THE JOURNAL OF PHYSICAL CHEMISTRY LETTERS
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCL

Letter

Uniting Nonempirical and Empirical Density Functional Approximation Strategies Using Constraint-Based Regularization

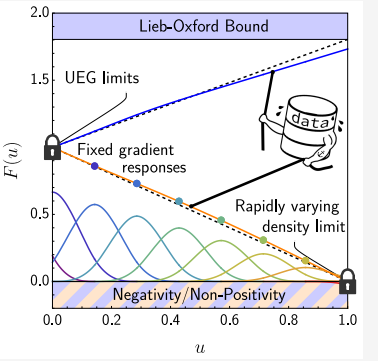
Zachary M. Sparrow, Brian G. Ernst, Trine K. Quady, and Robert A. DiStasio, Jr.*

Cite This: *J. Phys. Chem. Lett.* 2022, 13, 6896−6904

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: In this work, we present a general framework that unites the two primary strategies for constructing density functional approximations (DFAs): nonempirical (NE) constraint satisfaction and empirical (E) data-driven optimization. The proposed method employs B-splines, bell-shaped spline functions with compact support, to construct each inhomogeneity correction factor (ICF). This choice offers several distinct advantages over traditional polynomial expansions by enabling explicit enforcement of linear and nonlinear constraints as well as ICF smoothness using Tikhonov and penalized B-splines (P-splines) regularization. As proof-of-concept, we use the so-called CASE (constrained and smoothed empirical) framework to construct a constraint-satisfying and data-driven global hybrid that exhibits enhanced performance across a diverse set of chemical properties. We argue that the CASE approach can be used to generate DFAs that maintain the physical rigor and transferability of NE-DFAs while leveraging high-quality quantum-mechanical data to remove the arbitrariness of ansatz selection and improve performance.



Kohn−Sham density functional theory (KS-DFT) has become the *de facto* standard for electronic structure calculations in chemistry, physics, and materials science due to its favorable trade-off between accuracy and computational cost.[1] While there now exist hundreds of density functional approximations (DFAs) of varying complexity across all rungs of Perdew's popular Jacob's ladder,[2] most have been designed using either nonempirical (NE) or empirical (E) strategies.[1,3,4] NE-DFA strategies construct DFAs by proposing simple ansätze designed to satisfy well-defined physical constraints and norms (e.g., the uniform electron gas (UEG) limit,[5] second-order gradient responses,[6−9] Lieb−Oxford bound[10,11]). Resulting NE-DFAs (e.g., Perdew−Burke−Ernzerhof (PBE),[4] PBE0,[12] SCAN[13] (strongly constrained and appropriately normed)) tend to be more transferable across complex condensed-phase systems, making them more favored in the physics and materials science communities. E-DFA strategies construct DFAs by optimizing a physically motivated and flexible functional form to best reproduce reference quantum-chemical data. The resulting E-DFAs (e.g., Becke, three-parameter, Lee−Yang−Parr (B3LYP),[14] Minnesota functionals,[15−17] B97 family[1,3,18]) often perform quite well (typically exceeding NE-DFAs) on chemical systems and properties similar to the training data, and tend to be more popular for chemical applications.

When used independently, both of these strategies have shortcomings. For one, NE-DFA ansätze are somewhat arbitrary, and there is some flexibility when constructing a NE-DFA that satisfies a given set of constraints;[13,19] hence, there is no guarantee that the chosen ansatz will perform best in practice. The choice of constraints is also somewhat arbitrary or empirical;[20] for example, the correct series expansion of the exchange-correlation energy (E_{xc}) is sometimes ignored as it often results in inaccurate DFAs for real systems.[21] In the same breath, striving for the best-performing functional using an E-DFA strategy often goes hand-in-hand with sacrificing exact physical constraints, which is not ideal for transferability.[3,15,18,22] Furthermore, some E-DFAs suffer from nonphysical "bumps" or "wiggles" in the inhomogeneity correction factor (ICF), which violate an *implicit* smoothness constraint and can require significantly larger grids for accurate quadrature in practice.[23−26]

While both strategies provide useful information about the optimally performing DFA, neither suffices on its own. Hence, several groups have advocated for combining these strategies,[21,27] although constraint satisfaction during the data-driven optimization process has remained difficult to date. To address the E-DFA smoothness problem, the Bayesian error estimation functional (BEEF)[28−30] and Minnesota[31] functionals have adopted an explicit smoothness penalty in the regression procedure with reasonable success; the resulting ICFs are smoother than previous generations, albeit not always completely devoid of spurious features. Furthermore, the recent MCML (multi-purpose, constrained, and machine learned) approach[27] has made efforts to combine NE-DFA and E-DFA strategies by algebraically enforcing three linear constraints during the optimization process (expanding on an approach originally used by Truhlar and co-workers when constructing numerous Minnesota functionals[17,32,33]). While successful in enforcing the targeted constraints, the polynomial basis used in

Received: March 3, 2022
Accepted: April 4, 2022

https://doi.org/10.1021/acs.jpclett.2c00643
J. Phys. Chem. Lett. 2022, 13, 6896−6904

MCML (and all but a few[34] E-DFAs) prevents explicit enforcement of nonlinear constraints (such as inequalities) and makes satisfying any additional constraints nontrivial as each regression coefficient appears in every algebraic constraint.

In this work, we address this long-standing challenge of uniting NE-DFA and E-DFA strategies by presenting a general framework that seamlessly enables one to enforce exact physical constraints and ICF smoothness while simultaneously leveraging high-quality quantum-mechanical data during DFA construction. The proposed constrained and smoothed empirical (CASE) framework uses B-splines (i.e., compact bell-shaped piecewise functions[35]) during ICF construction, which allows for a tunable trade-off between ICF smoothness and flexibility via penalized B-splines (P-splines),[36] as well as explicit enforcement of both linear and nonlinear constraints via generalized Tikhonov regularization. As proof-of-concept, we use this framework to construct a global hybrid DFA that completely satisfies all but one constraint (and partially satisfies the remaining one) met by the PBE0 NE-DFA. When compared to PBE0 (and the popular B3LYP E-DFA), this CASE-generated DFA exhibits improved performance across a diverse set of chemical properties without sacrificing transferability or requiring large numerical quadrature grids. As such, we argue that the CASE framework can be used to construct next-generation DFAs that maintain the physical rigor and transferability of NE-DFAs while leveraging high-quality quantum-mechanical data to remove the arbitrariness of ansatz selection and improve performance.

Proof-of-Concept: Functional Form. As a proof-of-concept illustration of the CASE approach, we constructed and critically assessed a constraint-satisfying global hybrid generalized gradient approximation (GGA). The overall functional form for this CASE-generated DFA (hereafter referred to as CASE21) was chosen to be as simple as possible and was assembled from well-established ingredients used during DFA construction for over two decades now, that is, gradient corrections to the semilocal E_{xc} components in conjunction with a set fraction (25%) of exact exchange (E_{xx}), as generally recommended for global hybrid GGAs.[12,37] Our choice to construct a hybrid GGA (instead of a more complicated meta- or hybrid meta-GGA) was intentional, as the simpler functional form of GGA-based ICFs allows us to more clearly demonstrate how CASE can be used to enforce physical constraints and ICF smoothness during a data-driven DFA optimization procedure. However, the selection of such a relatively simple functional form is not a limitation of the CASE framework, which can be used to construct DFAs on every rung of Jacob's ladder. In fact, we expect that the full scope of this approach will be better realized when constructing more sophisticated functionals (e.g., meta- and hybrid meta-GGAs) that have the ability to satisfy more physical constraints and the flexibility to leverage larger amounts of benchmark data.

