Title
Designing and Assessing Density Functionals

Author
Mardirossian, Narbe

Publication Date
2016

DESIGNING AND ASSESSING DENSITY FUNCTIONALS

by

NARBE MARDIROSSIAN

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Chemistry

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Martin P. Head-Gordon, Chair
Professor William H. Miller
Professor Daryl C. Chrzan

Spring 2016

DESIGNING AND ASSESSING DENSITY FUNCTIONALS

Abstract

DESIGNING AND ASSESSING DENSITY FUNCTIONALS

by

NARBE MARDIROSSIAN

Doctor of Philosophy in Chemistry

University of California, Berkeley

Professor Martin P. Head-Gordon, Chair

This thesis is concerned with the development of minimally-parameterized and highly-transferable density functionals. A methodology for searching a given functional space is developed and used to parameterize three novel functionals: ωB97X-V – a 10-parameter, range-separated hybrid, generalized gradient approximation density functional with VV10 nonlocal correlation, B97M-V – a 12-parameter, local meta-generalized gradient approximation density funcitonal with VV10 nonlocal correlation, and ωB97M-V – a 12-parameter, range-separated hybrid, meta-generalized gradient approximation density functional with VV10 nonlocal correlation. These three functionals are validated by comparisons to the best existing density functionals in their class, and their proper usage (with respect to basis sets and integration grids) is documented to facilitate use in chemical applications.

Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I would like to thank Martin Head-Gordon for being an amazing adviser, teacher, and mentor. I consider myself truly lucky to have crossed paths with him as an undergrad while taking Chem 120A, as I had never been exposed to quantum mechanics or computational chemistry before then. I have learned a great deal from Martin both as an undergraduate and a graduate student, and will definitely miss our weekly Monday afternoon meetings. I am grateful to him for giving me the freedom to work on many interesting projects and for always being a source of encouragement and enthusiasm.

I would like to thank Gaia Tumanyan for being incredibly supportive, caring, and understanding throughout the past five years. Despite the 393 miles of separation, I have very fond memories from the past several years (such as our three wonderful trips to Yosemite and her visits to Berkeley), and look forward to making many more memories in the future (in Yosemite and elsewhere!). I am very excited to move back home so that we can finally see each other more often than just once a month. In addition, I must thank the UC Berkeley Office of Undergraduate Admissions for accepting both of us back in 2007, so that we could inevitably meet on January 24, 2008. The last 8+ years have been a pleasure.

I would like to thank my parents, Alec and Anita Mardirossian, for their unwavering love and support, and for always reminding me to eat fruits and vegetables (and to avoid processed meats). After nine years away from home, it is going to be very exciting for the five of us to finally reunite under one roof. I would like to thank Raffy Mardirossian for being a cool twin brother and Taline Mardirossian for getting accepted to UC Berkeley and giving me a valid reason to visit the best university in the world until at least 2020. I would also like to thank Arbi Derzakharian for checking in on me multiple times a week to make sure I was still alive.

I would like to thank Leslie Silvers for being very helpful with all sorts of administrative issues, including ordering desktops, nodes, Intel compilers, printers, paper, and ink. In addition, I thoroughly enjoyed the numerous conversations over the years about traveling to Yosemite, Alaska, and other interesting locations.

In my seven years (two as an undergraduate and five as a graduate) as part of the Head-Gordon Group, I have had the pleasure of meeting many undergraduate students, graduate students, and post-docs. John Parkhill was the first graduate student I met back in 2009, and I would like to thank him for being a fantastic mentor and for teaching me how to use Q-Chem, Python, UNIX, and for introducing me to DFT. I would like to thank Julian Azar for being a great friend, office mate, and lab mate, and for singing along to Father and Son with the maintenance guy that came to fix the lights in the summer of 2015, while I attempted to play the guitar. I am very lucky to have shared an office with him for six years and will cherish those memories forever. I would like to thank Eric Sundstrom for choosing me as his successor as the group's Systems Administrator, and for teaching me so much about computers and lamps. My experiences in Gilman Hall would have been much different (and much more mundane) had it not been for Paul Horn and Yuezhi Mao, who were always working along with me past midnight in the Pitzer Center. I would also like to thank Andreas Hauser for all of the interesting conversations about music, soccer, and Yosemite, for helping me set soccer ball juggling records in the Pitzer Center, and for continuing to keep in touch.

While it is difficult to mention all of the people that influenced me in the past five years at Berkeley, I would like to thank Samuel Manzer for helping me start the Head-Gordon Group DJ Khaled Fan Club, David Stück for making the infamous "combustion of H_2 in Cl_2" picture that I will hold on to forever, Jonathan Thirman for being the only person who was as determined as me about playing soccer throughout all five years, Jon Witte for being as enthusiastic as me about going to Yosemite, Alec White for giving me a lot of helpful tips about Portland, Jon Weisberg for helping me while I was trying to learn the guitar, Luke Bertels for occasionally falling asleep at his desk in the funniest positions possible, Rostam Razban for providing a loving home for the Logitech Z623 speakers I bought for my brother, Daniel Levine for engaging in enjoyable late-night conversations and venting sessions, Srimukh Prasad for putting up with all of my jokes and attempts at annoying him, Joonho Lee for trying (and failing) to get me sick on several occasions, Daniel Lambrecht for always giving me very valuable and helpful advice, Evgeny Epifanovsky for making my life much easier by integrating Q-Chem with CMake, Ehud Tsivion for using my functionals instead of the Minnesota functionals, and Susi Lehtola for many interesting conversations about DFT.

Despite spending a lot of my time in 48 Gilman, my graduate experience at UC Berkeley would not have been the same without Shant Torosean, Roupen Khanjian, and Varand "Kobe" Onany. I would like to thank Shant for being a great roommate during my first year at graduate school, and for finally returning to the Bay Area so that we could go hiking in Yosemite, road biking in Palo Alto, and mountain biking in the rain, which turned out to be a very painful experience for one of us. I would like to thank Roupen for keeping in touch despite moving to Portland from Berkeley, and for making my two trips to his new hometown very enjoyable. I would also like to thank Varand for a very memorable two months in 2015 when he showed up outside of my apartment door without warning and stayed for eight weeks.

Last, but not least, I would like to thank Sam Juha from Cheese n' Stuff for providing me with delicious, discounted lunches for the past five years.

Finally, I would like to thank Gund for always being there when I needed a friend.

Chapter 1

Introduction

1.1 The Schrödinger Equation

The Schrödinger equation, established in the mid-1920s, describes all aspects of the microscopic world, governing the behavior of electrons and nuclei in atoms, molecules, solids, liquids, and gases. However, its complexity allows exact treatment only for the simplest of systems, and approximations, which diminish its universality and accuracy, must be made for practical applications. Nevertheless, to determine the ground state energy (or an excited state energy) of a system of electrons and nuclei, it is necessary to solve the time-independent Schrödinger equation, given in Equation 1.1 in the form of an eigenvalue problem.

$$\hat{H}\left|\Psi_i\right\rangle = E_i\left|\Psi_i\right\rangle \tag{1.1}$$

The concept of an eigenvalue problem can be understood by considering Equation 1.2. In this simplified case, the operator, $\hat{O}\left(y\right) = \frac{d^2}{dy^2}$, is the second derivative, and the function, $f\left(y\right) = \sin\left(y\right)$, is the trigonometric sine function. The effect of the operator acting on the function is to return the original function unchanged, save for a multiplicative constant. Thus, $\sin\left(y\right)$ is an eigenfunction of the operator, $\frac{d^2}{dy^2}$, with an eigenvalue of -1.

$$\frac{d^2}{dy^2}\sin\left(y\right) = -\sin\left(y\right) \tag{1.2}$$

Similarly, in Equation 1.1, $\left|\Psi_i\right\rangle$ is an eigenfunction of the Hamiltonian operator, $\hat{H}$, with an eigenvalue, E_i, that is the energy of the state characterized by the wave function, $\left|\Psi_i\right\rangle$. The eigenfunctions and eigenvalues in Equation 1.1 are labeled by the index, i, because there are an infinite number of exact solutions to the time-independent Schrödinger equation, with $E_0 \leq E_1 \leq E_2 \leq E_3 \leq \cdots \leq E_i \leq \cdots$.

The lowest-energy solution, $E_0 = \langle\Psi_0|\,\hat{H}\,|\Psi_0\rangle$, is called the ground state energy, and corresponds to the ground state wave function, $|\Psi_0\rangle$.

The Hamiltonian operator can be written as a sum of five terms,

$$\begin{aligned}\hat{H} = &\left[-\sum_{i=1}^{n}\frac{1}{2}\nabla_i^2\right] + \left[-\sum_{A=1}^{M}\frac{1}{2m_A}\nabla_A^2\right] + \\ &\left[\sum_{i=1}^{n}\sum_{j>i}^{n}\frac{1}{|\mathbf{r}_i-\mathbf{r}_j|}\right] + \left[-\sum_{i=1}^{n}\sum_{A=1}^{M}\frac{Z_A}{|\mathbf{r}_i-\mathbf{R}_A|}\right] + \left[\sum_{A=1}^{M}\sum_{B>A}^{M}\frac{Z_A Z_B}{|\mathbf{R}_A-\mathbf{R}_B|}\right],\end{aligned} \tag{1.3}$$

where i and r_i represent the indices and coordinates of the n electrons, A and r_A represent the indices and coordinates of the M nuclei, Z is the atomic number of a given nucleus, and m is the ratio of the mass of a given nucleus to the mass of an electron. The first term in Equation 1.3, $\hat{T}_e$ (or simply, $\hat{T}$), represents the kinetic energy associated with the movement of the electrons, the second term, $\hat{T}_n$, represents the kinetic energy associated with the movement of the nuclei, the third term, $\hat{V}_{ee}$, represents the potential energy associated with the repulsion between the electrons, the fourth term, $\hat{V}_{en}$, represents the potential energy associated with the attraction between the electrons and the nuclei, and the fifth term, $\hat{V}_{nn}$, represents the potential energy associated with the repulsion between the nuclei. Since $\hat{H}$ and $|\Psi_i\rangle$ depend on the coordinates of both the electrons and the nuclei, solving the time-independent Schrödinger equation involves solving a partial differential equation in $3n + 3M$ dimensions.

In order to reduce the dimensionality of the problem at hand, it is necessary to invoke the Born–Oppenheimer approximation. Since the protons and neutrons that comprise the nuclei are at least 1800 times heavier than an electron, the Born–Oppenheimer approximation makes the assumption that the electrons are moving in a field of fixed nuclei. This simplification gives rise to the electronic Hamiltonian (Equation 1.4). Since the nuclei are “clamped”, the second term from Equation 1.3 vanishes, and the fifth term becomes an additive constant to the electronic energy.

$$\hat{H}_e = \hat{T} + \hat{V}_{ee} + \hat{V}_{en} = \left[-\sum_{i=1}^{n}\frac{1}{2}\nabla_i^2\right] + \left[\sum_{i=1}^{n}\sum_{j>i}^{n}\frac{1}{|\mathbf{r}_i-\mathbf{r}_j|}\right] + \left[-\sum_{i=1}^{n}\sum_{A=1}^{M}\frac{Z_A}{|\mathbf{r}_i-\mathbf{R}_A|}\right] \tag{1.4}$$

The new eigenvalue problem associated with the electronic Hamiltonian, which involves an electronic wave function that depends explicitly on the coordinates of the electrons and parametrically on the coordinates of the nuclei, is given in Equation 1.5, and the ground state energy of a system within the Born–Oppenheimer approximation is given in Equation 1.6.

$$\hat{H}_e \left|\Psi_{e,i}\right\rangle = E_{e,i} \left|\Psi_{e,i}\right\rangle \tag{1.5}$$

$$E_{tot,0} = E_{e,0} + \sum_{A=1}^{M} \sum_{B>A}^{M} \frac{Z_A Z_B}{\left|\mathbf{R}_A - \mathbf{R}_B\right|} \tag{1.6}$$

Despite these simplifications, Equation 1.5 is still incredibly difficult to solve, as it is a $3n$-dimensional partial differential equation. In order to make useful approximations to $\left|\Psi_{e,i}\right\rangle$, it is essential to incorporate the statistical properties of electrons into the electronic wave function (henceforth referred to without the "electronic" descriptor). The permutation operator, $\hat{P}_{12}$, which swaps the coordinates of electrons 1 and 2, is relevant in this respect. Because $\hat{P}_{12}$ and $\hat{H}_e$ commute, the wave function is also an eigenfunction of the permutation operator, with an eigenvalue of λ. Since swapping the coordinates of electrons 1 and 2 twice should have no effect, the value of λ can either be 1 or -1,

$$\hat{P}_{12}\hat{P}_{12} \left|\Psi_{e,i}\right\rangle = \lambda^2 \left|\Psi_{e,i}\right\rangle = \left|\Psi_{e,i}\right\rangle . \tag{1.7}$$

A value of $\lambda = 1$ is associated with bosons (e.g. photons), while a value of $\lambda = -1$ is associated with fermions (e.g. electrons). Thus, even the most basic approximation to $\left|\Psi_{e,i}\right\rangle$ must be antisymmetric with respect to the exchange of the spatial and spin coordinates of any two electrons.

The most naive approximation to an n-electron wave function is a product of n spin orbitals (Equation 1.8), each depending on the spatial and spin coordinates of a single electron.

$$\left|\Psi_{e,i}\left(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\right)\right\rangle \approx \left|\Phi_{\mathrm{H}}\left(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\right)\right\rangle = \prod_{i=1}^{n} \left|\phi_i\left(\mathbf{x}_i\right)\right\rangle \tag{1.8}$$

However, as demonstrated in Equation 1.9 for the case of two electrons, this Hartree product wave function does not satisfy the antisymmetry property.

$$\hat{P}_{12} \left|\Phi_{\mathrm{H}}\left(\mathbf{x}_1, \mathbf{x}_2\right)\right\rangle = \hat{P}_{12} \left(\left|\phi_1\left(\mathbf{x}_1\right)\right\rangle \left|\phi_2\left(\mathbf{x}_2\right)\right\rangle\right) = \left|\phi_1\left(\mathbf{x}_2\right)\right\rangle \left|\phi_2\left(\mathbf{x}_1\right)\right\rangle \neq - \left|\phi_1\left(\mathbf{x}_1\right)\right\rangle \left|\phi_2\left(\mathbf{x}_2\right)\right\rangle \tag{1.9}$$

On the other hand, the Slater determinant wave function given in Equation 1.10 is a valid antisymmetric wave function, since $\hat{P}_{12} \left|\Phi_{\mathrm{S}}\left(\mathbf{x}_1, \mathbf{x}_2\right)\right\rangle = - \left|\Phi_{\mathrm{S}}\left(\mathbf{x}_1, \mathbf{x}_2\right)\right\rangle$. Accordingly, $\left|\Psi_{e,i}\right\rangle$ can be approximated as a determinant of spin orbitals.

$$\left|\Phi_{\mathrm{S}}\left(\mathbf{x}_1, \mathbf{x}_2\right)\right\rangle = \frac{1}{\sqrt{2}} \begin{vmatrix} \left|\phi_1\left(\mathbf{x}_1\right)\right\rangle & \left|\phi_2\left(\mathbf{x}_1\right)\right\rangle \\ \left|\phi_1\left(\mathbf{x}_2\right)\right\rangle & \left|\phi_2\left(\mathbf{x}_2\right)\right\rangle \end{vmatrix} = \frac{1}{\sqrt{2}} \left[\left|\phi_1\left(\mathbf{x}_1\right)\right\rangle \left|\phi_2\left(\mathbf{x}_2\right)\right\rangle - \left|\phi_2\left(\mathbf{x}_1\right)\right\rangle \left|\phi_1\left(\mathbf{x}_2\right)\right\rangle\right] \tag{1.10}$$

A normalized Slater determinant with n spin orbitals is shown in Equation 1.11.

$$\left|\Phi_{\mathrm{S}}\left(\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_n\right)\right\rangle = \left(n!\right)^{-1/2}\begin{vmatrix} \left|\phi_1\left(\mathbf{x}_1\right)\right\rangle & \left|\phi_2\left(\mathbf{x}_1\right)\right\rangle & \cdots & \left|\phi_n\left(\mathbf{x}_1\right)\right\rangle \\ \left|\phi_1\left(\mathbf{x}_2\right)\right\rangle & \left|\phi_2\left(\mathbf{x}_2\right)\right\rangle & \cdots & \left|\phi_n\left(\mathbf{x}_2\right)\right\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \left|\phi_1\left(\mathbf{x}_n\right)\right\rangle & \left|\phi_2\left(\mathbf{x}_n\right)\right\rangle & \cdots & \left|\phi_n\left(\mathbf{x}_n\right)\right\rangle \end{vmatrix} \tag{1.11}$$

Since electrons are spin-$\frac{1}{2}$ particles, the spin orbitals, $\left|\phi\left(\mathbf{x}\right)\right\rangle$, that are chosen to represent them have both a spatial component, $\left|\psi\left(\mathbf{r}\right)\right\rangle$, and a spin component, which can either be spin up, $\left|\alpha\left(s\right)\right\rangle$, or spin down, $\left|\beta\left(s\right)\right\rangle$. Therefore, two spin orbitals can be formed from a single spatial orbital, as shown in Equations 1.12 and 1.13.

$$\left|\phi_1\left(\mathbf{x}_1\right)\right\rangle = \left|\psi_1\left(\mathbf{r}_1\right)\right\rangle\left|\alpha\left(s_1\right)\right\rangle \tag{1.12}$$

$$\left|\phi_2\left(\mathbf{x}_2\right)\right\rangle = \left|\psi_1\left(\mathbf{r}_2\right)\right\rangle\left|\beta\left(s_2\right)\right\rangle \tag{1.13}$$

In practice, the spatial orbitals are expanded (Equation 1.14) in a finite basis set of K atomic orbitals, $\left|\omega\left(\mathbf{r}\right)\right\rangle$. For closed-shell systems, a basis set of K atomic orbitals gives $n/2$ occupied spatial orbitals and $K - n/2$ unoccupied (or virtual) spatial orbitals (and equivalently, n occupied spin orbitals and $2K - n$ virtual spin orbitals).

$$\left|\psi_i\left(\mathbf{r}\right)\right\rangle = \sum_{\mu=1}^{K}\left|\omega_\mu\left(\mathbf{r}\right)\right\rangle c_{\mu i} \tag{1.14}$$

Approximating the wave function as a single Slater determinant of spin orbitals, $\left|\Psi_{e,i}\right\rangle \approx \left|\Phi_{\mathrm{S}}\right\rangle$, and finding the Slater determinant that satisfies the variational principle shown in Equation 1.15, is known as the Hartree–Fock (HF) approximation. The Slater determinant that minimizes the expectation value, $\left\langle\Phi_{\mathrm{S}}\right|\hat{H}_e\left|\Phi_{\mathrm{S}}\right\rangle$, is the Hartree–Fock wave function, $\left|\Phi_{\mathrm{HF}}\right\rangle$.

$$E_e^{\mathrm{HF}}\left[\Phi_{\mathrm{HF}}\right] = \min_{\Phi_{\mathrm{S}}}\left\{\left\langle\Phi_{\mathrm{S}}\right|\hat{H}_e\left|\Phi_{\mathrm{S}}\right\rangle\right\} \tag{1.15}$$

The Hartree–Fock approximation treats the interactions between electrons in a mean-field manner, meaning that the interaction of each electron with the rest of the electrons is handled as a single averaged interaction. In order to improve upon the Hartree–Fock wave function, a linear combination of singly- through n-tuply-excited determinants generated from the Hartree–Fock reference can be utilized, in addition to the Hartree–Fock wave function. This is known as the full configuration interaction (FCI) wave function (Equation 1.16), and can be truncated at all singly- and doubly-excited determinants to give the CISD wave function, all singly-, doubly-, and triply-excited determinants to give the CISDT wave function, and so on. Despite appearing as a straightforward path to the exact answer, the FCI method scales factorially with the number of spin orbitals, making it prohibitively expensive for systems with more than a few electrons.

For a system with n electrons and $2K$ spin orbitals, the total number of determinants in the FCI wave function is the binomial coefficient, $T = \binom{2K}{n} = \frac{(2K)!}{(2K-n)!n!}$. This can be broken down into a sum over the number of determinants of each excitation type (zero, single, double,..., n-tuple): $\binom{2K}{n} = \sum_{i=0}^{n} \binom{n}{i}\binom{2K-n}{i}$. As an example, the FCI wave function for methane ($n = 10$) in a moderately-sized atomic orbital basis set ($K = 42$) contains 2,761,025,887,620 determinants!

$$|\Phi_{\text{FCI}}\rangle = \sum_{i=1}^{T} |\Phi_{\text{FCI},i}\rangle = c_0 |\Phi_{\text{HF}}\rangle + \sum_{i=1}^{n} \sum_{a=1}^{2K-n} c_i^a |\Phi_i^a\rangle + \sum_{i=1}^{n} \sum_{j>i}^{n} \sum_{a=1}^{2K-n} \sum_{b>a}^{2K-n} c_{ij}^{ab} |\Phi_{ij}^{ab}\rangle + \cdots \tag{1.16}$$

At the complete basis set (CBS) limit (i.e. $K = \infty$), the FCI wave function is a linear combination of an infinite number of determinants (or wave functions), much like the exact electronic wave function, $|\Psi_e\rangle = \sum_{i=0}^{\infty} |\Psi_{e,i}\rangle$ (Equation 1.5). The basis formed by the determinants in the FCI wave function at the CBS limit is an example of a basis in which the electronic Schrödinger equation can be solved exactly. Unfortunately, it is an infinite basis and T must be finite in practice. Truncating the size of the FCI wave function can be achieved either by decreasing the size of the atomic orbital basis set (K) or by only considering certain types of excitations (CISD, CISDT, etc.). Solving the electronic Schrödinger equation with a truncated FCI wave function is accomplished (most primitively) by diagonalizing a matrix that corresponds to the representation of the electronic Hamiltonian in the basis of the T' selected determinants. Thus, to solve for the ground state and excited state energies of the aforementioned methane molecule, it would be necessary to not only diagonalize a 2,761,025,887,620 x 2,761,025,887,620 matrix, but to first determine the non-vanishing values of its roughly 10^{24} matrix elements.

Solving Equation 1.5 with the full FCI wave function gives the exact electronic ground state energy, $E_{e,0}^{\text{FCI}}$, within a given atomic orbital basis set (as well as $T - 1$ exact electronic excited state energies). At the CBS limit, $E_{e,0}^{\text{FCI}}$ is equal to the exact electronic ground state energy, $E_{e,0} = E_{e,0}^{\text{FCI}}$. The difference between this energy and the electronic Hartree–Fock energy (at the CBS limit) is called the correlation energy, $E_{corr} = E_{e,0}^{\text{FCI}} - E_e^{\text{HF}}$. A brief derivation of the Hartree–Fock equations as well as an introduction to the procedure used to solve for the Hartree–Fock energy and wave function will be given in Section 1.2. In order to account for the correlation energy that is missing from the Hartree–Fock approximation by definition, an alternate method known as density functional theory will be introduced in Section 1.3.

1.2 Hartree–Fock

The first step in deriving the Hartree–Fock equations is to determine an expression for $E_e^{\mathrm{S}}\left[\Phi_{\mathrm{S}}\right] = \langle\Phi_{\mathrm{S}}|\,\hat{H}_e\,|\Phi_{\mathrm{S}}\rangle$ as a function of the spin orbitals. Since the electronic Hamiltonian is composed of a pair of one-electron operators ($\hat{T}$ and $\hat{V}_{en}$) and a two-electron operator ($\hat{V}_{ee}$), it can be rewritten in terms of the two distinct types of operators described in Equations 1.17 and 1.18, as $\hat{H}_e = \hat{O}_1 + \hat{O}_2$.

$$\hat{O}_1 = \sum_{i=1}^{n} \hat{h}\left(\mathbf{r}_i\right) \qquad \hat{h}\left(\mathbf{r}_i\right) = \hat{t}\left(\mathbf{r}_i\right) + \hat{v}\left(\mathbf{r}_i\right) = -\frac{1}{2}\nabla_i^2 - \sum_{A=1}^{M} \frac{Z_A}{\left|\mathbf{r}_i - \mathbf{R}_A\right|} \tag{1.17}$$

$$\hat{O}_2 = \sum_{i=1}^{n}\sum_{j>i}^{n} \hat{g}\left(\mathbf{r}_i, \mathbf{r}_j\right) \qquad \hat{g}\left(\mathbf{r}_i, \mathbf{r}_j\right) = \frac{1}{\left|\mathbf{r}_i - \mathbf{r}_j\right|} = \frac{1}{r_{12}} \tag{1.18}$$

A set of guidelines known as the Slater–Condon rules can be used to work out the expression for $E_e^{\mathrm{S}}\left[\{\phi_i\}\right]$, which is given in Equation 1.19.

$$E_e^{\mathrm{S}}\left[\{\phi_i\}\right] = \sum_{i=1}^{n} \langle\phi_i|\,\hat{h}\,|\phi_i\rangle + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\langle\phi_i\phi_j|\phi_i\phi_j\rangle - \langle\phi_i\phi_j|\phi_j\phi_i\rangle\right] \tag{1.19}$$

Before continuing, it is necessary to explain the notation that has been used in Equation 1.19. This notation (known as "bra-ket" notation) is generally used in quantum mechanics and was introduced by Paul Dirac in the late 1930s. For the purposes of this introduction, the term in the first summation represents a single integral over one variable, while the terms in the second summation represent double integrals over two variables. The notation is further explained in Equations 1.20-1.22.

$$\langle\phi_i|\phi_i\rangle = \int \left[\phi_i^*\left(\mathbf{x}_1\right)\phi_i\left(\mathbf{x}_1\right)\right] \mathrm{d}\mathbf{x}_1 \tag{1.20}$$

$$\langle\phi_i|\,\hat{h}\,|\phi_i\rangle = \int \left[\phi_i^*\left(\mathbf{x}_1\right)\hat{h}\left(\mathbf{r}_1\right)\phi_i\left(\mathbf{x}_1\right)\right] \mathrm{d}\mathbf{x}_1 \tag{1.21}$$

$$\langle\phi_i\phi_j|\phi_i\phi_j\rangle = \int\int \left[\phi_i^*\left(\mathbf{x}_1\right)\phi_j^*\left(\mathbf{x}_2\right)\frac{1}{r_{12}}\phi_i\left(\mathbf{x}_1\right)\phi_j\left(\mathbf{x}_2\right)\right] \mathrm{d}\mathbf{x}_1\mathrm{d}\mathbf{x}_2 \tag{1.22}$$

The task at hand is to minimize the expression given in Equation 1.19 with respect to the spin orbitals, under the constraint that the spin orbitals remain orthogonal, $\langle\phi_i|\phi_j\rangle = \delta_{ij}$. The method of Lagrange multipliers can be used to this effect, where the first variation in the Lagrangian given in Equation 1.23 is set to zero, $\delta\mathcal{L} = 0$.

$$\mathcal{L}\left[\{\phi_i\}\right] = E_e^{\mathrm{S}}\left[\{\phi_i\}\right] - \sum_{i=1}^{n}\sum_{j=1}^{n}\epsilon_{ij}\left[\langle\phi_i|\phi_j\rangle - \delta_{ij}\right] \tag{1.23}$$

Eventually, the stationary condition leads to, $\hat{f}\left|\phi_i\right\rangle = \sum_{j=1}^{n} \epsilon_{ij}\left|\phi_j\right\rangle$, which can be further simplified by choosing a special set of (canonical) spin orbitals that zero the off-diagonal Lagrange multipliers to give the eigenvalue equation, $\hat{f}\left|\phi_i\right\rangle = \epsilon_i\left|\phi_i\right\rangle$. Since the Fock operator, $\hat{f}\left(\mathbf{x}_1\right) = \hat{h}\left(\mathbf{r}_1\right) + \sum_{i=1}^{n}\int\left[\phi_i^*\left(\mathbf{x}_2\right)\frac{1}{r_{12}}\left[1-\hat{P}_{12}\right]\phi_i\left(\mathbf{x}_2\right)\right]\mathrm{d}\mathbf{x}_2$, depends on the spin orbitals, an iterative method must be used to solve the Hartree–Fock equations. The spin-dependence of the Fock operator can be integrated out to give Equation 1.24, which is written in terms of spatial orbitals.

$$\hat{f}\left(\mathbf{r}_1\right) = \hat{h}\left(\mathbf{r}_1\right) + \sum_{i=1}^{n/2}\int\left[\psi_i^*\left(\mathbf{r}_2\right)\frac{1}{r_{12}}\left[2-\hat{P}_{12}\right]\psi_i\left(\mathbf{r}_2\right)\right]\mathrm{d}\mathbf{r}_2 \qquad \hat{f}\left|\psi_i\right\rangle = \epsilon_i\left|\psi_i\right\rangle \tag{1.24}$$

By expanding the spatial orbitals in the eigenvalue equation above in an atomic orbital basis set, $\left|\psi_i\right\rangle = \sum_{\nu=1}^{K}\left|\omega_\nu\right\rangle c_{\nu i}$, and left-projecting with $\left\langle\omega_\mu\right|$, a generalized eigenvalue problem, $\sum_{\nu=1}^{K} F_{\mu\nu}c_{\nu i} = \sum_{\nu=1}^{K} S_{\mu\nu}c_{\nu i}\epsilon_i$, emerges. This can be written in matrix form as $\mathbf{FC} = \mathbf{SC}\boldsymbol{\epsilon}$. While the elements of the overlap matrix, $S_{\mu\nu} = \left\langle\omega_\mu|\omega_\nu\right\rangle$, are straightforward to compute, the elements of the Fock matrix, $F_{\mu\nu} = H_{\mu\nu}^{core} + G_{\mu\nu}$ are given in Equations 1.26 and 1.27, where $P_{\mu\nu} = 2\sum_{i=1}^{n/2} c_{\mu i}c_{\nu i}^*$ is the one-particle density matrix and the Hartree–Fock electron density is given in Equation 1.25.

$$\rho^{\mathrm{HF}}\left(\mathbf{r}\right) = 2\sum_{i=1}^{n/2}\left|\psi_i\left(\mathbf{r}\right)\right|^2 = \sum_{\mu=1}^{K}\sum_{\nu=1}^{K} P_{\mu\nu}\omega_\mu\left(\mathbf{r}\right)\omega_\nu^*\left(\mathbf{r}\right) \tag{1.25}$$

$H_{\mu\nu}^{core}$ represents the elements of the core Hamiltonian matrix and corresponds to one-electron contributions, while $G_{\mu\nu}$ corresponds to two-electron contributions and involves the computation of the expensive four-center, two-electron repulsion integrals (ERI), $\left\langle\omega_\mu\omega_\sigma|\omega_\nu\omega_\lambda\right\rangle$. As shown in Equation 1.27, $G_{\mu\nu}$ can be further separated into two terms: $J_{\mu\nu}$ and $K_{\mu\nu}$. The first term represents the elements of the Coulomb matrix and corresponds to classical electron-electron repulsions, while the second term represents the elements of the exchange matrix and corresponds to non-classical exchange effects between same-spin electrons that arise from antisymmetrization.

$$H_{\mu\nu}^{core} = T_{\mu\nu} + V_{\mu\nu} = \left\langle\omega_\mu\right|\hat{t}\left|\omega_\nu\right\rangle + \left\langle\omega_\mu\right|\hat{v}\left|\omega_\nu\right\rangle \tag{1.26}$$

$$G_{\mu\nu} = J_{\mu\nu} + K_{\mu\nu} = \sum_{\lambda=1}^{K}\sum_{\sigma=1}^{K} P_{\lambda\sigma}\left\langle\omega_\mu\omega_\sigma|\omega_\nu\omega_\lambda\right\rangle - \frac{1}{2}\sum_{\lambda=1}^{K}\sum_{\sigma=1}^{K} P_{\lambda\sigma}\left\langle\omega_\mu\omega_\sigma|\omega_\lambda\omega_\nu\right\rangle \tag{1.27}$$

The total Hartree–Fock energy (Equation 1.28) can be expressed in terms of the elements of the density matrix, the core Hamiltonian matrix, and the Fock matrix.

$$E_{tot}^{HF} = \frac{1}{2}\sum_{\mu=1}^{K}\sum_{\nu=1}^{K} P_{\nu\mu}\left[H_{\mu\nu}^{core} + F_{\mu\nu}\right] + \sum_{A=1}^{M}\sum_{B>A}^{M}\frac{Z_A Z_B}{\left|\mathbf{R}_A - \mathbf{R}_B\right|} \tag{1.28}$$

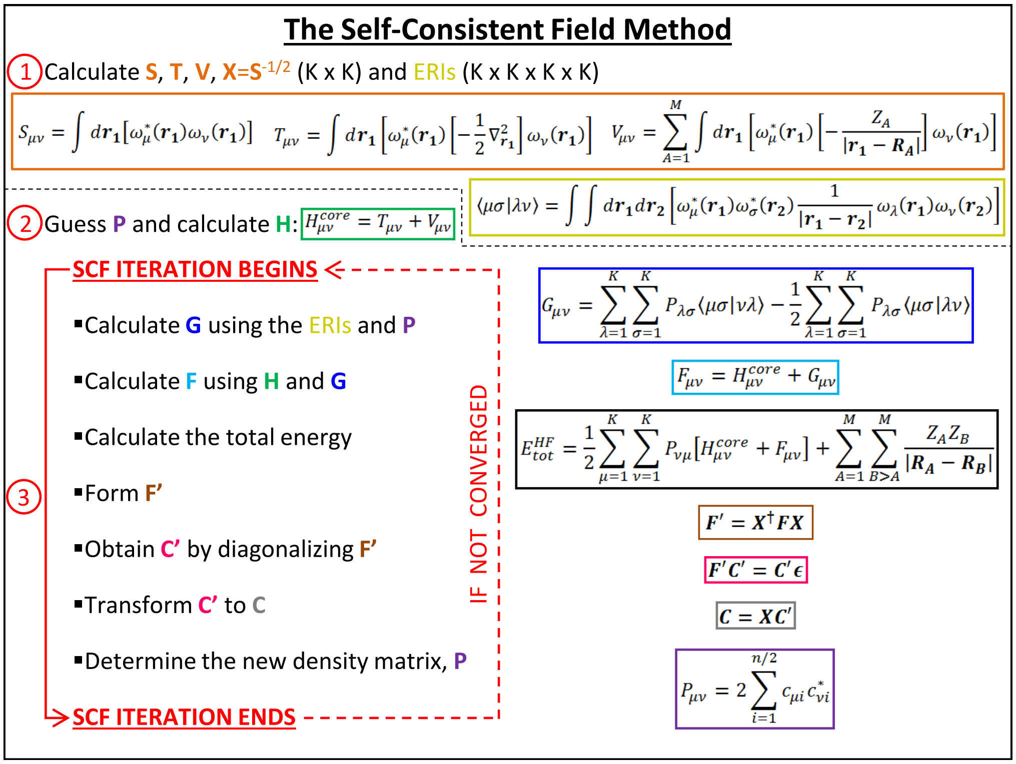


Figure 1.1: A graphical representation of the self-consistent field (SCF) method. K is the number of atomic orbital basis functions.

