v(\mathbf{r})\, d\mathbf{r}. \tag{27}$$

This is clearly an explicit density functional, and it is exact for any one-electron system. One should not be surprised, however, that this functional gives dismal results for many-electron systems.

Although Eq. (27) is useless for practical purposes, it tells us something about the true functional. First, for any one-electron system with a constant external potential, the true $E[\rho]$ should reduce to Eq. (27). Second, the fact that Eq. (27) is exact for some systems but not for others suggests that the true $E[\rho]$ and hence $E_{\rm xc}[\rho]$ cannot be written as a single analytic expression valid for all electron numbers. When a second electron is added to the system, the true $E[\rho]$ must switch discontinuously from Eq. (27) to something else. Such sudden switching is not a property of analytic functionals.

Another model system which gives rise to a more useful nonempirical functional is a *uniform electron gas*, also called the jellium model. The uniform electron gas is a system of many interacting electrons moving in the field of a uniform positive background charge of the same density as the averaged electron density. The latter requirement ensures overall electric neutrality. The total volume of this system is assumed to be large but finite, so that Kohn–Sham orbitals can be normalized. For a uniform electron gas, $\rho(\mathbf{r}) = \text{const}$.

One can show (see, for instance, Section 6.1 in Ref. 3) that for a clot of spin-unpolarized uniform electron gas of volume V the exchange energy is given *exactly* by the expression

$$E_{\rm x}^{\rm LDA}[\rho] = -C_{\rm x}\int \rho^{4/3}(\mathbf{r})\, d\mathbf{r}, \tag{28}$$

where $C_{\rm x} = (3/4)(3/\pi)^{1/3} \approx 0.73856$ and the integration is over V. The exact correlation functional for a uniform electron gas is not known (except in the high- and low-density limits), but the correlation energy of this system has been studied numerically and parametrized in the form of analytic functionals such as [10]

$$E_{\rm c}^{\rm LDA}[\rho] = -A\int \rho(1+\alpha_1 r_s) \times \ln\left[1 + \frac{1}{A(\beta_1 r_s^{1/2} + \beta_2 r_s + \beta_3 r_s^{3/2} + \beta_4 r_s^2)}\right] d\mathbf{r}, \tag{29}$$

where $r_s = (3/4\pi\rho)^{1/3}$ and A, α_1, β_1, β_2, β_3, and β_4 are fixed parameters.

In real atoms and molecules, the electron density is far from uniform (it is approximately piecewise exponential), so Eqs. (28) and (29) are no longer exact. Despite this, the sum of Eqs. (28) and (29) gives a reasonably accurate approximation to the true exchange-correlation energy. The LDA predicts fairly accurate bond lengths and lattice constants, but severely overestimates atomization energies of molecules and solids. For comparison, the Hartree–Fock method, which is computationally *more* expensive than the LDA, predicts bond lengths much less accurately than LDA and underestimates atomization energies with a mean absolute error which is twice as large as the overbinding error of LDA. This is remarkable: a basic DFT method outperforms a basic wavefunction method. Good as LDA is, it is still not good enough for most chemical applications. As we shall see in the following section, attempts to derive exact density functionals for nonuniform densities by formal density-gradient expansions do not yield better general-purpose approximations. This compels one to seek different, less formulaic procedures for going beyond LDA.

One way to improve the LDA is to relax the requirement that this functional be exact for a uniform electron gas and instead demand better performance for chemically relevant systems. For the exchange component, this can be achieved by treating the constant C_x in Eq. (28) as an empirical parameter—the technique is known as Slater's $X\alpha$ method [11]. For correlation, one can start with some LDA expression and reparametrize it by fitting to the exact correlation energies of a few atoms. This strategy is represented by the Brual–Rothstein functional [12]. The gains in accuracy made in this manner, however, are modest.

A third method for deriving density functionals is to start with a model for the coupling-constant-averaged exchange-correlation hole, $\rho_\mathrm{xc}(\mathbf{r}_1, \mathbf{r}_2)$. Once the hole is specified, we insert it into Eq. (9) and integrate over $\mathbf{r}_2$ to obtain a density functional. For a density functional that is not explicitly derived from an exchange-correlation hole, one assumes that there is a model hole underlying it. The implied hole may be hard or even impossible to recover from a given $E_\mathrm{x}[\rho]$ or $E_\mathrm{c}[\rho]$, but it strongly influences the performance of the functional. Unfortunately, approximation of exchange-correlation hole densities is as difficult as direct approximation of functionals themselves, so this method does not by itself lead to more accurate results.

Semilocal functionals beyond the local density approximation

The most natural way to account for the nonuniformity of electron density in atoms and molecules is to construct an approximate functional in terms of ρ and its gradient $\nabla\rho$ or, rather, the gradient norm $|\nabla\rho|$. Because density-functional approximations must satisfy certain dimensionality requirements, it is convenient to make the energy density e_xc depend on $|\nabla\rho|$ through the so-called reduced density gradient,

$$s = \frac{|\nabla\rho|}{\rho^{4/3}}. \tag{30}$$

The reduced gradient s is a dimensionless quantity since the dimensions of ρ and $|\nabla\rho|$ are length^{-3} and length^{-4}, respectively. Now one can attempt to improve upon the LDA by devising a functional of the form

$$E_\mathrm{xc}[\rho] = \int e_\mathrm{xc}^\mathrm{LDA}(\rho)\,\left[1 + \mu(\rho)s^2 + \ldots\right] d\mathbf{r}, \tag{31}$$

where $\mu(\rho)$ is a function of the density which reduces to a constant for the exchange component. Approximations of this type are called density-gradient expansions. The coefficients of the lowest powers of s in Eq. (31) can be rigorously derived for two extreme cases: the slowly varying density limit and the high-density limit [5]. Since the leading gradient correction terms are nonempirical, one might assume that Eq. (31) cannot be worse than the LDA. But DFT often confounds expectations. It turns out that truncated density-gradient expansions are *less* accurate than the LDA for atoms and molecules. In particular, addition of the $\mu(\rho)s^2$ term to the LDA energy density makes total correlation energies positive [5], which is an unphysical result.

The failure of truncated density-gradient expansions for $E_\mathrm{xc}[\rho]$ was analyzed and explained by Perdew and coworkers [5]. They showed that the exchange-correlation hole underlying the second-order gradient expansion exhibits spurious undamped oscillations as $|\mathbf{r}_1 - \mathbf{r}_2| \to \infty$ and so violates two important conditions, namely, the negativity constraint for the exchange hole charge,

$$\rho_\mathrm{x}(\mathbf{r}_1, \mathbf{r}_2) < 0, \tag{32}$$

and the requirement that the exchange-correlation hole charge be normalized to -1 for every reference point $\mathbf{r}_1$,

$$\int \rho_\mathrm{xc}(\mathbf{r}_1, \mathbf{r}_2)\, d\mathbf{r}_2 = -1. \tag{33}$$

The incorrect behavior of the function $\rho_\mathrm{x}(\mathbf{r}_1, \mathbf{r}_2)$ associated with second-order truncated density-gradient expansions translates via Eq. (9) into large errors in energy for real atoms and molecules.

Another problem with truncated density-gradient expansions is that the corresponding exchange potential, $v_\mathrm{x}(\mathbf{r}) = \delta E_\mathrm{x}[\rho]/\delta\rho(\mathbf{r})$, has a pathological divergence in the exponential density tails found in all atomic and molecular charge distributions. This divergence is caused by the density-gradient correction term which is proportional to $\rho^{1/3}s^2$ and so diverges asymptotically for an exponential density. To see this, we substitute $\rho(r) = e^{-br}$ into Eq. (30) and obtain

$$s = \frac{|\nabla\rho|}{\rho^{4/3}} = \frac{|\partial\rho/\partial r|}{\rho^{4/3}} = \frac{be^{-br}}{e^{-4br/3}} = be^{br/3}. \tag{34}$$

This shows that $\rho^{1/3}s^2 \sim e^{br/3} \to \infty$ as $r \to \infty$.

In order to remedy the unphysical behavior of the exchange hole and exchange potential associated with density-gradient expansions, Perdew, Becke, and others proposed to replace the truncated series in square brackets in Eq. (31) with a damping function $F_{\rm xc}(\rho, s)$, such that it remains finite as $r \to \infty$. This leads to density-functional approximations of the form

$$E_{\rm xc}[\rho] = \int e_{\rm xc}^{\rm LDA}(\rho) F_{\rm xc}(\rho, s)\, d\mathbf{r}, \tag{35}$$

which are called generalized gradient approximations (GGAs). The analytic form of the function $F_{\rm xc}$ varies from case to case. For example, Becke's 1986 exchange functional [13] and the Perdew–Burke–Ernzerhof (PBE) GGA [14] employ damping functions of the form

$$F_{\rm x}^{\rm PBE}(s) = 1 + \frac{as^2}{1+bs^2}, \tag{36}$$

whereas Becke's 1988 exchange functional (B88) uses

$$F_{\rm x}^{\rm B88}(s) = 1 + \frac{as^2}{1+bs\ln(s+\sqrt{1+s^2})}, \tag{37}$$

In both cases, a and b are functional-specific constants that are either determined from known exact properties of $E_{\rm x}[\rho]$ or are fitted to experimental data. Generalized gradient approximations for the correlation energy have a more complicated form but also use damping functions to ensure that the correlation energy density has proper behavior in various physically relevant limits.