With these points in mind, we write CASE21 as the following sum of exchange and correlation contributions:

$$E_{xc}^{\mathrm{CASE21}} = \frac{3}{4}E_x[\rho_\uparrow, \rho_\downarrow] + \frac{1}{4}E_{xx} + E_c[\rho, \zeta] \tag{1}$$

The semilocal exchange is defined using the exchange spin scaling relationship:[38]

$$E_x[\rho_\uparrow, \rho_\downarrow] = \frac{1}{2}(E_x[2\rho_\uparrow] + E_x[2\rho_\downarrow]) \tag{2}$$

in which

$$E_x[\rho_\sigma] = \int \rho_\sigma \epsilon_x^{\mathrm{LDA}}(\rho_\sigma) F_x(u_{x,\sigma})\, \mathrm{d}\boldsymbol{r} \tag{3}$$

ρ_σ is the spin density (with spin $\sigma \in \{\uparrow, \downarrow\}$), $\epsilon_x^{\mathrm{LDA}}$ is the exchange energy density per particle within the local density approximation (LDA), and $F_x(u_{x,\sigma})$ is the yet to be determined CASE21 exchange ICF. We employ $0 \le u_{x,\sigma} = (\gamma_x s_\sigma^2)/(1 + \gamma_x s_\sigma^2) < 1$ (as originally proposed by Becke[39]) as the finite-domain representation of the PBE dimensionless spin density gradient, $s_\sigma = |\nabla\rho_\sigma|/[\pi^{2/3}(2\rho_\sigma)^{4/3}]$. Here, we note that the PBE exchange ICF can be written as a linear function of $u_{x,\sigma}$ if $\gamma_x = \mu/\kappa \approx 0.273022$ (where μ and κ are the NE parameters in PBE), which we denote by $\overline{F}_x(u_{x,\sigma}) \equiv 1 + \kappa u_{x,\sigma}$. Hence, we argue that this is an appropriate choice for γ_x since the UEG exchange limit,[5] UEG linear response,[4] and Lieb−Oxford bound[10,11] can still be straightforwardly enforced in this smooth limiting form (*vide infra*).

We construct $E_c[\rho,\zeta]$ in analogy to $E_x[\rho_\sigma]$, namely,

$$E_c[\rho, \zeta] = \int \rho \epsilon_c^{\mathrm{LDA}}(\rho, \zeta) F_c(u_c)\, \mathrm{d}\boldsymbol{r} \tag{4}$$

in which $\epsilon_c^{\mathrm{LDA}}(\rho, \zeta)$ is the PW92[40] LDA correlation energy density per particle, $\rho = \rho_\uparrow + \rho_\downarrow$ is the total density, $\zeta = (\rho_\uparrow - \rho_\downarrow)/\rho$ is the relative spin polarization, and $F_c(u_c)$ is the yet to be determined CASE21 correlation ICF. As with exchange, we suggest a form for u_c such that a linear ICF, that is, $\overline{F}_c(u_c) \equiv 1 - u_c$, would satisfy the UEG correlation limit,[5] rapidly varying density limit,[4] and second-order gradient expansion for correlation.[6−9] Namely, we propose $0 \le u_c \equiv (-\phi^3 t^2)/(-\phi^3 t^2 + \gamma_c \epsilon_c^{\mathrm{LDA}}) < 1$, where $\phi = 1/2[(1+\zeta)^{2/3} + (1-\zeta)^{2/3}]$ is a spin scaling factor,[6] $\gamma_c = 1/\beta \approx 14.986886$ (where β is another NE parameter in PBE), and t is the following dimensionless spin-separated density gradient:

$$t \equiv \sqrt{a_0}\left(\frac{\pi}{3}\right)^{1/6} \frac{|\nabla\rho_\uparrow| + |\nabla\rho_\downarrow|}{4\rho^{7/6}\phi} \tag{5}$$

This quantity reduces to the PBE dimensionless density gradient (t^{PBE}, which has $|\nabla\rho|$ instead of $|\nabla\rho_\uparrow| + |\nabla\rho_\downarrow|$ in the numerator) when $|\nabla\zeta| = 0$, which was assumed during the construction of PBE correlation and is a relationship that allows DFAs based on t to satisfy PBE correlation constraints. We note in passing that the use of t^{PBE} yields qualitatively similar results to t (which might be expected, given that t and t^{PBE} are equivalent for closed-shell systems), although t slightly outperforms t^{PBE} quantitatively. With the above definition of u_c, eq 4 does not fully satisfy uniform scaling to the high-density limit for correlation;[41] however, it does completely cancel the $\epsilon_c^{\mathrm{LDA}}$ logarithmic singularity[42] and allows for satisfaction of all other PBE correlation constraints. Since the functional form described above was chosen for its simplicity, partial satisfaction of this constraint is not a restriction of the CASE approach; in principle, a functional form (albeit more complex) that completely satisfies all PBE correlation constraints could have been used.

The CASE Framework. CASE exchange and correlation ICFs are written as linear combinations of N_{sp} compact piecewise bell-shaped cubic ($k = 3$) uniform B-spline basis functions ($\{B_i\}$),[35] that is,

$$F_x(u_{x,\sigma}) = \sum_i^{N_{sp}} c_{x,i} B_i(u_{x,\sigma}) = \mathbf{c}_x \cdot \mathbf{B}_{x,\sigma}$$

$$F_c(u_c) = \sum_i^{N_{sp}} c_{c,i} B_i(u_c) = \mathbf{c}_c \cdot \mathbf{B}_c \tag{6}$$

https://doi.org/10.1021/acs.jpclett.2c00643
J. Phys. Chem. Lett. 2022, 13, 6896−6904

which is equivalent to constructing each ICF using a cubic spline[43] (see Supporting Information (SI) for more details). With an appropriate uniformly spaced knot vector,[35,36] the $B_i(u_{x,\sigma})$ and $B_i(u_c)$ are also uniformly spaced with all points in $0 \le u_{x,\sigma} \le 1$ and $0 \le u_c \le 1$ supported by three nonzero B-splines. As depicted in Figure 1a, setting $\mathbf{c}_x = 1 = \mathbf{c}_c$ in eq 6 results in $F_x(u_{x,\sigma}) = 1$ (LSDA exchange) and $F_c(u_c) = 1$ (LDA correlation).

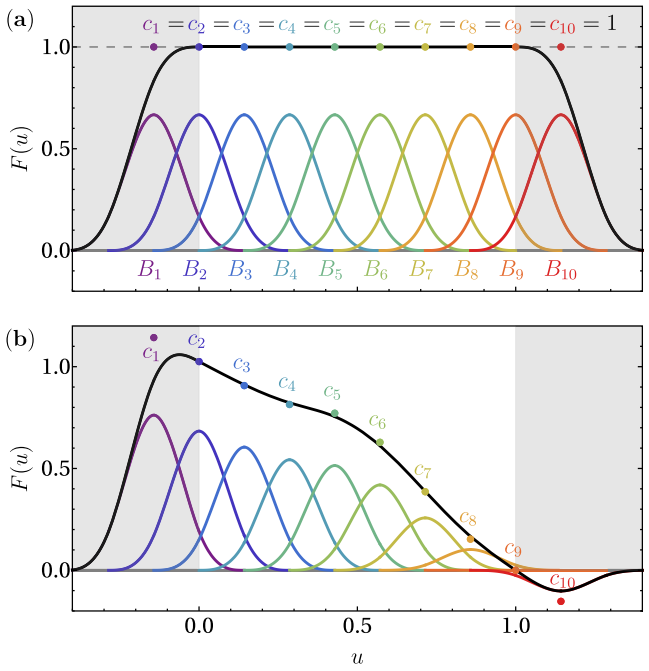


Figure 1. (a) B-spline basis functions ($\{B_i\}_{i=1,10}$, rainbow) used to represent exchange and correlation ICFs in the CASE approach. When all expansion coefficients are set to unity, the B-spline curve ($F(u) = \sum_i c_i B_i(u)$, black) is uniform in $0 \le u \le 1$ and recovers the LSDA/LDA limit. (b) B-spline curve with nonuniform coefficients. Note how the coefficients again closely align with the curve for $0 \le u \le 1$.

To seamlessly unite the NE-DFA and E-DFA strategies, the CASE approach uses generalized Tikhonov regularization[44] to determine $\mathbf{c} = (\mathbf{c}_x, \mathbf{c}_c)$, that is, ICF coefficients are found by minimizing the following loss function:

$$\mathcal{L} = \|\mathbf{X}\mathbf{c} - \mathbf{y}\|_{\mathbf{W}}^2 + \lambda \|\mathbf{c}\|_{\mathbf{A}}^2 + \eta \sum_i \|\mathbf{c} - \mathbf{c}_0\|_{\mathbf{Q}_i}^2 \tag{7}$$

wherein $\|\mathbf{v}\|_{\mathbf{M}}^2 = \mathbf{v}^T\mathbf{M}\,\mathbf{v}$ is the matrix norm of the vector $\mathbf{v}$ using the matrix $\mathbf{M}$, the sum is over the enforced constraints, and all other quantities will be defined below. Hence, the key to determining $\mathbf{c}$ lies in appropriate matrix norm choices in each term in $\mathcal{L}$.