In order to solve the generalized eigenvalue problem, $\mathbf{FC} = \mathbf{SC}\boldsymbol{\epsilon}$, it is necessary to define a transformation matrix, $\mathbf{X} = \mathbf{S}^{-1/2}$, such that $\mathbf{X}^\dagger\mathbf{SX} = \mathbf{1}$, and define a matrix $\mathbf{C}^{'}$, such that $\mathbf{C} = \mathbf{XC}^{'}$. With the help of this scheme (called symmetric orthogonalization), it is possible to convert the generalized eigenvalue problem into an ordinary eigenvalue problem, $\mathbf{F}^{'}\mathbf{C}^{'} = \mathbf{C}^{'}\boldsymbol{\epsilon}$, by following the steps in Equation 1.29.

$$\begin{aligned}
\mathbf{FC} &= \mathbf{SC}\boldsymbol{\epsilon} \\
\mathbf{FXC}^{'} &= \mathbf{SXC}^{'}\boldsymbol{\epsilon} \\
\mathbf{X}^\dagger\mathbf{FXC}^{'} &= \mathbf{X}^\dagger\mathbf{SXC}^{'}\boldsymbol{\epsilon} \\
\mathbf{F}^{'} &= \mathbf{X}^\dagger\mathbf{FX} \\
\mathbf{F}^{'}\mathbf{C}^{'} &= \mathbf{C}^{'}\boldsymbol{\epsilon}
\end{aligned} \tag{1.29}$$

Since the Fock matrix depends on its eigenvectors, the minimization of the total Hartree–Fock energy uses an iterative procedure known as the self-consistent field (SCF) method. A graphical representation of the SCF method is given in Figure 1.1.

1.3 Density Functional Theory

1.3.1 Theoretical Aspects

In the Hartree–Fock approximation, the electronic ground state energy is determined by minimizing the expectation value, $\langle\Phi_{\mathrm{S}}|\,\hat{H}_e\,|\Phi_{\mathrm{S}}\rangle$. In general, the exact electronic ground state energy is governed by the variational principle,

$$E_{e,0}\left[\Psi_{e,0}\right] = \min_{\Psi_{e,i}} \left\{ \langle\Psi_{e,i}|\,\hat{H}_e\,|\Psi_{e,i}\rangle \right\}. \tag{1.30}$$

Evidently, the electronic Hamiltonian contains all of the information necessary to resolve the electronic ground state properties of the associated system. Since the form of the electronic Hamiltonian is determined by the number of electrons, n, and the external potential, $\hat{v}\left(\mathbf{r}\right) = -\sum_{A=1}^{M} \frac{Z_A}{|\mathbf{r}-\mathbf{R}_A|}$, n and $\hat{v}\left(\mathbf{r}\right)$ completely fix $\hat{H}_e$ and determine the electronic ground state properties of a system (energy, wave function, etc.). The premise for this section is the notion that the (ground state electron) density, $\rho\left(\mathbf{r}\right)$, given in Equation 1.31, can alternatively be used to determine n and $\hat{v}\left(\mathbf{r}\right)$ and, consequently, the electronic ground state properties of a system. While the proofs that follow are extendable to degenerate ground states, their presentation in this section is limited solely non-degenerate ground states.

$$\rho\left(\mathbf{r}\right) = n \int\int\cdots\int \Psi_{e,0}\left(\mathbf{r}, s, \mathbf{x}_2, \cdots, \mathbf{x}_n\right) \Psi_{e,0}^{*}\left(\mathbf{r}, s, \mathbf{x}_2, \cdots, \mathbf{x}_n\right) ds\mathrm{d}\mathbf{x}_2 \cdots \mathrm{d}\mathbf{x}_n \tag{1.31}$$

Demonstrating that $\rho\left(\mathbf{r}\right)$ determines the number of electrons is straightforward, since $\int \rho\left(\mathbf{r}\right) \mathrm{d}\mathbf{r} = n$. On the other hand, demonstrating that the density *uniquely* defines the external potential is less trivial. It is certainly plausible that the density contains enough information to elucidate $\hat{v}\left(\mathbf{r}\right)$, since $\rho\left(\mathbf{r}\right)$ has sharp peaks at the coordinates of the nuclei, with a slope that is proportional to the corresponding atomic number. Unfortunately, this intuitive interpretation is not sufficient to prove that a 1:1 mapping between the density and the external potential exists (i.e. two external potentials that differ by more than a constant cannot correspond to the same density).

Fortunately, in 1964, Pierre Hohenberg and Walter Kohn (HK) proved two fundamental theorems[1] that established the existence of a formally exact theory that utilizes the density as the basic variable for determining the electronic ground state properties of a system: density functional theory (DFT). The first theorem, HKI, demonstrates a 1:1 mapping between the density and the external potential via contradiction. Suppose there exist two Hamiltonians ($\hat{H}_e = \hat{T} + \hat{V}_{ee} + \hat{V}_{en}$ and $\hat{H}_e^{'} = \hat{T} + \hat{V}_{ee} + \hat{V}_{en}^{'}$), corresponding to two ground state wave functions ($|\Psi_{e,0}\rangle$ and

$|\Psi'_{e,0}\rangle$) that integrate to the same density, $\rho(\mathbf{r})$. The proof (Equation 1.32) proceeds by demonstrating that two expressions of the variational principle produce contradictory results, indicating that two different external potentials cannot give rise to the same density.

$$\begin{array}{cc}
\langle\Psi'_{e,0}|\,\hat{H}_e\,|\Psi'_{e,0}\rangle > \langle\Psi_{e,0}|\,\hat{H}_e\,|\Psi_{e,0}\rangle & \langle\Psi_{e,0}|\,\hat{H}'_e\,|\Psi_{e,0}\rangle > \langle\Psi'_{e,0}|\,\hat{H}'_e\,|\Psi'_{e,0}\rangle \\[1em]
\langle\Psi'_{e,0}|\,\hat{T}+\hat{V}_{ee}\,|\Psi'_{e,0}\rangle + \int \hat{v}(\mathbf{r})\,\rho(\mathbf{r})\,\mathrm{d}\mathbf{r} > & \langle\Psi_{e,0}|\,\hat{T}+\hat{V}_{ee}\,|\Psi_{e,0}\rangle + \int \hat{v}'(\mathbf{r})\,\rho(\mathbf{r})\,\mathrm{d}\mathbf{r} > \\
\langle\Psi_{e,0}|\,\hat{T}+\hat{V}_{ee}\,|\Psi_{e,0}\rangle + \int \hat{v}(\mathbf{r})\,\rho(\mathbf{r})\,\mathrm{d}\mathbf{r} & \langle\Psi'_{e,0}|\,\hat{T}+\hat{V}_{ee}\,|\Psi'_{e,0}\rangle + \int \hat{v}'(\mathbf{r})\,\rho(\mathbf{r})\,\mathrm{d}\mathbf{r} \\[1em]
\langle\Psi'_{e,0}|\,\hat{T}+\hat{V}_{ee}\,|\Psi'_{e,0}\rangle > \langle\Psi_{e,0}|\,\hat{T}+\hat{V}_{ee}\,|\Psi_{e,0}\rangle & \langle\Psi_{e,0}|\,\hat{T}+\hat{V}_{ee}\,|\Psi_{e,0}\rangle > \langle\Psi'_{e,0}|\,\hat{T}+\hat{V}_{ee}\,|\Psi'_{e,0}\rangle
\end{array} \tag{1.32}$$

HKI proves that the electronic ground state energy of a system can be written as a functional of the density, $E_{e,0}[\rho(\mathbf{r})] = \langle\Psi_{e,0}|\,\hat{H}_e\,|\Psi_{e,0}\rangle$. This expression can be partitioned into a universal functional that is independent of the external potential, $F_{HK}[\rho(\mathbf{r})] = \langle\Psi_{e,0}|\,\hat{T}+\hat{V}_{ee}\,|\Psi_{e,0}\rangle$, and a system-specific functional that corresponds to the external potential, $\int \hat{v}(\mathbf{r})\,\rho(\mathbf{r})\,\mathrm{d}\mathbf{r} = \langle\Psi_{e,0}|\,\hat{V}_{en}\,|\Psi_{e,0}\rangle$, to give the expression,

$$E_{e,0}[\rho(\mathbf{r})] = F_{HK}[\rho(\mathbf{r})] + \int \hat{v}(\mathbf{r})\,\rho(\mathbf{r})\,\mathrm{d}\mathbf{r}. \tag{1.33}$$

Although $E_{e,0}[\rho(\mathbf{r})]$ has been separated into two components above, it is important not to lose sight of the fact that $\rho(\mathbf{r})$ corresponds to $|\Psi_{e,0}\rangle$, which corresponds to $\hat{v}(\mathbf{r})$. Thus, $F_{HK}[\rho(\mathbf{r})]$ is only valid for densities that come from an antisymmetric ground state wave function that corresponds to a valid external potential (a condition called v-representability). In general, Equation 1.33 can take any valid trial density, $\tilde{\rho}(\mathbf{r})$, as its input: $E_{e,v}[\tilde{\rho}(\mathbf{r})] = F_{HK}[\tilde{\rho}(\mathbf{r})] + \int \hat{v}(\mathbf{r})\,\tilde{\rho}(\mathbf{r})\,\mathrm{d}\mathbf{r}$. However, it is necessary to prove that $E_{e,v}[\tilde{\rho}(\mathbf{r})]$ returns the true electronic ground state energy when its input is the true ground state density. The second theorem, HKII, establishes a variational principle for $E_{e,v}[\rho(\mathbf{r})]$ by alluding to the variational principle mentioned earlier in Equation 1.30. The proof for HKII is reproduced in Equation 1.34 and establishes that $E_{e,v}[\tilde{\rho}(\mathbf{r})] > E_{e,v}[\rho(\mathbf{r})] = E_{e,0}[\rho(\mathbf{r})]$. $\tilde{\Psi}_{e,0}$ represents a ground state wave function that corresponds to a different external potential, $\tilde{\hat{v}}(\mathbf{r})$, and integrates to give $\tilde{\rho}(\mathbf{r})$. It is important to stress, once again, that this variational principle depends on the fact that $\tilde{\rho}(\mathbf{r})$ is associated with an antisymmetric ground state wave function that corresponds to a valid external potential (i.e. it must be v-representable).

$$\begin{gathered}
\langle\tilde{\Psi}_{e,0}|\,\hat{H}_e\,|\tilde{\Psi}_{e,0}\rangle > \langle\Psi_{e,0}|\,\hat{H}_e\,|\Psi_{e,0}\rangle \\
\langle\tilde{\Psi}_{e,0}|\,\hat{T}+\hat{V}_{ee}\,|\tilde{\Psi}_{e,0}\rangle + \langle\tilde{\Psi}_{e,0}|\,\hat{V}_{en}\,|\tilde{\Psi}_{e,0}\rangle > \langle\Psi_{e,0}|\,\hat{T}+\hat{V}_{ee}\,|\Psi_{e,0}\rangle + \langle\Psi_{e,0}|\,\hat{V}_{en}\,|\Psi_{e,0}\rangle \\
F_{HK}\left[\tilde{\rho}\left(\mathbf{r}\right)\right] + \int \hat{v}\left(\mathbf{r}\right)\tilde{\rho}\left(\mathbf{r}\right)\mathrm{d}\mathbf{r} > F_{HK}\left[\rho\left(\mathbf{r}\right)\right] + \int \hat{v}\left(\mathbf{r}\right)\rho\left(\mathbf{r}\right)\mathrm{d}\mathbf{r} \\
E_{e,v}\left[\tilde{\rho}\left(\mathbf{r}\right)\right] > E_{e,v}\left[\rho\left(\mathbf{r}\right)\right] = E_{e,0}\left[\rho\left(\mathbf{r}\right)\right]
\end{gathered} \tag{1.34}$$

HKII indicates that determining the electronic ground state energy of a system requires the minimization of $E_{e,v}\left[\tilde{\rho}\left(\mathbf{r}\right)\right]$. In other words, searching through all valid trial densities, $\tilde{\rho}\left(\mathbf{r}\right)$, the one that returns the lowest value for $E_{e,v}\left[\tilde{\rho}\left(\mathbf{r}\right)\right]$ is the true ground state density. Unfortunately, both the universal functional, $F_{HK}\left[\rho\left(\mathbf{r}\right)\right]$, and HKII require that the trial densities be v-representable. However, the criteria for demonstrating that a trial density is v-representable are unknown. The inability to check a trial density for v-representability presents a problem for practical implementations, since it is possible (during a variational search) to land on a density that is not v-representable, at which point the variational principle provided by HKII is no longer valid.

Fortunately, the v-representability dilemma was circumvented by Mel Levy in the late 1970s. Levy proposed a constrained search[2] that replaced the v-representability requirement with the looser n-representability requirement, namely, that the density must integrate to give the number of electrons, must be positive, must be continuous, and must come from an antisymmetric wave function. The idea behind Levy's constrained search can be understood by considering an n-representable density, $\bar{\rho}\left(\mathbf{r}\right)$. In principle, there are an infinite number of (ground or excited state) antisymmetric wave functions that can give $\bar{\rho}\left(\mathbf{r}\right)$ through integration (Equation 1.31). In order to distinguish the true ground state wave function, $|\Psi^{\bar{\rho}}_{e,0}\rangle$, from the rest, $|\Psi^{\bar{\rho}}_{e,i}\rangle$, it is necessary to invoke the variational principle, as shown in Equation 1.35.

$$\begin{gathered}
\langle\Psi^{\bar{\rho}}_{e,i}|\,\hat{H}_e\,|\Psi^{\bar{\rho}}_{e,i}\rangle > \langle\Psi^{\bar{\rho}}_{e,0}|\,\hat{H}_e\,|\Psi^{\bar{\rho}}_{e,0}\rangle \\
\langle\Psi^{\bar{\rho}}_{e,i}|\,\hat{T}+\hat{V}_{ee}\,|\Psi^{\bar{\rho}}_{e,i}\rangle + \int \hat{v}\left(\mathbf{r}\right)\bar{\rho}\left(\mathbf{r}\right)\mathrm{d}\mathbf{r} > \langle\Psi^{\bar{\rho}}_{e,0}|\,\hat{T}+\hat{V}_{ee}\,|\Psi^{\bar{\rho}}_{e,0}\rangle + \int \hat{v}\left(\mathbf{r}\right)\bar{\rho}\left(\mathbf{r}\right)\mathrm{d}\mathbf{r} \\
\langle\Psi^{\bar{\rho}}_{e,i}|\,\hat{T}+\hat{V}_{ee}\,|\Psi^{\bar{\rho}}_{e,i}\rangle > \langle\Psi^{\bar{\rho}}_{e,0}|\,\hat{T}+\hat{V}_{ee}\,|\Psi^{\bar{\rho}}_{e,0}\rangle
\end{gathered} \tag{1.35}$$

Accordingly, the true ground state wave function that integrates to $\bar{\rho}\left(\mathbf{r}\right)$ minimizes the sum of the kinetic and electron-electron repulsion energies,

$$F\left[\bar{\rho}\left(\mathbf{r}\right)\right] = \min_{\Psi^{\bar{\rho}}_{e,i}\to\bar{\rho}} \langle\Psi^{\bar{\rho}}_{e,i}|\,\hat{T}+\hat{V}_{ee}\,|\Psi^{\bar{\rho}}_{e,i}\rangle\,. \tag{1.36}$$

If $\bar{\rho}(\mathbf{r})$ is replaced with $\rho(\mathbf{r})$ in Equation 1.36, then the expression is equivalent to the HK universal functional, $F_{HK}[\rho(\mathbf{r})] = \min\limits_{\Psi_{e,i}\to\rho} \langle\Psi_{e,i}|\,\hat{T} + \hat{V}_{ee}\,|\Psi_{e,i}\rangle$. Thus, the HK universal functional can be reformulated as a constrained search over n-representable densities. With the v-representability constraint relaxed, the energy density functional can be written in terms of $F[\bar{\rho}(\mathbf{r})]$,

$$E_e[\bar{\rho}(\mathbf{r})] = F[\bar{\rho}(\mathbf{r})] + \int \hat{v}(\mathbf{r})\,\bar{\rho}(\mathbf{r})\,\mathrm{d}\mathbf{r}. \tag{1.37}$$

The ground state energy is given by,

$$E_{e,0}[\rho(\mathbf{r})] = \min_{\bar{\rho}} E_e[\bar{\rho}(\mathbf{r})] = \min_{\bar{\rho}} \left\{ F[\bar{\rho}(\mathbf{r})] + \int \hat{v}(\mathbf{r})\,\bar{\rho}(\mathbf{r})\,\mathrm{d}\mathbf{r} \right\}. \tag{1.38}$$

Thus, the Levy constrained search works by partitioning the Hilbert space (i.e. the wave function space) into sections that correspond to different n-representable densities. All of the wave functions that integrate to a given density are confined to a separate section. $F[\rho(\mathbf{r})]$ identifies the ground state wave function within each of these sections, while the minimization in Equation 1.38 takes place over all of these sections.

Ultimately, the electronic energy can be written as a functional of the density, $E_e[\rho(\mathbf{r})] = F[\rho(\mathbf{r})] + V_{en}[\rho(\mathbf{r})]$, where $F[\rho(\mathbf{r})] = T[\rho(\mathbf{r})] + V_{ee}[\rho(\mathbf{r})]$. The task that remains is to find an accurate approximation to $F[\rho(\mathbf{r})]$. $V_{ee}[\rho(\mathbf{r})]$ can be separated into two components, $J[\rho(\mathbf{r})]$, which corresponds to the classical electron-electron repulsion energy, and $Q[\rho(\mathbf{r})]$, which corresponds to the non-classical (quantum) electron-electron interaction energy. This gives the energy density functional,

$$E_e[\rho(\mathbf{r})] = T[\rho(\mathbf{r})] + V_{en}[\rho(\mathbf{r})] + J[\rho(\mathbf{r})] + Q[\rho(\mathbf{r})]. \tag{1.39}$$

The second and third terms in Equation 1.39 are known and can be computed according to Equations 1.40 and 1.41, respectively.

$$V_{en}[\rho(\mathbf{r})] = \int \hat{v}(\mathbf{r})\,\rho(\mathbf{r})\,\mathrm{d}\mathbf{r} \tag{1.40}$$

$$J[\rho(\mathbf{r})] = \frac{1}{2}\int\!\!\int \frac{\rho(\mathbf{r}_1)\,\rho(\mathbf{r}_2)}{r_{12}}\mathrm{d}\mathbf{r}_1\mathrm{d}\mathbf{r}_2 \tag{1.41}$$

Thus, it is only necessary to find accurate approximations to $T[\rho(\mathbf{r})]$ and $Q[\rho(\mathbf{r})]$. Since the kinetic energy contribution is the largest unknown term, the corresponding kinetic energy functional must be approximated very accurately. The simplest approximation for $T[\rho(\mathbf{r})]$ is the Thomas-Fermi model (Equation 1.42), which is exact for the uniform electron gas (UEG) and has been available since the 1930s. Although it is possible to make slight improvements to this form by introducing gradient corrections and nonlocality, the best existing approximations are only applicable to systems with nearly-uniform densities (such as certain alloys and semiconductors) and cannot properly describe chemical bonds in molecules. Accordingly, designing accurate kinetic energy functionals for orbital-free DFT is a very difficult task and has yet to be accomplished.

$$T\left[\rho\left(\mathbf{r}\right)\right]=\frac{3}{10}\left(3\pi^{2}\right)^{2/3}\int\rho\left(\mathbf{r}\right)^{5/3}\mathrm{d}\mathbf{r} \tag{1.42}$$

Fortunately, this obstacle was circumvented by Kohn and Sham (KS) in 1965 with the realization[3] that the kinetic energy could be accurately approximated by introducing a single Slater determinant of spin orbitals that describes a fictitious system of non-interacting electrons but integrates (via Equation 1.31) to the same density as that of the exact electronic wave function. In principle, KS-DFT is an exact theory. Since the non-interacting kinetic energy, $T_s\left[\{\phi_i\}\right]$, is not equal to the exact kinetic energy, $T\left[\rho\left(\mathbf{r}\right)\right]$, the difference between these two terms is combined with $Q\left[\rho\left(\mathbf{r}\right)\right]$ to define the exchange-correlation functional,

$$E_{xc}\left[\rho\left(\mathbf{r}\right)\right]=T\left[\rho\left(\mathbf{r}\right)\right]-T_s\left[\{\phi_i\}\right]+Q\left[\rho\left(\mathbf{r}\right)\right]=E_x\left[\rho\left(\mathbf{r}\right)\right]+E_c\left[\rho\left(\mathbf{r}\right)\right] \tag{1.43}$$

The only unknown term in KS-DFT is the form of the exchange-correlation functional, $E_{xc}\left[\rho\left(\mathbf{r}\right)\right]$, which can be further partitioned into an exchange functional, $E_x\left[\rho\left(\mathbf{r}\right)\right]$, and a correlation functional, $E_c\left[\rho\left(\mathbf{r}\right)\right]$. Since KS-DFT involves spin orbitals (like Hartree–Fock), a minor modification to the SCF method shown in Figure 1.1 enables the calculation of KS-DFT energies. First, the elements in $G_{\mu\nu}$ that correspond to the exchange matrix must be replaced with the elements of the exchange-correlation matrix, $F_{\mu\nu}^{XC}$, which corresponds to the potential of the exchange-correlation functional. Second, the exchange-correlation energy must be appended to the total energy. In the next section, practical aspects of KS-DFT will be discussed. The goal of this thesis is to design better approximations to the exchange-correlation functional.

1.3.2 Practical Aspects

Although Kohn and Sham published their landmark paper in 1965, it took nearly 30 years for KS-DFT (henceforth synonymous with DFT) to gain traction and achieve the popularity it enjoys today. In the past 30 years, hundreds of ab initio and semi-empirical density functionals have been developed by hundreds of chemists and physicists – so many, in fact, that it would be impractical to cover them all in this brief introduction. Nevertheless, it would be near blasphemy not to acknowledge John Perdew's continued efforts in the area of ab initio density functional development and Axel Becke's impactful contributions to the development of semi-empirical density functionals. Nearly all of the popular density functionals currently in use can be traced back to their groundbreaking ideas. This section will begin with an introduction to the components of local exchange-correlation functionals. A brief review of three major hurdles that these functionals face will follow, along with a summary of several existing methods that can ameliorate some of these difficulties. Finally, practical computational considerations (such as the choice of basis set and grid) will be addressed, and the current status of the accuracy of DFT for thermochemistry, non-covalent interactions, isomerization energies, and barrier heights will be discussed.

All exchange functionals that will be discussed can be described by Equation 1.44 (or simplifications thereof) and all correlation functionals that will be discussed can be described by Equation 1.45 (or simplifications thereof). In the exchange functional expression, e_x is the exchange energy density per unit volume, while in the correlation functional expression, ϵ_c, is the correlation energy density per electron.

$$E_x\left[\rho_\alpha, \rho_\beta, |\nabla\rho_\alpha|^2, |\nabla\rho_\beta|^2, \tau_\alpha, \tau_\beta\right] = \sum_\sigma^{\alpha,\beta} \int e_{x,\sigma}\left(\rho_\sigma, |\nabla\rho_\sigma|^2, \tau_\sigma\right) d\mathbf{r} \tag{1.44}$$

$$\begin{gathered} E_c\left[\rho_\alpha, \rho_\beta, |\nabla\rho_\alpha|^2, |\nabla\rho_\beta|^2, \nabla\rho_\alpha \cdot \nabla\rho_\beta, \tau_\alpha, \tau_\beta\right] = \\ \int (\rho_\alpha + \rho_\beta)\, \epsilon_c\left(\rho_\alpha, \rho_\beta, |\nabla\rho_\alpha|^2, |\nabla\rho_\beta|^2, \nabla\rho_\alpha \cdot \nabla\rho_\beta, \tau_\alpha, \tau_\beta\right) d\mathbf{r} \end{gathered} \tag{1.45}$$

The simplest exchange-correlation functionals depend only on the density and define a class (or rung[4]) known as the local spin-density approximation (LSDA). These functionals are exact for the uniform electron gas (UEG), but are highly inaccurate for molecular properties, since most realistic systems have inhomogeneous density distributions. The LSDA exchange functional has an exact analytic form and dates back to days of Slater and Dirac (Equation 1.46). On the other hand, there is no exact analytic form for the LSDA correlation functional, and the three most popular parameterizations[5–7] (PZ81, VWN5, and PW92) resort to fits to accurate Monte-Carlo data[8] from the late 1970s. Henceforth, any mention of the LSDA exchange-correlation functional (SPW92) will refer to the combination of the Slater-Dirac exchange functional (abbreviated as S) and the PW92 correlation functional.

$$E_x^S\left[\rho_\alpha, \rho_\beta\right] = -\frac{3}{2}\left(\frac{3}{4\pi}\right)^{1/3} \sum_\sigma^{\alpha,\beta} \int \rho_\sigma^{4/3} \mathrm{d}\mathbf{r} \tag{1.46}$$

In order to improve upon the systematic overbinding inherent to the LSDA, it is necessary to introduce an ingredient into the exchange-correlation functional that can account for inhomogeneities in the density: the density gradient, $\nabla\rho_\sigma$. These generalized gradient approximation (GGA) functionals tend to improve significantly upon the LSDA. Popular GGA exchange functionals[9–12] include B86, B88, PW91, and PBE, while popular GGA correlation functionals[11–14] include P86, LYP, PW91, and PBE. These components can be combined to define GGA exchange-correlation functionals, and the PBE exchange-correlation functional (PBE exchange + PBE correlation) is perhaps the most popular ab initio GGA functional, while BLYP (B88 exchange + LYP correlation) is perhaps the most popular semi-empirical GGA functional. As an example of the improvements offered by GGAs over the LSDA, the mean signed error (MSE) for SPW92 across 124 atomization energies is -58.11 kcal/mol, while PBE affords an MSE of -12.17 kcal/mol. The same trend can be seen for non-bonded interactions as well, with SPW92 having an MSE of -30.74 kcal/mol for the binding energies of 38 water clusters, while PBE has an MSE of only -1.26 kcal/mol. In general, GGAs improve upon the LSDA by reducing the overestimation of interaction energies.

Two additional ingredients that can be used to further improve the exchange-correlation functional are either the Laplacian of the density, $\nabla^2\rho_\sigma$, or the kinetic energy density, $\tau_\sigma = \sum_i^{n_\sigma} |\nabla\psi_{i,\sigma}|^2$. Since these two ingredients are related, only one or the other tends to appear in a given functional form (although there are exceptions[15]). The kinetic energy density is by far the more popular ingredient and has been used in many modern functionals to add flexibility to the functional form with respect to both constraint satisfaction (ab initio) and least-squares fitting (semi-empirical). Popular non-empirical meta-GGA exchange-correlation functionals are almost exclusively from Perdew and coworkers, and include[16–18] PKZB, TPSS, and revTPSS, as well as the newer (and slightly empirical) MGGA_MS0, MGGA_MS1, MGGA_MS2, MGGA_MVS, and SCAN functionals[19–22]. Empirical meta-GGA functionals are much more abundant, mainly due to the endless parameterizations provided by the Truhlar group, and include functionals[23–26] such as M06-L, M11-L, MN12-L, and most recently, and most disastrously, MN15-L. However, some earlier notable attempts at semi-empirical meta-GGAs include efforts by Scuseria[27] (VSXC) and Handy[28] (τ-HCTH). In general, meta-GGAs tend to perform better than GGAs, and the semi-empirical ones can even incorporate some "medium-range" dispersion (e.g. M06-L binds the sandwich benzene dimer, while both PBE and TPSS predict entirely unbound potential energy curves (PEC)). For the set of 124 atomization energies, TPSS, with an MSE of -2.72 kcal/mol, significantly improves upon the overbinding of PBE, but manages to underbind the water clusters with an MSE of 3.21 kcal/mol.