After generalized gradient approximations were perfected by the late 1980s, they were found to perform not only much better than the LDA, but also quite well relative to medium-level wavefunction methods. The latter fact is especially significant if we recall that GGAs have a much lower computational cost than wavefunction methods. As soon as all that came to light around 1991, many quantum chemists who had been previously skeptical about DFT finally became converts.

Constraint satisfaction

Although we do not know the exact exchange-correlation functional, we do know quite a few of its mathematical properties. Suppose we identify several such properties, adopt them as constraints, and then construct a density-functional approximation that satisfies those constraints. With respect to these mathematical properties, the resulting approximation will mimic the exact functional. We might also expect that the more properties our approximation shares with the exact functional, the more accurate and transferable it will be. This strategy of density-functional design, called constraint satisfaction [15], has produced some of the most successful density-functional approximations available today.

What properties of the exact functional are known? First of all, we know that for any admissible electron density, the exact exchange energy is strictly negative,

$$E_{\rm x}[\rho] < 0, \tag{38}$$

while the exact correlation energy is nonpositive,

$$E_{\rm c}[\rho] \leqslant 0. \tag{39}$$

The equality in Eq. (39) holds for all one-electron systems and only for such systems. Lieb and Oxford [16] showed that the exchange-correlation energy in Coulombic systems of electrons is also bounded from below:

$$E_{\rm x}[\rho] \geqslant E_{\rm xc}[\rho] \geqslant -C \int \rho^{4/3}(\mathbf{r})\, d\mathbf{r}, \tag{40}$$

where $C = 1.68$.

For any one-electron density $\rho_1(\mathbf{r})$, the exact $E_{\rm x}[\rho]$ cancels out the spurious Coulomb self-repulsion energy. This means that for any one-electron density ρ_1, the exact functionals should satisfy the relations

$$E_{\rm xc}[\rho_1] = E_{\rm x}[\rho_1] = -U[\rho_1], \tag{41}$$

where $U[\rho_1]$ is given by Eq. (6) with $\rho = \rho_1$. Notice that when this constraint applies, the Kohn–Sham functional of Eq. (2) correctly reduces to the exact one-electron density functional of Eq. (27).

For uniform electron densities, every exchange density-functional approximation should reduce to the known exact expression for a uniform electron gas,

$$E_{\rm x}[\rho] = E_{\rm x}^{\rm LDA}[\rho] \quad \text{if} \quad \rho(\mathbf{r}) = \text{const}, \tag{42}$$

where $E_{\rm x}^{\rm LDA}[\rho]$ is given by Eq. (28).

Mel Levy [17] deduced many properties of the exact exchange and correlation functionals under various coordinate scaling transformations of the density. The most important of these transformations is the *uniform* scaling of the density, defined by

$$\rho_\lambda(\mathbf{r}) = \lambda^3 \rho(\lambda \mathbf{r}), \tag{43}$$

where λ is a constant. The name "uniform" refers to the fact that all three Cartesian components of $\mathbf{r} = (x, y, z)$ are scaled by the same λ. As λ is varied, the density either contracts or becomes more diffuse, but the integral of $\rho_\lambda(\mathbf{r})$ over the entire space remains independent of λ:

$$\begin{aligned}
&\int \rho_\lambda(\mathbf{r})\, d\mathbf{r} = \int \lambda^3 \rho(\lambda \mathbf{r})\, d\mathbf{r} \\
&= \int dx \int dy \int dz\, \lambda^3 \rho(\lambda x, \lambda y, \lambda z) \\
&= \int d(\lambda x) \int d(\lambda y) \int d(\lambda z)\, \rho(\lambda x, \lambda y, \lambda z) \\
&= \int dx' \int dy' \int dz'\, \rho(x', y', z') = \int \rho(\mathbf{r}')\, d\mathbf{r}' = N.
\end{aligned} \tag{44}$$

The key property of the exact exchange functional is that it obeys the simple scaling law:

$$E_{\rm x}[\rho_\lambda] = \lambda E_{\rm x}[\rho]. \tag{45}$$

The exact correlation functional does not have a simple scaling behavior, but it is known that

$$\lim_{\lambda\to\infty} E_{\rm c}[\rho_\lambda] > -\infty. \tag{46}$$

It is also known that in a finite many-electron system, the true exchange-correlation potential $v_{\rm xc}(\mathbf{r})$, defined as the functional derivative of $E_{\rm xc}[\rho]$ with respect to ρ, has the following asymptotic behavior:

$$v_{\rm xc}(\mathbf{r}) \equiv \frac{\delta E_{\rm xc}[\rho]}{\delta\rho(\mathbf{r})} \xrightarrow[r\to\infty]{} -\frac{1}{r}. \tag{47}$$

The asymptotic behavior of the exchange-correlation energy density is as follows:

$$e_{\rm xc}(\mathbf{r}) \xrightarrow[r\to\infty]{} -\frac{\rho(r)}{2r}. \tag{48}$$

The list can be continued, but the message is clear: (i) density-functional approximations should reproduce known properties of the exact exchange-correlation functional; (ii) any approximation that violates a known exact constraint should be suspect. To illustrate the method of constraint satisfaction, we will explain how it was used to eliminate one embarrassing artifact of early density-functional approximations.

The hydrogen atom is one of the few systems of chemical interest for which the Schödinger equation can be solved analytically. The exact ground-state density of the H atom is $\rho(\mathbf{r}) = \frac{1}{\pi}e^{-2r}$ and the corresponding exact total energy is $E = -\frac{1}{2}$ hartree. The LDA and most GGAs fail to give these results because these functionals incorrectly predict nonzero correlation energies for one-electron systems, in violation of the constraint

$$E_{\rm c}[\rho_1] = 0, \tag{49}$$

where ρ_1 is a one-electron density. For the same reason, LDA and GGA give nonzero correlation energies for other one-electron systems such as H_2^+. One notable exception is the Lee-Yang–Parr (LYP) correlation GGA in which the correlation energy density is proportional to the product of spin-up and spin-down densities, $\rho_\alpha\rho_\beta$. As a result, LYP predicts $E_{\rm c} = 0$ for *any* N-electron system where all electrons have parallel spins. That is, LYP happens to be correct for $N = 1$, but is wrong for $N \geqslant 2$.

To satisfy the constraint of Eq. (49), Becke devised an indicator function which distinguishes one-electron densities from all others. This function is based on certain properties of the kinetic energy density $\tau(\mathbf{r})$ and its interplay with other density-functional ingredients. To understand Becke's reasoning, we consider the quantity

$$\tau_W = \frac{1}{8}\frac{|\nabla\rho|^2}{\rho}, \tag{50}$$

called the Weizsäcker gradient correction to the Thomas–Fermi kinetic energy density. The property of τ_W that we need is the following double inequality

$$0 \leqslant \tau_W \leqslant \tau - \frac{1}{2}\frac{|\mathbf{j}|^2}{\rho}, \tag{51}$$

where $\mathbf{j}$ is the current density defined by Eq. (19). The first part of this inequality, $\tau_W \geqslant 0$, is obvious from the definition of τ_W. Proof of the second part of Eq. (51) requires some work.

Let us consider first closed-shell systems. For such systems, the gradient of the density is given by

$$\nabla\rho = \nabla\left(2\sum_{k=1}^{N/2}\phi_k^*\phi_k\right) = 2\sum_{k=1}^{N/2}\left(\phi_k^*\nabla\phi_k + \phi_k\nabla\phi_k^*\right). \tag{52}$$

Here $2\sum_{k=1}^{N/2}\phi_k^*\nabla\phi_k$ is a complex-valued vector quantity which we can rewrite as

$$\begin{aligned} 2\sum_{k=1}^{N/2}\phi_k^*\nabla\phi_k &= \sum_{k=1}^{N/2}(\phi_k^*\nabla\phi_k + \phi_k\nabla\phi_k^*) \\ &\quad + \sum_{k=1}^{N/2}(\phi_k^*\nabla\phi_k - \phi_k\nabla\phi_k^*) \\ &= \frac{1}{2}\nabla\rho + i\mathbf{j}, \end{aligned} \tag{53}$$

where we used Eqs. (52) and (19) (the latter without the factor of 1/2 because we are summing over $N/2$ orbitals). Since $\frac{1}{2}\nabla\rho$ and $\mathbf{j}$ are always real, we can think of them, respectively, as the real and imaginary parts of $2\sum_{k=1}^{N/2}\phi_k^*\nabla\phi_k$. Since for $z = x + iy$ we have $|z|^2 = [\mathrm{Re}(z)]^2 + [\mathrm{Im}(z)]^2$, we can write

$$\begin{aligned} \left|2\sum_{k=1}^{N/2}\phi_k^*\nabla\phi_k\right|^2 &= 4\left|\sum_{k=1}^{N/2}\phi_k^*\nabla\phi_k\right|^2 \\ &= \frac{1}{4}|\nabla\rho|^2 + |\mathbf{j}|^2 = 2\rho\tau_W + |\mathbf{j}|^2. \end{aligned} \tag{54}$$