Goodness of Fit. In the goodness of fit term (i.e., the first term in $\mathcal{L}$), we construct the design matrix $\mathbf{X}$ by first noting that substitution of eq 6 into eqs 3 and 4 (with fixed orbitals) casts $E_x[\rho_\sigma]$ and $E_c[\rho, \zeta]$ into linear forms in $\mathbf{c}_x$ and $\mathbf{c}_c$:

$$E_x[\rho_\sigma] = \sum_i^{N_{sp}} c_{x,i} \int \rho_\sigma \epsilon_x^{LDA}(\rho_\sigma) B_i(u_{x,\sigma})\, d\mathbf{r} \equiv \mathbf{c}_x \cdot \boldsymbol{\xi}_{x,\sigma}$$

$$E_c[\rho, \zeta] = \sum_i^{N_{sp}} c_{c,i} \int \rho \epsilon_c^{LDA}(\rho, \zeta) B_i(u_c)\, d\mathbf{r} \equiv \mathbf{c}_c \cdot \boldsymbol{\xi}_c \tag{8}$$

Hence, linear combinations of $\boldsymbol{\xi}_{x,\sigma}$ and $\boldsymbol{\xi}_c$ can be used to construct *semilocal* xc contributions to energy differences ΔE_{xc} (e.g., atomization energies, reaction energies, barrier heights) in a form amenable to linear regression using reference data. That is, defining $\boldsymbol{\xi} \equiv (\boldsymbol{\xi}_x, \boldsymbol{\xi}_c)$, with $\boldsymbol{\xi}_x$ obtained after applying eqs 1 and 2 to $\boldsymbol{\xi}_{x,\uparrow}$ and $\boldsymbol{\xi}_{x,\downarrow}$ in eq 8, allows us to define $\mathbf{x}$ (a single row of $\mathbf{X}$) as

$$\Delta E_{xc} = \sum_j \nu_j (\mathbf{c} \cdot \boldsymbol{\xi}_j) = \mathbf{c} \cdot \sum_j \nu_j \boldsymbol{\xi}_j \equiv \mathbf{c} \cdot \mathbf{x} \tag{9}$$

in which ν_j is the stoichiometric coefficient for the j-th component in ΔE_{xc} (i.e., the energy of a molecule or atom), $\mathbf{y}$ is the corresponding vector of reference energy differences ΔE_{xc}^{ref}, and our choice for $\mathbf{W}$ (a square diagonal matrix of weights $w_i \equiv \min[1, 1/\Delta E_{xc,i}^{ref}]$) is motivated by the fact that the $\mathbf{c}$ minimizing the goodness of fit term only (i.e., weighted least-squares) is the best linear unbiased estimator (under some common assumptions) if the w_i are inversely proportional to the variance in each measurement.[45] Since E_{xc} is the only nonexact term in KS-DFT, both bias- and variance-type DFA errors should scale linearly with E_{xc},[46−48] making this a natural choice for $\mathbf{W}$. Here, we argue that the piecewise nature of the B-spline curves used in CASE offers more flexibility than the low-order polynomial expansions often used to represent E-DFA ICFs (e.g., the B97 family[1,3,18]); with the ability to conform to more subtle shapes, a B-spline ICF should be able to better leverage the reference data.

ICF Smoothness. For the second term in $\mathcal{L}$, we note that B-splines can be regularized by explicitly penalizing deviations from smoothness (i.e., ICF "bumps" or "wiggles") using P-splines, a regularization technique suggested by Eilers and Marx[35,36] based on the observation that B-spline coefficients closely resemble the B-spline curve (see Figure 1b). As such, ICF smoothness can be explicitly enforced in the CASE framework via a finite-difference penalty on $\mathbf{c}$; in this work, we interpret nonsmoothness as nonlinearity in the ICF, and construct $\mathbf{A}$ from the second-derivative finite-difference matrix (see SI). λ is a hyperparameter that governs the relative importance of the smoothness and goodness of fit contributions to $\mathcal{L}$, and interpolates (assuming $\eta \gg 1$, *vide infra*) between linear ICFs (i.e., $\bar{F}_x(u_{x,\sigma})$ and $\bar{F}_c(u_c)$) that are completely constraint-driven ($\lambda \to \infty$) and wiggly ICFs that are data-driven to the *maximum* amount possible in this framework ($\lambda \to 0$). As such, any nonlinearity in the final optimized ICFs can be attributed to the data. Here, we note that alternative interpretations of smoothness would result in penalizing other derivatives (e.g., $F'''(u)$). Separately penalizing the exchange and correlation ICFs (i.e., using two λ-hyperparameters) is also possible if the ICF smoothness contributions to $\mathcal{L}$ from $F_x(u_{x,\sigma})$ and $F_c(u_c)$ strongly differ. In this work, we found that P-spline regularization yields ICFs devoid of any spurious "wiggles" via single-λ penalization of $F''(u)$ (*vide infra*). In contrast, an excessively large penalty (which results in decreased performance) is usually required to remove all nonphysical "bumps" or "wiggles" in polynomial ICFs regularized via Tikhonov (or ridge) regression.[28,49] Furthermore, although such polynomial-based smoothness penalties are somewhat effective in reducing DFA grid dependence,[24,31] these approaches have been largely ineffective when enforced alongside constraints.[28,50] On the other hand, we find no issues when simultaneously enforcing ICF smoothness in conjunction with numerous linear and nonlinear constraints using the CASE approach.

Constraint Satisfaction. In the constraint satisfaction term in $\mathcal{L}$, the $\{\mathbf{Q}_i\}$ are chosen to measure constraint-specific

https://doi.org/10.1021/acs.jpclett.2c00643
J. Phys. Chem. Lett. 2022, 13, 6896−6904

deviations of $\mathbf{c}$ from $\mathbf{c}_0$, the coefficients corresponding to $\overline{F}_x(u_{x,\sigma})$ and $\overline{F}_c(u_c)$. Each $\mathbf{Q}_i$ corresponds to a constraint on $F(u)$ or $F'(u)$ and is constructed such that any constraint-satisfying $\mathbf{c}$ yields $\|\mathbf{c}-\mathbf{c}_0\|^2_{\mathbf{Q}_i}=0$ (see SI for $\mathbf{Q}_i$ construction details). η is a hyperparameter that governs the relative importance of the constraint satisfaction contribution to $\mathcal{L}$, and should be chosen to be large enough for strict constraint satisfaction but small enough to avoid conditioning issues. Since each B-spline has compact support, each $\mathbf{Q}_i$ only enforces a constraint on a small subset of $\mathbf{c}$ (e.g., those corresponding to the nonzero B-splines at $u=0$); in contrast, each constraint would generally involve every parameter in a polynomial-based ICF (e.g., MCML[27]). Another important consequence of this local support is that the B-spline curve itself will lie within the range of $\mathbf{c}$ (*cf.* Figure 1b). Hence, inequality constraints can be enforced via an iterative update to the corresponding $\mathbf{Q}_i$ using the shape constraint algorithm (SCA) of Bollaerts et al.,[51] which fixes all inequality-violating c_i to the constraint boundary. In contrast, there is no straightforward way to apply inequality constraints on a polynomial-based ICF, as each basis function is uniquely shaped and has global support.

CASE21: Training Procedure. Now we will demonstrate how the CASE framework described above can be used to train a fully self-consistent DFA, that is, the proof-of-concept CASE21 functional. Our self-consistent training procedure (Scheme 1, see Computational Methods for more details) leverages three distinct data sets: training ($\mathbf{X}_{train}$, $\mathbf{y}_{train}$), validation ($\mathbf{X}_{val}$, $\mathbf{y}_{val}$), and testing ($\mathbf{X}_{test}$, $\mathbf{y}_{test}$). For CASE21, we fully enforce the following 10 physical constraints satisfied by PBE: exchange spin scaling,[38] uniform density scaling for exchange,[52] UEG exchange limit,[5] UEG linear response,[4] Lieb–Oxford bound,[10,11] exchange energy negativity, UEG correlation limit,[5] second-order gradient expansion for correlation,[6−9] rapidly varying density limit for correlation,[4] and correlation energy nonpositivity.[4] We also partially enforce uniform scaling to the high-density limit for correlation (*vide supra*).[41,42] In a given iteration, the training set (a single database of heavy atom transfer reaction energies, HAT707[1,53]) is used to initially determine $\mathbf{c}$ by minimizing $\mathcal{L}$ (in conjunction with the SCA for satisfying inequality constraints) for a range of λ and a given set of orbitals $\{\psi_i\}$ (with initial $\{\psi_i^0\}$ generated using $\overline{F}_x(u_{x,\sigma})$ and $\overline{F}_c(u_c)$). With $\mathbf{c}(\lambda)$, a weighted root-mean-square error,