Despite the "systematic" improvement offered by additional physically-meaningful ingredients, there are three major limitations to the local exchange-correlation functionals described above: 1). sensitivity to self-interaction error (SIE), 2). inability to describe long-range correlation (dispersion), and 3). failures for strong correlation.

The simplest way to demonstrate self-interaction (i.e. electrons interacting with themselves) is to consider the Hartree–Fock description of the hydrogen atom. Since the hydrogen atom contains only one electron, the contribution from $\hat{V}_{ee} = \hat{J} + \hat{K}$ should be exactly zero. At the CBS limit, the Hartree–Fock energy for the hydrogen atom is -0.5 Hartrees (Eh), which comes from summing the contributions from $\hat{T}$ (0.5 Eh), $\hat{V}_{ne}$ (-1.0 Eh), $\hat{J}$ (0.3125 Eh), and $\hat{K}$ (-0.3125 Eh). Thus, the classical and non-classical electron-electron contributions cancel each other exactly, making Hartree–Fock one-electron SIE-free. In KS-DFT, since $\hat{K}$ is replaced by the exchange-correlation functional, most functionals are not one-electron SIE-free (i.e. $J\left[\rho\left(\mathbf{r}\right)\right] + E_{xc}\left[\rho\left(\mathbf{r}\right)\right] \neq 0$ for the hydrogen atom).

A possible workaround for this issue is to replace the local exchange functional with the exact exchange functional (Hartree–Fock), while keeping the local correlation functional. Early attempts at combining exact exchange with local correlation were unsuccessful, and the idea was abandoned until the early 1990s, when Becke introduced the idea of mixing a global fraction of exact exchange with the exchange-correlation functional. Most primitively (i.e. for the LSDA), these global hybrid (GH) functionals take the form given in Equation 1.47 and can be theoretically justified with the adiabatic connection[29–32]. A more empirical way to motivate global hybrid functionals is to consider the mean signed error (MSE) of Hartree–Fock and SPW92 for the aforementioned 124 atomization energies. While the former underbinds the atomization energies with an MSE of 112.79 kcal/mol, the latter overbinds them with an MSE of -58.11 kcal/mol. Although an equal mixing of the two ($c_x = 0.50$) would still potentially lead to underbinding, a value of $c_x = 0.34$ will lead to these MSEs adding to zero.

$$E_{xc} = c_x E_x^{HF} + \left(1 - c_x\right) E_x^{S} + E_c^{PW92} \tag{1.47}$$

The first global hybrid functional[33], B3PW91, was developed by Becke in 1993 by fitting three linear parameters to 56 atomization energies. B3PW91 is a global hybrid GGA density functional that takes the form given in Equation 1.48, where $c_x = 0.20$, $a_x = 0.72$, and $a_c = 0.81$. Most GH GGA density functionals have an exact exchange mixing parameter between 20% and 25%, including the most popular density functional, B3LYP (20%). The most popular ab initio[34] GH GGA is PBE0 (25%), while the most popular semi-empirical GH GGA is Becke's B97, which is the first systematically-parameterized density functional[35]. For semi-empirical GH GGAs, the value of c_x can change significantly depending on the type of systems in the training set. For example, the B97-K and SOGGA11-X density functionals[36, 37] have 42% and 40.15% exact exchange, respectively, because barrier heights were heavily emphasized in their training sets. In general, the inclusion of exact exchange helps counter the overbinding of GGA and even meta-GGA functionals. PBE0 eradicates the overbinding of PBE for atomization energies, with an MSE of only 1.12 kcal/mol (compared to -12.17 kcal/mol), and further improves the performance of its local counterpart for the water clusters, with an MSE of 0.20 kcal/mol (compared to -1.26 kcal/mol).

$$E_{xc} = c_x E_x^{HF} + (1 - c_x - a_x) E_x^S + a_x E_x^{B88} + (1 - a_c) E_c^{PW92} + a_c E_c^{PW91} \tag{1.48}$$

Naturally, the formula in Equation 1.48 can be extended to meta-GGAs to give GH meta-GGA density functionals. From the ab initio side, TPSSh, revTPSSh, and MGGA_MS2h have about 10% exact exchange[20, 38, 39], while the latest MGGA_MVSh functional[21] has an uncharacteristically large value of $c_x = 0.25$. Since 2004, Truhlar has published at least 12 global hybrid meta-GGA density functionals[40–46], including MPW1B95, MPWB1K, PW6B95, PWB6K, M05, M05-2X, M06, M06-2X, M06-HF, M08-HX, M08-SO, and MN15. The fraction of exact exchange across these functionals varies from 27% (M06) all the way to 100% (M06-HF).

While global hybrid functionals significantly improve upon their local counterparts for bonded interactions and kinetics, they only partially address the self-interaction issue. A more rigorous approach to this problem is through range-separation[47], where the exact exchange contribution is split into a short-range component ($E_{x,sr}^{HF}$) and a long-range component ($E_{x,lr}^{HF}$). The Coulomb operator of the short-range component is attenuated by the complementary error function, $\mathrm{erfc}\,(r_{12})$, while the Coulomb operator of the long-range component is attenuated by the error function, $\mathrm{erf}\,(r_{12})$. $E_{x,sr}^{HF}$ can be optionally scaled to give a non-zero fraction of short-range exact exchange, while the scaling coefficient of $E_{x,lr}^{HF}$ is usually set to one to ensure that the exchange functional is one-electron SIE-free in the long range. The corresponding local exchange functional should also be partitioned in the same way, but the short range component is scaled by $1 - c_x$ and the long-range component is discarded. The functional form for a typical range-separated hybrid (RSH) functional is shown in Equation 1.49.

$$E_{xc} = c_x E_{x,sr}^{HF} + E_{x,lr}^{HF} + (1 - c_x) E_{x,sr}^{DFT} + E_c^{DFT} \tag{1.49}$$

Notable range-separated hybrid GGA functionals include the semi-empirical ωB97 and ωB97X functionals from Chai and Head-Gordon[48], which have 0% and 15.77% short-range exact exchange, respectively, and tend to 100% exact exchange in the long-range. The functionals have empirical values for ω as well, with the former having a value of $\omega = 0.4$ and the latter, a value of $\omega = 0.3$. The CAM-B3LYP RSH GGA functional[49] by Handy and coworkers is also noteworthy. While it has 19% short-range exact exchange, its long-range coefficient is not 1, but rather, 0.65.

Finally, it is possible to entirely remove $E_{x,lr}^{HF}$ to form long-range screened functionals, such as HSE06 and N12-SX, that are suitable for molecular, as well as solid-state, calculations[50, 51]. RSH meta-GGA functionals are the least common of the functional types mentioned thus far, with M11 and MN12-SX by Truhlar serving as two examples[51, 52].

The second weakness of local exchange-correlation functionals is their inability to properly account for long-range correlation. The most straightforward way to demonstrate this concept is to consider the potential energy curve of a dispersion-bound system, such as the sandwich benzene dimer. For this system, LSDA, GGA, and meta-GGA functionals exhibit an exponential long-range decay, instead of the proper r_{12}^{-6} decay. Fortunately, the past decade has seen a vast amount of work dedicated to improving the description of non-covalent interactions within KS-DFT. The two most popular methods are the DFT-D tail by Grimme[53–56] and the VV10 nonlocal correlation functional by Vydrov and Van Voorhis[57].

Grimme's DFT-D method is a damped, atom-atom empirical potential that can be trained onto any of the aforementioned functionals. Three generations of DFT-D tails have been developed by Grimme thus far: DFT-D1, DFT-D2, and DFT-D3. The latter can be used either with the original damping function, DFT-D3(0), or the newer Becke-Johnson damping function, DFT-D3(BJ). While it is trivial to train a dispersion correction onto an existing ab initio or semi-empirical functional, simultaneously training a semi-empirical functional with a dispersion correction is more involved. The first successful attempt was Grimme's B97-D functional[54], a local GGA functional utilizing the DFT-D2 tail. Recently, both the DFT-D3(0) and DFT-D3(BJ) tails have been refit to the existing local exchange-correlation functional of B97-D to give B97-D3(0) and B97-D3(BJ), with the latter being one of the best dispersion-corrected local GGA functionals available. Other examples include the ωB97X-D (GGA) and ωM05-D (meta-GGA) RSH functionals by Chai and coworkers[58, 59], which use a slightly modified version of the DFT-D2 tail. Finally, the most recent functionals by Chai and coworkers, ωB97X-D3 (GGA) and ωM06-D3 (meta-GGA), are range-separated hybrids[60] that utilize the DFT-D3(0) tail. In general, dispersion corrections such as DFT-D should be used with functionals that tend to underbind non-covalent interactions (both strong and weak). However, fitting a dispersion correction to a functional like PBE is tricky, since PBE is already overbound for the binding energies of the 38 water clusters mentioned earlier. Consequently, PBE-D3(BJ) increases the MSE of PBE by a factor of almost 5 for these clusters (-5.86 vs. -1.26 kcal/mol). On the other hand, PBE-D3(BJ) drastically improves upon PBE for dispersion-bound systems.

The popular nonlocal correlation functional by Vydrov and Van Voorhis, VV10, is far less empirical than the DFT-D approach and only contains two empirical parameters. VV10 can also be either trained onto an existing functional, or simultaneously trained onto a semi-empirical functional. The functional that will be discussed in Chapter 3, ωB97X-V, is a range-separated hybrid, GGA density functional that has been simultaneously trained with the VV10 nonlocal correlation functional. In Chapter 4, the same concept will be applied to develop a local meta-GGA functional with nonlocal correlation, B97M-V, while in the last chapter, the ideas from the Chapters 3 and 4 will be combined to develop an RSH meta-GGA density functional with nonlocal correlation (ωB97M-V).

Finally, all of the aforementioned density functionals fail to some extent at describing multireference systems that exhibit strong correlation. The reason for this is simple: HF and KS-DFT are both methods that make use of a single determinant. In order to accurately describe a multireference system, it is necessary to include multiple determinants. Addressing the issue of strong correlation within KS-DFT is the least solved problem of the three presented thus far, and requires much further exploration. However, local functionals tend to outperform hybrids for these types of systems.

The two most important practical considerations that pertain to KS-DFT calculations are the choice of the atomic orbital basis set and the integration grid. Non-empirical density functionals should perform best at the basis set limit, while most semi-empirical density functionals are trained either at the basis set limit (or close to the basis set limit). Since local functionals are easier to converge to the basis set limit, their basis set requirements are not as strict as for hybrids. For bonded interactions, quadruple-zeta basis sets are preferred in general, but triple-zeta basis sets can act as substitutes for larger systems. For non-covalent interactions, reaching the basis set limit is easier due to the Boys-Bernardi counterpoise correction scheme. With the use of counterpoise corrections, a triple-zeta basis set is the CBS limit for most systems, and even some double-zeta basis sets will suffice (particularly for local functionals). On the other hand, if counterpoise corrections are unavailable, it is preferable to use a quadruple-zeta basis set, and minimally, a triple-zeta basis set. It is worth mentioning, however, that certain semi-empirical density functionals, primarily from the Truhlar group, are not compatible with counterpoise corrections and should be utilized without corrections for basis set superposition error (BSSE).

As far as the integration grid is concerned, LSDA and GGA functionals tend to require similar integration grids and are much less sensitive to the choice than meta-GGA functionals. For bonded interactions, LSDA and GGA functionals can be used with grids such as the (50,194) grid that has 50 radial shells per atom, with 194 points per shell. For non-bonded interactions, it is advisable to use the (75,302) grid for LSDA and GGA functionals. On the other hand, meta-GGA functionals, particularly empirical ones, require at least the (75,302) for bonded interactions. For non-bonded interactions use of the (99,590) grid is advised (although some meta-GGA functionals will require even finer grids for very weak interactions).

The types of interactions considered in this thesis can be divided into four main categories: thermochemistry (TC), non-covalent interactions (NC), isomerization energies (I), and barrier heights (BH). The TC category can be further partitioned into two subcategories: TCE (easy) and TCD (difficult). TCE covers single-reference systems and includes atomization energies, bond dissociation energies, electron affinities, ionization potentials, isodesmic reactions, reaction energies, heavy-atom transfer energies, and nucleophilic substitution energies. For TCE, the best-performing density functionals tend to be hybrid meta-GGA functionals. The TCD category is mostly composed of atomization energies, bond dissociation energies, and heavy-atom transfer energies that are notoriously multi-reference in nature. For these systems, either local functionals or hybrid functionals with small fractions of exact exchange (30% or less) tend to perform best.

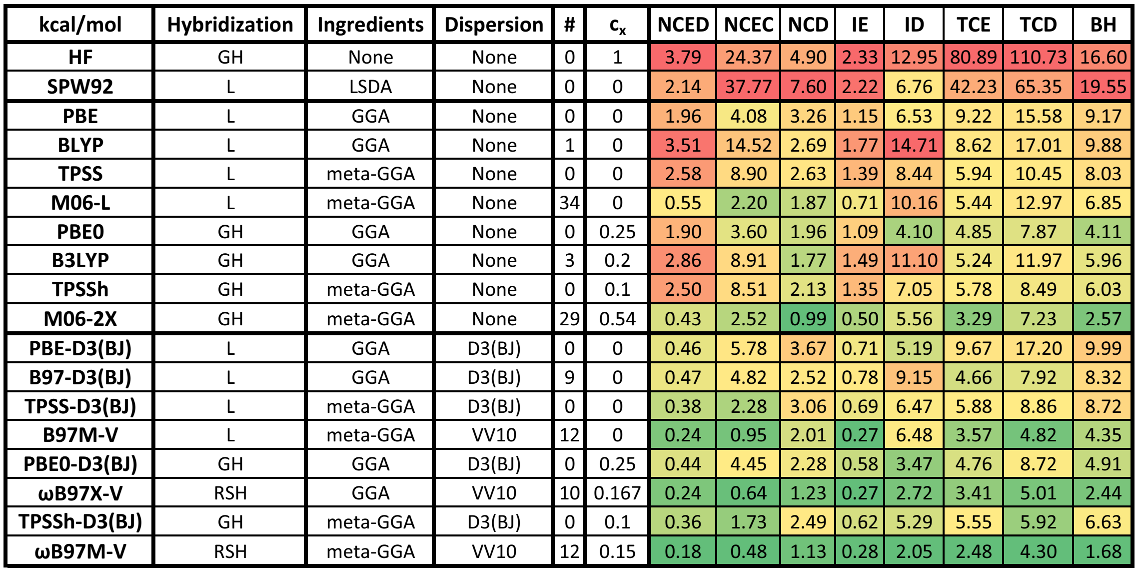
The NC category can be divided into three subcategories: NCED (easy dimers), NCEC (easy clusters), NCD (difficult dimers). The first subcategory, NCED, contains dimer binding energies that are not prone to self-interaction error. Typical systems include the methane dimer, the benzene dimer, the water dimer, almost 100 potential energy curves of hydrogen-bonded and dispersion-bound systems, as well as a handful of charged interactions. The NCEC subcategory contains binding energies of large clusters, including water clusters from 3-mers to 20-mers, ammonia clusters, hydrogen fluoride clusters, sulfate-water clusters, and water clusters with anions in the center (either fluoride or chloride). For the NCED and NCEC subcategories, the most important factor is the inclusion of a dispersion correction. Both local functionals and hybrid functionals with dispersion corrections perform very well, most notably the three functionals that will be presented later in this thesis: ωB97X-V, B97M-V, and ωB97M-V. On the other hand, the NCD subcategory contains non-covalent interactions that are very sensitive to self-interaction error, such as dimers involving radicals, as well as halogen-bonded systems. For these interactions, exact exchange is absolutely vital, and the best functionals contain at least 40% exact exchange.

The I category can be divided into two subcategories: IE (easy) and ID (difficult). IE contains easy isomerization energies that are not sensitive to delocalization error, such as the isomerization energies of alkanes, organic molecules, sulfate-water clusters, amino acids, and large water clusters. As with NCED and NCEC, dispersion corrections tend to be vital. However, certain functionals with large fractions of exact exchange (namely M05-2X and M06-2X) tend to perform well. The ID subcategory, much like the NCD subcategory, involves relative energies of isomers that are sensitive to SIE. This includes isomerization energies of enecarbonyles, styrene, conjugated dienes, as well as a notoriously difficult case, fullerenes. For this subcategory, exact exchange is very important, although too much exact exchange tends to be detrimental.

Finally, the BH category contains barrier heights of pericyclic reactions, cycloreversion reactions, proton exchange reactions, in addition to hydrogen transfer and non-hydrogen transfer barrier heights. Interestingly, a few of the datasets in the category tend to prefer local functionals, but most of the data points are better described by hybrid functionals. The best-performing functionals overall are either global hybrids with large fractions of exact exchange, or range-separated hybrids.

In order to demonstrate the current state of the performance of density functionals for a wide variety of systems, RMSDs for a total of 4399 data points partitioned across the 8 datatypes mentioned thus far (NCED, NCEC, NCD, IE, ID, TCE, TCD, and BH) are summarized in Figure 1.2. The results are calculated with the def2-QZVPPD basis set and use the (99,590) grid (for the local exchange-correlation functional) and the SG-1 grid[61] (for the VV10 nonlocal correlation functional). The first two functionals that are included are Hartree–Fock and SPW92, and are chosen because they represent the most basic versions of the two theories contained in Sections 1.2 and 1.3. The remaining 16 functionals can be separated into those that do not contain a dispersion correction, and those that are corrected by either the DFT-D or VV10 approaches. Within each of these realms, two local GGA, two local meta-GGA, two hybrid GGA, and two hybrid meta-GGA functionals are tested. The two functionals within each category are specifically chosen such that one is ab initio and the other is semi-empirical. This results in a total of 18 methods.

The results in Figure 1.2 clearly indicate that most density functionals significantly improve upon HF and SPW92. In most cases, semi-empirical functionals tend to outperform the ab initio functionals. For example, M06-L is better than TPSS for 6 of the 8 datatypes, M06-2X is better than TPSSh for all 8 datatypes, B97-D3(BJ) is equal to or better than PBE-D3(BJ) for 7 of the 8 datatypes, B97M-V is better than TPSS-D3(BJ) for all 8 datatypes, ωB97X-V outperforms PBE0-D3(BJ) across all 8 datatypes, and ωB97M-V outperforms TPSSh-D3(BJ) across all 8 datatypes. In two cases, however, the ab initio functional outperforms the semi-empirical functional overall: PBE vs. BLYP and PBE0 vs. B3LYP. In general, the dispersion-corrected funtionals shown at the bottom of the figure tend to perform better than the functionals that are lacking dispersion correction, especially for the NCED, NCEC, and IE datatypes. A much more thorough assessment of these types of functionals will be conducted in the upcoming chapters.



kcal/mol	Hybridization	Ingredients	Dispersion	#	c_x	NCED	NCEC	NCD	IE	ID	TCE	TCD	BH
HF	GH	None	None	0	1	3.79	24.37	4.90	2.33	12.95	80.89	110.73	16.60
SPW92	L	LSDA	None	0	0	2.14	37.77	7.60	2.22	6.76	42.23	65.35	19.55
PBE	L	GGA	None	0	0	1.96	4.08	3.26	1.15	6.53	9.22	15.58	9.17
BLYP	L	GGA	None	1	0	3.51	14.52	2.69	1.77	14.71	8.62	17.01	9.88
TPSS	L	meta-GGA	None	0	0	2.58	8.90	2.63	1.39	8.44	5.94	10.45	8.03
M06-L	L	meta-GGA	None	34	0	0.55	2.20	1.87	0.71	10.16	5.44	12.97	6.85
PBE0	GH	GGA	None	0	0.25	1.90	3.60	1.96	1.09	4.10	4.85	7.87	4.11
B3LYP	GH	GGA	None	3	0.2	2.86	8.91	1.77	1.49	11.10	5.24	11.97	5.96
TPSSh	GH	meta-GGA	None	0	0.1	2.50	8.51	2.13	1.35	7.05	5.78	8.49	6.03
M06-2X	GH	meta-GGA	None	29	0.54	0.43	2.52	0.99	0.50	5.56	3.29	7.23	2.57
PBE-D3(BJ)	L	GGA	D3(BJ)	0	0	0.46	5.78	3.67	0.71	5.19	9.67	17.20	9.99
B97-D3(BJ)	L	GGA	D3(BJ)	9	0	0.47	4.82	2.52	0.78	9.15	4.66	7.92	8.32
TPSS-D3(BJ)	L	meta-GGA	D3(BJ)	0	0	0.38	2.28	3.06	0.69	6.47	5.88	8.86	8.72
B97M-V	L	meta-GGA	VV10	12	0	0.24	0.95	2.01	0.27	6.48	3.57	4.82	4.35
PBE0-D3(BJ)	GH	GGA	D3(BJ)	0	0.25	0.44	4.45	2.28	0.58	3.47	4.76	8.72	4.91
ωB97X-V	RSH	GGA	VV10	10	0.167	0.24	0.64	1.23	0.27	2.72	3.41	5.01	2.44
TPSSh-D3(BJ)	GH	meta-GGA	D3(BJ)	0	0.1	0.36	1.73	2.49	0.62	5.29	5.55	5.92	6.63
ωB97M-V	RSH	meta-GGA	VV10	12	0.15	0.18	0.48	1.13	0.28	2.05	2.48	4.30	1.68

Figure 1.2: The performance of a selection of density functionals for 8 different datatypes. NCED stands for non-covalent dimers (easy), NCEC stands for non-covalent clusters (easy), NCD stands for non-covalent dimers (difficult), IE stands for isomerization energies (easy), ID stands for isomerization energies (difficult), TCE stands for thermochemistry (easy), TCD stands for thermochemistry (difficult), and BH stands for barrier heights. The partitioning of the 4399 data points contained in this figure into the 8 datatypes is: 1744, 243, 91, 755, 155, 947, 258, and 206. The values in the last 8 columns are RMSDs in kcal/mol. L stands for local, GH stands for global hybrid, and RSH stands for range-separated hybrid. The column labeled "#" contains the total number of parameters optimized on a training set for the corresponding functional. For the functionals with the DFT-D correction, the number of empirical parameters in the tail is not counted. c_x lists the fraction of exact exchange.

1.4 Outline

The goal of this work is to develop accurate exchange-correlation functionals that are minimally-parameterized and maximally-transferable. In order to achieve this goal, the combinatorial space of GGA density functionals is fully searched in the second chapter, with a total of nine types of functionals: local, global hybrid, or range-separated hybrids utilizing either no dispersion corrections, the DFT-D2 dispersion correction, or the VV10 nonlocal correlation functional. The results of this search indicate that of the nine types, the range-separated hybrid functional that utilizes the VV10 nonlocal correlation functional is the most accurate and transferable functional. In the third chapter, the functional that resulted from the search in Chapter 2 is introduced and thoroughly tested. This functional is called ωB97X-V. The fourth chapter deals with a much more difficult problem: searching the meta-GGA functional space. The scale of this search is on the order of 10^{40} possibilities and is thus impossible to complete without approximations. Nevertheless, using appropriate truncations, a local meta-GGA density functional, called B97M-V, is developed and thoroughly tested for transferability and other properties. In the final chapter, the best ingredients (range-separation and meta-GGA) are combined to create the best available Rung 1-4 density functional to date: ωB97M-V.

1.4.1 Chapter 2

The limit of accuracy for semi-empirical generalized gradient approximation (GGA) density functionals is explored by parameterizing a variety of local, global hybrid (GH), and range-separated hybrid (RSH) functionals. The training methodology employed differs from conventional approaches in 2 main ways: 1). Instead of uniformly truncating the exchange, same-spin correlation, and opposite-spin correlation functional inhomogeneity correction factors, all possible fits up to fourth order are considered, and 2). Instead of selecting the optimal functionals based solely on their training set performance, the fits are validated on an independent test set and ranked based on their overall performance on the training *and* test sets. The 3 different methods of accounting for exchange are trained both with and without dispersion corrections (DFT-D2 and VV10), resulting in a total of 491508 candidate functionals. For each of the 9 functional classes considered, the results illustrate the trade-off between improved training set performance and diminished transferability. Since all 491508 functionals are uniformly trained and tested, this methodology allows the relative strengths of each type of functional to be consistently compared and contrasted. The range-separated hybrid GGA functional paired with the VV10 nonlocal correlation functional emerges as the most accurate form for the present training and test sets, which span thermochemical energy differences, reaction barriers, and intermolecular interactions involving lighter main group elements. This work[62] has been published in *The Journal of Chemical Physics* in the *SPECIAL TOPIC: ADVANCES IN DENSITY FUNCTIONAL THEORY* issue.

1.4.2 Chapter 3

A 10-parameter, range-separated hybrid (RSH), generalized gradient approximation (GGA) density functional with nonlocal correlation (VV10) is presented. Instead of truncating the B97-type power series inhomogeneity correction factors (ICF) for the exchange, same-spin correlation, and opposite-spin correlation functionals uniformly, all 16383 combinations of the linear parameters up to fourth order ($m = 4$) are considered. These functionals are individually fit to a training set and the resulting parameters are validated on a primary test set in order to identify the 3 optimal ICF expansions. Through this procedure, it is discovered that the functional that performs best on the training *and* primary test sets has 7 linear parameters, with 3 additional nonlinear parameters from range-separation and nonlocal correlation. The resulting density functional, ωB97X-V, is further assessed on a secondary test set, the parallel-displaced coronene dimer, as well as several geometry datasets. Furthermore, the basis set dependence and integration grid sensitivity of ωB97X-V are analyzed and documented in order to facilitate the use of the functional. This work[63] has been published in *Physical Chemistry Chemical Physics* in the *PCCP's 15th anniversary* issue.

1.4.3 Chapter 4

A meta-generalized gradient approximation (meta-GGA) density functional paired with the VV10 nonlocal correlation functional is presented. The functional form is selected from more than 10 billion choices carved out of a functional space of almost 10^{40} possibilities. Raw data comes from training a vast number of candidate functional forms on a comprehensive training set of 1095 data points and testing the resulting fits on a comprehensive primary test set of 1153 data points. Functional forms are ranked based on their ability to reproduce the data in both the training and primary test sets with minimum empiricism, and filtered based on a set of physical constraints and an often-overlooked condition of satisfactory numerical precision with medium-sized integration grids. The resulting optimal functional form has 4 linear exchange parameters, 4 linear same-spin correlation parameters, and 4 linear opposite-spin correlation parameters, for a total of 12 fitted parameters. The final density functional, B97M-V, is further assessed on a secondary test set of 212 data points, applied to several large systems including the coronene dimer and water clusters, tested for the accurate prediction of intramolecular and intermolecular geometries, verified to have a readily attainable basis set limit, and checked for grid sensitivity. Compared to existing density functionals, B97M-V is remarkably accurate for non-bonded interactions and very satisfactory for thermochemical quantities such as atomization energies, but inherits the demonstrable limitations of existing local density functionals for barrier heights. This work[64] has been published in *The Journal of Chemical Physics*.

1.4.4 Chapter 5

A combinatorially-optimized, range-separated hybrid, meta-GGA density functional with VV10 nonlocal correlation is presented. The final 12-parameter functional form is selected from approximately 10 billion candidate fits that are trained on a training set of 870 data points and tested on a primary test set of 2965 data points. The resulting density functional, ωB97M-V, is further tested for transferability on a secondary test set of 1152 data points. For comparison, ωB97M-V is benchmarked against 10 leading density functionals including M06-2X, ωB97X-D, M08-HX, M11, ωM05-D, and ωB97X-V. Encouragingly, the overall performance of ωB97M-V on nearly 5000 data points clearly surpasses all of the tested density functionals. In order to facilitate the use of ωB97M-V, its basis set dependence and integration grid sensitivity are thoroughly assessed, and recommendations that take into account both efficiency and accuracy are provided. This work has been submitted to *The Journal of Chemical Physics*.

Chapter 2

Exploring the GGA functional space

2.1 Introduction

While empirical parameters have been used in density functionals since the 1950s, the first systematic optimization of a density functional was performed by Axel Becke in 1997[35]. However, this breakthrough would not have been possible without several significant developments that took place in the preceding decades. Firstly, Frank Herman's extension[65] of John Slater's Xα method[66] (Equations 2.1 and 2.2) to the X$\alpha\beta$ method (Equation 2.3) introduced a gradient-based correction, $s_\sigma = \frac{|\nabla\rho_\sigma|}{\rho_\sigma^{4/3}}$, to the X$\alpha$ exchange energy density based on dimensional arguments. A major drawback of the semi-empirical X$\alpha\beta$ method was the asymptotic divergence of its exchange potential. A solution to this was proposed by Becke in 1986, when he modified the inhomogeneity correction factor introduced by Herman in order to produce the divergence free X$\alpha\beta\gamma$ (B86) exchange functional[9] (Equation 2.4).

$$E_{x\alpha} = -\sum_{\sigma}^{\uparrow,\downarrow} \int C_{x\alpha}\rho_\sigma^{4/3} d\mathbf{r} \tag{2.1}$$

$$C_{x\alpha} = \frac{9}{4}\alpha\left(\frac{3}{4\pi}\right)^{1/3} \tag{2.2}$$

$$E_{x\alpha\beta} = -\sum_{\sigma}^{\uparrow,\downarrow} \int C_{x\alpha}\rho_\sigma^{4/3}\left[1 + \frac{\beta}{C_{x\alpha}} s_\sigma^2\right] d\mathbf{r} \tag{2.3}$$

$$E_{x\alpha\beta\gamma} = -\sum_{\sigma}^{\uparrow,\downarrow} \int C_{x\alpha}\rho_\sigma^{4/3}\left[1 + \frac{\beta}{C_{x\alpha}\gamma_x}\frac{\gamma_x s_\sigma^2}{1+\gamma_x s_\sigma^2}\right] d\mathbf{r} \tag{2.4}$$

11 years later[35], Becke generalized the inhomogeneity correction factor of the B86 exchange functional with an m^{th} order power series in the dimensionless variable, $u_{x,\sigma} = \frac{\gamma_x s_\sigma^2}{1+\gamma_x s_\sigma^2}$:

$$E_x^{B97} = -\sum_{\sigma}^{\uparrow,\downarrow} \int C_{x\alpha} \rho_\sigma^{4/3} \left[\sum_{i=0}^{m} c_{x,i} u_{x,\sigma}^i \right] d\mathbf{r} \tag{2.5}$$

This scheme was applied to both the exchange functional and the spin-decomposed same-spin and opposite-spin correlation functionals to produce the B97 density functional (Section 2.3). The original B97 functional truncated the expansions at $m = 2$, and included a fraction of exact exchange, leaving 10 undetermined linear parameters for fitting to thermochemical data.