But according to the Cauchy–Schwarz inequality,

$$4\left|\sum_{k=1}^{N/2}\phi_k^*\nabla\phi_k\right|^2 \leqslant 4\left(\sum_{k=1}^{N/2}|\phi_k|^2\right)\left(\sum_{k=1}^{N/2}|\nabla\phi_k|^2\right) = 2\rho\tau. \tag{55}$$

Comparing Eqs. (54) and (55) we see that $2\rho\tau_W + |\mathbf{j}|^2 \leqslant 2\rho\tau$ or, equivalently,

$$\tau_W \leqslant \tau - \frac{1}{2}\frac{|\mathbf{j}|^2}{\rho}. \tag{56}$$

This concludes the proof of Eq. (51). Note that for real orbitals, where $\mathbf{j}$ is identically zero, Eq. (51) reduces to

$$0 \leqslant \tau_W \leqslant \tau. \tag{57}$$

The next step is an important observation that the equality in Eq. (55) holds only if the number of occupied Kohn–Sham orbitals is one. In that case,

$$\tau = \tau_W + \frac{1}{2}\frac{|\mathbf{j}|^2}{\rho} \tag{58}$$

or simply $\tau = \tau_W$ if the orbital is real.

For spin-polarized system (when $\rho_\alpha \neq \rho_\beta$), Eq. (51) branches into two separate inequalities, one for each spin:

$$0 \leqslant \frac{|\nabla\rho_\sigma|^2}{8\rho_\sigma} \leqslant \tau_\sigma - \frac{1}{2}\frac{|\mathbf{j}_\sigma|^2}{\rho_\sigma}, \tag{59}$$

where $\sigma = \alpha$ or β. The quantities ρ_σ, τ_σ, and $\mathbf{j}_\sigma$ are given by equations similar to Eqs. (4), (17), and (19) in which only singly-occupied σ-spin orbitals are included. Again, if only one σ-spin orbital is occupied, the second inequality in Eq. (59) becomes a strict equality,

$$\tau_\sigma = \frac{1}{8}\frac{|\nabla\rho_\sigma|^2}{\rho_\sigma} + \frac{1}{2}\frac{|\mathbf{j}_\sigma|^2}{\rho_\sigma}. \tag{60}$$

Following Becke [18, 19], we now introduce the function

$$\eta_\sigma = \frac{1}{\tau_\sigma}\left(\tau_\sigma - \frac{1}{8}\frac{|\nabla\rho_\sigma|^2}{\rho_\sigma} - \frac{1}{2}\frac{|\mathbf{j}_\sigma|^2}{\rho_\sigma}\right). \tag{61}$$

As explained above, $\eta_\sigma(\mathbf{r})$ vanishes identically for any one-electron system and is strictly positive in systems which contain two or more σ-spin electrons. Consider now the meta-GGA correlation functional

$$E_{\rm c}[\rho] = \int \left[e_{\rm c}^{\alpha\beta}(\mathbf{r}) + \sum_\sigma e_{\rm c}^{\sigma\sigma}(\mathbf{r})\eta_\sigma(\mathbf{r})\right] d\mathbf{r}, \tag{62}$$

where $e_{\rm c}^{\alpha\beta}$ and $e_{\rm c}^{\sigma\sigma}$ are some GGA-type expressions for the opposite-spin and parallel-spin correlation energy densities, respectively. Because of the presence of $\eta_\sigma(\mathbf{r})$ in Eq. (62), every functional of this form will correctly yield zero for the $\sigma\sigma$-spin correlation energy in any system with a single σ-spin electron, and a nonzero energy in any system with two or more σ-spin electrons. This is now a standard trick for constructing correlation functionals that are free from the one-electron self-interaction error. Density-functional approximations that use it include Bc88 [20], Bc95 [21], B98 [22], τ-HCTH [23], TPSS [24], VS98 [25], M06 [26], and others.

Although constraint satisfaction is currently the most rigorous practical method of constructing density-functional approximations, it has its limitations. Enforcement of any particular constraint does not by itself guarantee that the resulting functional will be better. This is because by imposing one known constraint we may unwittingly violate other—unknown—constraints which may be more important. In fact, better performance is sometimes achieved when an exact constraint is relaxed. For example, any GGA can and should reduce to the LDA functional of Eq. (28) when $\rho(\mathbf{r}) = \text{const}$ because LDA is the proper functional for a uniform density. Some of the most successful density functionals in chemistry sacrifice this property in favor of better performance for non-uniform densities. In particular, the LYP correlation functional in not exact for a uniform electron gas, yet predicts highly accurate correlation energies for atoms. BLYP, B3LYP, and other exchange-correlation functionals that include LYP also fail to yield correct energies for a uniform electron gas, but this has little effect on their performance in chemical applications.

Another example of beneficial and even intentional constraint violation involves GGA functionals. For a slowly varying density (that is, for $s \to 0$), any exchange GGA should reproduce the known low-order terms in the exact density-gradient expansion:

$$E_{\rm x}[\rho] = -C_{\rm x}\int \rho^{4/3}(1 + \mu s^2 + \ldots)\, d\mathbf{r}, \tag{63}$$

where the theoretical value of μ is $10/81 \approx 0.1235$. When this constraint is enforced, GGAs predict accurate bond lengths in molecules and lattice constants in solids, but give poor atomization energies. Perdew and coworkers [27] showed that a GGA can produce accurate atomization energies only if it strongly violates Eq. (63) and has an enhanced gradient dependence. The PBE GGA in particular was designed to give accurate atomization energies, and so it has $\mu = 0.2195$, which is about twice as large as required by the density-gradient expansion. All attempts to construct an accurate GGA face the dilemma [28]: using the theoretical value of μ leads to accurate bond lengths but yields poor atomization energies; increasing μ improves atomization energies but worsens bond lengths and lattice constants. Being a very restrictive form, GGAs cannot simultaneously perform well for both properties. Thus, for calculations of atomization energies, one should use the PBE GGA or its hybrid versions with $\mu = 0.2195$. For bulk properties of solids, one should use a modified version called PBEsol (PBE revised for solids) which restores the nonempirical value $\mu = 10/81$.

The comeback of exact exchange: Global and local hybrids

A decade ago, Peter Gill [29] published an "obituary" for DFT in the *Australian Journal of Chemistry*. According to his account, DFT was born in 1927 and passed away in 1993. The cause of her demise was an unsuccessful operation performed on her by "an eminent Canadian surgeon", a follower of Dr. Frankenstein, who attempted to cure DFT by blending her with wavefunction theory into a "grisly hybrid". It would be instructive for us here to understand what prompted the famous surgeon to recommend such a drastic treatment.

As we discussed earlier, semilocal correlation functionals do not work well in combination with the exact exchange functional of Eq. (8), but good performance is eas-

ily achieved if both exchange and correlation approximations are semilocal. In 1993, however, Becke showed [30] that one can go beyond the accuracy of GGAs by representing the exchange contribution with a *mixture* of the exact exchange functional and a semilocal approximation. This discovery led to many so-called *hybrid* functionals such as B3PW91, B3LYP, and PBEh.

The basic form of hybrid functionals is

$$E_{\rm xc} = aE_{\rm x}^{\rm exact} + (1-a)E_{\rm x} + E_{\rm c}, \tag{64}$$

where $E_{\rm x}$ and $E_{\rm c}$ are some semilocal density-functional approximations and a ($0 \leqslant a \leqslant 1$) is a universal parameter called a mixing fraction. The value of a is usually determined by empirical fitting of Eq. (64) to reproduce experimental atomization energies, exact nonrelativistic energies, reaction barrier heights, and other data. Fitting to atomization energies typically gives $a \approx 0.2$ for GGAs and $a \approx 0.1$ for meta-GGAs, while fitting to reaction barrier heights yields $a \approx 0.5$.

Mixing exact and approximate exchange functionals is not an empirical cookbook recipe. The hybrid scheme has a theoretical underpinning which not only explains why hybrid functionals work better than GGAs but also predicts the optimal value of a in various situations [31].

From the point of view of computational chemists, hybrid functionals were a smashing success because they represented the first quantum-mechanical method that was simultaneously accurate, reliable, and computationally cheap. Ironically, it is the "grisly hybrids" that made DFT so effective and popular.