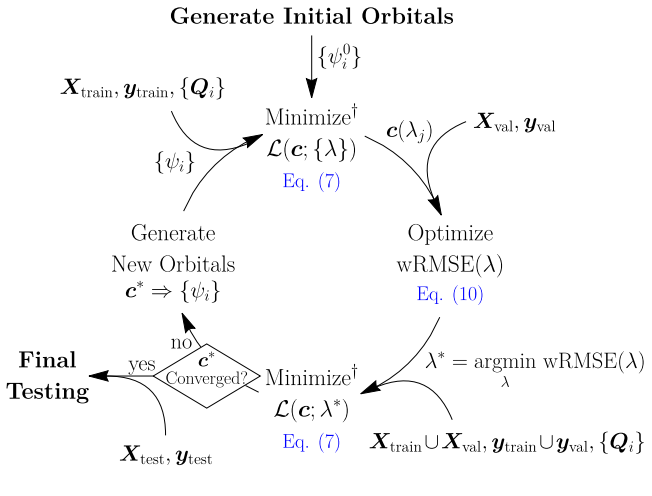
$$\mathrm{wRMSE}(\lambda)=\sqrt{\mathrm{diag}(\mathbf{W})\cdot\mathbf{r}(\lambda)^2/\mathrm{Tr}(\mathbf{W})} \tag{10}$$

in which $\mathbf{r}(\lambda)=\mathbf{X}_{val}\mathbf{c}(\lambda)-\mathbf{y}_{val}$ is the error vector and $\mathbf{r}(\lambda)^2$ is the element-wise square of $\mathbf{r}(\lambda)$, is computed on the validation set (which contains absolute energies of H−O from AE18[1,54] and all atomization energies in TAE203[55,56]). Using $\lambda^*=\mathrm{argmin}_\lambda\,\mathrm{wRMSE}(\lambda)$, $\mathbf{c}^*$ is determined by reoptimizing $\mathcal{L}$ (in conjunction with the SCA) over the training and validation sets. New $\{\psi_i\}$ are then generated using $\mathbf{c}^*$, and the entire cycle is repeated until $\mathbf{c}^*$ is stationary. At this point, the testing set (which contains more diverse chemical properties than the training and validation sets, *vide infra*) is used to assess performance and transferability. During initial minimization of eq 7, we found that $\mathbf{c}^*$ was fairly insensitive to the choice of training data, and the determination of λ^* was most robust if the training and validation sets contained distinct chemical properties. Hence, we limited the databases used for the training and validation sets to only a few chemical properties (i.e., reaction, atomization, and absolute energies), as this emphasizes transferability when evaluating the more diverse set of properties in the testing set. In particular, HAT707,[1,53] TAE203,[55,56] and AE18[1,54] were chosen because they are fairly large and reliable databases comprised of distinct chemical properties that collectively quantify the energies of covalent bonds (HAT707) relative to the energies of atoms (TAE203 and AE18).

We used $N_{sp}=10$ in Scheme 1 to generate the self-consistently optimized CASE21 DFA (six iterations; convergence criterion of $|\Delta\mathbf{c}|<10^{-5}$; see SI for $\mathbf{c}^*$). Even with a finite η value ($\eta=10^8$), CASE21 nearly exactly satisfies all enforced constraints, that is, $F_x(0)$, $F_c(0)$, $F'_x(0)$, and $F'_c(0)$ differ from their corresponding exact values by $\sim10^{-5}$, $F_c(1)$ differs by $\sim10^{-6}$, and all other constraints are exactly satisfied. When nonzero, the deviations were similar in magnitude to the convergence criterion and negligible in practical calculations. We therefore conclude that the proposed CASE framework successfully enforced all constraints without sacrificing smoothness, which still remains a challenge for other DFA training procedures.[28,50] To confirm that CASE21 remains representative of DFAs trained with other N_{sp} values (and to investigate the dependence of the CASE framework on N_{sp}), we non-self-consistently optimized $\mathbf{c}$ for select $N_{sp}\in[6,40]$ using the CASE21 orbitals. As depicted in Figure 2, the resulting ICFs and their first derivatives were all smooth and very similar (particularly for $N_{sp}\geq10$), and the number of *effective* degrees of freedom[57] (DoF, see SI for derivation and more details) change slowly for $N_{sp}\geq10$. We therefore expect little dependence on N_{sp} for any DFA constructed with 10−40 B-splines and use this observation as an *a posteriori* justification for our choice of $N_{sp}=10$. From the piecewise nature of $F'(u)$ in this plot, one can also see that the CASE21 ICFs (DoF = 1.22) subtly deviate from linearity in ways that cannot be precisely obtained using low-order polynomial expansions. Here, we also note that the CASE21 xc enhancement factors (see SI) are noncrossing for different r_s values, which is a consequence of satisfying uniform density scaling for correlation at the GGA level;[52] although this additional constraint was not explicitly enforced during the construction of CASE21 or PBE, both of these DFAs have this property.

CASE21: Final Testing. Having demonstrated that the CASE approach is able to enforce physical constraints and ICF smoothness in conjunction with a data-driven optimization

Scheme 1. Self-Consistent Training Procedure for Generating DFAs in the CASE Framework



†Subject to inequality constraints enforced by the SCA.

https://doi.org/10.1021/acs.jpclett.2c00643
J. Phys. Chem. Lett. 2022, 13, 6896−6904

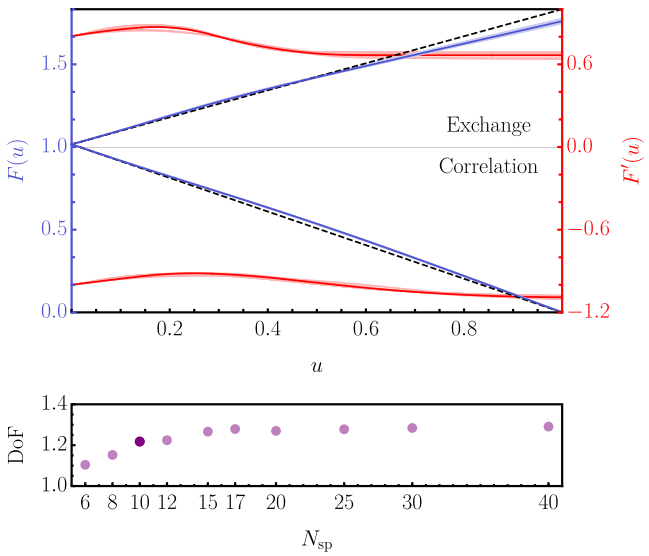


Figure 2. (top) Exchange and correlation ICFs (blue) and first derivatives (red) for select $N_{sp} \in [6,40]$. Highlighted curves (dark blue and dark red) correspond to the self-consistently optimized CASE21 DFA (with $N_{sp} = 10$). Dashed lines represent the parameter-free linear ICFs ($\bar{F}_x(u_{x,\sigma})$ and $\bar{F}_c(u_c)$) designed to satisfy the same constraints as CASE21. (bottom) Effective degrees of freedom (DoF) for select $N_{sp} \in [6,40]$, with the dark purple point corresponding to CASE21.

procedure, we now compare the performance of CASE21 to the PBE0 and B3LYP hybrid DFAs across a diverse set of chemical properties in Figure 3. Since correlation is treated semilocally in CASE21, PBE0, and B3LYP, only databases containing small-to-

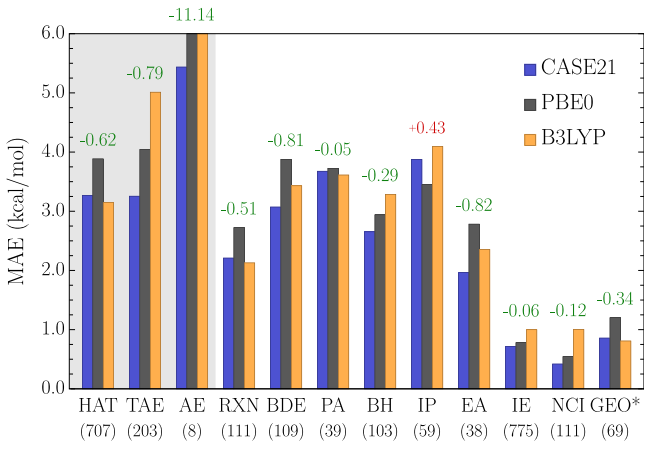


Figure 3. Mean absolute errors of CASE21 (blue), PBE0 (gray), and B3LYP (orange) in the training/validation (shaded region) and testing (white region) sets. Bar labels indicate the relative performance of CASE21 and PBE0, with green and red numbers representing increased ($MAE^{CASE21} < MAE^{PBE0}$) and decreased ($MAE^{CASE21} > MAE^{PBE0}$) performance, respectively. Properties (number of data points) include HAT (heavy atom transfer reaction energies),[1,53] TAE (total atomization energies),[55,56] AE (absolute energies),[1,54] RXN (reaction energies),[1,53,56,58−64] BDE (bond dissociation energies),[1,53,56,61,65] PA (proton affinities),[56,61,66,67] BH (barrier heights),[1,56,59−61] IP (ionization potentials),[1,56,58,61,68,69] EA (electron affinities),[1,58,68,69] IE (isomerization energies),[1,15,16,53,56,58,61,67,70−80] NCI (noncovalent interaction energies),[1,81−87] and GEO (geometry energy offsets).[88−90] See SI for more details regarding the databases used in this work. The asterisk (*) indicates that all three GEO MAEs were scaled by 10× for clarity.

medium molecules with minimal contributions from nonlocal correlation were selected for the testing (as well as training and validation) set; this also applies to the noncovalent interaction (NCI) databases used here, which are mostly comprised of hydrogen- and halogen-bonded dimers (instead of primarily dispersion-bound systems). In this work, we report the robust mean absolute error (MAE) metric for the selected DFAs on databases grouped by chemical property, as this avoids the use of arbitrarily weighted error metrics (which would otherwise be needed to account for the differences in frequency and magnitude of the various properties in the training, validation, and test sets).