As an approach to GGA density functionals, B97 has unparalleled flexibility. As a result, it is not surprising that at least 15 B97-based density functionals have been parameterized since 1997. These include local functionals[54, 67–69] (HCTH/93, HCTH/120, HCTH/147, HCTH/407, B97-D), global hybrid functionals[35, 67, 70, 71] (B97, B97-1, B97-2, B97-3), range-separated hybrid functionals[48, 58, 60, 63] (ωB97, ωB97X, ωB97X-D, ωB97X-D3, ωB97X-V), and even double hybrid functionals[72] (ωB97X-2).

The purpose of this work is to use the flexibility of the B97 form to attempt to systematically explore the accuracy attainable with different possible GGA functionals that build upon the basic B97 framework with different augmentations to exchange and correlation. Table 2.1 lists a variety of ingredients that can be incorporated into a B97-based density functional. To adhere to the functional form of the local component of B97, it is necessary to restrict the local exchange and correlation functionals to depend solely on the density and its gradient. However, the options for nonlocal exchange range from global hybrid exchange to range-separated exchange to no nonlocal exchange at all. These 3 options can be seamlessly integrated into the B97 functional form. From the perspective of dispersion corrections, options[54, 55, 57, 73–75] such as DFT-D2, DFT-D3, vdW-DF-04, vdW-DF-10, VV09, VV10, MP2, RPA, and beyond, exist. All of these methods can be easily appended to the B97 functional form as well.

B97-based semi-empirical density functionals have typically been optimized using uniformly truncated inhomogeneity correction factors (ICF) for the exchange, same-spin correlation, and opposite-spin correlation functionals. One method of approaching the limit of accuracy for GGA-based functionals is to try uniform expansions between $m = 0$ and a large m-value in order to select the optimal m-value based on a “goodness-of-fit” index[76] that is related to the training set performance. This approach can differentiate between uniformly truncated ICFs, and whether by

Exchange		Correlation	
Local	Nonlocal	Local	Nonlocal
1). $\underline{\rho}$	1). **None**	1). $\underline{\rho}$	1). **None**
2). $\underline{\nabla\rho}$	2). **GH**	2). $\underline{\nabla\rho}$	2). **DFT-D2**
3). τ	3). **RSH**	3). τ	3). **VV10**

Table 2.1: Ingredients that can be incorporated into a density functional. GH stands for global hybrid and RSH stands for range-separated hybrid. DFT-D2 refers to Grimme's dispersion tail and VV10 refers to the VV10 nonlocal correlation (NLC) functional. The underlined ingredients were not varied, while the ingredients in bold were varied, resulting in a total of 9 candidate functional forms. While the kinetic energy density, τ, is a valid candidate for inclusion in the local parts of both the exchange and correlation functionals, this chapter focuses exclusively on GGA functionals.

this approach, or by careful inspection, B97-based ICFs are usually truncated at either $m = 2$, $m = 3$, or $m = 4$. One functional that is not based on uniform truncation is ωB97X-V[63], which was developed based on a variation of the following methodology.

In contrast to uniform truncation, the most general approach is to perform all possible optimizations up to a certain power of m, including even those that skip powers (equivalent to setting the skipped coefficient to zero). This approach leads to thousands of competing fits (i.e. thousands of competing functional forms). It is difficult to differentiate between so many possible functionals using any inspection of training set results, including the "goodness-of-fit" index. Yet, it will be essential to face this complexity since it is likely that the simplest functional capable of yielding good accuracy on the training set data will perform best in applications.

While the goal of fitting to a training set is to minimize the training set root-mean-square-deviation (RMSD), it is even more desirable for a parameterized functional to be transferable. In order to differentiate between the thousands of resulting functionals *and* assure transferability, it is essential to take into account the performance of a given fit on both the training set *and* an independent test set. The test set is not used to determine any parameters, but will instead guide the choice of how many (and which) coefficients should be included in the least-squares fit. Taking the conventional approach of solely considering training set performance, it is guaranteed that the fit with the most linear parameters will have the smallest training set RMSD. Thus, if the training set RMSD is plotted with respect to the number of linear parameters, the resulting figure resembles the plots contained in Figure 2.1. However, if both the training set performance *and* the test set performance are taken into account, the plots begin to resemble parabolas (Figure 2.3). Thus, it is much easier to pick out an "optimal" functional with this methodology.

In this work, we parameterize 9 flavors of B97-based density functionals by varying the nonlocal exchange and dispersion correction (nonlocal correlation) components in bold in Table 2.1. While 14 of the 15 aforementioned B97-based density functionals have uniformly truncated inhomogeneity correction factors, all possible combinations of the exchange, same-spin correlation, and opposite-spin correlation expansion coefficients up to fourth order are tested. Using this methodology, an optimal functional from each category is selected, and the 9 resulting optimal functionals are compared to determine the optimal pairing of nonlocal exchange and dispersion.

2.2 Computational Details

An integration grid of 99 radial points and 590 angular points, (99,590), was used to evaluate local exchange-correlation (xc) functionals, while the SG-1 grid[61] was used for the VV10 nonlocal correlation (NLC) functional[57]. For the rare-gas dimers and the absolute atomic energies, a (500,974) integration grid was used to evaluate local xc functionals, along with a (99,590) grid for the VV10 NLC functional. The aug-cc-pVQZ [aQZ] basis set[77, 78] was used for all thermochemistry (TC) datapoints except the second-row absolute atomic energies (aug-cc-pCVQZ)[77, 78], while the aug-cc-pVTZ [aTZ] basis set[77, 78] was used for all noncovalent interactions (NC) datapoints except the rare-gas dimers (aug-cc-pVQZ). Furthermore, the noncovalent interactions were computed *without* counterpoise corrections. For B97-D2, Grimme's DFT-D2 dispersion tail was used with an s_6 coefficient[79] of 0.75. Grimme's B97-D functional[54] uses the DFT-D2 dispersion tail as well, with an s_6 coefficient of 1.25. All of the calculations were performed with a development version of Q-Chem 4.0[80].

2.3 Theory

The complete functional form for all of the trained functionals is given by Equations 2.6-2.8. The components of the exchange functional and correlation functional are described in Sections 2.3.1 and 2.3.2, respectively. The acronyms used in Equations 2.6-2.8 (and henceforth) are: exchange-correlation (xc), exchange (x), correlation (c), short-range (sr), long-range (lr), same-spin (ss), opposite-spin (os), and dispersion (disp).

$$E_{xc} = E_x + E_c \tag{2.6}$$

$$E_x = E_x^{B97} + c_x E_{x,sr}^{exact} + d_x E_{x,lr}^{exact} \tag{2.7}$$

$$E_c = E_{c,ss}^{B97} + E_{c,os}^{B97} + E_{disp} \tag{2.8}$$

For local (exchange) functionals, $c_x = d_x = 0$, while for global hybrid functionals, $c_x = d_x$, where c_x is the (global) fraction of exact exchange. For range-separated hybrid functionals, $d_x = 1$, $E_x^{B97} = E_{x,sr}^{B97}$ (Section 2.3.1), and c_x is the fraction of short-range exact exchange.

2.3.1 Exchange Functional Form

The local exchange component of the B97 functional form is given by Equations 2.9 and 2.10:

$$E_x^{B97} = \sum_{\sigma}^{\alpha,\beta} \int e_{x,\sigma}^{UEG}(\rho_\sigma) g_x\left(u_{x,\sigma}\right) d\mathbf{r} \tag{2.9}$$

$$g_x\left(u_{x,\sigma}\right) = \sum_{i=0}^{m_x} c_{x,i} u_{x,\sigma}^{i} = \sum_{i=0}^{m_x} c_{x,i} \left[\frac{\gamma_x s_\sigma^2}{1+\gamma_x s_\sigma^2}\right]^i \tag{2.10}$$

where the dimensionless variable, $u_{x,\sigma} \in [0,1]$, is a finite domain transformation of the reduced spin-density gradient, $s_\sigma = \frac{|\nabla \rho_\sigma|}{\rho_\sigma^{4/3}} \in [0,\infty)$. In Equation 2.9, $e_{x,\sigma}^{UEG}(\rho_\sigma)$ is the exchange energy density per unit volume of a uniform electron gas (UEG) and $g_x\left(u_{x,\sigma}\right)$ is the exchange functional inhomogeneity correction factor (ICF). The linear local exchange parameters, $c_{x,i}$, will be determined by least-squares fitting to a training set in Section 2.5, while $\gamma_x = 0.004$ is a nonlinear local exchange parameter that was fit to the Hartree–Fock exchange energies of 20 atoms in 1986 by Becke[9]. For range-separated hybrid functionals, the conventional Coulomb operator in the local exchange component is attenuated by the complementary error function (erfc), resulting in an additional multiplicative factor, $F\left(a_\sigma\right)$, in the integrand of Equation 2.9:

$$\begin{aligned} F\left(a_\sigma\right) = 1 - \frac{2}{3} a_\sigma \Bigg[2\sqrt{\pi}\,\mathrm{erf}\left(\frac{1}{a_\sigma}\right) - 3a_\sigma + \\ a_\sigma^3 + \left[2a_\sigma - a_\sigma^3\right] \exp\left(-\frac{1}{a_\sigma^2}\right) \Bigg] \end{aligned} \tag{2.11}$$

where $a_\sigma = \frac{\omega}{k_{F\sigma}}$, ω is the nonlinear range-separation parameter that controls the transition from local exchange to nonlocal exact exchange with respect to interelectronic distance, and $k_{F\sigma} = \left[6\pi^2 \rho_\sigma\right]^{1/3}$ is the spin-polarized Fermi wave vector. The inclusion of $F\left(a_\sigma\right)$ in the integrand of Equation 2.9 gives $E_{x,sr}^{B97}$.

When considering both global hybrid and range-separated hybrid functionals, the most general way to deal with nonlocal exchange is to split the Coulomb operator in the conventional expression for exact exchange into a short-range component ($E_{x,sr}^{exact}$) and a long-range component ($E_{x,lr}^{exact}$) with the erfc and erf Coulomb functions, respectively:

$$\begin{aligned} E_{x,sr}^{exact} &= -\frac{1}{2}\sum_{\sigma}^{\alpha,\beta}\sum_{i,j}^{occ.}\int\int \psi_{i\sigma}^{*}(\mathbf{r}_1)\,\psi_{j\sigma}^{*}(\mathbf{r}_2)\,\frac{\mathrm{erfc}\,(\omega r_{12})}{r_{12}} \\ &\times \psi_{j\sigma}(\mathbf{r}_1)\,\psi_{i\sigma}(\mathbf{r}_2)\,d\mathbf{r}_1 d\mathbf{r}_2 \end{aligned} \tag{2.12}$$

$$\begin{aligned} E_{x,lr}^{exact} &= -\frac{1}{2}\sum_{\sigma}^{\alpha,\beta}\sum_{i,j}^{occ.}\int\int \psi_{i\sigma}^{*}(\mathbf{r}_1)\,\psi_{j\sigma}^{*}(\mathbf{r}_2)\,\frac{\mathrm{erf}\,(\omega r_{12})}{r_{12}} \\ &\times \psi_{j\sigma}(\mathbf{r}_1)\,\psi_{i\sigma}(\mathbf{r}_2)\,d\mathbf{r}_1 d\mathbf{r}_2 \end{aligned} \tag{2.13}$$

where $\psi_{i\sigma}$ and $\psi_{j\sigma}$ are the occupied Kohn–Sham spatial orbitals. Since $\frac{\mathrm{erfc}(\omega r_{12})}{r_{12}} + \frac{\mathrm{erf}(\omega r_{12})}{r_{12}} = \frac{1}{r_{12}}$, $E_x^{exact} = E_{x,sr}^{exact} + E_{x,lr}^{exact}$ for global hybrids, where $c_x = d_x$ is the fraction of (global) exact exchange. For range-separated hybrids, instead of setting the percentage of exact-exchange at $r = 0$ to zero, an (optional) optimizable parameter, c_x, controls the amount of short-range exact exchange. Additionally, the value of ω is fixed at 0.3 for all of the range-separated hybrid functionals.

2.3.2 Correlation Functional Form

The local correlation component of the B97 functional form is given by Equations 2.14-2.17:

$$E_{c,ss}^{B97} = \sum_{\sigma}^{\alpha,\beta}\int e_{c,\sigma\sigma}^{PW92}\, g_{c,ss}\,(u_{c,\sigma\sigma})\, d\mathbf{r} \tag{2.14}$$

$$g_{c,ss}\,(u_{c,\sigma\sigma}) = \sum_{i=0}^{m_{css}} c_{css,i}\, u_{c,\sigma\sigma}^{i} = \sum_{i=0}^{m_{css}} c_{css,i}\left[\frac{\gamma_{css} s_{\sigma}^2}{1+\gamma_{css} s_{\sigma}^2}\right]^i \tag{2.15}$$

$$E_{c,os}^{B97} = \int e_{c,\alpha\beta}^{PW92}\, g_{c,os}\,(u_{c,\alpha\beta})\, d\mathbf{r} \tag{2.16}$$

$$g_{c,os}\,(u_{c,\alpha\beta}) = \sum_{i=0}^{m_{cos}} c_{cos,i}\, u_{c,\alpha\beta}^{i} = \sum_{i=0}^{m_{cos}} c_{cos,i}\left[\frac{\gamma_{cos} s_{\alpha\beta}^2}{1+\gamma_{cos} s_{\alpha\beta}^2}\right]^i \tag{2.17}$$

where $s_{\alpha\beta}^2 = \frac{1}{2}\left(s_{\alpha}^2 + s_{\beta}^2\right)$, and $e_{c,\sigma\sigma}^{PW92}$ and $e_{c,\alpha\beta}^{PW92}$ are the PW92[7] same-spin and opposite-spin correlation energy densities per unit volume[81]. The linear local correlation parameters, $c_{css,i}$ and $c_{cos,i}$, will be determined by least-squares fitting to a training set in Section 2.5, while $\gamma_{css} = 0.2$ and $\gamma_{cos} = 0.006$ are nonlinear local correlation parameters that were fit to the correlation energies of helium and neon in 1997 by Becke[35].

Since functionals are trained both with and without dispersion corrections, the E_{disp} term requires further elaboration. When dispersion corrections are not used, $E_{disp} = 0$. Two different dispersion correction methods are used in combination with the local, GH, and RSH functionals: one dispersion tail (DT) and one nonlocal correlation (NLC) functional.

The dispersion tail (DFT-D2) has the following form:

$$E_{disp}^{DFT-D2} = -s_6 \sum_{i=1}^{N_{at}-1} \sum_{j=i+1}^{N_{at}} \frac{C_6^{ij}}{R_{ij}^6} f_{damp}^{DFT-D2}(R_{ij}) \tag{2.18}$$

$$f_{damp}^{DFT-D2}(R_{ij}) = \frac{1}{1 + e^{-d(R_{ij}/R_{r,ij}-1)}} \tag{2.19}$$

In the damping function, $R_{r,ij} = R_{0,i} + R_{0,j}$ is the sum of the van der Waals (vdW) radii of a pair of atoms, $C_6^{ij} = \sqrt{C_6^i C_6^j}$ is the dispersion coefficient of a pair of atoms, and $d = 20$. In training the DFT-D2 dispersion tail onto the density functionals, the linear s_6 coefficient is optimized and counted as a linear parameter. The empirical C_6 parameters and vdW Radii, R_0, can be found in Table 1 of Reference 54.

The nonlocal correlation functional that is considered is VV10[57]:

$$E_{disp}^{VV10} = \int \rho(\mathbf{r}) \left[\frac{1}{32} \left[\frac{3}{b^2} \right]^{3/4} + \frac{1}{2} \int \rho(\mathbf{r}') \Phi(\mathbf{r}, \mathbf{r}', \{b, C\}) \, d\mathbf{r}' \right] d\mathbf{r} \tag{2.20}$$

where $\Phi(\mathbf{r}, \mathbf{r}', \{b, C\})$ is the nonlocal correlation kernel defined in Reference 57. The VV10 NLC functional introduces 2 nonlinear parameters: b, which controls the short-range damping of the $1/r^6$ asymptote, and C, which controls the accuracy of the asymptotic C_6 coefficients. Since it is much more difficult to train the nonlinear parameters of the VV10 NLC functional, the parameters that were optimized for ωB97X-V ($b = 6$ and $C = 0.01$) are used here without further optimization.

2.4 Datasets

In total, the training and test sets used for the parameterization and validation of the candidate functionals contain 2301 datapoints, requiring 1961 single-point calculations. Of the 2301 datapoints, 1108 belong to the training set and 1193 belong to the test set. Furthermore, the training and test sets contain both thermochemistry (TC) data as well as noncovalent interactions (NC) data. The training set contains

787 TC datapoints and 321 NC datapoints, while the test set contains 146 TC datapoints and 1047 NC datapoints (for an overall total of 933 TC datapoints and 1368 NC datapoints). The partitioning of the training and test sets was carried out with the quality of the reference values in mind, such that the training set contains the highest confidence data. Table 2.2 lists the 36 datasets that form the training and test sets. The references for the datasets are given in the rightmost column of Table 2.2, while additional details can be found in Reference 63. In addition to the general division into TC and NC data for the training and test sets, we will report the results for the 3 rare-gas (RG) potential energy curves (PEC) separately, as a delicate diagnostic of the balance between Pauli repulsion and attractive dispersion interactions.

2.5 Training Methodology

In order to train and test the candidate functionals, single-point calculations are carried out with the unoptimized functionals ($g_x = g_{c,ss} = g_{c,os} = 1$) for the 1961 geometries that correspond to the 2301 datapoints in the training and test sets. In order to gather all of the data that will be used for the upcoming analysis, 4 sets of calculations must be carried out: 1). LSDA without VV10, 2). LSDA with VV10, 3). SR-LSDA without VV10, and 4). SR-LSDA with VV10. The VV10 calculations must be carried out separately because the VV10 NLC functional is implemented within the SCF procedure. Conveniently, however, running the LSDA (local spin density approximation) functional is sufficient for gathering data for both the local and global hybrid variants, since a global hybrid functional with an initial guess of $c_x = 0$ is a local functional. Following the single-point calculations, the resulting 4 sets of densities are saved to disk and used to calculate the values that the expansion coefficients in the power series ($c_{x,i}$, $c_{css,i}$, and $c_{cos,i}$) multiply, up to fourth order ($i \in [0, 4]$). In addition to these 15 contributions, the value of E_x^{exact} is required for GH functionals and the value of $E_{x,sr}^{exact}$ is required for RSH functionals.

The calculated values are used to form a (# of Datapoints) x (# of Linear Parameters) matrix, A, that contains the appropriate contributions for a given datapoint. In addition to the A matrix, a column of values corresponding to the errors in the unoptimized functional ($y = E_{REF} - E_{DFT}$) is computed. Since weights are used during training, a diagonal (# of Datapoints) x (# of Datapoints) training set weight matrix (W_{Train}) is required as well. The diagonal elements corresponding to the training set data contain the appropriate weights, while the remaining diagonal elements corresponding to the test set data are set to zero. The change in the linear parameters, Δb, is found by a weighted least-squares fit:

Name	#	Description	Ref.
HAT707	505	Heavy-atom transfer reaction energies	82
BDE99	83	Bond dissociation reaction energies	82
TAE_nonMR124	124	Total atomization energies	82
SN13	13	Nucleophilic substitution reaction energies	82
ISOMER20	18	Isomerization reaction energies	82
DBH24	24	Diverse barrier heights	83,84
EA6	6	Electron affinities of atoms	85
IP6	6	Ionization potentials of atoms	85
AE8	8	Absolute atomic energies	86
SW49Rel345	28	${SO_4}^{2-}(H_2O)_n$ $(n = 3 - 5)$ relative energies	87
SW49Bind345	30	${SO_4}^{2-}(H_2O)_n$ $(n = 3 - 5)$ binding energies	87
NBC10A2	37	Methane dimer and benzene-methane dimer PECs	88,89
HBC6A	118	Formic acid, formamide acid, and formamidine acid dimer PECs	90,89
BzDC215	108	Benzene and first- and second-row hydride PECs	91
EA7	7	Electron affinities of small molecules	85
IP7	7	Ionization potentials of small molecules	85
Gill12	12	Neutral, radical, anionic, and cationic isodesmic reaction energies	92
AlkAtom19	19	$n = 1 - 8$ alkane atomization energies	93
AlkIsomer11	11	$n = 4 - 8$ alkane isomerization energies	93
AlkIsod14	14	$n = 3 - 8$ alkane isodesmic reaction energies	93
HTBH38	38	Hydrogen transfer barrier heights	94
NHTBH38	38	Non-hydrogen transfer barrier heights	95
SW49Rel6	17	${SO_4}^{2-}(H_2O)_n$ $(n = 6)$ relative energies	87
SW49Bind6	18	${SO_4}^{2-}(H_2O)_n$ $(n = 6)$ binding energies	87
NNTT41	41	Neon-neon PEC	96
AATT41	41	Argon-argon PEC	96
NATT41	41	Neon-argon PEC	96
NBC10A1	53	Parallel-displaced (3.4 Å), sandwich, and T-shaped benzene dimer PECs	88,89
NBC10A3	39	S2 and T3 configuration pyridine dimer PECs	97,89
WATER27	23	Neutral and charged water interactions	98,99
HW30	30	Hydrocarbon and water dimers	100
NCCE31	18	Noncovalent complexation energies	101
S22x5	110	Hydrogen-bonded and dispersion-bonded complex PECs	102
S66x8	528	Biomolecular structure complex PECs	103
S22	22	Equilibrium geometries from S22x5	104,89
S66	66	Equilibrium geometries from S66x8	103,105

Table 2.2: Summary of the datasets found in the training and test sets. The datasets above the thick black line are in the training set and the datasets below the thick black line are in the test set. Within the training and test sets, datasets above the thin black line contain thermochemistry datapoints, while datasets below the thin black line contain noncovalent interactions datapoints. PEC stands for potential energy curve.

$$\Delta b = (A^T W_{Train} A)^{-1} (A^T W_{Train} y) \tag{2.21}$$

and the training set RMSD is calculated by:

$$RMSD_{Train} = \sqrt{\frac{\operatorname{diag}(W_{Train}) \cdot (y - A\Delta b)^2}{\#_{Train}}} \tag{2.22}$$

Additional statistical measures are calculated using Equation 2.22 with the appropriate weight matrix and #.

In total, 10 quantities will be used to gauge the performance of the resulting functionals: the training set RMSD (RMSD$_{Train}$), the test set RMSD (RMSD$_{Test}$), the RMSD for the 3 rare-gas dimer PECs (RMSD$_{RG}$), the total RMSD (RMSD$_{Total}$), the thermochemistry (TC) RMSD (RMSD$_{TC}$), the noncovalent interactions (NC) RMSD (RMSD$_{NC}$), the training set TC RMSD (RMSD$_{TC,Train}$), the test set TC RMSD (RMSD$_{TC,Test}$), the training set NC RMSD (RMSD$_{NC,Train}$), and the test set NC RMSD (RMSD$_{NC,Test}$).

Since contributions are computed up to fourth order for the exchange, same-spin correlation, and opposite-spin correlation functionals, as many as 15 linear GGA parameters are available for optimization. The optional short-range exchange parameter that is unique to range-separated hybrid functionals adds a 16th parameter for the RSHs. The uniform electron gas (UEG) constraint for the same-spin and opposite-spin correlation functionals can be incorporated by making $c_{css,0}$ and $c_{cos,0}$ optional parameters, but the same cannot be done with the UEG constraint for exchange (except for local functionals). Thus, fits that violate the UEG limit for exchange are optimized separately from fits that do incorporate the UEG limit for exchange. As an example of the number of fits that result from this methodology, local functionals that are constructed to satisfy the UEG constraint for exchange have 14 optional parameters, giving a total of $\sum_{i=1}^{14} \binom{14}{i} = 2^{14} - 1 = 16383$ possible fits. Table 2.3 lists the total number of fits for local, GH, and RSH functionals with and without the UEG limit for exchange in place.

In order to refer to the thousands of resulting functionals with clarity, we will use a nomenclature that is fully specified in Table 2.4. As examples, “GN-012.012.012.Xn” would describe Becke’s 10-parameter B97 functional, “LD-012.012.012.0n” would describe Grimme’s 10-parameter B97-D functional, “RN-1234.1234.1234.Xy” would describe the 13-parameter ωB97X functional, and “RV-12.01.01.Xy” would describe the 7-parameter ωB97X-V functional. As can be seen with the descriptor for ωB97X-V, since the UEG limit for exchange was used as a constraint, “0” does not appear in the label for the exchange functional ICF (even though $c_{x,0} \neq 1$), because the

#	$c_{x,0} + c_x = 1$	$c_{x,0} + c_x \neq 1$
Local	$2^{14} - 1 = 16383$	$2^{15} - 2^{14} = 16384$
GH	$2^{14} - 1 = 16383$	$2^{15} - 1 = 32767$
RSH	$2^{15} - 1 = 32767$	$2^{16} - 2^{14} = 49152$

Table 2.3: Total number of least-squares fits (#) that can be performed when considering parameters up to fourth order in the power series inhomogeneity correction factors. While the type of dispersion correction used has no bearing on the total number of possible fits, whether or not the UEG constraint for exchange is enforced is important and is addressed in the second and third columns, respectively.

Symbol	Meaning	Allowed Values	Meaning
i	exchange	L	local
		G	global hybrid
		R	range-separated hybrid
j	dispersion correction	N	none
		D	DFT-D2 dispersion tail
		V	VV10 nonlocal correlation functional
$\{p_x\}$	linear exchange parameters	any subset of 01234	each included integer, m, is a single parameter multiplying u_x^m
$\{p_{css}\}$	linear same-spin correlation parameters	any subset of 01234	each included integer, m, is a single parameter multiplying $u_{c,\sigma\sigma}^m$
$\{p_{cos}\}$	linear opposite-spin correlation parameters	any subset of 01234	each included integer, m, is a single parameter multiplying $u_{c,\alpha\beta}^m$
k	(short-range) exact exchange	0	no (short-range) exact exchange included
		X	(short-range) exact exchange included
l	UEG for exchange	y	UEG limit for exchange is enforced
		n	UEG limit for exchange is not enforced

Table 2.4: Explanation of the nomenclature for the descriptors that refer to the thousands of optimized functionals. A given descriptor takes on the following form: "ij-$\{p_x\}$.$\{p_{css}\}$.$\{p_{cos}\}$.kl". If none of the coefficients of a given ICF are optimized, $\emptyset$ is used as a placeholder. As an example, "GN-012.012.012.Xn" would describe Becke's 10-parameter B97 functional.

4th label, "Xy", implies that $c_{x,0} = 1 - c_x$. In addition, nonlinear parameters are not counted when considering the number of parameters corresponding to a given fit , since the nonlinear parameters were not varied in this work. Henceforth, any mention of the number of parameters implicitly refers to the number of linear parameters. As a more complicated example, if a Local+DFT-D2 functional requires the optimization of $\{c_{x,1}, c_{x,3}, c_{x,4}, c_{cos,0}, c_{cos,2}\}$, "LD-134.$\emptyset$.02.0y" will be used as its descriptor. Henceforth, quotations will not be used for the descriptors.

As far as weights are concerned, thermochemistry datapoints in the training and test sets are given weights of 1 and 2.5 respectively (except for datapoints in EA6 and IP6 which are weighted by 5), noncovalent interactions datapoints in the training and test sets are given weights of 10 and 25, respectively, and datapoints corresponding to the 3 rare-gas dimer PECs in the test set are given weights of 2500. Even though the 3 rare-gas dimer PECs are technically in the test set, they are not included in the calculation of RMSD_{Test}. However, they are included in the calculation of RMSD_{Total}. The 3 rare-gas dimer PECs are included in the NC and test set NC RMSDs because their unweighted contributions are very small and do not contribute significantly. Of the 10 RMSDs, only the first 4 are weighted, while the latter 6 are unweighted.

2.6 Training Results

It is important to point out that the selection procedure utilized to identify the optimal functionals is not (and cannot be) unique. However, as we shall see, it recovers the self-consistently optimized ωB97X-V functional, even though a slightly different selection procedure was used in Reference 63. In addition, the resulting optimal functionals are usually significantly better than existing functionals of the same type, as will be discussed in Section 2.7.

While a variety of selection procedures were initially explored, the one that was finally chosen is quite simple. First, the total (weighted) RMSDs are computed and plotted. Next, a screening process rejects fits that predict rare-gas dimer equilibrium bond lengths that are too long or too short by more than 0.1 Å. Since the plots are still overflowing with data points, all of the points for a fixed number of linear parameters are removed, except for the point that corresponds to the lowest total RMSD with and without the UEG constraint for exchange. The resulting plots (Figure 2.3) are much simpler to analyze and contain filled circles (satisfy the UEG constraint for exchange) and unfilled circles (do not satisfy the UEG constraint for exchange).

Starting at the fewest number of linear parameters, an additional empirical parameter is accepted if the improvement in the total RMSD is more than 0.05 kcal/mol. This final stage does not take into account whether or not the UEG constraint for exchange is enforced. The 9 optimal functionals are chosen in this manner and will be discussed and compared to existing functionals in Section 2.7. Since the optimal functionals are chosen based on their total RMSDs, the corresponding training and test set RMSDs of the optimal functionals are shown in red in Figures 2.1 and 2.2.

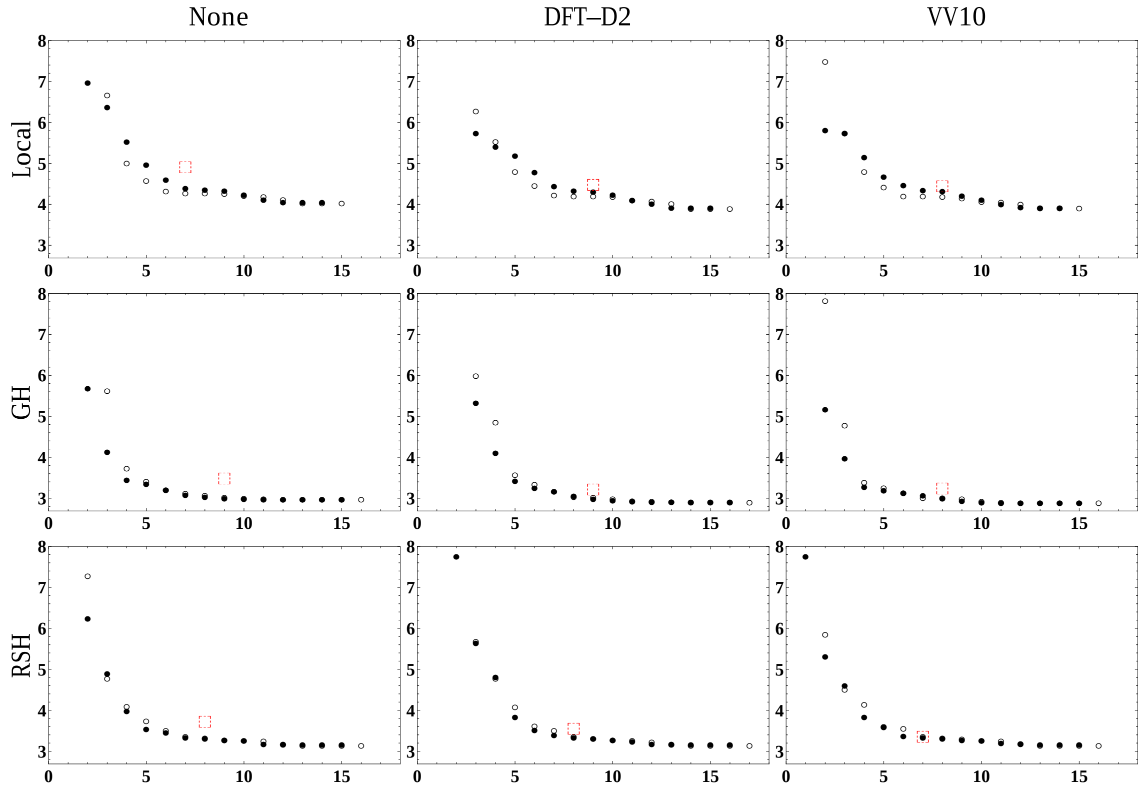


Figure 2.1: Plots showing the lowest training set RMSD (in kcal/mol) for a fixed number of linear parameters for all 9 candidate functional forms considered. Filled circles correspond to fits which satisfy the UEG limit for exchange and unfilled circles indicate that the UEG limit for exchange was allowed to relax. The red box indicates the training set RMSD of the optimal functional, which is usually not the best for the training set data alone.