The term "hybrid" in relation to functionals such as B3LYP is now often used with the qualifier *global* to indicate that the value of a in Eq. (64) is position-independent. This can be emphasized by rewriting Eq. (64) in terms of energy densities:

$$E_{\rm xc} = \int \left[a e_{\rm x}^{\rm exact}(\mathbf{r}) + (1-a)e_{\rm x}(\mathbf{r}) + e_{\rm c}(\mathbf{r}) \right] \, d\mathbf{r}. \tag{65}$$

The fact that the optimal value of a has large system-dependent variations suggests a generalization of Eq. (65) by turning the mixing fraction a into a function of $\mathbf{r}$,

$$E_{\rm xc} = \int \left\{ a(\mathbf{r}) e_{\rm x}^{\rm exact}(\mathbf{r}) + [1-a(\mathbf{r})]e_{\rm x}(\mathbf{r}) + e_{\rm c}(\mathbf{r}) \right\} \, d\mathbf{r}, \tag{66}$$

Such forms are called *local hybrids*. In the local hybrid scheme, the objective is to devise a mixing fraction $a(\mathbf{r})$ that adapts to the local chemical environment. The basic requirements for the mixing fraction $a(\mathbf{r})$ are that it be restricted to the range of values between 0 and 1, and reduce to 1 for any one-electron density.

The first mixing fraction was suggested by Becke [19],

$$a(\mathbf{r}) = \frac{\tau_W(\mathbf{r})}{\tau(\mathbf{r})}, \tag{67}$$

and implemented in a local hybrid functional by Jaramillo *et al.* [32]. This choice gives accurate reaction barriers but produces disappointing results for atomization energies [32]. More recently, Kaupp and coworkers [33] constructed and implemented self-consistently several local hybrid functionals with various mixing fractions. One of those is given by

$$a(\mathbf{r}) = \sum_{m=1}^{M} b_m \left[\frac{\tau_W(\mathbf{r})}{\tau(\mathbf{r})} \right]^m, \tag{68}$$

where M is a small integer and b_m are fractional coefficients. Another is

$$a(\mathbf{r}) = \left[\frac{s(\mathbf{r})}{b + s(\mathbf{r})} \right]^2, \tag{69}$$

where s is the reduced gradient of Eq. (30) and b is a positive parameter. It was found that a local hybrid functional using the mixing fraction of Eq. (68) with $M = 1$ and $b_1 \approx 0.5$ predicts simultaneously accurate atomization energies and reaction barriers.

In view of the resounding success of global hybrid functionals, the local hybrid scheme was initially thought to hold great promise. However, finding a mixing fraction $a(\mathbf{r})$ which would decisively beat global hybrid functionals proved more difficult than anticipated. As a result, the overall accuracy of the best local hybrid functionals proposed to date is not significantly higher than that of the global hybrid scheme with an optimal mixing constant. Attempts to develop better local-hybrid approximations continue despite these setbacks.

The best of both worlds: Range-separated hybrids

Interaction of opposite-spin electrons at close range (small $r_{12} \equiv |\mathbf{r}_1 - \mathbf{r}_2|$) is adequately described by semilocal exchange-correlation approximations, but not by the exact (Hartree–Fock-type) exchange functional. In fact, the Hartree–Fock method does not correlate the motion of electrons with opposite spins at all. That is why molecular properties for which short-range electron interactions are dominant (e.g., equilibrium geometries and atomization energies) are predicted by approximate DFT much better than by the Hartree–Fock method. Conversely, when two electrons are far apart (large r_{12}), their interaction is better described with the exact exchange functional than with semilocal density-functional approximations. Consequently, properties determined by long-range interactions (e.g., electronic Rydberg excitations, polarizabilities, charge transfer processes) require a large fraction of exact exchange (50% or more). The physical insight arising from these observations suggests a hybrid scheme in which short-range interactions are treated by density-functional approximations while long-range interactions are described by the exact exchange. This is precisely the idea of so-called *range-separated* or *screened* hybrid functionals, and it proved to be one of the DFT's biggest successes of the past decade.

In the range-separated hybrid scheme, the electron-electron Coulomb repulsion operator is partitioned into a short-range (SR) and a long-range (LR) component

$$\frac{1}{r_{12}} = \underbrace{\frac{1-f(r_{12})}{r_{12}}}_{\text{SR}} + \underbrace{\frac{f(r_{12})}{r_{12}}}_{\text{LR}}, \tag{70}$$

where $f(r_{12})$ is a "screening function" that satisfies the following requirements: (a) $0 \leqslant f \leqslant 1$; (b) $f \to 0$ when $r_{12} \to 0$; and (c) $f \to 1$ when $r_{12} \to \infty$. The short-range component of a given exchange functional can be obtained by replacing the Coulomb operator $1/r_{12}$ in Eq. (11) with its short-range part to give

$$E_{\text{x}}^{\text{SR}}[\rho] = \frac{1}{2}\int d\mathbf{r}_1\, \rho(\mathbf{r}_1) \int d\mathbf{r}_2\, \frac{1-f(r_{12})}{r_{12}} \rho_{\text{x}}(\mathbf{r}_1, \mathbf{r}_2), \tag{71}$$

where $\rho_{\text{x}}(\mathbf{r}_1, \mathbf{r}_2)$ is the exchange hole density corresponding to the functional. Similarly, the long-range exchange component of a functional may be defined by

$$E_{\text{x}}^{\text{LR}}[\rho] = \frac{1}{2}\int d\mathbf{r}_1\, \rho(\mathbf{r}_1) \int d\mathbf{r}_2\, \frac{f(r_{12})}{r_{12}} \rho_{\text{x}}(\mathbf{r}_1, \mathbf{r}_2). \tag{72}$$

For instance, the long-range part of the exact exchange functional [whose exchange hole is given by Eq. (13)] is

$$E_{\text{x}}^{\text{exact,LR}}[\rho] = -\sum_{k,l=1}^{N/2} \int d\mathbf{r}_1 \int d\mathbf{r}_2 \times \phi_k(\mathbf{r}_1)\phi_k^*(\mathbf{r}_2)\frac{f(r_{12})}{r_{12}}\phi_l^*(\mathbf{r}_1)\phi_l(\mathbf{r}_2). \tag{73}$$

The two popular choices for the screening function are the exponential function

$$f(r_{12}) = 1 - e^{-\omega r_{12}}, \tag{74}$$

where ω is a positive constant, and the Gauss error function,

$$f(r_{12}) = \text{erf}(\omega r_{12}) = \frac{2}{\sqrt{\pi}} \int_0^{\omega r_{12}} e^{-t^2}\, dt, \tag{75}$$

where ω is also a positive parameter. The error function is convenient in calculations employing Gaussian-type basis sets because all necessary two-electron integrals in this case can be evaluated efficiently.

To separate a functional into a long-range and a short-range parts by Eqs. (71) and (72), one needs the associated exchange hole. Aside from the exact exchange functional, exchange holes are known for only a handful of density-functional approximations such as LDA, Becke–Roussel [34], PBE, and TPSS. (In the case of LDA, the short- and long-range parts can be derived in closed form [35, 36]; in the cases of PBE and TPSS, exchange holes were reverse-engineered from the corresponding functionals.) To circumvent this restriction, Hirao and coworkers [37, 38] proposed a different definition of the screened components which does not require the exchange hole and so is applicable to any GGA.

Screened hybrid functionals that combine the long-range part of exact exchange with the short-range part of a semilocal density-functional approximation have been proposed by several researchers [37–39]. In particular, Vydrov and Scuseria [39] combined the short-range PBE exchange with the long-range exact exchange into a long-range-corrected PBE hybrid functional called LC-ωPBE. This functional is given by

$$E_{\text{xc}}^{\text{LC-}\omega\text{PBE}}(\omega) = E_{\text{x}}^{\text{exact,LR}}(\omega) + E_{\text{x}}^{\text{PBE,SR}}(\omega) + E_{\text{c}}^{\text{PBE}}, \tag{76}$$

where the recommended value of the screening parameter is $\omega = 0.40$ bohr^{-1}. Long-range-corrected functionals such as LC-ωPBE have excellent performance for a wider range of properties than other types of density-functional approximations.

A different way of combining short- and long-range parts of exchange functionals has found use in condensed-matter physics. It had long been suspected that certain properties of solids should be better described with hybrid functionals than with semilocal approximations. Unfortunately, the exact exchange energy is difficult to evaluate accurately for metallic and weakly insulating solids using conventional techniques. This is due to the unphysically slow spatial decay of exact-exchange interactions in systems with vanishing band gaps, which itself is a consequence of the essentially nonlocal character of the exact-exchange energy density. To make hybrid DFT calculations on solids possible, Heyd, Scuseria, and Ernzerhof (HSE) [40] proposed to replace the long-range portion of exact exchange in a global hybrid functional with a long-range part of a semilocal density functional. This is equivalent to taking a semilocal functional and hybridizing the short-range part of exchange. If the starting functional is the PBE GGA, this construction yields the HSE functional,

$$E_{\text{xc}}^{\text{HSE}}(\omega) = aE_{\text{x}}^{\text{exact,SR}}(\omega) + (1-a)E_{\text{x}}^{\text{PBE,SR}}(\omega) + E_{\text{x}}^{\text{PBE,LR}}(\omega) + E_{\text{c}}^{\text{PBE}}, \tag{77}$$

where the parameter ω ($0 \leqslant \omega < \infty$) is adjusted to achieve the best possible accuracy for the problem of interest. Observe that smaller values of ω cause the mixing to be switched on at shorter interelectron distances. The HSE functional can be viewed as an interpolation between pure PBE and the global hybrid PBE functional (PBEh): When $a = 0.25$ and $\omega = 0$, HSE reduces to PBEh, while in the limit $\omega \to \infty$ it reduces to PBE. For solids, computational cost of HSE is much closer to that of PBE than of PBEh. The main practical advantage of the HSE hybrid is that it predicts much more accurate lattice constants and band gaps than any standard semilocal functional including LDA, PBE, and TPSS [41].