With these points in mind, we find that CASE21 outperforms the PBE0 NE-DFA on 11 of 12 properties, with improvements as large as 0.81 and 0.82 kcal/mol for bond dissociation energies (BDE) and electron affinities (EA), respectively. In the testing set, CASE21 decreases the PBE0 MAE in 8 of 9 properties by an average of 0.34 kcal/mol. On the other hand, PBE0 outperforms CASE21 for ionization potentials (IP), which may be attributed to incomplete or partial satisfaction of uniform scaling to the high-density limit for correlation in CASE21,[42] as this results in slightly less accurate (but still reasonable) correlation energies in the He isoelectronic series (see SI for more details). CASE21 also outperforms B3LYP (a popular E-DFA for chemical applications) on 8 of 12 properties; in the testing set, CASE21 decreases the B3LYP MAE in 6 of 9 properties by an average of 0.41 kcal/mol (while B3LYP only offers a marginal ∼0.05 kcal/mol improvement on the remaining 3 of 9). In this assessment, we also measured the performance of CASE21 for molecular structure optimization using the geometry energy offset (GEO) metric of Vukovic and Burke,[88] and found that CASE21 improves upon PBE0 and performs on par with B3LYP, which is recognized as one of the best DFAs for predicting molecular geometries. To put these results into context, the more advanced ωB97X (ωB97X-V) E-DFA[18,91] improves upon the PBE0 MAE for 8 of 9 (9 of 9) properties in the testing set by an average of 0.58 kcal/mol (0.59 kcal/mol). Despite the fact that CASE21 is a hybrid GGA that does not include range-separated exact exchange and nonlocal correlation, the improved performance of CASE21 in the testing data set is 58% (57%) on average of that achieved by ωB97X (ωB97X-V). Encouraged by these results, we also considered how the performance of CASE21 depends on the fraction of exact exchange, but ultimately found that the initial value of 1/4 was essentially optimal (see SI for more details). Although our focus to this point has been on molecular properties, CASE21 calculations of the lattice constants, bulk moduli, and cohesive energies of bulk Si and C (diamond) also showed promising preliminary results for solid-state properties (see SI). For these systems, the CASE21 predictions were significantly better than B3LYP and slightly worse than PBE0, suggesting that further studies into the performance of CASE21 on solid-state properties (as well as the inclusion of such properties in the training and validation sets) are warranted. Taken together, this analysis demonstrates that CASE21 is largely able to preserve the physical rigor and transferability of the PBE0 NE-DFA while offering a noteworthy increase in performance on chemical systems (even when compared to the B3LYP E-DFA), despite having only 1.22 effective DoF (compared to the 3.0 effective DoF in B3LYP; see SI for derivation).

Although the CASE21 ICFs are clearly smooth (cf. Figure 2), we also investigated the grid dependence of this DFA for completeness. Since Lebedev−Treutler grids[92] with 50 radial

https://doi.org/10.1021/acs.jpclett.2c00643
J. Phys. Chem. Lett. 2022, 13, 6896−6904

and 194 angular grid points (i.e., (50, 194)) are typically large enough to obtain accurate energetics with standard hybrid GGAs (such as PBE0),[23] we compared the performance of CASE21 using this grid to the larger grids employed during the training procedure (see Computational Methods). Using all points in the training, validation, and testing data sets ($N = 2263$; excluding GEO), we found nearly identical mean absolute deviations of 1.84×10^{-2} kcal/mol for CASE21 and 1.83×10^{-2} kcal/mol for PBE0, thereby indicating that CASE21 does not require larger quadrature grids than PBE0 for accurate integration and demonstrating the effectiveness of the P-spline regularization central to the CASE framework.

Closing Remarks. In this work, we presented the CASE (constrained and smoothed empirical) framework for uniting the NE-DFA and E-DFA construction paradigms. By employing a B-spline representation for the ICFs, this approach has several distinct advantages over the historical choice of a polynomial basis, namely, explicit enforcement of linear and nonlinear constraints (using Tikhonov regularization) as well as penalization of nonphysical ICF "bumps" or "wiggles" (using P-splines), that are seamlessly integrated with data-driven optimization. As proof-of-concept, we used this approach to construct a global hybrid GGA that completely satisfies all but one constraint (and partially satisfies the remaining one) met by the PBE0 NE-DFA. Despite being trained on only a handful of properties, this CASE-generated DFA outperforms PBE0 and B3LYP (arguably the most popular E-DFA for chemical applications) across a diverse set of chemical properties. As such, we argue that the CASE framework solves the long-standing problem of uniting these seemingly disparate DFA strategies, and can be used to design next-generation DFAs that maintain the physical rigor and transferability of NE-DFAs while leveraging benchmark quantum-mechanical data to remove the arbitrariness of ansatz selection and improve performance. Alternatively, the CASE framework can also be used to enforce ICF smoothness in conjunction with physical constraints during NE-DFA construction (i.e., without leveraging data) or enforce ICF smoothness in conjunction with data-driven optimization during E-DFA construction (i.e., without requiring constraint satisfaction). Future work will extend this approach to more sophisticated DFAs (e.g., meta-GGAs, range-separated hybrids, DFAs with nonlocal correlation) that have the ability to satisfy more physical constraints and the flexibility to leverage larger amounts of data, where we expect that the larger function space made accessible by a B-spline ICF expansion will provide even more significant advantages over traditional low-order polynomials. Future work will also explore the performance of CASE-generated DFAs when treating condensed-phase systems as well as the use of B-splines for constructing robust features for machine-learning chemical properties.

Computational Methods. All gas-phase electronic structure calculations were performed using in-house versions of `Psi4 (v1.3.2)`[93] and `LibXC (v4.3.4)`[94] modified with a self-consistent implementation of the CASE21 DFA (including functional derivatives analytically computed using `Mathematica v12.1`). Self-consistent field (SCF) calculations were performed using density fitting (DF) in conjunction with the def2-QZVPPD[95,96] and def2-QZVPP-JKFIT[97,98] basis sets and an energy convergence threshold of `e_convergence = 1e-12`. During DFA training, all calculations employed (99, 590) Lebedev–Treutler grids[92] except for the calculations of the absolute energies in AE18,[1,54] which used (500, 974). Minimization of $\mathcal{L}$ in eq 7 and optimization of wRMSE(λ) in eq 10 were performed in `Mathematica v12.1`. All solid-state electronic structure calculations were performed using the `PWscf` package in `Quantum ESPRESSO`,[99] in conjunction with norm-conserving HSCV-PBE pseudopotentials[100,101] and converged planewave kinetic energy cutoffs (40 Ry and 120 Ry for Si and C, respectively) and k-point grids ($4 \times 4 \times 4$ and $8 \times 8 \times 8$ for Si and C, respectively). All solid-state properties were determined by fitting the Murnaghan equation of state[102] to 10 points centered around the expected equilibrium lattice constant.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpclett.2c00643.