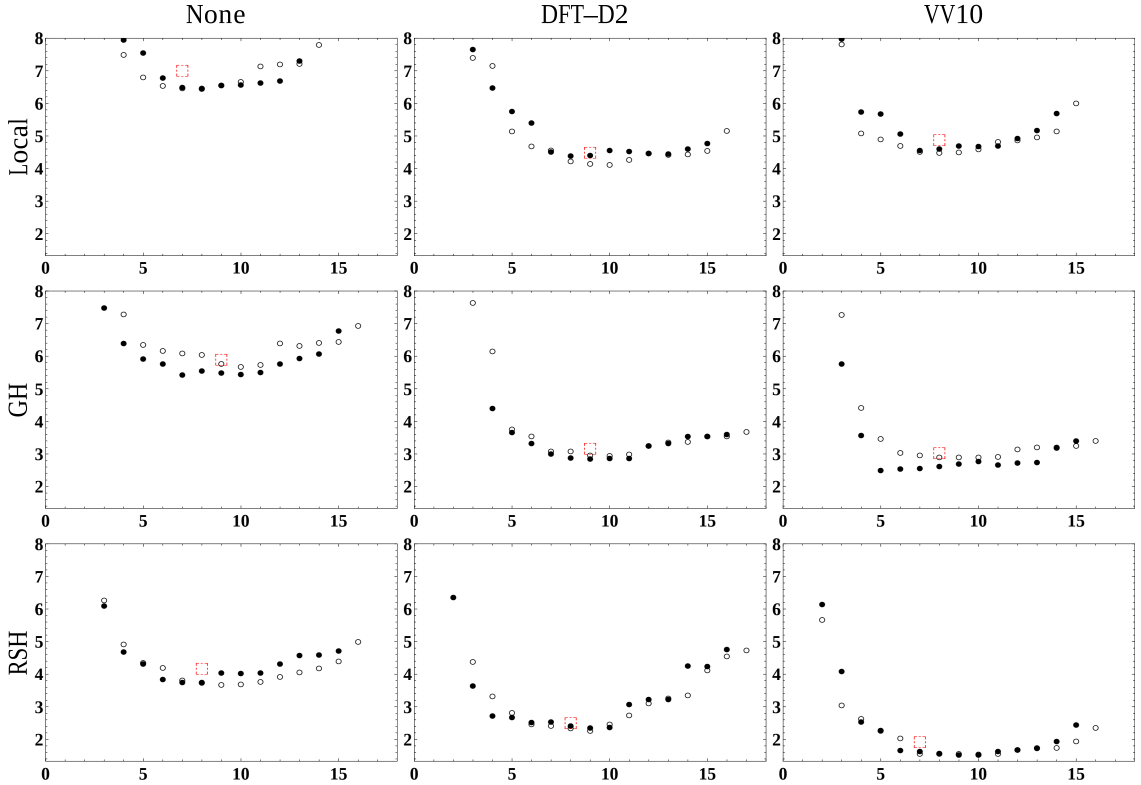


Figure 2.2: Plots showing the lowest test set RMSD (in kcal/mol) for a fixed number of linear parameters for all 9 candidate functional forms considered. Filled circles correspond to fits which satisfy the UEG limit for exchange and unfilled circles indicate that the UEG limit for exchange was allowed to relax. The red box indicates the test set RMSD of the optimal functional.

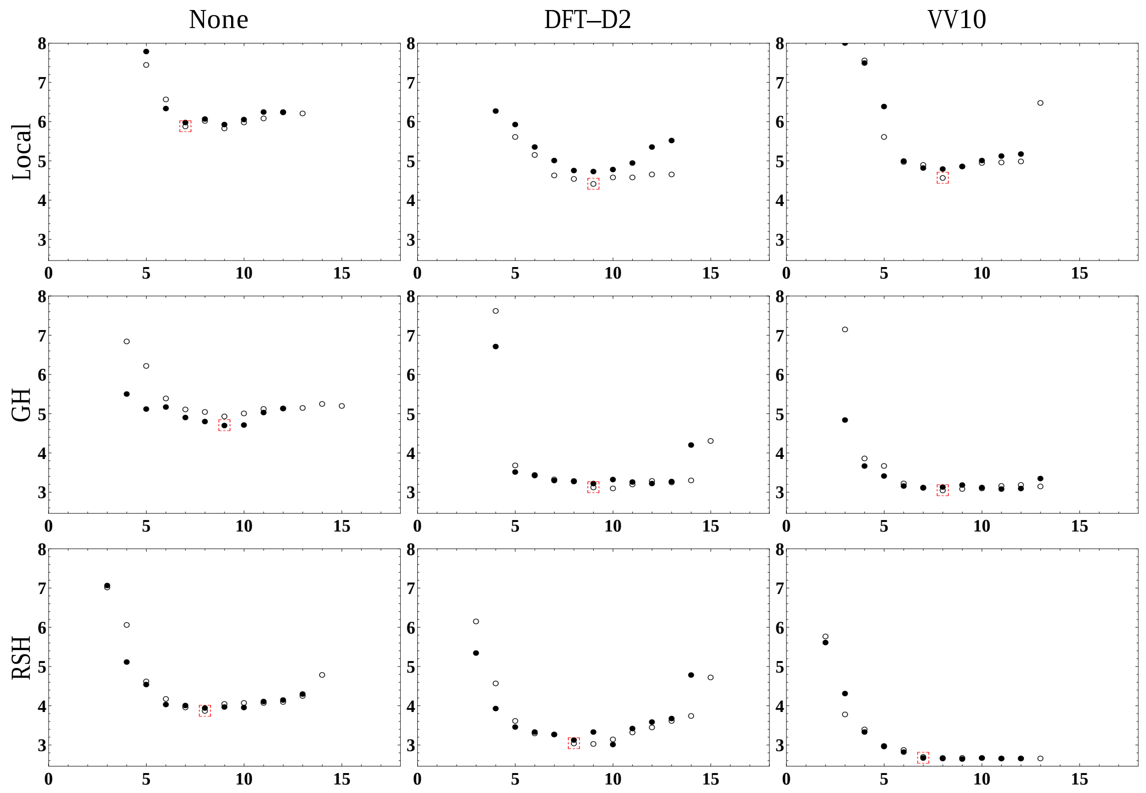


Figure 2.3: Plots showing the lowest total RMSD (in kcal/mol) for a fixed number of linear parameters for all 9 candidate functional forms considered. Filled circles correspond to fits which satisfy the UEG limit for exchange and unfilled circles indicate that the UEG limit for exchange was allowed to relax. The red box indicates the total RMSD of the optimal functional. Due to the screening process described in Section 2.6, points that correspond to fits that predict rare-gas dimer equilibrium bond lengths that are too long or too short by more than 0.1 Å have been removed.

All of the RMSDs considered in this section are generated using Equation 2.22 (with the appropriate weight matrix and #) and are least-squares fit RMSDs. While none of the functionals are self-consistently optimized, the recent self-consistent optimization of ωB97X-V indicated that the least-squares fit RMSDs generally differ from the actual RMSDs of the self-consistently optimized functional by 0.05 kcal/mol on average. While it would be impractical to self-consistently optimize thousands of functionals, we firmly believe that this procedure is effective in predicting the quality of a functional based on least-squares fit RMSDs.

Since the parameters that are obtained from all of these fits are not self-consistently optimized, it is not immediately obvious how much they will differ from the final set of parameters. Thus, it is difficult to comment on the usefulness of the parameters of the 9 resulting optimal functionals. However, the parameters for Becke's B97 functional were optimized in the same post-LSDA manner as all of the functionals considered in this chapter, and comparing the parameters of B97 and B97-1, or alternatively considering Table 3 from Reference 63, indicates that the self-consistently optimized parameters do not differ drastically from those from the end of the first optimization cycle. While it is best to self-consistently optimize the parameters of a semi-empirical density functional, the parameters for the 9 optimal functionals are provided in Table 2.5.

Tables 2.6, 2.7, and 2.8 contain data for range-separated hybrid, global hybrid, and local functionals, respectively. Each method of accounting for exchange was trained both with dispersion corrections (DFT-D2 and VV10) and without dispersion corrections (None). For each pairing, the columns labeled "Minimum" contain the best possible value for a given RMSD category, while the columns labeled "Optimal" contain the results for the functionals that were selected from Figure 2.3. For the remainder of this section, the least-squares fit RMSDs will simply be referred to as RMSDs.

We begin the analysis with the RSH+VV10 category, since our newest density functional, ωB97X-V, belongs to this class. Generally speaking, the interesting comparisons in Table 2.6 are between the best possible result in a given row for *any* candidate RSH+VV10 functional (i.e. the Minimum VV10 column) and the corresponding result obtained with the optimal functional (i.e. the Optimal VV10 column). The optimal RSH+VV10 functional coincides with the ωB97X-V functional, and, as summarized in Table 2.5, has 7 linear parameters. Compared to the smallest training set RMSD possible (3.14 kcal/mol), a value of 3.36 kcal/mol is certainly reasonable for a functional with 9 fewer linear parameters. Similarly, most other comparisons show that the functional chosen by our selection method yields results for the other reported RMSDs that are competitive with the best values attainable. The largest difference is for the RMSD$_{RG}$, where it is possible to do nearly twice as

	None	DFT-D2	VV10
Local	LN-012.∅.0134.0n (7) {1.07, -0.94, 5.04, 0.45, 8.83, -65.31, 39.84} N/A	LD-012.0.0123.0n (9) {1.09, -0.89, 4.96, 0.25, 0.43, 12.06, -27.85, -1.57} s_6=0.71	LV-012.0.0123.0n (8) {1.09, -0.79, 4.74, 0.43, 0.41, 11.90, -27.19, -1.52} N/A
GH	GN-24.1234.13.Xy (9) {1.95, 2.70, -6.54, 33.21, -49.17, 22.23, 3.33, -24.59} c_x=0.23	GD-0234.0.04.Xn (9) {0.79, 2.46, -3.89, 4.73, 0.33, 1.10, -13.90} c_x=0.24; s_6=0.64	GV-02.0.0234.Xn (8) {0.81, 2.00, 0.51, 1.05, 9.19, -39.37, 22.86} c_x=0.22
RSH	RN-14.34.012.Xn (8) {0.58, 11.25, -9.17, 9.08, 1.09, 2.67, -10.32} c_x=0.02	RD-02.12.01.Xn (8) {0.87, 2.24, -3.62, 3.30, 1.35, -2.36} c_x=0.18; s_6=0.71	RV-12.01.01.Xy (7) {0.61, 1.18, 0.58, -0.27, 1.22, -1.87} c_x=0.16

Table 2.5: Characteristics of the 9 optimal functionals. Within a cell, the first row lists the descriptor (Table 2.4 with the number of associated linear parameters in parentheses, the second row lists the non-self-consistently optimized GGA parameters, and the third row (when applicable) lists the values for the (short-range) exact exchange parameter, c_x, and the linear DFT-D2 dispersion coefficient, s_6.

RSH	Minimum			Optimal		
kcal/mol	None	DFT-D2	VV10	None	DFT-D2	VV10
Train	3.14	3.14	3.14	3.73	3.55	3.36
Test	3.68	2.28	1.53	4.18	2.50	1.92
RG	0.88	0.80	0.51	1.97	2.50	0.95
Total	3.87	3.03	2.66	3.87	3.05	2.68
TC	3.49	3.47	3.44	4.09	3.80	3.62
NC	0.58	0.36	0.24	0.71	0.43	0.32
TC,Train	3.55	3.56	3.56	4.28	4.02	3.86
TC,Test	2.19	1.76	1.66	2.85	2.26	1.81
NC,Train	0.25	0.26	0.23	0.42	0.42	0.31
NC,Test	0.62	0.33	0.23	0.77	0.43	0.32

Table 2.6: RMSDs in kcal/mol for range-separated hybrid functionals. The data in the training and test sets consists of thermochemical (TC) and noncovalent (NC) energy differences. The 3 rare-gas (RG) dimer PEC RMSDs are reported separately. The "Minimum" columns contain the smallest possible RMSD value for the particular entry from *all* trained functionals of that class. Hence, each entry within a column generally corresponds to a *different* functional. The "Optimal" columns contain the RMSD value for the best overall functional selected from that class. Hence, each entry within a column corresponds to the *same* functional. Details regarding the optimal functional are provided in Table 2.5.

GH	Minimum			Optimal		
kcal/mol	None	DFT-D2	VV10	None	DFT-D2	VV10
Train	2.97	2.90	2.89	3.49	3.22	3.25
Test	5.44	2.86	2.51	5.90	3.17	3.03
RG	0.76	0.67	0.31	2.01	1.82	0.65
Total	4.72	3.11	3.06	4.72	3.14	3.06
TC	3.45	3.31	3.31	3.87	3.48	3.61
NC	0.91	0.42	0.38	0.97	0.54	0.48
TC,Train	3.24	3.21	3.18	3.78	3.62	3.70
TC,Test	3.13	2.60	2.59	4.28	2.60	3.06
NC,Train	0.41	0.37	0.35	0.48	0.48	0.42
NC,Test	1.01	0.40	0.38	1.08	0.56	0.50

Table 2.7: RMSDs in kcal/mol for global hybrid functionals. The format is explained in the caption of Table 2.6.

Local	Minimum			Optimal		
kcal/mol	None	DFT-D2	VV10	None	DFT-D2	VV10
Train	4.03	3.89	3.91	4.91	4.48	4.44
Test	6.46	4.13	4.49	7.00	4.50	4.88
RG	0.97	1.34	0.72	2.09	3.15	2.57
Total	5.84	4.43	4.57	5.89	4.43	4.57
TC	5.03	4.69	4.85	5.72	5.23	5.27
NC	0.95	0.51	0.54	1.08	0.61	0.65
TC,Train	4.41	4.30	4.32	5.55	5.13	5.11
TC,Test	5.83	4.87	5.50	6.57	5.71	6.06
NC,Train	0.59	0.56	0.53	0.65	0.62	0.55
NC,Test	1.03	0.48	0.54	1.18	0.61	0.68

Table 2.8: RMSDs in kcal/mol for local functionals. The format is explained in the caption of Table 2.6.

kcal/mol	Train	Test	RG	Total	TC	NC	TC,Train	TC,Test	NC,Train	NC,Test
RV-12.01.01.Xy	3.36	1.92	0.95	2.68	3.62	0.32	3.86	1.81	0.31	0.32
RV-1.1.1.Xy	3.84	3.86	1.34	3.76	4.64	0.39	4.42	5.68	0.36	0.40
RV-12.12.12.Xy	3.60	2.25	1.82	2.96	3.92	0.37	4.17	2.09	0.31	0.38
RV-123.123.123.Xy	3.50	2.78	7.28	3.51	3.77	0.49	4.01	2.04	0.41	0.51
RV-1234.1234.1234.Xy	3.34	1.94	7.91	3.24	3.63	0.31	3.84	2.15	0.32	0.30
RV-01.01.01.Xn	3.59	2.72	1.87	3.13	3.83	0.48	4.09	1.86	0.40	0.50
RV-012.012.012.Xn	3.34	2.17	1.62	2.78	3.61	0.36	3.83	1.99	0.31	0.37
RV-0123.0123.0123.Xn	3.30	2.37	1.28	2.82	3.56	0.38	3.76	2.22	0.31	0.40
RV-01234.01234.01234.Xn	3.14	2.36	14.00	4.22	3.47	0.33	3.57	2.91	0.28	0.34

Table 2.9: RMSDs for the optimal RSH+VV10 functional, as well as functionals that would be considered if the present methodology was not being utilized. While the nomenclature is explained in Section 2.5, the first functional corresponds to the optimal RSH+VV10 fit (which coincides with the functional form of ωB97X-V), the next 4 functionals are uniformly truncated $m = 1$ through $m = 4$ fits with all of the UEG constraints enforced, while the last 4 are uniformly truncated $m = 1$ through $m = 4$ fits with none of the UEG constraints enforced. The fraction of short-range exact exchange is optimized for all of the fits in this table.

well (of course at the expense of TC results) as our chosen functional. Nonetheless, the performance of the chosen functional for the 3 rare-gas dimer PECs is actually much better than virtually all existing functionals, as will be seen in Section 2.7.

For the RSH+VV10 category only, we include additional data in Table 2.9 for functionals that would be considered if the present methodology was not being utilized, to demonstrate that our procedure for selecting the optimal functional is effective. Since the 16-parameter RV-01234.01234.01234.Xn functional has the lowest training set RMSD (3.14 kcal/mol), it is useful to compare the test set RMSD of this functional (2.36 kcal/mol) with that of the optimal 7-parameter RSH+VV10 functional (1.92 kcal/mol). While the training set RMSD of the optimal RSH+VV10 functional is 0.22 kcal/mol larger than that of the RV-01234.01234.01234.Xn functional, its test set RMSD is smaller by more than 0.40 kcal/mol. The optimal functional's performance on the TC data in the test set is more than 1.5 times better than RV-01234.01234.01234.Xn, and its RMSD$_{RG}$ is smaller by a factor of 15. These results demonstrate the improved transferability of the optimal 7-parameter functional against a 16-parameter alternative, which comes at the necessary expense of slightly poorer training set performance.

Considering the 4 functionals in Table 2.9 that satisfy the UEG limits for exchange and correlation, the lowest total RMSD is attained by the RV-12.12.12.Xy functional (7 linear parameters). Since this functional is equivalent to the optimal RSH+VV10 functional with respect to the number of linear parameters, comparing the two highlights the advantages of the present scheme. The optimal RSH+VV10 functional beats the RV-12.12.12.Xy functional in all 10 RMSD categories, and by considerable margins for most. Applying the same analysis to the functionals that do not satisfy the UEG limits, the RV-012.012.012.Xn functional with 10 linear parameters emerges as the one with the lowest total RMSD. However, the optimal RSH+VV10 functional still beats this functional with respect to 7 of the 10 RMSDs. Thus, it is clear that this training, testing, and selection procedure allows us to pick a "best of both worlds" functional that fits well to the training set data, yet is highly transferable.

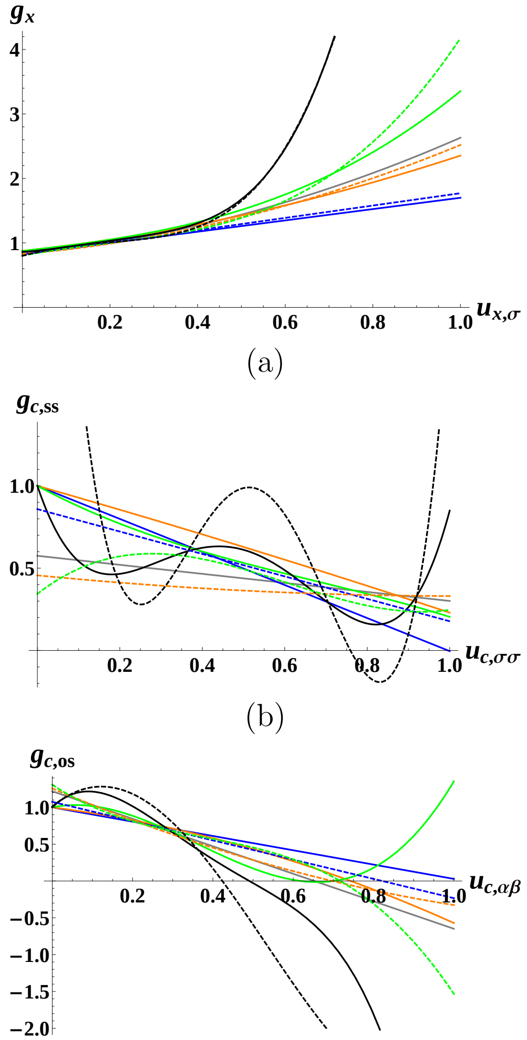
Of the 8 conventional functionals considered in Table 2.9, the RV-012.012.012.Xn functional has the lowest total RMSD. Another cross-check to consider is training an RV-012.012.012.Xn functional by fitting it to everything in both the training and test sets. The resulting RV-012.012.012.Xn* functional has a TC RMSD of 3.61 kcal/mol and an NC RMSD of 0.30 kcal/mol. However, we have no guarantee that it is transferable. In comparison, the optimal RSH+VV10 functional, has a TC RMSD of 3.62 kcal/mol and an NC RMSD of 0.32 kcal/mol. For the optimal RSH+VV10 functional, the resulting parameters, $\{c_{x,1}, c_{x,2}, c_{css,0}, c_{css,1}, c_{cos,0}, c_{cos,1}, c_x\}$, are $\{0.61, 1.18, 0.58, -0.27, 1.22, -1.87, 0.16\}$. An interesting test is to compare these parameters with the ones that result from training the optimal RSH+VV10 functional on both the training and test sets. The resulting parameters from such a fit are $\{0.60, 1.29, 0.58, -0.32, 1.24, -1.94, 0.16\}$. Since the parameters do not change significantly, this suggests that the training set on its own is sufficiently large for properly determining the parameters.

The inhomogeneity correction factor (ICF) plots associated with the 9 functionals from Table 2.9 are shown in Figure 2.4. The optimal RSH+VV10 functional is indicated by the gray lines, which are smooth and well-behaved in all 3 cases. The uniformly truncated $m = 1$ to $m = 4$ functionals (with a non-zero fraction of short-range exact exchange) are indicated by blue, orange, green, and black lines, respectively. Solid lines indicate satisfaction of all 3 UEG constraints, while dashed lines indicate that none of the UEG constraints are satisfied. Since it is preferable to have well-behaved ICFs for transferability, Figure 2.4 serves as another motivation for the functional selection procedure that is being used. Starting with the exchange functional ICF plots, the optimal RSH+VV10 functional and $m = 2$ plots are quite similar (both are quadratic), while the $m = 4$ plots are similar to the rest between $u_{x,\sigma} = 0$ and $u_{x,\sigma} = 0.5$, but shoot up sharply at $u_{x,\sigma} = 0.5$. While it has been shown[63] that most of the chemically relevant grid points lie between $u_{x,\sigma} = 0$ and $u_{x,\sigma} = 0.5$, it is still preferable to have a curve that looks like the gray one than either of the black ones. Moving on to the same-spin correlation functional ICFs, the quartic $m = 4$ ICFs are oscillatory and seem unphysical, particularly the black dashed curve that does not preserve the UEG limit. The remaining same-spin correlation ICFs are generally well-behaved. The functionals which relax the UEG limit reduce the amount of LSDA same-spin correlation at $u_{c,\sigma\sigma} = 0$ by as much as a factor of 2. For the opposite-spin correlation functional ICFs, the cubic $m = 3$ and quartic $m = 4$ functionals are the outliers, while the remaining functionals behave similarly. Most of the non-UEG functionals increase the amount of LSDA opposite-spin correlation at $u_{c,\alpha\beta} = 0$ by a factor of 1.2.

There are certainly alternatives to the procedure that is used to find the optimal functional for a given exchange/dispersion pairing. For example, if one considers the top 2 functionals with 7 linear parameters from the RSH+VV10 optimization, they are virtually indistinguishable as far as their 10 RMSDs are concerned, and differ only with respect to the same-spin correlation component. Thus, while the best RSH+VV10 functional with 7 linear parameters is of the RV-12.01.01.Xy form, the second best is of the RV-12.02.01.Xy form. Therefore, we note that the functionals presented here as optimal could be slightly different if a different selection procedure was used. However, after experimenting with various possible options, we can claim that the optimal functional either remains the same or is only very slightly different and that the RMSDs of the optimal functionals are representative of the level of accuracy achievable by the given functional form. In reality, if one were to choose to self-consistently optimize a functional from a certain category, it would certainly be beneficial to consider the top 10 or 20 functionals from a variety of selection procedures in order to assure that the absolute best functional has been chosen.

Before moving on to the remaining 8 categories, it is interesting to consider whether the relaxation of the UEG constraint for exchange is beneficial for the RSH+VV10 category. According to Figure 2.3, it is clear that for a majority of the points, the relaxation of this constraint leads to no improvements. In fact, the best RSH+VV10 functional that results with the constraint in place is the RV-12.01.01.Xy functional with a value of $c_x = 0.163$, while the best RSH+VV10 functional that results without the constraint is an 8-parameter RV-012.01.01.Xn functional with values of $c_{x,0} = 0.845$ and $c_x = 0.161$, resulting in $c_{x,0} + c_x = 1.006$. Thus, there is absolutely no reason to select the RV-012.01.01.Xn functional over the RV-12.01.01.Xy functional, especially since the total RMSD of the RV-12.01.01.Xy functional is slightly lower than that of the RV-012.01.01.Xn functional.

Moving on to the RSH+DFT-D2 category (Table 2.6), it is clear that the DFT-D2 dispersion tail is inferior to the VV10 NLC functional when coupled with RSH exchange. The optimal RSH+DFT-D2 functional is an 8-parameter functional with a total RMSD of 3.05 kcal/mol (compared to 2.68 kcal/mol for the optimal RSH+VV10 functional). In comparison to the optimal RSH+VV10 functional, the optimal RSH+DFT-D2 functional is worse with respect to all 10 RMSD categories. Compar-



(a)

(b)

(c)

Figure 2.4: Exchange, same-spin correlation, and opposite-spin correlation inhomogeneity correction factors for 9 functionals from the RSH+VV10 category. The optimal functional from the RSH+VV10 category is shown in gray. The remaining 8 lines belong to uniformly truncated $m = 1$ through $m = 4$ functionals (blue, orange, green, black), with the solid lines indicating satisfaction of all 3 UEG constraints and the dashed lines indicating that none of the UEG constraints are satisfied. The fraction of short-range exact exchange is optimized for all of the fits that are plotted.

kcal/mol	None	DFT-D2	VV10
Local	5.89	4.43	4.57
GH	4.72	3.14	3.06
RSH	3.87	3.05	2.68

Table 2.10: Total RMSDs in kcal/mol for the optimal functionals from all 9 categories.

ing the minimum RMSDs possible by the 2 types of functionals, they are equivalent only with respect to the training set RMSD and the training set TC RMSD, while the RSH+VV10 functional form outperforms the RSH+DFT-D2 functional form on the remaining 8 RMSDs. In addition, the RMSD$_{RG}$ of the optimal RSH+DFT-D2 functional is more than 2.5 times larger than that of the optimal RSH+VV10 functional.

While the functionals in both the RSH+DFT-D2 and RSH+VV10 categories are able to account for dispersion, it is interesting to compare them to the RSH+None category without dispersion corrections. From this category, the optimal functional that emerges is an 8-parameter functional that maintains $c_{x,0} = 1$ but violates the UEG limit ever so slightly for exchange by optimizing $c_x = 0.02$. As expected, the total RMSD of the optimal RSH functional is larger than that of the optimal RSH+DFT-D2 functional, and even larger than that of the optimal RSH+VV10 functional.

Instead of performing similar comparisons for the 6 remaining local and GH functionals, it is easiest to compare the total RMSDs of all 9 optimal functionals with the help of Table 2.10. This table confirms that the best overall performance is achieved by the optimal RSH+VV10 functional. Keeping the dispersion component constant, the RSH functionals outperform the GH functionals, while the GH functionals outperform the local functionals. As far as dispersion corrections are concerned, it is obvious that the functionals without dispersion corrections (None) perform worse than those with either DFT-D2 or VV10. However, it is less obvious which of the dispersion corrections is better. For the local exchange category, the optimal functional with the DFT-D2 dispersion tail slightly outperforms the one with the VV10 NLC functional, while the reverse is true for the GH exchange category. Ultimately, it is clear that as far as performance is concerned, the RSH+VV10 functional form is the best from the 9 variants considered.

While Figure 2.3 in its present form has already been stripped of thousands of datapoints for clarity, it still contains a great deal of information. It is very interesting that from the 9 categories, 7 of the optimal functionals do not satisfy the UEG constraint for exchange, while only the GH+None and RSH+VV10 optimal functionals satisfy this limit. In certain cases, as in the Local+DFT-D2 case, the difference between the total RMSDs of the optimal 9-parameter functional (4.43 kcal/mol) and the best possible 9-parameter functional that satisfies the UEG constraint for exchange (4.76 kcal/mol) is more than 0.30 kcal/mol. For other cases, like for the RSH+VV10 category, the difference is very small.

To convey an idea of what the plots in Figure 2.3 would look like if points had not been removed, Figure 2.5 shows all of the points corresponding to the RSH+VV10 fits for values of RMSD_{Total} between 2.65 and 3.00 kcal/mol. The filled red circles correspond to fits that do not skip orders in any of the dimensionless variables *and* satisfy the UEG constraint for exchange, while the unfilled red circles belong to similar non-skipping fits that do not satisfy the UEG constraint for exchange. Considering only the filled red circles, it is clear that going from 5 to 6 to 7 linear parameters results in large decreases in the total RMSD, at a rate of 0.15 kcal/mol per parameter. As the 7-parameter point is reached, the total RMSDs completely flatten out, and the quality of the fits begins to deteriorate after 9 linear parameters. The lowest 7-parameter filled red circle corresponds to the optimal RSH+VV10 functional that has been selected from considering Figure 2.3.

Once fits that skip orders in u are introduced (black dots), it is possible to slightly reduce the total RMSD of the optimal 7-parameter fit by going to the best 9-parameter fit, but by our selection criteria, the additional 0.02 kcal/mol improvement is not worth the 2 additional parameters. Finally, 2 special points on this plot corresponding to conventional uniform truncations are indicated by filled cyan triangles. The upright triangle corresponds to the 7-parameter $m = 2$ functional from the fourth row of Table 2.9 that satisfies all 3 UEG constraints (RV-12.12.12.Xy), while the downright triangle is the related 10-parameter RV-012.012.012.Xn functional that violates all 3 UEG constraints (the GGA portion is identical to that of Becke's B97 functional). Comparing these 2 functionals to the optimal 7-parameter functional again shows the ability of our selection procedure to reveal the best functional possible for the least number of empirical parameters. In fact, Figure 2.5 indicates that it is possible to considerably outperform the $m = 2$ functional that does not satisfy any of the UEG constraints with 3 less empirical parameters.