Empirical fits

So far we have discussed methods of density-functional design that avoid empiricism as much as possible. New density-functional approximations were obtained either by rigorous derivations for exactly solvable models or by devising phenomenological mathematical expressions that were consistent with known properties of the exact functional. At some point in this process it was necessary to introduce one or more parameters whose values were *a priori* unknown. These values were found by fitting computed properties to high-quality experimental data. It is because of this step that DFT is sometimes regarded as a semiempirical method.

Since there is no hope of deriving the exact exchange-correlation functional, while the method of constraint satisfaction is arduous and slow, it is hard to resist the pragmatism of fully empirical constructions. In the empirical approach, one starts by postulating a flexible analytic representation for the energy density and then tunes it by minimizing discrepancies between theoretical predictions and experimental observations. For example, on the basis of analysis of the density matrix expansion, Van Voorhis and Scuseria (VS98) [25] proposed parametrizing the exchange functional in the following form

$$E_{\mathrm{x}}^{\mathrm{VS98}}[\rho] = \int \rho^{4/3}\left[\frac{b_0}{h(s,z)} + \frac{b_1 s^2 + b_2 z}{h^2(s,z)} + \frac{b_3 s^4 + b_4 s^2 z + b_5 z^2}{h^3(s,z)}\right] d\mathbf{r}, \qquad (78)$$

where s is defined by Eq. (30), $z = \tau\rho^{-5/3} - C_F$ with $C_F = \frac{3}{5}(3\pi^2)^{2/3}$, and $h(s,z) = 1 + c(s^2 + z)$, whereas b_0, b_1, b_2, b_3, b_4, b_5, and c are adjustable parameters.

Optimization of empirical functionals can be carried out on several levels. On the first level, one optimizes linear and nonlinear parameters appearing in the expression for the energy density. On the second level, one writes the density-functional approximation in the form

$$E_{\mathrm{xc}}[\rho] = \int \sum_m a_m e_{\mathrm{xc}}^{(m)}(\rho, s, \ldots)\, d\mathbf{r}, \qquad (79)$$

where $e_{\mathrm{xc}}^{(m)}$ are various representations of the exchange-correlation energy density and a_m are adjustable empirical coefficients. All global and local hybrid functionals belong to this type. If desired, one may proceed even further to the third level, called "external optimization" [42], and consider a linear combination of several "model chemistries",

$$E_{\mathrm{xc}}[\rho] = \sum_n d_n E_{\mathrm{xc}}^{(n)}[\rho], \qquad (80)$$

where the quantities $E_{\mathrm{xc}}^{(n)}[\rho]$ represent results of fully self-consistent Kohn–Sham calculations using different functionals; d_n are their weights fitted to a set of experimental data. Naturally, functionals that are optimized on two or three levels achieve a higher accuracy than functionals optimized on one level only.

The parameter optimization is usually accomplished by minimizing the root mean square (RMS) deviation of predictions from experiment,

$$\mathrm{RMS} = \sqrt{\frac{\sum_s \sum_p (x_{sp}^{\mathrm{calc}} - x_{sp}^{\mathrm{exp}})^2}{N_x}}, \qquad (81)$$

where x_{sp}^{calc} and x_{sp}^{exp} are, respectively, the calculated and experimental values of property p in system s, and N_x is total number of such data.

Most of the existing empirical density functionals are based on the analytic representations of exchange and correlation energy densities proposed, respectively, by Van Voorhis and Scuseria [25] and by Becke [43, 44]. These functionals include VS98 [25], Becke's 1997 exchange-correlation approximation (B97) [43], the 1998 hybrid GGA [45] and hybrid meta-GGA [22] of Schmider and Becke, the GGA of Hamprecht, Cohen, Tozer, and Handy [46] (HCTH) and its various reparametrizations. The most sophisticated empirical exchange-correlation functionals existing today are those of the Minnesota 2006 (M06) suite developed by Zhao and Truhlar [26, 47].

The M06 suite consists of four functionals: M06, M06-2X, M06-L, and M06-HF. All four have the same analytic form combining the functional forms of LDA, PBE, VS98, and B97, but differ by the values of more than 40 independent empirical parameters. The parameters are adjusted for optimal performance in four different types of chemical problems. M06 is a hybrid meta-GGA with 27% of exact exchange; it is designed to provide a consistently good accuracy for transition metals, main-group thermochemistry, medium-range correlation energy, and barrier heights. M06-2X has twice as much exact exchange as M06 ($a = 0.54$; other parameters are re-optimized) and is trained to give the best possible performance for main-group compounds, valence and Rydberg electronic excitation energies, and noncovalent interactions. The M06-2X parametrization, however, is not good for transition metals. M06-L (where L stands for local) is a reparametrization of M06 with no exact exchange, dropped to enable application of the functional to very large and periodic systems. M06-L is the most accurate for transition metal compounds, but not very accurate for reaction barrier heights which require a large fraction of exact exchange. Finally, M06-HF includes 100% of exact exchange to achieve good performance for charge transfer excited states. Although M06 functionals contain many empirical parameters, they also respect several important exact constraints including the uniform electron gas limit [Eq. (42)] and are free from the one-electron self-interaction error [Eq. (41)].

Development of empirical density functionals requires large databases of accurate experimental data. Early empirical functionals were trained on relatively small test sets of atomization energies. By contrast, functionals of the M06 suite rely on a truly massive set of data which

includes dozens of atomization energies, ionization potentials, electron and proton affinities; bond dissociation energies, isomerization energies, a variety of reactions barriers; hydrogen-bonded systems; charge-transfer, dipole-interaction, and π-π stacking complexes; valence and Rydberg vertical excitation energies; thermochemistry of transition metal reactions. The high flexibility combined with the unprecedented diversity of the training set enable the M06 functionals to predict chemical and physical properties with a reliability matching that of some high-level wavefunction methods.

Correlation functionals compatible with exact exchange

Perhaps the most sophisticated density functionals constructed to date are Becke's nondynamical correlation functional of 2005 (B05) [48] and the 2008 hyper-GGA of Perdew, Staroverov, Tao, and Scuseria (PSTS) [49]. Both functionals use the exact-exchange energy density of Eq. (18) as an ingredient in the *correlation* part. This makes the correlation functional compatible with the exact exchange functional. The complexity of the B05 and PSTS functionals reflects not only the difficulty of the problem, but also our growing understanding of the interplay between exchange and correlation.

The starting point for the B05 model is analysis of two types of electron correlation, called dynamical and nondynamical (static). Dynamical correlation is due to close-range Coulombic interactions and so is essentially local in character. In systems where most of the correlation energy is dynamical, the exact exchange and correlation holes are both localized around the reference electron. Such systems include atoms, molecules near their equilibrium geometry, and the uniform electron gas. Semilocal density-functional approximations (LDA, GGA, meta-GGA) work well for such systems precisely because the LDA, GGA, meta-GGA exchange and correlation holes are themselves localized. Nondynamical correlation arises in many-electron systems consisting of two of more fragments whose Coulombic interaction is weak or negligible. Each such fragment is effectively an independent system, so the exact exchange-correlation hole for electrons of any one fragment is contained entirely within that fragment. The exact exchange hole in such systems is split between all fragments, which means that the exact correlation hole must be delocalized as well. Semilocal density-functional approximation cannot recognize this delocalization and so they do not work well for systems with strong nondynamical correlation. A possible way to detect and account for nondynamical correlation is by using the real-space structure of the exact exchange hole as a diagnostic tool.