B-spline definitions; enforcement of ICF constraints; training, validation, and testing data sets; derivation of optimal coefficients and effective degrees of freedom for weighted generalized Tikhonov regularization; optimized CASE21 ICF coefficients; comparison of PBE and CASE21 xc enhancement factors; correlation energies in the helium isoelectronic series; assessment of the fraction of exact exchange in CASE21; and solid-state properties (PDF)

Reference/benchmark, CASE21, PBE0, and B3LYP energies for each energy difference in the training, validation, and testing data sets (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Robert A. DiStasio, Jr. − *Department of Chemistry and Chemical Biology, Cornell University, Ithaca, New York 14853, United States;* orcid.org/0000-0003-2732-194X; Email: distasio@cornell.edu

Authors

Zachary M. Sparrow − *Department of Chemistry and Chemical Biology, Cornell University, Ithaca, New York 14853, United States;* orcid.org/0000-0001-6163-2843

Brian G. Ernst − *Department of Chemistry and Chemical Biology, Cornell University, Ithaca, New York 14853, United States;* orcid.org/0000-0002-7900-9360

Trine K. Quady − *Department of Chemistry and Chemical Biology, Cornell University, Ithaca, New York 14853, United States;* orcid.org/0000-0001-7777-909X

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpclett.2c00643

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

All authors thank Kieron Burke, Garnet Chan, Alexandre Tkatchenko, and Don Truhlar for helpful scientific discussions and Susi Lehtola for implementing CASE21 in `LibXC (v5.1.7)`. All authors also thank Richard Kang and Dzmitry (Dima) Vaido for their help in assembling databases and making modifications to the `Psi4` and `LibXC` codes, Hsin-Yu Ko for help in computing the solid-state properties, and Yang Yang for assisting with basis set selection for the He isoelectronic series. This material is based upon work supported by the National Science Foundation under Grant No. CHE-1945676. This work

was supported in part by the Cornell Center for Materials Research with funding from the Research Experience for Undergraduates program (DMR-1757420 and DMR-1719875). R.A.D. also gratefully acknowledges financial support from an Alfred P. Sloan Research Fellowship. This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

REFERENCES

(1) Mardirossian, N.; Head-Gordon, M. Thirty years of density functional theory in computational chemistry: An overview and extensive assessment of 200 density functionals. *Mol. Phys.* **2017**, *115*, 2315−2372.
(2) Perdew, J. P.; Schmidt, K. Jacob's ladder of density functional approximations for the exchange-correlation energy. *AIP Conf. Proc.* **2000**, *577*, 1−20.
(3) Becke, A. D. Density-functional thermochemistry. V. Systematic optimization of exchange-correlation functionals. *J. Chem. Phys.* **1997**, *107*, 8554−8560.
(4) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **1996**, *77*, 3865−3868.
(5) Kurth, S.; Perdew, J. P.; Blaha, P. Molecular and solid-state tests of density functional approximations: LSD, GGAs, and meta-GGAs. *Int. J. Quantum Chem.* **1999**, *75*, 889−909.
(6) Wang, Y.; Perdew, J. P. Spin scaling of the electron-gas correlation energy in the high-density limit. *Phys. Rev. B* **1991**, *43*, 8911−8916.
(7) Ma, S.-K.; Brueckner, K. A. Correlation energy of an electron gas with a slowly varying high density. *Phys. Rev.* **1968**, *165*, 18−31.
(8) Geldart, D. J. W.; Rasolt, M. Exchange and correlation energy of an inhomogeneous electron gas at metallic densities. *Phys. Rev. B* **1976**, *13*, 1477−1488.
(9) Langreth, D. C.; Perdew, J. P. Theory of nonuniform electronic systems. I. Analysis of the gradient approximation and a generalization that works. *Phys. Rev. B* **1980**, *21*, 5469−5493.
(10) Lieb, E. H.; Oxford, S. Improved lower bound on the indirect Coulomb energy. *Int. J. Quantum Chem.* **1981**, *19*, 427−439.
(11) Perdew, J. P.; Ruzsinszky, A.; Sun, J.; Burke, K. Gedanken densities and exact constraints in density functional theory. *J. Chem. Phys.* **2014**, *140*, 18A533.
(12) Adamo, C.; Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **1999**, *110*, 6158−6170.
(13) Sun, J.; Ruzsinszky, A.; Perdew, J. P. Strongly constrained and appropriately normed semilocal density functional. *Phys. Rev. Lett.* **2015**, *115*, 036402.
(14) Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **1993**, *98*, 5648−5652.
(15) Peverati, R.; Truhlar, D. G. Quest for a universal density functional: The accuracy of density functionals across a broad spectrum of databases in chemistry and physics. *Philos. Trans. R. Soc. A* **2014**, *372*, 20120476.
(16) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. Exchange-correlation functional with broad accuracy for metallic and nonmetallic compounds, kinetics, and noncovalent interactions. *J. Chem. Phys.* **2005**, *123*, 161103.
(17) Zhao, Y.; Truhlar, D. G. 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.* **2008**, *120*, 215−241.
(18) Mardirossian, N.; Head-Gordon, M. ωB97X-V: A 10-parameter, range-separated hybrid, generalized gradient approximation density functional with nonlocal correlation, designed by a survival-of-the-fittest strategy. *Phys. Chem. Chem. Phys.* **2014**, *16*, 9904−9924.
(19) Perdew, J. P.; Ruzsinszky, A.; Tao, J.; Staroverov, V. N.; Scuseria, G. E.; Csonka, G. I. Prescription for the design and selection of density functional approximations: More constraint satisfaction with fewer fits. *J. Chem. Phys.* **2005**, *123*, 062201.
(20) Pedroza, L. S.; da Silva, A. J.; Capelle, K. Gradient-dependent density functionals of the Perdew-Burke-Ernzerhof type for atoms, molecules, and solids. *Phys. Rev. B* **2009**, *79*, 201106.
(21) Yu, H. S.; Li, S. L.; Truhlar, D. G. Perspective: Kohn-Sham density functional theory descending a staircase. *J. Chem. Phys.* **2016**, *145*, 130901.
(22) Medvedev, M. G.; Bushmarinov, I. S.; Sun, J.; Perdew, J. P.; Lyssenko, K. A. Density functional theory is straying from the path toward the exact functional. *Science* **2017**, *355*, 49−52.
(23) Dasgupta, S.; Herbert, J. M. Standard grids for high-precision integration of modern density functionals: SG-2 and SG-3. *J. Comput. Chem.* **2017**, *38*, 869−882.
(24) Mardirossian, N.; Head-Gordon, M. How accurate are the Minnesota density functionals for noncovalent interactions, isomerization energies, thermochemistry, and barrier heights involving molecules composed of main-group elements? *J. Chem. Theory Comput.* **2016**, *12*, 4303−4325.
(25) Wheeler, S. E.; Houk, K. N. Integration grid errors for meta-GGA-predicted reaction energies: Origin of grid errors for the M06 suite of functionals. *J. Chem. Theory Comput.* **2010**, *6*, 395−404.
(26) Bootsma, A. N.; Wheeler, S. E. Popular integration grids can result in large errors in DFT-computed free energies. Preprint *ChemRxiv* 2019, DOI: 10.26434/chemrxiv.8864204.v5.
(27) Brown, K.; Maimaiti, Y.; Trepte, K.; Bligaard, T.; Voss, J. MCML: Combining physical constraints with experimental data for multi-purpose meta-generalized gradient approximation. *J. Comput. Chem.* **2021**, *42*, 2004−2013.
(28) Wellendorff, J.; Lundgaard, K. T.; Møgelhøj, A.; Petzold, V.; Landis, D. D.; Nørskov, J. K.; Bligaard, T.; Jacobsen, K. W. Density functionals for surface science: Exchange-correlation model development with Bayesian error estimation. *Phys. Rev. B* **2012**, *85*, 235149.
(29) Wellendorff, J.; Lundgaard, K. T.; Jacobsen, K. W.; Bligaard, T. mBEEF: An accurate semi-local Bayesian error estimation density functional. *J. Chem. Phys.* **2014**, *140*, 144107.
(30) Lundgaard, K. T.; Wellendorff, J.; Voss, J.; Jacobsen, K. W.; Bligaard, T. mBEEF-vdW: Robust fitting of error estimation density functionals. *Phys. Rev. B* **2016**, *93*, 235162.
(31) Verma, P.; Truhlar, D. G. Status and challenges of density functional theory. *Trends Chem.* **2020**, *2*, 302−318.
(32) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. Design of density functionals by combining the method of constraint satisfaction with parametrization for thermochemistry, thermochemical kinetics, and noncovalent interactions. *J. Chem. Theory Comput.* **2006**, *2*, 364−382.
(33) Peverati, R.; Truhlar, D. G. Communication: A global hybrid generalized gradient approximation to the exchange-correlation functional that satisfies the second-order density-gradient constraint and has broad applicability in chemistry. *J. Chem. Phys.* **2011**, *135*, 191102.
(34) Chan, G. K.-L.; Handy, N. C. An extensive study of gradient approximations to the exchange-correlation and kinetic energy functionals. *J. Chem. Phys.* **2000**, *112*, 5639−5653.
(35) Eilers, P. H. C.; Marx, B. D. Flexible smoothing with B-splines and penalties. *Stat. Sci.* **1996**, *11*, 89−121.
(36) Eilers, P. H. C.; Marx, B. D.; Durbán, M. Twenty years of P-splines. *SORT* **2015**, *39*, 149−186.
(37) Ernzerhof, M.; Scuseria, G. E. Assessment of the Perdew−Burke−Ernzerhof exchange-correlation functional. *J. Chem. Phys.* **1999**, *110*, 5029−5036.
(38) Oliver, G. L.; Perdew, J. P. Spin-density gradient expansion for the kinetic energy. *Phys. Rev. A* **1979**, *20*, 397−403.
(39) Becke, A. D. Density functional calculations of molecular bond energies. *J. Chem. Phys.* **1986**, *84*, 4524−4529.
(40) Perdew, J. P.; Wang, Y. Accurate and simple analytic representation of the electron-gas correlation energy. *Phys. Rev. B* **1992**, *45*, 13244−13249.