2.7 Comparisons

All of the 9 types of functionals considered thus far have existing non-empirical and semi-empirical counterparts. We compare the 9 resulting optimal functionals to the following: PBE[12], B97-D[54], VV10[57], B97[35], B97-D2[79], B3LYP-NL[106],

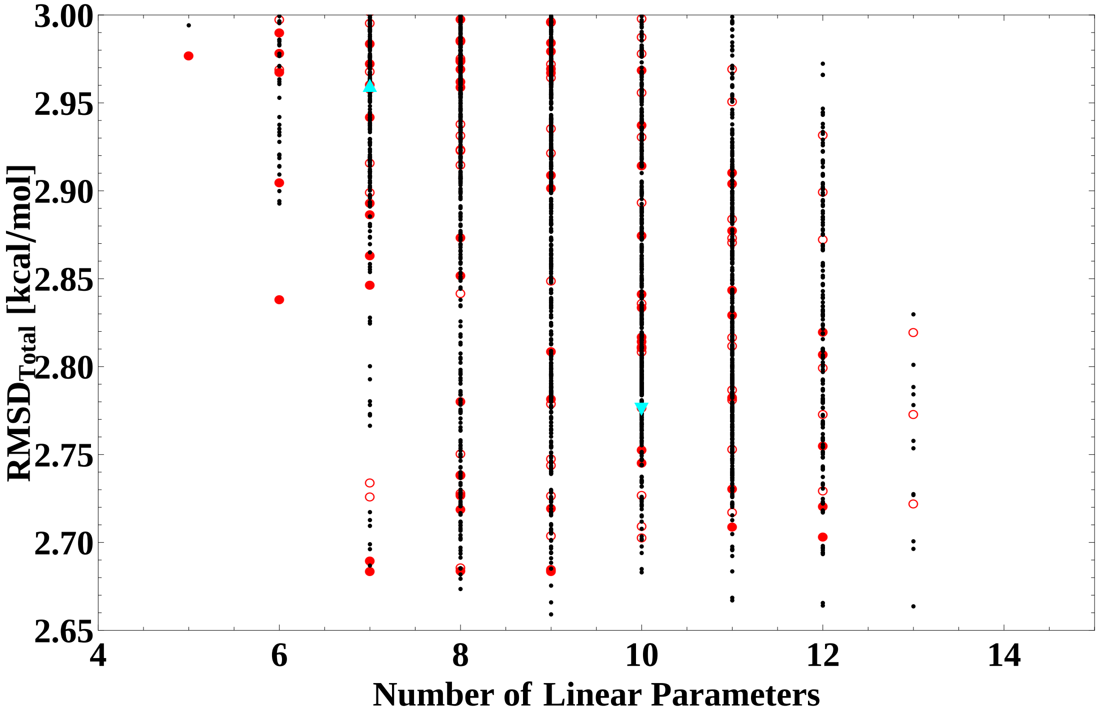


Figure 2.5: Total RMSDs plotted against the number of linear parameters for all 81919 possible RSH+VV10 fits. The filled red circles correspond to fits that do not skip orders in any of the dimensionless variables *and* satisfy the UEG constraint for exchange, while the unfilled red circles belong to similar non-skipping fits that do not satisfy the UEG constraint for exchange. The remaining points correspond to fits that skip orders in one or more of the ICFs. The filled upright cyan triangle corresponds to the total RMSD of the RSH+VV10 $m = 2$ functional that satisfies all 3 UEG constraints, while the filled downright cyan triangle corresponds to the total RMSD of the RSH+VV10 $m = 2$ functional that does not satisfy any of the UEG constraints. The optimal functional from the RSH+VV10 category is indicated by the lowest point on the vertical line that corresponds to 7 linear parameters.

Category	Local+None	Local+DFT-D2	Local+VV10	GH+None	GH+DFT-D2	GH+VV10	RSH+None	RSH+DFT-D2	RSH+VV10
kcal/mol	PBE	B97-D	VV10	B97	B97-D2	B3LYP-NL	ωB97X	ωB97X-D	LC-VV10
Train	8.85 (4.91)	4.78 (4.48)	9.10 (4.44)	4.36 (3.49)	3.36 (3.22)	4.19 (3.25)	3.67 (3.73)	3.42 (3.55)	6.01 (3.36)
Test	12.67 (7.00)	5.21 (4.50)	8.11 (4.88)	15.41 (5.90)	3.84 (3.17)	6.62 (3.03)	5.08 (4.18)	2.83 (2.50)	5.22 (1.92)
RG	3.76 (2.09)	5.55 (3.15)	1.61 (2.57)	5.82 (2.01)	2.65 (1.82)	1.73 (0.65)	1.82 (1.97)	8.79 (2.50)	2.36 (0.95)
Total	10.64 (5.89)	5.03 (4.43)	8.40 (4.57)	11.02 (4.72)	3.56 (3.14)	5.38 (3.06)	4.32 (3.87)	3.68 (3.05)	5.51 (2.68)
TC	10.19 (5.72)	5.67 (5.23)	10.09 (5.27)	4.85 (3.87)	4.04 (3.48)	4.99 (3.61)	3.89 (4.09)	3.64 (3.80)	6.93 (3.62)
NC	2.05 (1.08)	0.70 (0.61)	1.29 (0.65)	2.71 (0.97)	0.49 (0.54)	1.02 (0.48)	0.92 (0.71)	0.53 (0.43)	0.68 (0.32)
TC,Train	10.27 (5.55)	5.43 (5.13)	10.44 (5.11)	3.89 (3.78)	3.83 (3.62)	4.47 (3.70)	4.10 (4.28)	3.82 (4.02)	6.86 (3.86)
TC,Test	9.76 (6.57)	6.79 (5.71)	7.87 (6.06)	8.28 (4.28)	5.00 (2.60)	7.18 (3.06)	2.45 (2.85)	2.42 (2.26)	7.32 (1.81)
NC,Train	0.89 (0.65)	0.76 (0.62)	1.09 (0.55)	1.66 (0.48)	0.44 (0.48)	0.97 (0.42)	0.66 (0.42)	0.61 (0.42)	0.86 (0.31)
NC,Test	2.29 (1.18)	0.68 (0.61)	1.35 (0.68)	2.96 (1.08)	0.50 (0.56)	1.04 (0.50)	0.98 (0.77)	0.50 (0.43)	0.61 (0.32)

Table 2.11: RMSDs in kcal/mol for a variety of existing density functionals for comparison to the RMSDs of the 9 optimal functionals (shown in parentheses).

ωB97X[48], ωB97X-D[58], and LC-VV10[57]. A summary of how the optimal functionals obtained here compare with these selected existing functionals is given in Table 2.11.

2.7.1 Local Functionals

Beginning with the Local+None category, we can compare the resulting optimal functional to the non-empirical PBE functional. In general, the addition of 7 empirical parameters reduces the RMSDs by a factor of 2. The TC RMSD of PBE is reduced from 10.19 kcal/mol to 5.72 kcal/mol, while the NC RMSD of PBE is reduced from 2.05 kcal/mol to 1.08 kcal/mol. However, since both of these statistical measures contain datapoints from the training set, it is imperative to compare the performance of the 2 functionals on the test set. The $\text{RMSD}_{TC,Test}$ of the optimal Local+None functional is more than 3 kcal/mol lower than that of PBE, while its $\text{RMSD}_{NC,Test}$ is smaller by a factor of 2.

Moving on to the Local+DFT-D2 functionals, we can compare the resulting optimal functional to Grimme's B97-D functional, since the optimal functional is a reoptimization of this functional on a different training set (with a different set of ICF expansions). The 10-15% improvement in performance is not as drastic as in the Local+None category, confirming that the B97-D functional is near the limit of accuracy achievable by a Local+DFT-D2 GGA functional.

Finally, it is interesting to compare the performance of the existing VV10 exchange-correlation (xc) functional (rPW86 exchange[107] + PBE correlation + VV10 NLC) with the optimal Local+VV10 functional. As in the Local+None case, the optimal functional generally improves upon the performance of the VV10 xc functional by a factor of 2. However, it is interesting to point out that the performance of the VV10 xc functional is *better* for the 3 rare-gas dimer PECs, indicating that the weight of 2500 may be insufficient. For the optimization of the ωB97X-V functional, a weight of 25000 provided PECs that matched or improved upon those of the VV10 xc functional.

2.7.2 GH Functionals

Moving on to the GH+None category, we can compare against Becke's B97 functional. The largest improvements come from the noncovalent interactions, since B97 was only fit to TC data. Thus, there is a threefold improvement in both $\mathrm{RMSD}_{NC,Train}$ and $\mathrm{RMSD}_{NC,Test}$, while the thermochemistry improvements are less dramatic. However, the $\mathrm{RMSD}_{TC,Test}$ value for the optimal functional is smaller by a factor of 2, primarily due to its improved performance on the AlkAtom19 dataset.

While B97 was optimized only on thermochemistry, B97-D2 improves upon the NC RMSD of B97 by a factor of more than 5, with the help of only one additional linear parameter. While the optimal GH+DFT-D2 functional is 10-15% better for thermochemistry in general, it is 5-10% worse for noncovalent interactions. However, the performance of B97-D2 for the 3 rare-gas dimer PECs is worse by a factor of 1.5. Overall, it appears as though the B97-D2 functional is near the limit of accuracy achievable by a GH+DFT-D2 GGA functional.

Finally, the optimal GH+VV10 functional can be compared to Grimme's recent parameterization of the B3LYP-NL functional. The B3LYP-NL functional was developed by appending the VV10 NLC functional to the existing B3LYP functional and optimizing only the b parameter ($b = 4.8$). Compared to B3LYP-NL, the performance of the optimal GH+VV10 functional is generally better by a factor of 1.5-2. As yet another indication of transferability, while the $\mathrm{RMSD}_{TC,Train}$ of the optimal functional is only 20% better than that of B3LYP-NL, its $\mathrm{RMSD}_{TC,Test}$ value is smaller by a factor of 2.

2.7.3 RSH Functionals

Considering the RSH functionals, the first valid comparison is between the optimal RSH+None functional and ωB97X. Since ωB97X was trained primarily on thermochemistry, it is not surprising that it is 5-10% better than the optimal RSH+None functional for thermochemistry. Conversely, the performance of the optimal RSH+None functional is 15-20% better for noncovalent interactions. In addition, the performance of both functionals for the 3 rare-gas dimer PECs is almost identical. It appears that ωB97X is moderately close to the RSH+None performance limit, but employs significantly more parameters than our methodology establishes is necessary (13 vs. 8).

Moving on to the RSH+DFT-D2 category, the optimal functional is compared to ωB97X-D. As a reminder, the damping function that was used for ωB97X-D is slightly different from the one used in DFT-D2 and requires the optimization of a nonlinear parameter instead of a linear s_6 parameter. Nevertheless, ωB97X-D has TC, NC, and RG RMSDs of 3.64, 0.53, and 8.79 kcal/mol, compared to 3.80, 0.43, and 2.50 kcal/mol for the optimal RSH+DFT-D2 functional. As far as the 3 rare-gas dimer PECs are concerned, it is clear that the selection strategy has worked, since the RMSD$_{RG}$ of the optimal functional is 3.5 times smaller than that of ωB97X-D. Even though the optimal RSH+DFT-D2 functional has 5 less linear parameters than ωB97X-D, its performance on the noncovalent interactions in the test set is 15% better, as is its performance for the thermochemistry data in the test set.

Finally, we can compare the optimal RSH+VV10 functional to LC-VV10. The comparison between LC-VV10 and the optimal RSH+VV10 functional is interesting, because the main difference between the 2 functionals is that the GGA component of the optimal functional has been parameterized. The TC, NC, and RG RMSDs of LC-VV10 are 6.93, 0.68, and 2.36 kcal/mol, compared to 3.62, 0.32, and 0.95 kcal/mol for the optimal RSH+VV10 functional. Thus, by simply adding 7 empirical parameters, all 3 RMSDs are reduced by at least a factor of 2. In addition, Figure 1 from Reference 57 indicates that the VV10 xc functional (and thus the VV10 NLC functional) is very accurate for the argon dimer and krypton dimer PECs, so it is desirable to maintain this feature as empirical parameters are added. Largely due to the methodology employed here, the performance of the optimal RSH+VV10 functional is at least 1.5 times better than VV10 and LC-VV10 on the neon dimer, argon dimer, and neon-argon dimer PECs.

2.8 Conclusions

In developing new semi-empirical density functionals, there are numerous pitfalls on the road to achieving better performance than existing functionals. In this work, we have tried to address, within a limited scope, 2 of the principal issues: (a). “How does one assess the practical benefit of physical augmentation of a functional in a consistent way, including its transferability?”, and (b). “How can one determine when an optimal number of empirical parameters have been incorporated into a given functional form?”

To address the first question with manageable scope, we have compared 3 types of density functionals that are all built upon standard generalized gradient approximations of the Becke 97 form[35]: local, global hybrid, and range-separated hybrid. Each of these 3 basic forms are compared against augmented forms that include dispersion corrections via either Grimme's DFT-D2 dispersion tail or the VV10 nonlocal correlation functional. This defines a 3 by 3 grid of functional forms, each of which can be trained with an enormous variety of parameters.

To address the second question, as well as to complete the evaluation of the first question, we have developed a protocol for selecting the best functional of each type. This protocol involves training an enormous number of candidate functionals containing different numbers of linear parameters on 1108 pieces of training set data. The best such functional is selected based on an additional 1193 pieces of test set data, to assess transferability as well as overall performance. It should be noted that functionals are not trained self-consistently, but the RMSDs obtained are reliable indicators of self-consistent performance, as we have validated elsewhere for the most complicated form considered.

The first main outcome is the conclusion that the best functionals of each type considered contain significantly fewer linear parameters than many existing functionals in the literature. We believe this is largely because of the emphasis on transferability, rather than just training set performance. Typical "optimal" functionals involve between 7 and 9 linear empirical parameters. Functionals with larger numbers of linear parameters can train better but exhibit increasingly poor transferability. Of course there are fine differences between competing best choices in some cases, but this overall result is robust.

The second main outcome concerns the relative performance of the different functional forms within this consistent framework. We find that by far the best possible performance is obtained by the range-separated hybrid functional, coupled to the VV10 NLC functional. This is accordingly the best single candidate for self-consistent optimization, a topic that we have addressed elsewhere to define the ωB97X-V functional[63]. While the self-consistent optimization of a local GGA functional appended with VV10 is an interesting opportunity for a lower cost functional, it is unclear whether the resulting functional will perform significantly better than the best existing local GGAs with DFT-D2 corrections, such as B97-D.

The third main outcome concerns how the 9 optimized forms compare with existing literature functionals that fit within each of those 9 categories. In some cases, very significant improvements are evident, such as for a local functional (vs. PBE) and for a range-separated hybrid functional with VV10 (vs. LC-VV10), which are due largely to comparing against non-empirical (PBE) or relatively non-empirical (LC-VV10) functionals. In other cases, such as range-separated hybrids with a dispersion tail, modest improvements are possible while significantly reducing the number of linear parameters (vs. ωB97X-D), indicating that less semi-empiricism than existing choices can actually be advantageous.

Finally, there are interesting non-trivial opportunities to extend the present analysis beyond the GGA framework we have restricted ourselves to here. It is clearly very desirable to explore the question of how much additional improvement can be obtained by semi-empirical functionals that depend on the kinetic energy density. This will vastly increase the number of possible functionals to approximately 2^{75}, so it is unlikely to be possible to do it up to the $m = 4$ truncation we have employed here. However, the encouraging conclusions about the relatively low degree of optimal semi-empiricism suggest that this may in fact not be necessary. We hope to report on this problem in the near future.

Chapter 3

ωB97X-V: An RSH GGA density functional

3.1 Introduction

3.1.1 Background

In the early 1950s, John Slater introduced the first semi-empirical exchange-only density functional: the Xα method[66]. Since then, hundreds of parameterized density functionals have been developed, expanding upon the simplicity of Slater's functional. While the Xα method depended solely on the electron density, its successors have taken into account both the gradient and Laplacian of the electron density, the kinetic energy density, occupied orbitals through exact exchange, and even virtual orbitals through post-Hartree–Fock methods such as MP2. However, almost 50 years after the advent of Kohn–Sham DFT[1, 3], the exact exchange-correlation functional remains elusive. Whereas Slater's functional only had 1 optimizable parameter, today's density functionals can have more than 40 parameters.

While including more parameters in the functional optimization guarantees better performance on the training set, the most desirable attribute of a parameterized density functional is the promise of transferability, and a functional with *less* parameters is more likely to be transferable than a functional with *more* parameters. Consequently, additional empirical parameters should be included in the functional form only if they contribute to improving the performance of the functional on *both* the training and primary test sets.

The first systematic optimization of a density functional[35] was conducted by Axel Becke in 1997. The resulting global hybrid (GH) GGA density functional, B97, had 10 linear parameters that resulted from uniformly truncating the power series

ICFs for the exchange, same-spin correlation, and opposite-spin correlation functionals at second order ($m = 2$). Furthermore, Becke demonstrated that including more parameters into the linear fit negligibly enhanced the training set RMSD *and* introduced unphysically oscillatory character into the ICF plots[35].

Becke revolutionized the systematic parameterization of exchange-correlation functionals with B97[35]. Since then, multiple B97-based density functionals have been developed. In 1998, Handy and coworkers[67] self-consistently optimized the parameters of Becke's functional (B97-1) and developed a new local GGA density functional (HCTH/93). With HCTH/93, the ICFs were truncated at $m = 4$ (for a total of 15 parameters) instead of at $m = 2$ as in the hybrid B97 and B97-1 functionals (10 parameters each). While Becke's work had indicated that values of $m > 2$ resulted in unphysical functionals, the use of $m = 4$ in HCTH/93 was justified by using a larger training set that included nuclear gradients and ZMP exchange-correlation potentials[108]. The same training procedure was used to develop 2 additional 15-parameter functionals (HCTH/120 and HCTH/147)[68], with the number following the slash indicative of the size of the training set. Handy's development of local GGA functionals culminated with the HCTH/407 functional[69] (15 parameters).

Additional attempts[70, 71] at developing hybrid B97-based functionals were made by Tozer and coworkers with B97-2 and B97-3. B97-2 (10 parameters) kept the same value for m as B97 and B97-1, but included multiplicative potentials in its training set (following the example set by the HCTH family). Finally, B97-3 was parameterized with an even larger training set and an $m = 4$ truncation for the ICFs (16 parameters).

While the underlying ingredient that captures inhomogeneities in the density in B97-based density functionals is the gradient of the density, the first systematic optimization of a meta-GGA density functional was conducted almost simultaneously by Scuseria (VSXC)[27] and Becke (B98)[15]. Both of these functionals were introduced nearly 10 years after Becke had first motivated the use of the kinetic energy density (and the Laplacian of the density) with the BR89 exchange functional[109] and the Bc88 correlation functional[110]. While the 10-parameter B98 functional included a fraction of exact exchange and depended on the density, the gradient and Laplacian of the density, as well as the kinetic energy density, the 21-parameter VSXC functional contained no exact-exchange mixing and depended on the density, its gradient, and the kinetic energy density.

Handy's entry into the world of meta-GGAs came with the B97- and B98-based τ-HCTH (16 parameters) and (global) hybrid τ-HCTH (17 parameters) functionals[28]. While the correlation functionals of both τ-HCTH and hybrid τ-HCTH were based on the B97 functional form, the exchange functionals had a B97-type component *and* a B98-type component (with a slightly modified τ-dependent dimensionless parameter

(w_σ) introduced by Becke[111]).

In the spirit of non-empiricism, the meta-GGA counterpart of PBE, TPSS[17], was developed in 2003, followed by the 17-parameter, global hybrid meta-GGA functional, BMK[36]. The functional form of BMK was virtually identical to that of the hybrid τ-HCTH functional, but a primary goal of its parameterization was improved performance for kinetics.

Since 2005, Truhlar has developed 10 highly-parameterized local (M06-L[23], M11-L[24]), global hybrid (M05[41], M05-2X[43], M06[45], M06-2X[45], M06-HF[44], M08-HX[46], M08-SO[46]), and range-separated hybrid (M11[52]) meta-GGA density functionals with 20 to 50 parameters. The underlying parameterizable exchange functional component for the Minnesota functionals is a power series (in Becke's w_σ parameter) that multiplies the non-empirical GGA enhancement factor of the PBE exchange functional. Additional components for select functionals include VSXC- and RPBE-based[112] exchange functional ICFs and PBE-, B97-, VSXC-, and B98-based correlation functional ICFs.

While attempts to remedy the self-interaction error (SIE) inherent to most density functionals date back to the early 1980s[5], the elimination of SIE in the long-range (for the exchange functional) was facilitated by the development of range-separation, namely, the splitting of the Coulomb operator into short- and long-range components controlled by the erfc and erf functions, respectively. While this separation was initially used by Savin[113] to combine short-range DFT with long-range configuration-interaction, Hirao[114] successfully applied this scheme to Becke's B88 exchange functional[10] and combined it with his one-parameter progressive correlation functional[115, 116] to produce LRC-BOP. In addition to eliminating long-range self-interaction for the exchange functional, range-separated hybrid functionals come with further benefits including improved performance for Rydberg and charge transfer excitations within the TD-DFT approach[117].

While Hirao's approach to range separation involved using the analytic expression for the LSDA exchange hole[47] along with a modified Fermi wave vector that contained the exchange functional ICF to derive the expression for the range-separated enhancement factor for the LSDA exchange energy density, an alternate path[118–120] was pursued by Scuseria whereby a general model for the GGA exchange hole was developed and used to obtain the range-separated enhancement factor. While the efforts of Scuseria[118–121] and Herbert[122, 123] were directed towards combining the long-range-corrected (LRC) approach with existing exchange and correlation functionals, Chai and Head-Gordon used the flexible B97 functional form as the foundation for a series of semi-empirical, range-separated hybrid functionals, namely, ωB97[48], ωB97X[48], and ωB97X-D[58]. Although the ICFs of these functionals were uniformly truncated at $m = 4$, the uniform electron gas (UEG) limits were satisfied, resulting in a total of 13, 14, and 15 optimized parameters, respectively. Furthermore, these functionals used the LRC scheme of Hirao rather than Scuseria.

Since long-range electron correlations that account for van der Waals (vdW) interactions cannot be properly described by standard density functionals[124, 125], there has been an increased effort in the past decade to remedy this issue. A comprehensive review of various approaches to extending the applicability of DFT to dispersive interactions can be found in Reference 126.

The simplest and cheapest methods that account for dispersion are Grimme's empirical DFT-D methods[53–55]. Grimme's first attempt at an empirical dispersion tail was DFT-D1[53], which was only available for 6 elements (H, C, N, O, F, and Ne). With the atomic C_6 parameters and van der Waals Radii predetermined, the single linear optimizable parameter of the DFT-D1 dispersion tail (s_6) was trained onto 3 existing local GGA density functionals (BLYP, BP86, and PBE) and dramatically improved the ability of the parent functional to describe vdW interactions.

Following the success of DFT-D1, Grimme introduced the DFT-D2 dispersion tail along with an explicitly parameterized, B97-based, local GGA density functional called B97-D. B97-D maintained Becke's $m = 2$ truncation for the ICFs and had a total of 10 optimized parameters. While the form of the DFT-D2 dispersion correction term was identical to that of DFT-D1, atomic C_6 parameters and van der Waals Radii were made available for all elements through xenon, and the existing values from DFT-D1 were improved. Furthermore, s_6 parameters were determined for PBE, BLYP, BP86, TPSS, and B3LYP.

The latest addition to the DFT-D family is the DFT-D3 dispersion tail, which uses fractional coordination numbers to account for variations in atomic dispersion coefficients in different chemical environments and contains a 2-body *and* 3-body term. In addition, DFT-D3 uses an improved damping function (motivated by the work of Salahub[127] and Head-Gordon[58]) that has an additional nonlinear parameter in its denominator. Since the 2-body part of DFT-D3 includes both $1/r^6$ and $1/r^8$ terms, 2 linear (s_6 and s_8) and 2 nonlinear parameters ($s_{r,6}$ and $s_{r,8}$) are available for optimization. However, Grimme and coworkers demonstrated that s_6 and $s_{r,8}$ can be set to unity for non-double-hybrid density functionals, leaving 2 optimizable parameters. The DFT-D3 dispersion tail was trained onto more than 10 existing density functionals and generally improved upon its predecessors for describing dispersive interactions.

The Becke and Johnson (BJ) exchange-dipole moment (XDM) model[128–134] takes advantage of the fact that dispersive interactions can be accounted for via the exchange-hole dipole moments of the interacting species. The prominent term in such an (attractive) interaction is of the instantaneous dipole-induced dipole nature, and Becke and Johnson have motivated 2 variants of their method: XDM6, which only includes interatomic $1/r^6$ interactions, and XDM10, which additionally depends on interatomic $1/r^8$ and $1/r^{10}$ interactions. XDM6 and XDM10 have 1 and 2 optimizable parameters, respectively, and the methods have been recently implemented in both a self-consistent-field (SCF) and post-SCF manner[135].

While the DFT-D and XDM approaches rely on predetermined atomic parameters (C_6 coefficients, vdW Radii, atomic polarizabilities, etc.) to compute the dispersion interaction, several methods that account for dispersion through their dependence on the electron density have been developed in the past decade, including vdW-DF-04[73], vdW-DF-10[74], VV09[75], and VV10[57]. These nonlocal correlation (NLC) functionals rely on a double space integral over the density and a nonlocal correlation kernel, and are computationally more expensive than the DFT-D and XDM methods. Langreth, Lundqvist, and coworkers introduced the first NLC scheme (vdW-DF-04) that could be applied to overlapping densities in 2004. Several years later, Vydrov and Van Voorhis self-consistently implemented the vdW-DF-04 NLC functional for use with Gaussian basis sets[136], and proposed modifications (vdW-DF-09) to improve its compatibility with existing exchange functionals[137]. The VV09[75, 138, 139] NLC functional of Vydrov and Van Voorhis adopted a simple analytic form for the nonlocal correlation kernel, instead of relying on a numerically tabulated kernel. Based on their experience garnered from VV09, Vydrov and Van Voorhis proposed an even simpler NLC functional, VV10, that improved upon its predecessor by employing a less elaborate function for the damping of the $1/r^6$ asymptote. In 2010, Langreth, Lundqvist, and coworkers proposed an improved NLC functional (vdW-DF-10) to account for the tendency of vdW-DF-04 to overestimate equilibrium bond lengths and underestimate the binding energies of hydrogen-bonded complexes. While the vdW-DF methods have no optimizable parameters, VV09 and VV10 have 1 and 2 optimizable parameters, respectively.

3.1.2 Design Goals and Strategy

Due to the flexibility provided by the B97 functional form, it is the foundation for the functional introduced in this chapter. 3 significant changes are made to the original B97 functional form: 1). the new functional is a range-separated hybrid functional instead of a global hybrid functional, 2). the VV10 NLC functional[57] is included to provide a sound description of nonlocal electron correlation, and 3). the ICF truncation orders for the exchange, same-spin correlation, and opposite-spin correlation functionals are determined individually. The resulting density functional, ωB97X-V, has 7 linear parameters (2 for local exchange, 4 for local correlation, and 1 for short-range exact exchange) and 3 nonlinear parameters (1 for range-separation and 2 for nonlocal correlation), for a total of 10 optimized parameters.

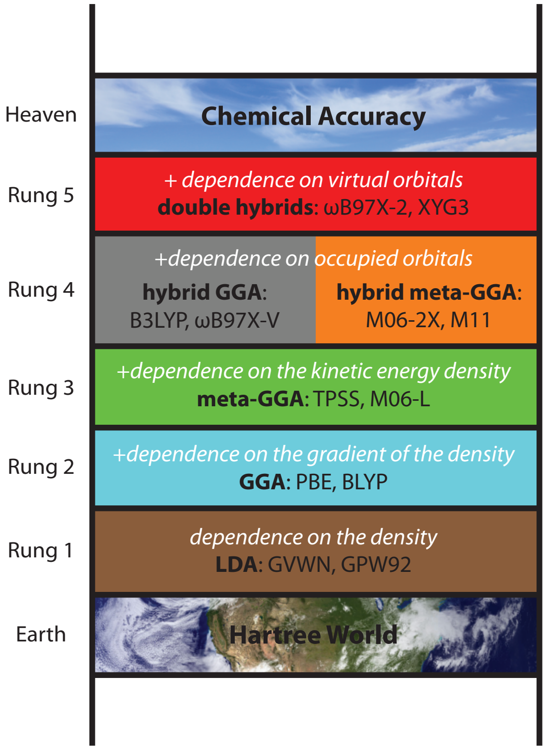


Figure 3.1: Perdew's "Jacob's Ladder".

The ingredients included in the functional form of ωB97X-V place it on the fourth rung of Perdew's "Jacob's Ladder"[4] (Figure 3.1). An alternate view of the ingredients that can be incorporated into a density functional is given in Table 3.1. When choosing the components of a density functional, the first consideration involves selecting the ingredients that will constitute the local exchange-correlation (xc) functional. While the density and its gradient were chosen for ωB97X-V, it is worthwhile to note that the kinetic energy density, $\tau_\sigma = \frac{1}{2}\sum_i^{occ.} |\nabla\psi_{i,\sigma}|^2$, was taken into consideration. However, studies[140–143] have indicated that functionals containing this ingredient require very fine integration grids to overcome oscillatory behavior in the potential energy curves (PEC) of weakly-bound systems. Furthermore, preliminary results from attempting to develop a range-separated hybrid meta-GGA density functional have indicated that the great freedom associated with parameterizing a $(\nabla\rho, \tau)$ surface requires further exploration, and will be addressed in a future publication.

Exchange		Correlation	
Local	Nonlocal	Local	Nonlocal
1). $\underline{\rho}$	1). None	1). $\underline{\rho}$	1). None
2). $\underline{\nabla\rho}$	2). GH	2). $\underline{\nabla\rho}$	2). DFT-D2
3). τ	3). $\underline{\text{RSH}}$	3). τ	3). DFT-D3
4). $\nabla^2\rho$		4). $\nabla^2\rho$	4). $\underline{\text{VV10}}$
			5). MP2

Table 3.1: Ingredients that can be incorporated into a density functional. GH stands for global hybrid and RSH stands for range-separated hybrid. The nonlocal correlation list is certainly not comprehensive, as it excludes functionals such as VV09, XDM6, XDM10, vDW-DF-04, and vDW-DF-10, as well as higher-scaling post-SCF methods like RPA, MP3, CCSD, and CCSD(T). The underlined ingredients define the functional form of ωB97X-V.

After finalizing the local xc functional form, the question of whether to include exact exchange arises. Since most density functionals that perform well on both bonded and non-bonded interactions have a fraction of exact exchange, the subsequent question is whether to go with the global hybrid (GH) functional form or the range-separated hybrid (RSH) functional form. While GH functionals have a fixed fraction of exact exchange for all interelectronic distances, the fraction of exact exchange varies from 0 (or a positive non-zero fraction) at short-range to 1 at long-range for RSH functionals, conveniently eliminating self-interaction error in the long-range for the exchange functional. Since GH and RSH functionals have the same computational scaling, the RSH functional form was chosen for ωB97X-V.