In an isolated hydrogen atom, the exact exchange hole around the reference electron is contained entirely within the vicinity of the nucleus and integrates to -1. In a highly stretched H_2 molecule, the exact exchange hole is divided between the two atoms. As a result, the effective normalization of the exact-exchange hole around each H atom in stretched H_2 is only $-\frac{1}{2}$. According to Becke [50], an effective normalization of $-\frac{1}{2}$ means that the reference electron excludes less than one opposite-spin electron from its immediate vicinity, which raises the energy of each H atom in the stretched H_2 molecule. Therefore, Becke argued, the effective hole in each half of stretched H_2 needs to be deepened to repel electrons of opposite spin. The deepening of the effective exchange hole amounts to introducing *nondynamical correlation* and is modeled as follows:

$$\rho^{\alpha}_{\rm xc}(\mathbf{r}_1,\mathbf{r}_2) = \rho^{\alpha}_{\rm x}(\mathbf{r}_1,\mathbf{r}_2) + f_{\rm c}(\mathbf{r}_1)\rho^{\beta}_{\rm x}(\mathbf{r}_1,\mathbf{r}_2), \tag{82}$$

$$\rho^{\beta}_{\rm xc}(\mathbf{r}_1,\mathbf{r}_2) = \rho^{\beta}_{\rm x}(\mathbf{r}_1,\mathbf{r}_2) + f_{\rm c}(\mathbf{r}_1)\rho^{\alpha}_{\rm x}(\mathbf{r}_1,\mathbf{r}_2). \tag{83}$$

Here $\rho^{\alpha}_{\rm x}$ and $\rho^{\beta}_{\rm x}$ are effective holes seen, respectively, by spin-up and spin-down electrons, while $f_{\rm c}$ is a position-dependent correlation parameter determined by two physical constraints: (i) $0 \leqslant f_{\rm c} \leqslant 1$; (ii) an exchange-correlation hole cannot contain more than one electron. The explicit form of this parameter proposed by Becke is

$$f_{\rm c}(\mathbf{r}) = \min\left[\frac{1-N^{\alpha}_{\rm x}(\mathbf{r})}{N^{\beta}_{\rm x}(\mathbf{r})}, \frac{1-N^{\beta}_{\rm x}(\mathbf{r})}{N^{\alpha}_{\rm x}(\mathbf{r})}, 1\right], \tag{84}$$

where $N^{\alpha}_{\rm x}(\mathbf{r})$ and $N^{\beta}_{\rm x}(\mathbf{r})$ are position-dependent integrals of the exchange hole charge over the atomic region when the reference electron is at $\mathbf{r}$. The exchange-correlation energy is obtained by substituting the above expressions for the exchange-correlation holes into Eq. (9) to give

$$E_{\rm xc}[\rho] = \frac{1}{2}\sum_{\sigma=\alpha,\beta}\int d\mathbf{r}_1\,\rho_{\sigma}(\mathbf{r}_1)\int d\mathbf{r}_2\,\frac{\rho^{\sigma}_{\rm xc}(\mathbf{r}_1,\mathbf{r}_2)}{r_{12}}. \tag{85}$$

Since the exact exchange hole of Eq. (13) cannot be integrated efficiently, the values of $N^{\alpha}_{\rm x}$ and $N^{\beta}_{\rm x}$ in the B05 model are found using the approximate Becke–Roussel model exchange hole [34] instead of the exact $\rho^{\sigma}_{\rm x}(\mathbf{r}_1,\mathbf{r}_2)$. Even with this simplification, one still needs to solve numerically a complicated nonlinear equation for each point $\mathbf{r}$. Another obstacle is that the piecewise definition of $f_{\rm c}(\mathbf{r})$ by Eq. (84) causes this function to have a discontinuous derivative which complicates self-consistent implementation of the B05 model. These difficulties were surmounted by Arbuznikov and Kaupp [51] and by Proynov and coworkers [52] who implemented the B05 functional in a fully self-consistent Kohn–Sham scheme, with minor modifications of Becke's original definitions. Although these researchers have so far reported only preliminary results, the numbers are encouraging: B05 does perform significantly better than B3LYP for difficult reaction barriers and gives excellent bond lengths [52].

The PSTS hyper-GGA also employs the exact-exchange energy density to model nondynamical correlation. PSTS is essentially a local hybrid with a very complicated mixing fraction designed to interpolate between two extreme types of density regions for which the

proper amount of exact exchange is known. The first type of density regions are called "normal". These are the regions where the exact exchange-correlation hole is spatially localized around an electron and integrates to -1 over a narrow range. As we saw above, this is the situation where semiempirical density-functional approximations for exchange and correlation work very well because of mutual error cancellation. Therefore, in normal regions the local fraction of exact exchange, $a(\mathbf{r})$, is designed to be small. "Abnormal" regions are those where the exact exchange-correlation hole is highly nonlocal and integrates to a value greater than -1 over the region. For example, all multicenter one-electron densities and regions with a fractional electron charge are abnormal in this sense. In abnormal regions, the mixing fraction $a(\mathbf{r})$ should be close to 1. For all intermediate situations, the mixing fraction adjusts the amount of exact exchange to some appropriate value between 0 and 1.

Construction of the PSTS mixing fraction is largely phenomenological and is guided by exact constraints. Nevertheless, the complexity of the problem requires a few empirical parameters. These parameters were determined by fitting to 97 molecular standard enthalpies of formation and 42 reaction barrier heights. Initial assessment of the PSTS functional showed that it performs much better than conventional global hybrids for reaction barrier heights, although there was no accuracy gain for atomization energies. As with the B05 functional, self-consistent implementation of the PSTS hyper-GGA is nontrivial and requires further simplifications [53].

Current trends and outlook for the future

One of the most fascinating topics in DFT that has come to prominence recently is the performance of density functionals for systems with fractional electron numbers and fractional spins [54–59]. It has been even argued [56] that all failures of present-day DFT can be understood by analyzing the errors of existing exchange-correlation approximations in such systems.

The story starts in 1982 when Perdew, Parr, Levy, and Balduz [60] published a seminal paper in which they analyzed behavior of the exact density functional in systems with a fractional number of electrons. How can a system have a "fractional number of electrons"? As far as real atoms and molecules are concerned, electrons are of course indivisible, so the total number of electrons in a real chemical system is always an integer. What is meant by a system with a fractional electron number is a linear combination ("ensemble") of wavefunctions representing systems with different integer electron numbers.

Consider an example. When the internuclear distance in an H_2^+ molecule is stretched to infinity, the electron is physically localized either on one nucleus or on the other. Let the wavefunctions representing these two states, $H\cdots H^+$ and $H^+\cdots H$, be ϕ_L and ϕ_R, respectively. The wavefunctions ϕ_L and ϕ_R are degenerate ground-state eigenfunctions of the Hamiltonian. By the fundamental quantum-mechanical principle of linear superposition, any normalized linear combination of these wavefunctions, $c_L\phi_L + c_R\phi_R$, where $|c_L|^2 + |c_R|^2 = 1$, is also a valid ground-state solution of the Schrödinger equation. This includes a half-and-half combination with $|c_L|^2 = |c_R|^2 = \frac{1}{2}$ in which each of the atoms has only half an electron. In this sense, any fractional electron number $q = |c_L|^2$ is possible on the left atom, giving rise to the supermolecule $H^{1-q}\cdots H^{+q}$. We say that the region around each proton in stretched H_2 is a system with a fractional electron number (or a fractional charge). In practice, fractional charges are found not only at infinite nuclear separation but even in a moderately stretched H_2^+ molecule where the internuclear distance is a little greater than at equilibrium. Similarly, a molecular ion of the general formula A_2^+ can be viewed as a supersystem composed of two many-electron systems with fractional electron numbers.

Suppose now that we have a system with $N = J + q$ electrons, where J is a positive integer and $0 \leqslant q \leqslant 1$. Within Kohn–Sham DFT, the electron density of this system is constructed in accordance with the *Aufbau* principle, that is, by filling each of the J lowest-energy Kohn–Sham spin-orbitals with one electron and placing the fraction q of an electron in the highest-occupied molecular (Kohn–Sham) orbital (HOMO). The reason for using the *Aufbau* rule is because our system belongs to a supersystem that is supposed to be in the ground state. Thus, we write

$$\rho(\mathbf{r}) = \sum_{k=1}^{J} |\phi_k(\mathbf{r})|^2 + q|\phi_{\mathrm{HOMO}}(\mathbf{r})|^2. \qquad (86)$$

Perdew and coworkers [60] showed that, in general, the exact ground-state energy of a $(J+q)$-electron system is a linear combination of the ground-state energies of the J- and $(J+1)$-electron systems:

$$E(J+q) = (1-q)E(J) + qE(J+1), \qquad 0 \leqslant q \leqslant 1 \quad (87)$$

This means that the plot of the exact E as a function of q between J and $J+1$ is a straight line.

It turns out that if we calculate the electronic energy using any existing density-functional approximation and then plot $E(J+q)$ as a function of q, the result will *not* be a straight line. Approximate density functionals are close to target at the end points J and $J+1$, but fail to reproduce the straight line in between: the actual plot is a curve that is usually bent downward. This means that in systems such as H_2^+, application of the variational principle to approximate density functionals yields the maximally delocalized density ($H^{+1/2}\cdots H^{+1/2}$) whose energy is much lower than it should be. This artificial lowering of the energy in systems with fluctuating electron number is known as the charge delocalization error.