https://doi.org/10.1021/acs.jpclett.2c00643
J. Phys. Chem. Lett. 2022, 13, 6896−6904

(41) Levy, M. Asymptotic coordinate scaling bound for exchange-correlation energy in density-functional theory. *Int. J. Quantum Chem.* **1989**, *36*, 617−619.
(42) Although the CASE21 correlation form is able to exactly cancel the LDA logarithmic singularity, the correlation energy completely vanishes in this limit. However, to fully satisfy uniform scaling to the high-density limit for correlation, the correlation energy should be nonzero in this limit, for example, $E_c = -0.0467$ hartree for a two-electron atom as $Z \rightarrow \infty$ (see SI for more details).[4]
(43) Prautzsch, H.; Boehm, W.; Paluszny, M. *Bezier and B-Spline Techniques*, 1st ed.; Springer: Berlin, 2002.
(44) Hansen, P. C. *Rank-Deficient and Discrete Ill-Posed Problems*; SIAM: Philadelphia, 1998.
(45) Strutz, T. *Data Fitting and Uncertainty: A Practical Introduction to Weighted Least Squares and Beyond*, 2nd ed.; Springer Vieweg: Wiesbaden, 2016.
(46) Sparrow, Z. M.; Ernst, B. G.; Joo, P. T.; Lao, K. U.; DiStasio, R. A., Jr. NENCI-2021. I. A large benchmark database of non-equilibrium non-covalent interactions emphasizing close intermolecular contacts. *J. Chem. Phys.* **2021**, *155*, 184303.
(47) Ernst, B. G.; Sparrow, Z. M.; DiStasio, R. A., Jr. NENCI-2021. II. Evaluating the performance of quantum chemical and density functional approximations on the NENCI-2021 benchmark database. Manuscript in preparation, **2022**.
(48) Kang, R.; Sparrow, Z. M.; Ernst, B. G.; DiStasio, R. A., Jr. NECI-2022: A large benchmark database of non-equilibrium covalent interactions. Manuscript in preparation, **2022**.
(49) Yu, H. S.; Zhang, W.; Verma, P.; He, X.; Truhlar, D. G. Nonseparable exchange−correlation functional for molecules, including homogeneous catalysis involving transition metals. *Phys. Chem. Chem. Phys.* **2015**, *17*, 12146−12160.
(50) Petzold, V.; Bligaard, T.; Jacobsen, K. W. Construction of new electronic density functionals with error estimation through fitting. *Top. Catal.* **2012**, *55*, 402−417.
(51) Bollaerts, K.; Eilers, P. H. C.; van Mechelen, I. Simple and multiple P-splines regression with shape constraints. *Br. J. Math. Stat. Psychol.* **2006**, *59*, 451−469.
(52) Levy, M.; Perdew, J. P. Hellmann-Feynman, virial, and scaling requisites for the exact universal density functionals. Shape of the correlation potential and diamagnetic susceptibility for atoms. *Phys. Rev. A* **1985**, *32*, 2010−2021.
(53) Karton, A.; Daon, S.; Martin, J. M. L. W4−11: A high-confidence benchmark dataset for computational thermochemistry derived from first-principles W4 data. *Chem. Phys. Lett.* **2011**, *510*, 165−178.
(54) Chakravorty, S. J.; Gwaltney, S. R.; Davidson, E. R.; Parpia, F. A.; Fischer, C. F. Ground-state correlation energies for atomic ions with 3 to 18 electrons. *Phys. Rev. A* **1993**, *47*, 3649−3670.
(55) Karton, A.; Sylvetsky, N.; Martin, J. M. L. W4−17: A diverse and high-confidence dataset of atomization energies for benchmarking high-level electronic structure methods. *J. Comput. Chem.* **2017**, *38*, 2063−2075.
(56) Morgante, P.; Peverati, R. ACCDB: A collection of chemistry databases for broad computational purposes. *J. Comput. Chem.* **2019**, *40*, 839−848.
(57) Ye, J. On measuring and correcting the effects of data mining and model selection. *J. Am. Stat. Assoc.* **1998**, *93*, 120−131.
(58) Goerigk, L.; Grimme, S. A general database for main group thermochemistry, kinetics, and noncovalent interactions − Assessment of common and reparameterized (*meta*-)GGA density functionals. *J. Chem. Theory Comput.* **2010**, *6*, 107−126.
(59) Zhao, Y.; González-García, N.; Truhlar, D. G. Benchmark database of barrier heights for heavy atom transfer, nucleophilic substitution, association, and unimolecular reactions and its use to test theoretical methods. *J. Phys. Chem. A* **2005**, *109*, 2012−2018.
(60) Zhao, Y.; Lynch, B. J.; Truhlar, D. G. Development and assessment of a new hybrid density functional model for thermochemical kinetics. *J. Phys. Chem. A* **2004**, *108*, 2715−2719.
(61) Goerigk, L.; Hansen, A.; Bauer, C.; Ehrlich, S.; Najibi, A.; Grimme, S. A look at the density functional theory zoo with the advanced GMTKN55 database for general main group thermochemistry, kinetics and noncovalent interactions. *Phys. Chem. Chem. Phys.* **2017**, *19*, 32184−32215.
(62) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. Assessment of Gaussian-2 and density functional theories for the computation of enthalpies of formation. *J. Chem. Phys.* **1997**, *106*, 1063−1079.
(63) Neese, F.; Schwabe, T.; Kossmann, S.; Schirmer, B.; Grimme, S. Assessment of orbital-optimized, spin-component scaled second-order many-body perturbation theory for thermochemistry and kinetics. *J. Chem. Theory Comput.* **2009**, *5*, 3060−3073.
(64) Karton, A.; O'Reilly, R. J.; Radom, L. Assessment of theoretical procedures for calculating barrier heights for a diverse set of water-catalyzed proton-transfer reactions. *J. Phys. Chem. A* **2012**, *116*, 4211−4221.
(65) Yu, H.; Truhlar, D. G. Components of the bond energy in polar diatomic molecules, radicals, and ions formed by group-1 and group-2 metal atoms. *J. Chem. Theory Comput.* **2015**, *11*, 2968−2983.
(66) Parthiban, S.; Martin, J. M. L. Assessment of W1 and W2 theories for the computation of electron affinities, ionization potentials, heats of formation, and proton affinities. *J. Chem. Phys.* **2001**, *114*, 6014−6029.
(67) Zhao, Y.; Truhlar, D. G. Assessment of density functionals for π systems: Energy differences between cumulenes and poly-ynes; proton affinities, bond length alternation, and torsional potentials of conjugated polyenes; and proton affinities of conjugated shiff bases. *J. Phys. Chem. A* **2006**, *110*, 10478−10486.
(68) Curtiss, L. A.; Raghavachari, K.; Trucks, G. W.; Pople, J. A. Gaussian-2 theory for molecular energies of first- and second-row compounds. *J. Chem. Phys.* **1991**, *94*, 7221−7230.
(69) Lynch, B. J.; Zhao, Y.; Truhlar, D. G. Effectiveness of diffuse basis functions for calculating relative energies by density functional theory. *J. Phys. Chem. A* **2003**, *107*, 1384−1388.
(70) Gruzman, D.; Karton, A.; Martin, J. M. L. Performance of ab initio and density functional methods for conformational equilibria of C_nH_{2n+2} alkane isomers (n = 4−8). *J. Phys. Chem. A* **2009**, *113*, 11974−11983.
(71) Wilke, J. J.; Lind, M. C.; Schaefer, H. F., III; Császár, A. G.; Allen, W. D. Conformers of gaseous cysteine. *J. Chem. Theory Comput.* **2009**, *5*, 1511−1523.
(72) Yu, L.-J.; Sarrami, F.; Karton, A.; O'Reilly, R. J. An assessment of theoretical procedures for π-conjugation stabilisation energies in enones. *Mol. Phys.* **2015**, *113*, 1284−1296.
(73) Lao, K. U.; Herbert, J. M. Accurate and efficient quantum chemistry calculations for noncovalent interactions in many-body systems: The XSAPT family of methods. *J. Phys. Chem. A* **2015**, *119*, 235−252.
(74) Grimme, S.; Steinmetz, M.; Korth, M. How to compute isomerization energies of organic molecules with quantum chemical methods. *J. Org. Chem.* **2007**, *72*, 2118−2126.
(75) Luo, S.; Zhao, Y.; Truhlar, D. G. Validation of electronic structure methods for isomerization reactions of large organic molecules. *Phys. Chem. Chem. Phys.* **2011**, *13*, 13683−13689.
(76) Martin, J. M. L. What can we learn about dispersion from the conformer surface of *n*-pentane? *J. Phys. Chem. A* **2013**, *117*, 3118−3132.
(77) Zhao, Y.; Truhlar, D. G. A new local density functional for main-group thermochemistry, transition metal bonding, thermochemical kinetics, and noncovalent interactions. *J. Chem. Phys.* **2006**, *125*, 194101.
(78) Csonka, G. I.; French, A. D.; Johnson, G. P.; Stortz, C. A. Evaluation of density functionals and basis sets for carbohydrates. *J. Chem. Theory Comput.* **2009**, *5*, 679−692.
(79) Mardirossian, N.; Lambrecht, D. S.; McCaslin, L.; Xantheas, S. S.; Head-Gordon, M. The performance of density functionals for sulfate−water clusters. *J. Chem. Theory Comput.* **2013**, *9*, 1368−1380.
(80) Kesharwani, M. K.; Karton, A.; Martin, J. M. L. Benchmark *ab initio* conformational energies for the proteinogenic amino acids through explicitly correlated methods. Assessment of density functional methods. *J. Chem. Theory Comput.* **2016**, *12*, 444−454.