Finally, it is necessary to decide whether a dispersion correction should be appended to the functional. Since a range-separated hybrid GGA functional cannot accurately bind dispersion-bound systems on its own, a dispersion correction is required. As a result of the poor performance of ωB97X-D on a recent benchmark of dianionic sulfate-water clusters[87] (a failure that can be attributed to the fact that dispersion tails cannot differentiate between neutral and charged atoms), the DFT-D methods were not considered. The NLC functional of Vydrov and Van Voorhis (VV10) was selected over MP2 because it does not carry the additional weight of fifth-order computational scaling and is a simple functional of the electron density. These considerations ultimately led to the functional form containing the underlined components in Table 3.1.

The transferability of a semi-empirical density functional is an attribute that must be assured during its development. In order to obtain a maximally-transferable functional, the 2301 datapoints that were initially set aside for training were divided into a training set of 1108 datapoints and a primary test set of 1193 datapoints. The motivation for this partitioning initially came from the decision to determine the 3 ICF expansions individually, instead of truncating them uniformly. During the optimization of B97, Becke uniformly varied the truncation order (m) from 0 to 8, and chose to truncate uniformly at $m = 2$ based on the 9 resulting RMSDs. In this work, the decision to individually determine the ICF expansions means that there are $2^{m_x+m_{css}+m_{cos}+2} - 1$ possible candidate functionals (where m_x, m_{css}, and m_{cos} are the largest orders included in the exchange, same-spin correlation, and opposite-spin correlation functional power series ICFs). Since there is no literature on whether truncating the ICFs individually and/or skipping orders in the ICFs will produce transferable functionals, it was deemed necessary to train the linear parameters of the candidate functionals on the training set and check the transferability of the fits on the primary test set. Ultimately, this survival-of-the-fittest strategy will be used to answer the following basic questions: 1). "Does having more parameters necessarily imply better overall performance for a density functional?", 2). "Is there a point at which the inclusion of additional parameters becomes detrimental to the overall performance of a density functional?", and 3). "Can we achieve better overall performance by selectively optimizing certain parameters in the ICFs, or is the conventional scheme of uniform truncation sufficient?".

3.2 Computational Details

An integration grid of 99 radial points and 590 angular points, (99,590), was used to evaluate local xc functionals, while the SG-1 grid[61] was used for the VV10 NLC functional[57]. For the rare-gas dimers and the absolute atomic energies, a (500,974) integration grid was used to evaluate local xc functionals, along with a (99,590) grid for the VV10 NLC functional. For M06-L[23] and M11-L[24], calculations in the training and primary test set were carried out with the (250,590) grid. All calculations on the coronene dimer were carried out with a (75,302) grid for local xc functionals and the SG-1 grid for the VV10 NLC functional. The aug-cc-pVQZ [aQZ] basis set[77, 78] was used for all thermochemistry datapoints except the second-row absolute atomic energies (aug-cc-pCVQZ)[77, 78], while the aug-cc-pVTZ [aTZ] basis set[77, 78] was used for all noncovalent interactions datapoints except the rare-gas dimers (aug-cc-pVQZ). For the X40 dataset[144] in the secondary test set, the def2-TZVPPD basis set[145, 146] was used. Furthermore, the noncovalent interactions were computed *without* counterpoise corrections unless otherwise noted. For PBE-D2, B3LYP-D2, and B97-D2, Grimme's DFT-D2 dispersion tail was used with the following[54, 79] s_6 coefficients: 0.75, 1.05, and 0.75. Grimme's B97-D functional uses the DFT-D2 dispersion tail as well, with an s_6 coefficient of 1.25. All of the calculations were performed with a development version of Q-Chem 4.0[80].

3.3 Theory

The complete functional form for ωB97X-V is given by Equations 3.1-3.3. The components of the exchange functional and correlation functional are described in Sections 3.3.1 and 3.3.2, respectively. The acronyms used in Equations 3.1-3.3 (and henceforth) are: exchange-correlation (xc), exchange (x), correlation (c), short-range (sr), long-range (lr), same-spin (ss), opposite-spin (os), and nonlocal (nl).

$$E_{xc}^{\omega B97X-V} = E_{x}^{\omega B97X-V} + E_{c}^{\omega B97X-V} \tag{3.1}$$

$$E_{x}^{\omega B97X-V} = E_{x,sr}^{B97} + c_x E_{x,sr}^{exact} + E_{x,lr}^{exact} \tag{3.2}$$

$$E_{c}^{\omega B97X-V} = E_{c,ss}^{B97} + E_{c,os}^{B97} + E_{c,nl}^{VV10} \tag{3.3}$$

3.3.1 Exchange Functional Form

The local spin-density approximation (LSDA) for exchange can be expressed in terms of the first-order spinless reduced density matrix for a uniform electron gas (UEG):

$$E_x^{LSDA} = -\frac{1}{2}\sum_{\sigma}^{\alpha,\beta}\int\int \frac{1}{s}|\rho_\sigma^{UEG}(\mathbf{r},s)|^2 d\mathbf{r}d\mathbf{s} \tag{3.4}$$

$$\rho_\sigma^{UEG}(\mathbf{r},s) = 3\rho_\sigma(\mathbf{r})\left[\frac{\sin(k_{F\sigma}s) - k_{F\sigma}s\cos(k_{F\sigma}s)}{[k_{F\sigma}s]^3}\right] \tag{3.5}$$

where $\mathbf{s} = \mathbf{r}_1 - \mathbf{r}_2$, $\mathbf{r} = \frac{1}{2}[\mathbf{r}_1 + \mathbf{r}_2]$, and $k_{F\sigma} = [6\pi^2\rho_\sigma]^{1/3}$ is the spin-polarized Fermi wave vector. Integration of Equation 3.4 over $\mathbf{s}$ gives the well-known expression for the LSDA exchange energy in terms of the exchange energy density per unit volume of a uniform electron gas:

$$E_x^{LSDA} = \sum_{\sigma}^{\alpha,\beta}\int e_{x,\sigma}^{UEG}(\rho_\sigma)d\mathbf{r} \tag{3.6}$$

$$e_{x,\sigma}^{UEG}(\rho_\sigma) = -\frac{3}{2}\left(\frac{3}{4\pi}\right)^{1/3}\rho_\sigma^{4/3} \tag{3.7}$$

Transforming E_x^{LSDA} to its short-range counterpart, $E_{x,sr}^{LSDA}$, is accomplished by replacing $\frac{1}{s}$ in Equation 3.4 with $\frac{\mathrm{erfc}(\omega s)}{s}$ and carrying out the same integration. The resulting SR-LSDA exchange functional:

$$E_{x,sr}^{LSDA} = \sum_{\sigma}^{\alpha,\beta} \int e_{x,\sigma}^{UEG}(\rho_\sigma) F\left(a_\sigma\right) d\mathbf{r} \tag{3.8}$$

is conveniently identical to its unattenuated counterpart, with the exception of a multiplicative attenuation function, $F\left(a_\sigma\right)$:

$$\begin{gathered} F\left(a_\sigma\right) = 1 - \frac{2}{3} a_\sigma \left[2\sqrt{\pi} \text{erf}\left(\frac{1}{a_\sigma}\right) - 3a_\sigma + \right. \\ \left. a_\sigma^3 + \left[2a_\sigma - a_\sigma^3\right] \exp\left(-\frac{1}{a_\sigma^2}\right) \right] \end{gathered} \tag{3.9}$$

where $a_\sigma = \frac{\omega}{k_{F\sigma}}$ and ω is the nonlinear range-separation parameter that controls the transition from local DFT exchange to nonlocal exact exchange with respect to the interelectronic distance.

Accounting for inhomogeneities in the electron density is achieved by multiplying the integrand of the SR-LSDA exchange functional by the power series inhomogeneity correction factor[35], $g_x\left(u_{x,\sigma}\right)$, resulting in the SR-B97 exchange functional:

$$E_{x,sr}^{B97} = \sum_{\sigma}^{\alpha,\beta} \int e_{x,\sigma}^{UEG}(\rho_\sigma) F\left(a_\sigma\right) g_x\left(u_{x,\sigma}\right) d\mathbf{r} \tag{3.10}$$

$$g_x\left(u_{x,\sigma}\right) = \sum_{i=0}^{m_x} c_{x,i} u_{x,\sigma}^i = \sum_{i=0}^{m_x} c_{x,i} \left[\frac{\gamma_x s_\sigma^2}{1 + \gamma_x s_\sigma^2}\right]^i \tag{3.11}$$

where the dimensionless variable, $u_{x,\sigma} \in [0,1]$, is a finite domain transformation of the reduced spin-density gradient, $s_\sigma = \frac{|\nabla \rho_\sigma|}{\rho_\sigma^{4/3}} \in [0, \infty)$. The linear DFT exchange parameters, $c_{x,i}$, will be determined by least-squares fitting to a training set in Section 3.5, while $\gamma_x = 0.004$ is a nonlinear local exchange parameter that was fit to the Hartree–Fock exchange energies of 20 atoms in 1986 by Becke[9].

Nonlocal exchange is introduced by splitting the Coulomb operator in the conventional expression for exact exchange into a short-range component ($E_{x,sr}^{exact}$) and a long-range component ($E_{x,lr}^{exact}$) with the erfc and erf Coulomb functions, respectively:

$$\begin{aligned} E_{x,sr}^{exact} = & -\frac{1}{2} \sum_{\sigma}^{\alpha,\beta} \sum_{i,j}^{occ.} \int \int \psi_{i\sigma}^*\left(\mathbf{r}_1\right) \psi_{j\sigma}^*\left(\mathbf{r}_2\right) \frac{\text{erfc}\left(\omega r_{12}\right)}{r_{12}} \\ & \times \psi_{j\sigma}\left(\mathbf{r}_1\right) \psi_{i\sigma}\left(\mathbf{r}_2\right) d\mathbf{r}_1 d\mathbf{r}_2 \end{aligned} \tag{3.12}$$

$$\begin{aligned} E_{x,lr}^{exact} = & -\frac{1}{2} \sum_{\sigma}^{\alpha,\beta} \sum_{i,j}^{occ.} \int \int \psi_{i\sigma}^*\left(\mathbf{r}_1\right) \psi_{j\sigma}^*\left(\mathbf{r}_2\right) \frac{\text{erf}\left(\omega r_{12}\right)}{r_{12}} \\ & \times \psi_{j\sigma}\left(\mathbf{r}_1\right) \psi_{i\sigma}\left(\mathbf{r}_2\right) d\mathbf{r}_1 d\mathbf{r}_2 \end{aligned} \tag{3.13}$$

where $\psi_{i\sigma}$ and $\psi_{j\sigma}$ are the occupied Kohn–Sham spatial orbitals. Instead of setting the percentage of exact-exchange at $r = 0$ to zero, an optimizable parameter, c_x, controls the amount of short-range exact exchange.

3.3.2 Correlation Functional Form

Closed-form expressions for the correlation energy density per particle of a uniform electron gas, $\epsilon_c^{UEG}(\rho)$, are only known for the low- and high-density limits of the paramagnetic and ferromagnetic cases of the uniform electron gas (UEG). Using the Monte-Carlo data of Ceperley and Alder[8], Perdew and Wang developed an analytic spin-compensated representation[7], $\epsilon_c^{PW92}(\rho)$, for $\epsilon_c^{UEG}(\rho)$. Combined with the spin-polarization interpolation formula of Vosko, Wilk, and Nusair[6], the spin-polarized PW92 correlation energy density per particle, $\epsilon_c^{PW92}(\rho_\alpha, \rho_\beta)$, is the starting point for the ωB97X-V correlation functional:

$$E_c^{LSDA} = \int \rho \epsilon_c^{PW92}(\rho_\alpha, \rho_\beta)\, d\mathbf{r} \tag{3.14}$$

Using the spin decomposition technique of Hermann Stoll and coworkers[81], the LSDA correlation energy functional above can be separated into same-spin and opposite-spin components:

$$\begin{aligned} E_{c,ss}^{LSDA} = \sum_\sigma^{\alpha,\beta} \int e_{c,\sigma\sigma}^{PW92} d\mathbf{r} = \int \rho_\alpha \epsilon_c^{PW92}(\rho_\alpha, 0)\, d\mathbf{r} + \\ \int \rho_\beta \epsilon_c^{PW92}(0, \rho_\beta)\, d\mathbf{r} \end{aligned} \tag{3.15}$$

$$\begin{aligned} E_{c,os}^{LSDA} = \int e_{c,\alpha\beta}^{PW92} d\mathbf{r} = \int \rho \epsilon_c^{PW92}(\rho_\alpha, \rho_\beta)\, d\mathbf{r} - \\ \int \rho_\alpha \epsilon_c^{PW92}(\rho_\alpha, 0)\, d\mathbf{r} - \int \rho_\beta \epsilon_c^{PW92}(0, \rho_\beta)\, d\mathbf{r} \end{aligned} \tag{3.16}$$

where $e_{c,\sigma\sigma}^{PW92}$ and $e_{c,\alpha\beta}^{PW92}$ are the PW92 same-spin and opposite-spin correlation energy densities per unit volume. Extending Equations 3.15 and 3.16 to account for inhomogeneities in the electron density is straightforward, since the same approach used for the exchange functional can be applied to the correlation functional:

$$E_{c,ss}^{B97} = \sum_\sigma^{\alpha,\beta} \int e_{c,\sigma\sigma}^{PW92} g_{c,ss}(u_{c,\sigma\sigma})\, d\mathbf{r} \tag{3.17}$$

$$g_{c,ss}(u_{c,\sigma\sigma}) = \sum_{i=0}^{m_{css}} c_{css,i} u_{c,\sigma\sigma}^i = \sum_{i=0}^{m_{css}} c_{css,i} \left[\frac{\gamma_{css} s_\sigma^2}{1 + \gamma_{css} s_\sigma^2} \right]^i \tag{3.18}$$

$$E_{c,os}^{B97} = \int e_{c,\alpha\beta}^{PW92} g_{c,os}\left(u_{c,\alpha\beta}\right) d\mathbf{r} \tag{3.19}$$

$$g_{c,os}\left(u_{c,\alpha\beta}\right) = \sum_{i=0}^{m_{cos}} c_{cos,i} u_{c,\alpha\beta}^{i} = \sum_{i=0}^{m_{cos}} c_{cos,i} \left[\frac{\gamma_{cos} s_{\alpha\beta}^{2}}{1+\gamma_{cos} s_{\alpha\beta}^{2}}\right]^{i} \tag{3.20}$$

where $s_{\alpha\beta}^2 = \frac{1}{2}\left(s_\alpha^2 + s_\beta^2\right)$. The linear DFT correlation parameters, $c_{css,i}$ and $c_{cos,i}$, will be determined by least-squares fitting to a training set in Section 3.5, while $\gamma_{css} = 0.2$ and $\gamma_{cos} = 0.006$ are nonlinear local correlation parameters that were fit to the correlation energies of helium and neon in 1986 by Becke[9].

Nonlocal correlation is taken into account via the VV10 NLC functional[57]:

$$E_{c,nl}^{VV10} = \int \rho\left(\mathbf{r}\right) \left[\frac{1}{32}\left[\frac{3}{b^2}\right]^{3/4} + \frac{1}{2}\int \rho\left(\mathbf{r}'\right) \Phi\left(\mathbf{r},\mathbf{r}',\{b,C\}\right) d\mathbf{r}'\right] d\mathbf{r} \tag{3.21}$$

where $\Phi\left(\mathbf{r},\mathbf{r}',\{b,C\}\right)$ is the nonlocal correlation kernel defined in Reference 57. The VV10 NLC functional introduces 2 nonlinear parameters: b, which controls the short-range damping of the $1/r^6$ asymptote, and C, which controls the accuracy of the asymptotic C_6 coefficients.

3.4 Datasets

In total, the training, primary test, and secondary test sets used for the parameterization, validation, and assessment of ωB97X-V contain 2486 datapoints, requiring 2455 single-point calculations. Of the 2486 datapoints, 1108 belong to the training set, 1193 belong to the primary test set, and 185 belong to the secondary test set. Furthermore, the training, primary test, and secondary test sets contain both thermochemistry (TC) data as well as noncovalent interactions (NC) data. The training set contains 787 TC datapoints and 321 NC datapoints, the primary test set contains 146 TC datapoints and 1047 NC datapoints, and the secondary test set contains 69 TC datapoints and 116 NC datapoints (for an overall total of 1002 TC datapoints and 1484 NC datapoints). Table 3.2 lists the 47 datasets that form the training, primary test, and secondary test sets. Details regarding the datasets will be discussed in this section. The references for the datasets are given in the rightmost column of Table 3.2 and will not be repeated in the text unless specific values from a table are being referenced.

The first 5 thermochemistry datasets in the training set are from Jan Martin's W4-11 dataset. All datapoints that involve multireference systems were removed from these datasets.

The reference values for DBH24 were taken from the second column ("TAE_e") of Table 1 in Reference 84 and the geometries were taken from the online Minnesota databases.

EA6 and EA7 (as well as IP6 and IP7) are subsets of Truhlar's EA13 and IP13 datasets. For the atoms in EA6 and IP6 (C, O, Si, P, S, and Cl), the reference values were recomputed at the CCSD(T)/aug-cc-pwC(Q5)Z[78, 147] level, while for the remaining molecules in EA7 and IP7, the 14 reference values (as well as the geometries) were taken from the online Minnesota databases (Column "REF1" in EA13/03 and IP21).

AE8 contains the absolute energies of the following atoms: H, He, B, C, N, O, F, and Ne. The energies for the latter 6 atoms were taken from Table XI (rightmost entry) in Reference 86, while the exact energies of the hydrogen atom (-0.5 hartree) and the helium atom[148] (-2.90372 hartrees) were used.

Moving on to the NC data in the training set, 3 potential energy curves (PEC) were removed from Sherrill's NBC10A dataset (parallel-displaced benzene dimer (3.2 Å and 3.6 Å) and benzene-H_2S dimer), and the rest of the PECs were divided into 3 subdatasets (NBC10A1, NBC10A2, and NBC10A3), while all 6 PECs from HBC6A were used.

From Crittenden's BzDC215 dataset, the following interactions were included in the training set: C_6H_6-{HF, H_2O, NH_3, CH_4, HCl}.

The reference values for AlkAtom19, AlkIsomer11, and AlkIsod14 were taken from Tables S7, S8, and S9 of the Supporting Information of Reference 93 (along with the B3LYP/pc-2 optimized geometries).

The geometries and reference values for HTBH38 and NHTBH38 were taken from the online Minnesota databases (Column "REF1" in HTBH38/08 and NHTBH38/08).

For the rare-gas dimers, the PECs each have 41 points with the following ranges in increments of 0.1 Å: Ne_2 (2.59 to 6.59 Å), Ar_2 (3.26 to 7.26 Å), and NeAr (2.98 to 6.98 Å). The reference values were taken from the Tang-Toennies potential model.

From WATER27, the 4 datapoints corresponding to the water 20-mers were excluded, and the rest of the reference values and geometries were taken from Grimme's online GMTKN30 database.

For HW30, the reference values were taken from the fourth column ("$E_{int}^{CCSD(T)/CBS}$") of Table 1 in Reference 100.

For NCCE31, only the 18 interactions from Table 1 of Reference 101 were used, and the reference values in the last column ("final D_e") of this table were used. The geometries were taken from the online Minnesota databases.

The revised S22B values from Sherrill were used for the S22 dataset, while Hobza's revised S66 reference values were used for the S66 dataset.

The reference values and geometries for G21EA, G21IP, and CYCONF were taken from Grimme's online GMTKN30 database. The reference values and geometries for PA8 were taken from the online Minnesota databases (Column "REF1" in PA8/06). For A24, the reference values were computed by adding the values in the "CCSD(T)/CBS" and "core correlation" columns in Table 1 of Reference 149.

The MP2/haTZ optimized geometries for the 8 water hexamers (bag, book1, book2, cage, cyclicboat1, cyclicboat2, cyclicchair (ring), prism) in H2O6Rel and H2O6Bind were taken from References 150 and 151. The term, haTZ, indicates that the aug-cc-pVTZ basis set was used for oxygen and the cc-pVTZ basis set was used for hydrogen. The binding energies in H2O6Bind were taken from the last column ("CCSD(T)/CBS - relaxation") of Table S5 in the Supporting Information of Reference 152, and the relative energies in H2O6Rel were computed using the binding energies in H2O6Bind (i.e. not the hexamer absolute energies).

The reference values for HW6F and HW6Cl were taken from the third column ("RI-CCSD(T)/CBS") of Tables S6 and S7 in the Supporting Information of Reference 152.

The reference values for DS14 were taken from the third column ("CBS (ΔaTZ)") of Table 1 in Reference 153.

3.5 Training

With a training set of 1108 datapoints, it is clear that a comprehensive 3-parameter nonlinear optimization is impractical. As a result, the 3 nonlinear parameters were determined from fits to a subset of the training set over a 3-dimensional set of points. The resulting values for ω, b, and C were 0.3, 6.0, and 0.01, respectively. ω was optimized in increments of 0.1, b was optimized in increments of 0.25, and C was optimized in increments of 0.0025. In comparison to existing functionals, the value of $\omega = 0.3$ is identical to the value that was determined for ωB97X[48] by Chai and Head-Gordon. In addition, Vydrov and Van Voorhis[57] found that the best b and C values were 5.9 and 0.0093 for rPW86+PBE+VV10 [VV10] and 6.3 and 0.0089 for ωPBE+PBE+VV10 [LC-VV10]. Thus, the nonlinear values that have been determined are reasonable. Any inaccuracies in these parameters will be accounted for during the optimization of the linear parameters.

With the nonlinear parameters determined, the 1961 required single-point calculations (corresponding to the data in the training and primary test sets) were carried out with the unoptimized ωB97X-V functional ($g_x = g_{c,ss} = g_{c,os} = 1$) in order to acquire the data necessary to perform least-squares fits to the training set *and* check the accuracy of the fits on the primary test set. Contributions to the total energy

Name	Description	#	Ref.
HAT707	Heavy-atom transfer reaction energies	505	82
BDE99	Bond dissociation reaction energies	83	82
TAE_nonMR124	Total atomization energies	124	82
SN13	Nucleophilic substitution reaction energies	13	82
ISOMER20	Isomerization reaction energies	18	82
DBH24	Diverse barrier heights	24	83,84
EA6	Electron affinities of atoms	6	85
IP6	Ionization potentials of atoms	6	85
AE8	Absolute atomic energies	8	86
SW49Rel345	$SO_4{}^{2-}(H_2O)_n$ ($n = 3-5$) relative energies	28	87
SW49Bind345	$SO_4{}^{2-}(H_2O)_n$ ($n = 3-5$) binding energies	30	87
NBC10A2	Methane dimer and benzene-methane dimer PECs	37	88,89
HBC6A	Formic acid, formamide acid, and formamidine acid dimer PECs	118	90,89
BzDC215	Benzene and first- and second-row hydride PECs	108	91
EA7	Electron affinities of small molecules	7	85
IP7	Ionization potentials of small molecules	7	85
Gill12	Neutral, radical, anionic, and cationic isodesmic reaction energies	12	92
AlkAtom19	$n = 1-8$ alkane atomization energies	19	93
AlkIsomer11	$n = 4-8$ alkane isomerization energies	11	93
AlkIsod14	$n = 3-8$ alkane isodesmic reaction energies	14	93
HTBH38	Hydrogen transfer barrier heights	38	94
NHTBH38	Non-hydrogen transfer barrier heights	38	95
SW49Rel6	$SO_4{}^{2-}(H_2O)_n$ ($n = 6$) relative energies	17	87
SW49Bind6	$SO_4{}^{2-}(H_2O)_n$ ($n = 6$) binding energies	18	87
NNTT41	Neon-neon PEC	41	96
AATT41	Argon-argon PEC	41	96
NATT41	Neon-argon PEC	41	96
NBC10A1	Parallel-displaced (3.4 Å), sandwich, and T-shaped benzene dimer PECs	53	88,89
NBC10A3	S2 and T3 configuration pyridine dimer PECs	39	97,89
WATER27	Neutral and charged water interactions	23	98,99
HW30	Hydrocarbon and water dimers	30	100
NCCE31	Noncovalent complexation energies	18	101
S22x5	Hydrogen-bonded and dispersion-bonded complex PECs	110	102
S66x8	Biomolecular structure complex PECs	528	103
S22	Equilibrium geometries from S22x5	22	104,89
S66	Equilibrium geometries from S66x8	66	103,105
G21EA	Adiabatic electron affinities	25	154,99
G21IP	Adiabatic ionization potentials	36	154,99
PA8	Adiabatic proton affinities	8	155,156
A24	Small noncovalent complexes	24	149
X40	Noncovalent interactions of halogenated molecules	40	144
H2O6Rel	Relative energies of water hexamers	8	152
H2O6Bind	Binding energies of water hexamers	8	152
HW6F	Binding energies of $F^-(H_2O)_n$ ($n = 1-6$)	6	152
HW6Cl	Binding energies of $Cl^-(H_2O)_n$ ($n = 1-6$)	6	152
CYCONF	Relative energies of cysteine conformers	10	157,99
DS14	Binding energies for complexes containing divalent sulfur	14	153

Table 3.2: Summary of the datasets found in the training, primary test, and secondary test sets. The datasets above the first thick black line are in the training set, the datasets between the first and second thick black lines are in the primary test set, while the datasets below the second thick black line are in the secondary test set. Within the training, primary test, and secondary test sets, datasets above the thin black line contain thermochemistry datapoints, while datasets below the thin black line contain noncovalent interactions datapoints. PEC stands for potential energy curve.

from $c_{x,i}$, $c_{css,i}$, and $c_{cos,i}$ for $i \in [0, 4]$ were saved, along with the contribution from c_x.

The only constraint that was deliberately enforced was the uniform electron gas limit for exchange ($c_{x,0} + c_x = 1$), since a review of the parameters of existing density functionals indicated that relaxing this constraint usually results in $c_{x,0} + c_x \approx 1$. For example, for B97[35], $c_{x,0} + c_x = 1.0037$, and for B97-1[67], $c_{x,0} + c_x = 0.999518$. On the other hand, $\{c_{css,0},\ c_{cos,0}\}$ for B97 and B97-1 are $\{0.17, 0.95\}$ and $\{0.08, 0.96\}$, respectively. Furthermore, since past experience with ωB97X[48] and ωB97X-D[58] indicated that c_x optimizes to a non-zero value (0.158 and 0.222, respectively), this parameter was always included in the least-squares fits.

With 4 free parameters from the exchange functional ICF, and 5 free parameters from each of the correlation functional ICFs, the total number of least-squares fits amounts to $\sum_{i=1}^{14} \binom{14}{i} = 2^{14} - 1 = 16383$.

Thermochemistry datapoints in the training and primary test sets are given weights of 1 and 2.5 respectively (except for datapoints in EA6 and IP6 which are weighted by 5), noncovalent interactions datapoints in the training and primary test sets are given weights of 10 and 25, respectively, and datapoints corresponding to the rare-gas dimer PECs in the primary test set are given weights of 25000. The total RMSD is defined as a weighted RMSD of all 1961 datapoints in the training and primary test sets, with the aforementioned weights.

Using the least-squares fits data, preliminary training set, primary test set, and total RMSDs were generated for all 16383 possible functional forms. Based on this data, it was concluded that fits that skip powers in the dimensionless variable, u, tend to perform comparably or worse than functionals with the same number of parameters that do not skip powers in u. After disregarding functionals that skip powers in u, a total of 500 unique functional forms remained.

The training set and primary test set RMSDs for these 500 least-squares fits plotted with respect to the number of linear parameters are shown in Figure 3.2. The data contained in Figure 3.2 is crucial for assessing the extent to which our design goal of a highly-accurate, transferable functional can be accomplished with minimal empiricism. Considering the figure on the left (which contains the training set RMSDs), it is evident that the best functional with a given number of linear parameters improves rapidly with each additional parameter up to 6 or 7 parameters. Subsequent additional parameters provide only small improvements to the training set RMSD. Thus, the curve corresponding to the lowest training set RMSD for a fixed number of linear parameters resembles a hockey stick.

On the right side of Figure 3.2 is the corresponding data for the primary test set RMSD. The most important attribute of this plot is that the curve defined by the lowest primary test set RMSD for a fixed number of linear parameters looks roughly similar to the plot on the left for up to 7 parameters, but the addition of more parameters leads to no improvement. In fact, the optimal primary test set RMSD values for functionals with more than 7 linear parameters gradually worsen, indicating reduced transferability. This plot depicts a useful guideline for determining the appropriate "extent of empiricism".

In order to easily identify the functionals in Figure 3.2, the following nomenclature will be utilized: 3 numbers (0-4) will indicate at which order the exchange, same-spin correlation, and opposite-spin correlation functional ICFs are truncated, while 3 letters (y(es) or n(o)) will indicate whether or not the UEG limit is enforced. For example, 3y0n2n indicates that the exchange functional ICF is truncated at $m_x = 3$, while the same-spin and opposite-spin correlation functional ICFs are truncated at $m_{css} = 0$ and $m_{cos} = 2$, respectively. Furthermore, it indicates that the UEG limit for exchange is satisfied (by construction), while neither of the UEG limits for correlation are enforced.

The primary test set RMSD plot on the right in Figure 3.2 was used to identify outstanding functionals. This figure clearly indicates that increasing the number of linear parameters past 7 is either ineffective or detrimental towards the goal of minimizing the primary test set RMSD. The 2 7-parameter fits that have virtually the same primary test set RMSDs are: 2y1n1n (green square, gray check mark, 2.05 kcal/mol) and 2y2y1n (downright orange triangle, magenta check mark, 2.03 kcal/mol). In order to differentiate between the two, it was necessary to consider their performance on the training set as well. Accordingly, the training set RMSD plot on the left in Figure 3.2 indicates that the 2y1n1n functional (3.36 kcal/mol) has a slightly lower training set RMSD than the 2y2y1n functional (3.40 kcal/mol). To isolate the winner, the total RMSD was plotted in the same manner. Figure 3.3 shows the total RMSD for all 16383 least-squares fits and it is clear that the 2y1n1n functional is the optimal choice, even when functionals that skip orders in u are considered. Ultimately, the 2y1n1n functional with 7 linear parameters was selected for self-consistent optimization.

With a training set RMSD of 3.36 kcal/mol, a primary test set RMSD of 2.05 kcal/mol, and a total RMSD of 2.76 kcal/mol, the 2y1n1n functional compares very well to the "best" functional from each RMSD category. The functional with the lowest training set RMSD is the 15-parameter 4y4n4n functional at 3.16 kcal/mol (with a primary test RMSD of 15.57 kcal/mol and a total RMSD of 11.42 kcal/mol), while the functional with the lowest primary test set RMSD is the aforementioned 2y2y1n functional at 2.03 kcal/mol (with a total RMSD of 2.78 kcal/mol). Finally, the functional with the lowest total RMSD is the 8-parameter 2y2n1n functional.