Similar analysis of spin-up and spin-down degeneracies in electrically neutral open-shell systems leads to the

concept of fractional spin [56]. Consider an isolated hydrogen atom. In the absence of an external magnetic field, the ground state of this system is doubly degenerate: the spin-up eigenstate $\psi_\uparrow$ has the same energy as the spin-down eigenstate $\psi_\downarrow$, that is, $E_\uparrow = E_\downarrow$. Since the Hamiltonian is spin-independent, any normalized linear combination of these eigenfunctions, $c_\uparrow\psi_\uparrow + c_\downarrow\psi_\downarrow$, with $|c_\uparrow|^2 + |c_\downarrow|^2 = 1$ is also an eigenfunction of the Hamiltonian. Assuming that this linear combination is normalized, we can interpret it as a wavefunction of an H atom with a fraction $\gamma = |c_\uparrow|^2$ of α-spin and a fraction $|c_\downarrow|^2 = 1 - \gamma$ of β-spin. The exact energy of an isolated H atom is independent of γ, so we should have

$$E(\gamma) = \gamma E_\uparrow + (1-\gamma)E_\downarrow = \text{const}, \qquad 0 \leqslant \gamma \leqslant 1. \quad (88)$$

Therefore, the plot of $E(\gamma)$ for an H atom should be a horizontal line segment. Weitao Yang and coworkers found that, instead of a horizontal line, all approximate density functionals predict a curve that is bent upward [58]. They also found that the maximum deviation from linearity, which occurs at the midpoint, coincides with the magnitude of the nondynamical correlation error in an infinitely stretched H_2 molecule. This discovery revealed an intimate connection between the fractional spin error and nondynamical correlation.

The practical implications of the charge and spin delocalization errors are enormous. Binding energy curves for dissociating neutral molecules predicted with approximate DFT have massive positive errors at large internuclear distances. This occurs because neutral molecules dissociate into fractional-spin fragments for which approximate density functionals predict too high energies. In calculations of reaction barriers, theoretical energies of reactants are fairly accurate, but the energies of transition states are too low because transition states often consist of weakly interacting fractionally charged fragments for which approximate functionals predict too low energies. As a result, reaction barriers are severely underestimated. At the same time, molecular polarizability (a measure of the responsiveness of the electron density to an applied electric field) predicted by approximate density functionals is too high because fractional charges are artificially driven toward the edges of the molecule. In short, LDA, GGA, and meta-GGA fail to predict accurately many molecular properties because these approximations violate the important exact constraints of Eqs. (87) and (88).

The second fundamental result that follows from the analysis of Perdew, Parr, Levy, and Balduz [60] is that the slope of the exact function $E(N)$, where N is a continuous electron number, changes discontinuously when N passes through an integer value. The significance of this fact will come to light once we reveal the physical meaning of the slope of $E(N)$.

Suppose first that N approaches the nearest integer J from above, that is, $N = J + q$, where q is a fractional electron number ($0 \leqslant q \leqslant 1$). From Eq. (87) we obtain

$$\frac{dE}{dN} = \frac{dE}{dq} = E(J+1) - E(J), \qquad N = J + q \quad (89)$$

The quantity $E(J+1) - E(J)$ is the negative ionization potential of the $(J+1)$-electron system or, equivalently, the negative electron affinity of the J-electron system. Now let N approach J from below, that is, let us take $N = J - q$, where $q \geqslant 0$. We rewrite Eq. (87) as

$$E(J-q) = qE(J-1) + (1-q)E(J), \qquad 0 \leqslant q \leqslant 1 \quad (90)$$

Differentiation of this equation with respect to N yields

$$\frac{dE}{dN} = -\frac{dE}{dq} = E(J) - E(J-1), \qquad N = J - q \quad (91)$$

This is the negative ionization potential of the J-electron system or, equivalently, the negative electron affinity of the $(J-1)$-electron system. It is instructive to rewrite these relations in terms of one-sided limits:

$$\left.\frac{dE}{dN}\right|_{N\to J^+} = \lim_{\delta\to 0}\left.\frac{dE}{dN}\right|_{J+\delta} = -I_{J+1} = -A_J, \quad (92)$$

$$\left.\frac{dE}{dN}\right|_{N\to J^-} = \lim_{\delta\to 0}\left.\frac{dE}{dN}\right|_{J-\delta} = -I_J = -A_{J-1}, \quad (93)$$

where I_J and A_J are, respectively, the ionization potential and electron affinity of the J-electron system. The last two equations mean that the exact derivative dE/dN jumps by a constant when the number of electrons passes through an integer J. This constant is equal to

$$\left.\frac{dE}{dN}\right|_{N\to J^+} - \left.\frac{dE}{dN}\right|_{N\to J^-} = I_J - A_J. \quad (94)$$

Let us summarize. The exact ground-state energy of an N-electron system (i.e., a system with a continuous electron number N), plotted as a function of N, is a linkage of straight-line segments. The function $E(N)$ is itself continuous, but its first derivative, dE/dN, is discontinuous at all integer values of N. When N approaches an integer J from below, dE/dN is the exact negative ionization potential of the J-electron system. When N approaches an integer J from above, dE/dN is the exact negative electron affinity of the J-electron system.

Equations (87), (88), and (94) represent fundamental properties of the exact density functional. All semilocal density-functional approximations tend to violate these equations in many ways. The function $E(N)$ in approximate DFT no longer consists of straight line segments, but is a linkage of curves. Discontinuities of dE/dN are observed only when a fraction of electron is added to a new orbital shell or subshell, whereas at other integer values of J, the curve $E(N)$ is smooth. Since the slopes of $E(N)$ are incorrect in approximate DFT, many physical properties including total energies, ionization potentials, electron affinities, band gaps, polarizabilities are predicted with large errors.

Long-range-corrected hybrid density functionals and functionals that combine exact exchange with compatible nonlocal correlation violate the exact constraints of Eqs. (87), (88), and (94) to a lesser extent than the older (semilocal) approximations. For this reason, the newer functionals exhibit significantly better performance for a wider range of molecular properties than LDAs, GGAs, and meta-GGAs. Nevertheless, there is currently no approximate density functional that is entirely free from the fractional-charge and fractional-spin errors, or which has correct derivative discontinuities at every integer electron number. Finding a way to construct functionals that respect the constraints of Eqs. (87), (88), and (94) would be a crucial step in overcoming the limitations of present-day DFT. If such a method is found, it will take DFT to the next level of predictive capability.

This brings us to an optimistic conclusion. If the history of DFT teaches us anything, it is that breakthroughs in density functional development are usually preceded by years of scrutiny and introspection. This is why successful functionals tend to arrive in waves. The waves that have come ashore so far are LDAs, GGAs, global hybrids, and range-separated hybrids. The latest advances in our understanding of the limitations of existing density-functional approximations open exciting new opportunities for theorists and give us reasons to hope that the future of density-functional theory is secure.

[1] Kohn, W. (1999) Nobel lecture: Electronic structure of matter—wave functions and density functionals. *Rev. Mod. Phys.*, **71**, 1253.

[2] Scuseria, G. E. and Staroverov, V. N. (2005) Progress in the development of exchange-correlation functionals. In *Theory and Applications of Computational Chemistry: The First Forty Years*; Dykstra, C. E., Frenking, G., Kim, K. S., and Scuseria, G. E. (eds.), Elsevier, Amsterdam, pp. 669–724.

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

[4] Szabo, A. and Ostlund, N. S. (1982) *Modern Quantum Chemistry*. Macmillan, New York.

[5] Perdew, J. P. and Kurth, S. (2003) Density functionals for non-relativistic Coulomb systems in the new century. In *A Primer in Density Functional Theory*; Fiolhais, C., Nogueira, F., and Marques, M. (eds.), Springer, Berlin, pp. 1–55.

[6] Baerends, E. J. and Gritsenko, O. V. (1997) A quantum chemical view of density functional theory. *J. Phys. Chem. A*, **101**, 5383.

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

[8] Perdew, J. P. and Schmidt, K. (2001) Jacob's ladder of density functional approximations for the exchange-correlation energy. In *Density Functional Theory and Its Application to Materials*; Van Doren, V., Van Alsenoy, C., and Geerlings, P. (eds.), AIP, Melville, NY, pp. 1–20.

[9] Dewar, M. J. S. (1969) *The Molecular Orbital Theory of Organic Chemistry*. McGraw-Hill, New York.

[10] Perdew, J. P. and Wang, Y. (1992) Accurate and simple analytic representation of the electron-gas correlation energy. *Phys. Rev. B*, **45**, 13244.

[11] Slater, J. C. (1972) Statistical exchange-correlation in the self-consistent field. *Adv. Quantum Chem.*, **6**, 1.

[12] Brual, G. and Rothstein, S. M. (1978) Rare gas interactions using an improved statistical method. *J. Chem. Phys.*, **69**, 1177.

[13] Becke, A. D. (1986) Density functional calculations of molecular bond energies. *J. Chem. Phys.*, **84**, 4524.

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

[15] Perdew, J. P., Ruzsinszky, A., Tao, J., Staroverov, V. N., Scuseria, G. E., and Csonka, G. I. (2005) Prescription for the design and selection of density functional approximations: More constraint satisfaction with fewer fits. *J. Chem. Phys.*, **123**, 062201.

[16] Lieb, E. H. and Oxford, S. (1981) Improved lower bound on the indirect Coulomb energy. *Int. J. Quantum Chem.*, **19**, 427.