https://doi.org/10.1021/acs.jpclett.2c00643
J. Phys. Chem. Lett. 2022, 13, 6896−6904

(81) Řezáč, J.; Hobza, P. Describing noncovalent interactions beyond the common approximations: How accurate is the "gold standard," CCSD(T) at the complete basis set limit? *J. Chem. Theory Comput.* **2013**, *9*, 2151−2155.
(82) Řezáč, J.; Hobza, P. Advanced corrections of hydrogen bonding and dispersion for semiempirical quantum mechanical methods. *J. Chem. Theory Comput.* **2012**, *8*, 141−151.
(83) Boese, A. D. Assessment of coupled cluster theory and more approximate methods for hydrogen bonded systems. *J. Chem. Theory Comput.* **2013**, *9*, 4403−4413.
(84) Boese, A. D. Basis set limit coupled-cluster studies of hydrogen-bonded systems. *Mol. Phys.* **2015**, *113*, 1618−1629.
(85) Boese, A. D. Density functional theory and hydrogen bonds: Are we there yet? *ChemPhysChem* **2015**, *16*, 978−985.
(86) Smith, D. G. A.; Jankowski, P.; Slawik, M.; Witek, H. A.; Patkowski, K. Basis set convergence of the post-CCSD(T) contribution to noncovalent interaction energies. *J. Chem. Theory Comput.* **2014**, *10*, 3140−3150.
(87) Kozuch, S.; Martin, J. M. L. Halogen bonds: Benchmarks and theoretical analysis. *J. Chem. Theory Comput.* **2013**, *9*, 1918−1931.
(88) Vuckovic, S.; Burke, K. Quantifying and understanding errors in molecular geometries. *J. Phys. Chem. Lett.* **2020**, *11*, 9957−9964.
(89) Piccardo, M.; Penocchio, E.; Puzzarini, C.; Biczysko, M.; Barone, V. Semi-experimental equilibrium structure determinations by employing B3LYP/SNSD anharmonic force fields: Validation and application to semirigid organic molecules. *J. Phys. Chem. A* **2015**, *119*, 2058−2082.
(90) Penocchio, E.; Piccardo, M.; Barone, V. Semiexperimental equilibrium structures for building blocks of organic and biological molecules: The B2PLYP route. *J. Chem. Theory Comput.* **2015**, *11*, 4689−4707.
(91) Chai, J.-D.; Head-Gordon, M. Systematic optimization of long-range corrected hybrid density functionals. *J. Chem. Phys.* **2008**, *128*, 084106.
(92) Treutler, O.; Ahlrichs, R. Efficient molecular numerical integration schemes. *J. Chem. Phys.* **1995**, *102*, 346−354.
(93) Parrish, R. M.; Burns, L. A.; Smith, D. G. A.; Simmonett, A. C.; DePrince, A. E.; Hohenstein, E. G.; Bozkaya, U.; Sokolov, A. Y.; Di Remigio, R.; Richard, R. M.; Gonthier, J. F.; James, A. M.; McAlexander, H. R.; Kumar, A.; Saitow, M.; Wang, X.; Pritchard, B. P.; Verma, P.; Schaefer, H. F., III; Patkowski, K.; King, R. A.; Valeev, E. F.; Evangelista, F. A.; Turney, J. M.; Crawford, T. D.; Sherrill, C. D. Psi4 1.1: An open-source electronic structure program emphasizing automation, advanced libraries, and interoperability. *J. Chem. Theory Comput.* **2017**, *13*, 3185−3197.
(94) Lehtola, S.; Steigemann, C.; Oliveira, M. J. T.; Marques, M. A. L. Recent developments in libxc−A comprehensive library of functionals for density functional theory. *SoftwareX* **2018**, *7*, 1−5.
(95) Weigend, F.; Furche, F.; Ahlrichs, R. Gaussian basis sets of quadruple zeta valence quality for atoms H-Kr. *J. Chem. Phys.* **2003**, *119*, 12753−12762.
(96) Rappoport, D.; Furche, F. Property-optimized Gaussian basis sets for molecular response calculations. *J. Chem. Phys.* **2010**, *133*, 134105.
(97) Schuchardt, K. L.; Didier, B. T.; Elsethagen, T.; Sun, L.; Gurumoorthi, V.; Chase, J.; Li, J.; Windus, T. L. Basis set exchange: A community database for computational sciences. *J. Chem. Inf. Model.* **2007**, *47*, 1045−1052.
(98) Pritchard, B. P.; Altarawy, D.; Didier, B.; Gibson, T. D.; Windus, T. L. New basis set exchange: An open, up-to-date resource for the molecular sciences community. *J. Chem. Inf. Model.* **2019**, *59*, 4814−4820.
(99) Giannozzi, P.; Andreussi, O.; Brumme, T.; Bunau, O.; Nardelli, M. B.; Calandra, M.; Car, R.; Cavazzoni, C.; Ceresoli, D.; Cococcioni, M.; Colonna, N.; Carnimeo, I.; Corso, A. D.; de Gironcoli, S.; Delugas, P.; DiStasio, R. A., Jr.; Ferretti, A.; Floris, A.; Fratesi, G.; Fugallo, G.; Gebauer, R.; Gerstmann, U.; Giustino, F.; Gorni, T.; Jia, J.; Kawamura, M.; Ko, H.-Y.; Kokalj, A.; Küçükbenli, E.; Lazzeri, M.; Marsili, M.; Marzari, N.; Mauri, F.; Nguyen, N. L.; Nguyen, H.-V.; Otero-de-la-Roza, A.; Paulatto, L.; Poncé, S.; Rocca, D.; Sabatini, R.; Santra, B.; Schlipf, M.; Seitsonen, A. P.; Smogunov, A.; Timrov, I.; Thonhauser, T.; Umari, P.; Vast, N.; Wu, X.; Baroni, S. Advanced capabilities for materials modelling with Quantum ESPRESSO. *J. Phys.: Condens. Matter* **2017**, *29*, 465901.
(100) Hamann, D. R.; Schlüter, M.; Chiang, C. Norm-conserving pseudopotentials. *Phys. Rev. Lett.* **1979**, *43*, 1494−1497.
(101) Vanderbilt, D. Optimally smooth norm-conserving pseudopotentials. *Phys. Rev. B* **1985**, *32*, 8412−8415.
(102) Murnaghan, F. D. The compressibility of media under extreme pressures. *Proc. Natl. Acad. Sci. U.S.A.* **1944**, *30*, 244−247.

https://doi.org/10.1021/acs.jpclett.2c00643
J. Phys. Chem. Lett. 2022, 13, 6896−6904