[17] Levy, M. (1995) Coordinate scaling requirements for approximating exchange and correlation. In *Density Functional Theory*; Gross, E. K. U. and Dreizler, R. M. (eds.), Plenum, New York, pp. 11–31.

[18] Becke, A. D. (1996) Current-density dependent exchange-correlation functionals. *Can. J. Chem.*, **74**, 995.

[19] Becke, A. D. (1998) A new inhomogeneity parameter in density-functional theory. *J. Chem. Phys.*, **109**, 2092.

[20] Becke, A. D. (1988) Correlation energy of an inhomogeneous electron gas: A coordinate-space model. *J. Chem. Phys.*, **88**, 1053.

[21] Becke, A. D. (1996) Density-functional thermochemistry. IV. A new dynamical correlation functional and implications for exact-exchange mixing. *J. Chem. Phys.*, **104**, 1040.

[22] Schmider, H. L. and Becke, A. D. (1998) Density functional from the extended G2 test set: Second-order gradient functionals. *J. Chem. Phys.*, **109**, 8188.

[23] Boese, A. D. and Handy, N. C. (2002) New exchange-correlation density functionals: The role of the kinetic-energy density. *J. Chem. Phys.*, **116**, 9559.

[24] Tao, J., Perdew, J. P., Staroverov, V. N., and Scuseria, G. E. (2003) Climbing the density functional ladder: Nonempirical meta-generalized gradient approximation designed for molecules and solids. *Phys. Rev. Lett.*, **91**, 146401.

[25] Van Voorhis, T. and Scuseria, G. E. (1998) A novel form for the exchange-correlation energy functional. *J. Chem. Phys.*, **109**, 400.

[26] Zhao, Y. and Truhlar, D. G. (2008) The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12

other functionals. *Theor. Chem. Acc.*, **120**, 215.
[27] Perdew, J. P., Constantin, L. A., Sagvolden, E., and Burke, K. (2006) Relevance of the slowly varying electron gas to atoms, molecules, and solids. *Phys. Rev. Lett.*, **97**, 223002.
[28] Perdew, J. P., Ruzsinszky, A., Csonka, G. I., Vydrov, O. A., Scuseria, G. E., Constantin, L. A., Zhou, X., and Burke, K. (2008) Restoring the density-gradient expansion for exchange in solids and surfaces. *Phys. Rev. Lett.*, **100**, 136406.
[29] Gill, P. M. W. (2001) Obituary: Density functional theory (1927–1993). *Aust. J. Chem.*, **54**, 661.
[30] Becke, A. D. (1993) Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.*, **98**, 5648.
[31] Burke, K., Perdew, J. P., and Ernzerhof, M. (1998) Mixing exact exchange with GGA: When to say when. In *Electronic Density Functional Theory: Recent Progress and New Directions*; Dobson, J. F., Vignale, G., and Das, M. P. (eds.), Plenum Press, New York, pp. 57–68.
[32] Jaramillo, J., Scuseria, G. E., and Ernzerhof, M. (2003) Local hybrid functionals. *J. Chem. Phys.*, **118**, 1068.
[33] Kaupp, M., Arbuznikov, A. V., and Bahmann, H. (2010) On occupied-orbital dependent exchange-correlation functionals: From local hybrids to Becke's B05 model. *Z. Phys. Chem.*, **224**, 545.
[34] Becke, A. D. and Roussel, M. R. (1989) Exchange holes in inhomogeneous systems: A coordinate-space model. *Phys. Rev. A*, **39**, 3761.
[35] Savin, A. (1996) On degeneracy, near-degeneracy and density functional theory. In *Recent Developments and Applications of Modern Density Functional Theory (Theoretical and Computational Chemistry, Vol. 4)*; Seminario, J. M. (ed.), Elsevier, Amsterdam, pp. 327–357.
[36] Gill, P. M. W., Adamson, R. D., and Pople, J. A. (1996) Coulomb-attenuated exchange energy density functionals. *Mol. Phys.*, **88**, 1005.
[37] Iikura, H., Tsuneda, T., Yanai, T., and Hirao, K. (2001) A long-range correction scheme for generalized-gradient-approximation exchange functionals. *J. Chem. Phys.*, **115**, 3540.
[38] Nakano, H., Nakajima, T., Tsuneda, T., and Hirao, K. (2005) Recent advances in *ab initio*, density functional theory, and relativistic electronic structure theory. In *Theory and Applications of Computational Chemistry: The First Forty Years*; Dykstra, C. E., Frenking, G., Kim, K. S., and Scuseria, G. E. (eds.), Elsevier, Amsterdam, pp. 507–557.
[39] Vydrov, O. A. and Scuseria, G. E. (2006) Assessment of a long-range corrected hybrid functional. *J. Chem. Phys.*, **125**, 234109.
[40] Heyd, J., Scuseria, G. E., and Ernzerhof, M. (2003) Hybrid functional based on a screened Coulomb potential. *J. Chem. Phys.*, **118**, 8207.
[41] Heyd, J. and Scuseria, G. E. (2004) Efficient hybrid density functional calculations in solids: Assessment of the Heyd–Scuseria–Ernzerhof screened Coulomb hybrid functional. *J. Chem. Phys.*, **121**, 1187.
[42] Adamson, R. D., Gill, P. M. W., and Pople, J. A. (1998) Empirical density functionals. *Chem. Phys. Lett.*, **284**, 6.
[43] Becke, A. D. (1997) Density-functional thermochemistry. V. systematic optimization of exchange-correlation functionals. *J. Chem. Phys.*, **107**, 8554.
[44] Becke, A. D. (1999) Exploring the limits of gradient corrections in density functional theory. *J. Comput. Chem.*, **20**, 63.
[45] Schmider, H. L. and Becke, A. D. (1998) Optimized density functionals from the extended G2 test set. *J. Chem. Phys.*, **108**, 9624.
[46] Hamprecht, F. A., Cohen, A. J., Tozer, D. J., and Handy, N. C. (1998) Development and assessment of new exchange-correlation functionals. *J. Chem. Phys.*, **109**, 6264.
[47] Zhao, Y. and Truhlar, D. G. (2008) Density functionals with broad applicability in chemistry. *Acc. Chem. Res.*, **41**, 157.
[48] Becke, A. D. (2005) Real-space post-Hartree–Fock correlation models. *J. Chem. Phys.*, **122**, 064101.
[49] Perdew, J. P., Staroverov, V. N., Tao, J., and Scuseria, G. E. (2008) Density functional with full exact exchange, balanced nonlocality of correlation, and constraint satisfaction. *Phys. Rev. A*, **78**, 052513.
[50] Becke, A. D. (2003) A real-space model for nondynamical correlation. *J. Chem. Phys.*, **119**, 2972.
[51] Arbuznikov, A. V. and Kaupp, M. (2009) On the self-consistent implementation of general occupied-orbital dependent exchange-correlation functionals with application to the B05 functional. *J. Chem. Phys.*, **131**, 084103.
[52] Proynov, E., Shao, Y., and Kong, J. (2010) Efficient self-consistent DFT calculation of nondynamic correlation based on the B05 method. *Chem. Phys. Lett.*, **493**, 381.
[53] Jiménez-Hoyos, C. A., Janesko, B. G., Scuseria, G. E., Staroverov, V. N., and Perdew, J. P. (2009) Assessment of a density functional with full exact exchange and balanced nonlocality of correlation. *Mol. Phys.*, **107**, 1077.
[54] Perdew, J. P. (1985) What do the Kohn–Sham orbital energies mean? How do atoms dissociate? In *Density Functional Methods in Physics*; Dreizler, R. M. and da Providência, J. (eds.), Plenum, New York, pp. 265–308.
[55] Perdew, J. P., Ruzsinszky, A., Csonka, G. I., Vydrov, O. A., Scuseria, G. E., Staroverov, V. N., and Tao, J. (2007) Exchange and correlation in open systems of fluctuating electron number. *Phys. Rev. A*, **76**, 040501(R).
[56] Cohen, A. J., Mori-Sánchez, P., and Yang, W. (2008) Insights into current limitations of density functional theory. *Science*, **321**, 792.
[57] Mori-Sánchez, P., Cohen, A. J., and Yang, W. (2008) Localization and delocalization errors in density functional theory and implications for band-gap predictions. *Phys. Rev. Lett.*, **100**, 146401.
[58] Cohen, A. J., Mori-Sánchez, P., and Yang, W. (2008) Fractional spins and static correlation error in density functional theory. *J. Chem. Phys.*, **129**, 121104.
[59] Mori-Sánchez, P., Cohen, A. J., and Yang, W. (2009) Discontinuous nature of the exchange-correlation functional in strongly correlated systems. *Phys. Rev. Lett.*, **102**, 066403.
[60] Perdew, J. P., Parr, R. G., Levy, M., and Balduz, Jr., J. L. (1982) Density-functional theory for fractional particle number: Derivative discontinuities of the energy. *Phys. Rev. Lett.*, **49**, 1691.