The 3 RMSDs of this fit match those of the 2y1n1n functional to the second decimal place, and the additional parameter is therefore unwarranted. Since Figure 3.3 plots the total RMSD for all 16383 fits, there are functional forms that skip orders in u that have slightly smaller total RMSDs than the 2y1n1n functional. The lowest total RMSD of 2.74 kcal/mol is achieved by a 9-parameter 2y4n1n functional that skips the first order linear parameter in $u_{c,\sigma\sigma}$. This fit has a training RMSD of 3.36 kcal/mol and a primary test set RMSD of 2.01 kcal/mol. However, its negligible improvement over the 2y1n1n functional form is not worth the 2 additional empirical parameters. Compared to the existing ωB97X and ωB97X-D functionals (13 linear parameters each), the new functional has 6 fewer linear parameters.

The training set RMSD of the unoptimized ωB97X-V functional is 25.84 kcal/mol, while the training set RMSD of the carefully selected 2y1n1n functional is 3.36 kcal/mol. With the data from Cycle 1, it is straightforward to produce a series of interesting results that will be discussed here. For example, if only the first-order gradient correction to the exchange functional is optimized (the resulting functional would be an RSH ($c_x = 0$) B86-type exchange functional with PW92 local correlation and VV10 nonlocal correlation), the resulting coefficient is $c_{x,1} = 0.718$. While it is not an apples-to-apples comparison, the corresponding coefficient that Becke determined in 1986 for the local B86 exchange-only functional by fitting to atomic exchange energies was 0.967. This single-parameter fit cuts the training set RMSD by more than a factor of 3 (7.76 kcal/mol). If only the short-range exact exchange parameter is allowed to vary (with the UEG limit for exchange enforced), the resulting value is $c_x = 0.475$, with a training set RMSD of 10.47 kcal/mol. Finally, the training set RMSD for the 2n2n2n "B97" functional form is 3.34 kcal/mol, with a primary test set RMSD of 2.62 kcal/mol and a total RMSD of 2.99 kcal/mol. For comparison, the training set, primary test set, and total RMSDs for Becke's B97 functional are 4.36, 15.75, and 11.74 kcal/mol, respectively, while appending Grimme's DFT-D2 dispersion tail to B97 gives the resulting B97-D2 functional corresponding RMSDs of 3.36, 4.52, and 4.00 kcal/mol, respectively.

Including the initial cycle (Cycle 1) with the unoptimized ωB97X-V functional, the self-consistent optimization of ωB97X-V required 3 cycles. For the first cycle, the datapoints in the training *and* primary test sets were evaluated in order to generate Figures 3.2 and 3.3 and determine the functional form that would be self-consistently optimized. For the latter 2 cycles, only the 1108 datapoints in the training set were required to fine-tune the parameters. The parameters from the beginning of all 3 cycles are listed in Table 3.3.

The RMSDs for the 14 datasets in the training set are shown in Table 3.4. The columns labeled "Cycle 1" and "Cycle 3" contain the actual RMSDs from the end of the respective cycle, while the column labeled "Cycle 1P" contains the least-squares

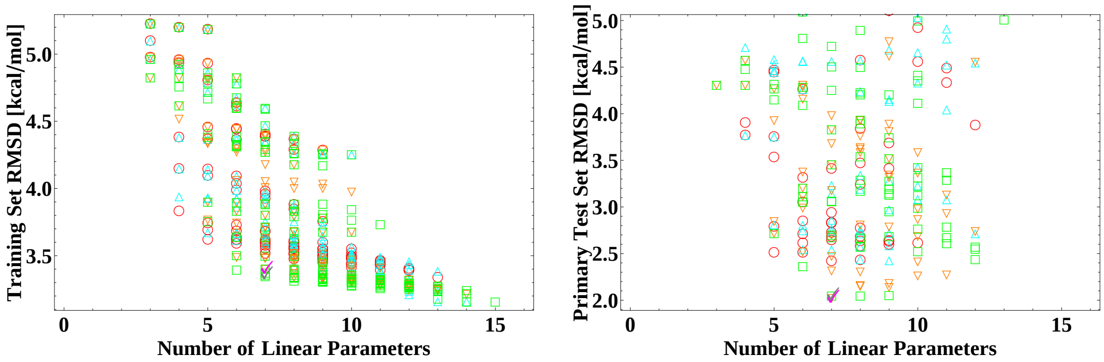


Figure 3.2: Training and primary test set RMSDs from 500 least-squares fits plotted against the number of linear parameters. Red circles indicate functionals that satisfy the UEG limits for exchange and correlation, upright cyan triangles indicate functionals that satisfy the UEG limits for exchange and opposite-spin correlation, downright orange triangles indicate functionals that satisfy the UEG limits for exchange and same-spin correlation, and green squares indicate functionals that satisfy the UEG limit for exchange only. The 2 checkmarked functionals are: 2y1n1n (gray) and 2y2y1n (magenta). The nomenclature is explained in Section 3.5.

fit RMSDs from the end of the first cycle. Figures 3.2 and 3.3 were generated with the same data that is used to produce the values in the "Cycle 1P" column. It is quite remarkable that the least-squares fit RMSDs from Cycle 1P so closely resemble the final RMSDs of the ωB97X-V functional from Cycle 3. For all of the datasets except AE8, the least-squares fit and final RMSDs differ by 0.05 kcal/mol at most, while for AE8 the difference is 0.20 kcal/mol, due to the large magnitude of the absolute atomic energies.

3.6 Characteristics of ωB97X-V

The final parameters of ωB97X-V can be found in the fourth column of Table 3.3 under the "Cycle 3" heading. While the uniform electron gas (UEG) limit for exchange was enforced from the beginning, the UEG limits for same-spin correlation *and* opposite-spin correlation were allowed to relax and are not satisfied. However, the first-order corrections to same-spin correlation and opposite-spin correlation should be (and are) negative, since the LSDA overestimates the correlation energies of atoms by a factor of 2. On the other hand, the LSDA underestimates the

Parameter	Cycle 1	Cycle 2	Cycle 3
$c_{x,1}$	1.000	0.614	**0.603**
$c_{x,2}$	0.000	1.181	**1.194**
$c_{css,0}$	1.000	0.575	**0.556**
$c_{css,1}$	0.000	-0.274	**-0.257**
$c_{cos,0}$	1.000	1.219	**1.219**
$c_{cos,1}$	0.000	-1.867	**-1.850**
c_x	0.000	0.163	**0.167**

Table 3.3: Linear parameters from the beginning of all 3 cycles of the self-consistent optimization of ωB97X-V. The corresponding nonlinear parameters are $\omega = 0.3$, $b = 6.0$, and $C = 0.01$. The final parameters are listed under Cycle 3. The value of $c_{x,0}$ is not listed because the uniform electron gas limit for exchange was enforced, requiring $c_{x,0} = 1 - c_x$. The initial guess ("Cycle 1") corresponding to the unoptimized ωB97X-V functional was attained by setting all of the linear parameters to zero, besides $c_{x,0} = c_{css,0} = c_{cos,0} = 1$.

Dataset	Cycle 1	Cycle 1P	Cycle 3
HAT707	8.88	4.26	4.28
BDE99	15.10	3.36	3.38
TAE_nonMR124	36.14	3.31	3.34
SN13	11.48	0.97	1.01
ISOMER20	2.98	1.59	1.64
DBH24	7.56	1.77	1.81
EA6	12.42	2.35	2.34
IP6	12.54	3.74	3.76
AE8	247.87	1.57	1.77
SW49Rel345	1.49	0.34	0.33
SW49Bind345	4.22	0.31	0.29
NBC10A2	0.08	0.08	0.09
HBC6A	2.71	0.39	0.39
BzDC215	0.53	0.26	0.27

Table 3.4: RMSDs in kcal/mol for the 14 datasets comprising the training set. The columns labeled "Cycle 1" and "Cycle 3" contain the actual RMSDs from the end of the respective cycle, while the column labeled "Cycle 1P" contains the least-squares fit RMSDs from the end of the first cycle.

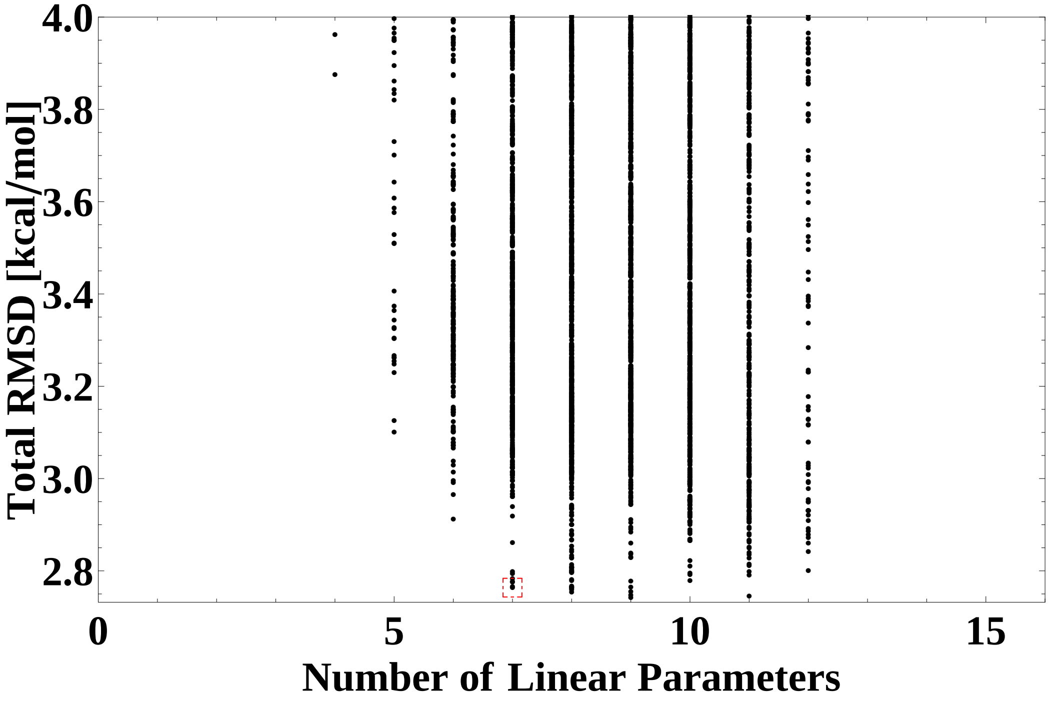


Figure 3.3: Total RMSDs from 16383 least-squares fits plotted against the number of linear parameters. The boxed point corresponds to the 2y1n1n functional with 7 linear parameters that was picked for self-consistent optimization. The nomenclature is explained in Section 3.5.

exchange energies of atoms by approximately 10%, and the first-order correction to exchange should be (and is) positive.

It is important to consider the behavior of the ICFs for large gradients in the electron density ($u \approx 1$). The ωB97X-V exchange functional ICF is well-behaved, with a value of 2.630 at $u_{x,\sigma} = 1$. Even though the Lieb-Oxford bound[158] (LOB) was not used as a constraint during the optimization of ωB97X-V, the resulting exchange functional ICF violates the LOB (2.273) by only 0.357. In comparison, the exchange functional ICFs of ωB97X and ωB97X-D have values of 10.189 and 8.396 at $u_{x,\sigma} = 1$, respectively. The ωB97X-V exchange functional ICF exceeds the LOB at $u_{x,\sigma} = 0.874$, which corresponds to $s_\sigma = 41.707$. According to Perdew and coworkers[159], only the range corresponding to $0 \leq s_\sigma \leq 18.562$ is important for the exchange-correlation energy of atoms. Thus, it is likely that the ωB97X-V exchange functional ICF violates the Lieb-Oxford bound outside of the realm that is considered important for systems of interest. It is important to point out that the LiebOxford bound is a constraint on the total integrated exchange energy of a density functional, not the exchange energy corresponding to a given grid point. However, if the local exchange functional ICF does not violate the LOB, then the local exchange functional will never violate the LOB. While the same-spin correlation ICF is positive for all values of $u_{c,\sigma\sigma}$, the opposite-spin correlation ICF takes on negative values for $u_{c,\alpha\beta} \geq 0.659$. Although this implies that certain grid points will contribute positive values to the overall correlation energy, Becke witnessed similar behavior for B97 at $u_{c,\alpha\beta} = 0.54$ and concluded that the effects from this feature were of little importance because the sign change occurred far out from the region of chemical relevance.

In order to further investigate the implications of the exchange functional ICF of ωB97X-V violating the LOB at a value of $u_{x,\sigma} = 0.874$, a molecule from the training set was selected and analyzed. The molecule that was chosen was one of the dianionic sulfate-water cluster isomers with 3 water molecules. With the slightly coarser (75,302) integration grid, this molecule requires 317,100 grid points for the integration of the local exchange energy (-60.2615 hartrees). Since there is a value of $u_{x,\sigma}$ that corresponds to each of these grid points, the goal of this analysis was to determine if a significant fraction of these grid points corresponded to values of $u_{x,\sigma} > 0.874$. Figure 3.4 plots the sum of the grid point exchange energies between $u_{x,\sigma} = 0$ and $u_{x,\sigma} < z$ for $z \in [0, 1]$ in increments of 0.01. The data indicates that the majority of the local exchange energy is recovered by $u_{x,\sigma} = 0.3$, and the points that lie between $u_{x,\sigma} = 0.3$ and $u_{x,\sigma} = 1.0$ contribute negligibly. In fact, the grid points between $u_{x,\sigma} = 0$ and $u_{x,\sigma} < 0.874$ contribute 99.9998% of the total local exchange energy, while the remaining grid points contribute only an additional -0.076 kcal/mol to the local exchange energy. Thus, we believe that it is safe to conclude that the violation of the Lieb-Oxford bound by the ωB97X-V exchange functional ICF has nearly negligible chemical implications. For the 1961 systems in the training and primary test sets, the ratio of the total exchange energy of ωB97X-V (local and nonlocal) to the exchange energy of the LSDA is at least 1.076, at most 1.193, and 1.107 on average.

Figure 3.5 shows the exchange, same-spin correlation, and opposite-spin correlation functional ICF plots for the final ωB97X-V functional, as well as several B97-based functionals, namely, ωB97X, ωB97X-D, and the original B97 functional by Becke. Compared to the ICFs of existing functionals, the exchange functional ICF of ωB97X-V is almost identical to that of B97. While the B97 and ωB97X-V exchange functional ICFs are quadratic, they are quartic for both ωB97X and ωB97X-D. The largest difference among the 4 functionals considered is seen in the same-spin correlation functional ICFs, where ωB97X and ωB97X-D are quartic and oscillatory, B97 is quadratic, and ωB97X-V is linear. In fact, the same-spin corre-

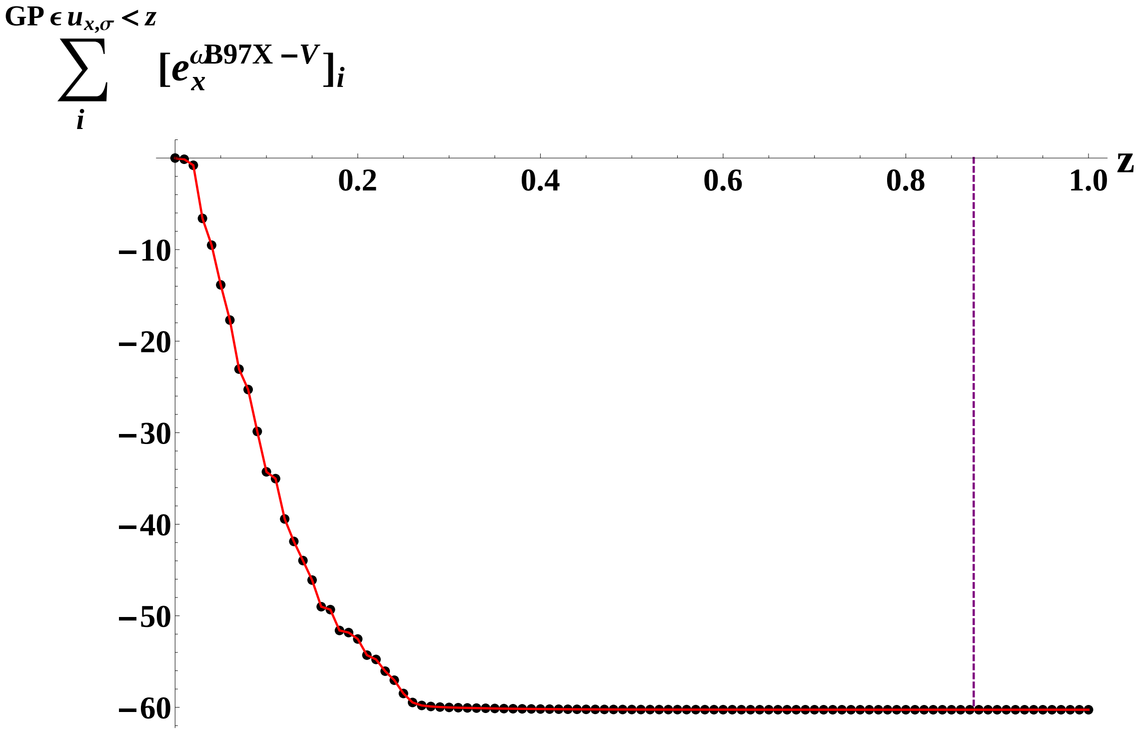


Figure 3.4: Sum of the local exchange energy contributions from all grid points (GP) between 0 and z for a $SO_4{}^{2-}(H_2O)_3$ isomer. The points are evaluated in increments of 0.01. For example, the sixth point from the left corresponds to the sum of the local exchange energy contributions from all grid points that fall between $u_{x,\sigma} = 0.0$ and $u_{x,\sigma} = 0.05$. The last point is equivalent to the total local exchange energy of the molecule, -60.2615 hartrees. The dashed purple line corresponds to the value of $u_{x,\sigma} = 0.874$ at which the exchange functional ICF of ωB97X-V violates the Lieb-Oxford bound. The grid points between $u_{x,\sigma} = 0$ and $u_{x,\sigma} < 0.874$ contribute 99.9998% of the total local exchange energy of the system.

System	PBE	B3LYP	B97	B97-D	VV10	LC-VV10	ωB97X-D	ωB97X-V	M06-L	M06	M06-2X	M11-L	M11
H-	2.11	-4.40	0.50	-5.36	-1.67	4.11	-2.69	1.05	1.13	2.49	4.07	-21.14	-3.03
He	6.77	-7.22	-3.94	-8.09	-8.01	4.13	-3.73	0.12	-7.46	-5.02	-1.02	-10.66	-3.46
Li	9.97	-9.35	-6.45	-6.20	-15.63	12.24	-8.29	-0.68	-8.05	-6.98	-3.92	-36.67	-6.70
Li-	13.26	-7.10	-3.19	-6.84	-11.51	17.43	-5.68	3.91	-7.42	-4.22	-0.13	-31.37	-6.01
Be	23.47	-3.75	0.14	1.47	-15.21	28.57	0.10	2.73	0.51	2.09	-0.07	-38.91	-2.29
Ne	44.67	-27.21	-2.67	1.79	-110.44	42.42	-11.13	0.58	-16.04	-11.50	-4.01	-17.51	-27.43
Ar	121.70	-17.44	34.72	-38.17	-211.12	129.85	-8.19	40.53	-6.11	2.23	-1.07	-40.45	-7.70

Table 3.5: Difference (in kcal/mol) between the DFT energy and the best available variational energy for 5 atoms and 2 anions. Entries that are negative correspond to non-variational values. The UGBS basis set was used with the (500,974) grid for local xc functionals and the (99,590) grid for the VV10 NLC functional.

lation functional ICF of ωB97X-V looks liked an averaged-out version of its ωB97X and ωB97X-D counterparts. Finally, the opposite-spin correlation functional ICFs are similar in the region between $u_{c,\alpha\beta} = 0$ and $u_{c,\alpha\beta} = 0.5$, and differ only for large values of $u_{c,\alpha\beta}$.

While DFT is a variational method in principle, non-empirical and semi-empirical exchange-correlation functionals that attempt to approximate the exact exchange-correlation functional are not guaranteed to provide variational energies. In order to assess the variational validity of ωB97X-V, the absolute energies of the 6 systems from Reference 160 (along with the absolute energy of the argon atom from Reference 86) were computed and compared to the best available variational values. Table 3.5 contains the difference (in kcal/mol) between the DFT energy and the best available variational energy, with negative entries corresponding to non-variational energies and positive entries corresponding to variational energies. The UGBS basis set[161] was used with the (500,974) grid for local xc functionals and the (99,590) grid for the VV10 NLC functional. The only functionals that provide variational energies for all 7 test systems are PBE and LC-VV10. ωB97X-V provides a non-variational absolute energy for the lithium atom, but it is lower by only -0.68 kcal/mol. The rest of the functionals have at least 4 non-variational entries. It is surprising that while LC-VV10 has variational energies for all 7 cases, its local version, VV10, has non-variational energies for all 7 cases.

3.7 Results and Discussion

3.7.1 Performance

Thus far, we have completed the training and selection of the ωB97X-V functional in a manner that is internally consistent, such that the resulting functional emerges

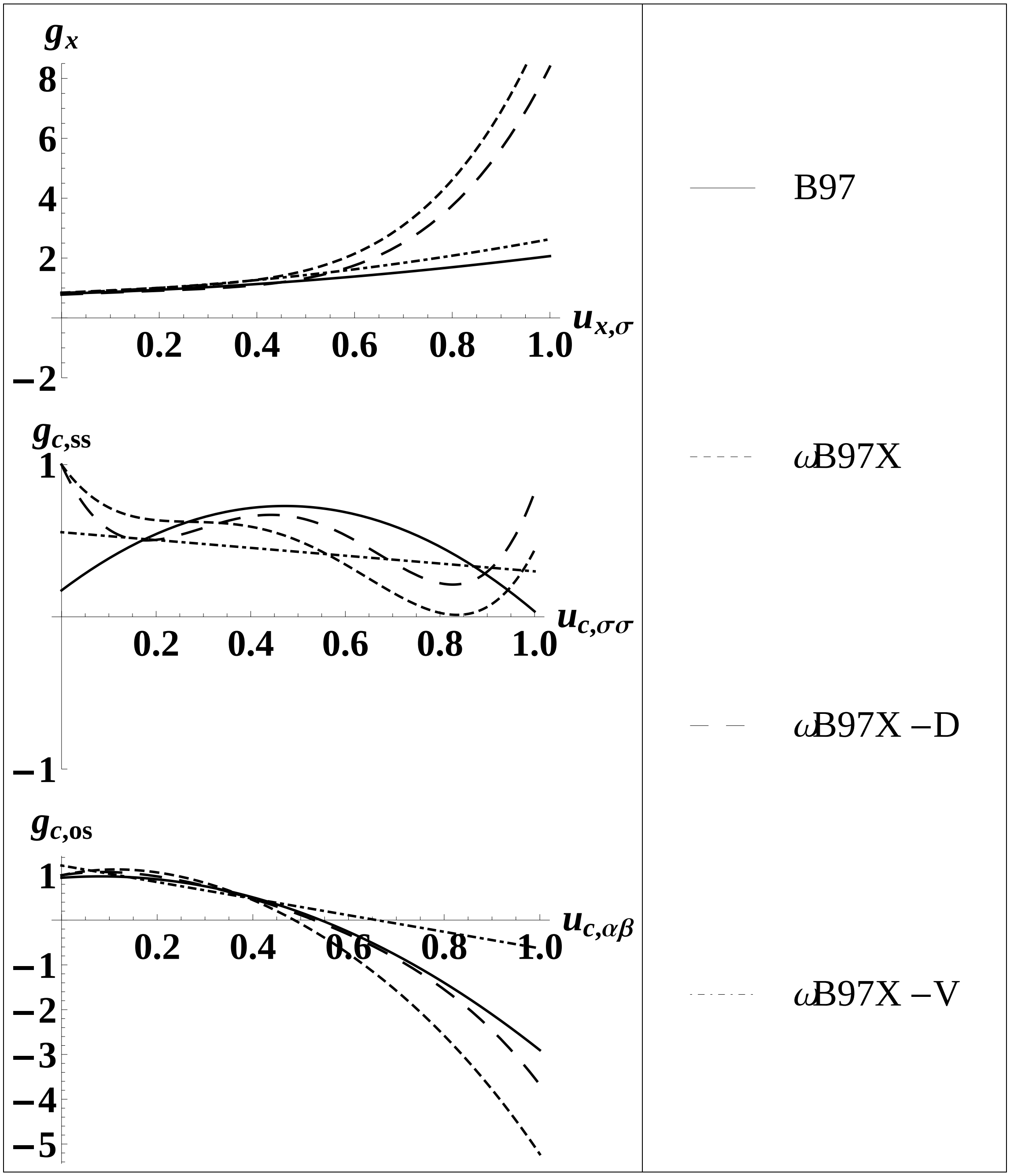


Figure 3.5: Inhomogeneity correction factors for the exchange, same-spin correlation, and opposite-spin correlation functionals of B97, ωB97X, ωB97X-D, and ωB97X-V.

as the fittest of over 16000 candidate functionals. The next step is to compare against a selection of existing density functionals across the training set, primary test set, and the hitherto unused secondary test set. The density functionals that were benchmarked along with ωB97X-V were selected for reasons that will be briefly described in this section.

PBE is arguably the best unparameterized density functional available, while B3LYP is certainly the most popular density functional to date. B97 is the forefather of dozens of parameterized density functionals (including ωB97X-V). Since the 3 functionals mentioned thus far are ill-equipped for describing weak dispersive interactions, dispersion-corrected variants are commonly employed as well. While the dispersion tails of PBE-D2, B3LYP-D2, and B97-D2 were optimized onto the existing parent functional, Grimme's B97-D functional was the first attempt at simultaneously parameterizing the linear parameters of a density functional with a dispersion tail. Since the VV10 nonlocal correlation functional is a vital component of ωB97X-V, the 2 existing density functionals that incorporate the VV10 NLC functional (VV10 and LC-VV10) must be benchmarked. With respect to dispersion-corrected RSH GGA density functionals, the direct predecessor to ωB97X-V is ωB97X-D, and it is important to confirm that ωB97X-V improves upon its older counterpart. Finally, from the 10 Minnesota functionals mentioned in Section 3.1, M06-L, M06, M06-2X, M11-L and M11 were chosen.

Table 3.6 contains the RMSDs for all of the datasets in the training, primary test, and secondary test sets for ωB97X-V and 15 existing density functionals (PBE, PBE-D2, B3LYP, B3LYP-D2, B97, B97-D2, B97-D, VV10, LC-VV10, ωB97X-D, M06-L, M06, M06-2X, M11-L, and M11). Table 3.7 contains information regarding the benchmarked functionals.

Of the 16 benchmarked density functionals, M06-2X has the best overall performance for thermochemistry (3.21 kcal/mol) and ωB97X-V has the best overall performance for noncovalent interactions (0.32 kcal/mol). After M06-2X, the next best density functionals for thermochemistry are ωB97X-V and ωB97X-D, with RMSDs of 3.60 and 3.61 kcal/mol, respectively. After ωB97X-V, the next best functionals for noncovalent interactions are M06-L and B97-D2, with RMSDs of 0.47 and 0.48 kcal/mol, respectively. Before the individual datasets are discussed, the overall performance of the functionals for thermochemistry (TC) and noncovalent interactions (NC) will be discussed.

Overall, the performance of ωB97X-V is identical to that of ωB97X-D for thermochemistry, but more than 0.20 kcal/mol better for noncovalent interactions. As a sanity check, it is important to verify that ωB97X-V drastically improves upon VV10. ωB97X-V reduces the overall thermochemistry RMSD of its counterpart by a factor of 2.7. Furthermore, it improves upon the overall NC RMSD of VV10 by a factor of 4.3. In addition to outperforming VV10, ωB97X-V should perform much better than LC-VV10 as well, since its local GGA components have been carefully parameterized. As expected, ωB97X-V reduces the overall TC and NC RMSDs of LC-VV10 by a factor of 2. The functional forms of ωB97X-V and M11 are similar, since both functionals are range-separated hybrids that have exchange-correlation functionals that depend on the density as well as its gradient. However, the exchange-correlation functional of M11 additionally depends on the kinetic energy density, while ωB97X-V includes the VV10 NLC functional instead. Furthermore, M11 has 4 times more parameters than ωB97X-V. Thus, it is quite surprising that ωB97X-V outperforms M11 with respect to both overall thermochemistry and overall noncovalent interactions. While M11-L is the newest local meta-GGA Minnesota functional, its older counterpart, M06-L, outperforms it with respect to both overall thermochemistry and overall noncovalent interactions. M06-2X is better than M06 by a factor of 1.3 for thermochemistry, while the reverse is true for noncovalent interactions. It is interesting to compare the performance of the functionals with the DFT-D2 dispersion tails to their parent functionals to assess the extent to which the addition of the tail enhances the performance of the parent functional for noncovalent interactions. Furthermore, it is desirable that the parent functional's performance for thermochemistry remains unaltered. The addition of DFT-D2 to B3LYP improves its overall NC RMSD by a factor of 3, and also slightly improves its performance for thermochemistry. With Becke's B97 functional, the improvement for noncovalent interactions is even more dramatic, as the DFT-D2 tail cuts the overall NC RMSD of B97 by more than a factor of 5. For PBE, however, the addition of the DFT-D2 tail enhances the description of noncovalent interactions very slightly, and worsens the TC RMSD.

Since it would be tedious to individually address the performance of the benchmarked density functionals on the 47 datasets in Table 3.6, only a handful of datasets will be discussed. The TAE_nonMR124 dataset is comprised of the atomization energies of small molecules computed with the Weizmann-4 (W4) theory. ωB97X-V has an RMSD of 3.34 kcal/mol on this dataset of 124 datapoints, and is second only to M06-2X (3.24 kcal/mol). Furthermore, the performance of ωB97X-V is more than 0.30 kcal/mol better than that of ωB97X-D, and more than 1 kcal/mol better than that of M11. The DBH24 dataset contains 24 forward and reverse barrier heights computed (at least) with the Weizmann-3.2 (W3.2) theory. M06-2X performs the best for this dataset (1.12 kcal/mol), followed by M11 (1.48 kcal/mol) and ωB97X-V (1.81 kcal/mol). The AlkAtom19 dataset contains the atomization energies of 19 alkanes ranging from methane to octane. Since these molecules are much larger than the ones found in TAE_nonMR124, it is important to assess the performance of ωB97X-V on this dataset to determine if the satisfactory performance for the atomization energies of smaller molecules is transferable to larger molecules. The RMSD of ωB97X-V on AlkAtom19 (0.71 kcal/mol) is 7 times smaller than that of M06-2X (5.27 kcal/mol) and 5 times smaller than that of M11 (3.94 kcal/mol). ωB97X-V beats ωB97X-D by a factor of 4 on the AlkAtom19 dataset, and is outperformed only by B3LYP-D2 (0.64 kcal/mol). However, it is clear that the surprisingly small RMSD of B3LYP-D2 is a fortuitous result, since the dispersion tail contributes only to the absolute energies of the alkanes (and not the atoms). Thus, while B3LYP drastically underestimates the atomization energies of the alkanes in AlkAtom19, the dispersion tail manages to lower the absolute energies of the alkanes to overcome this underestimation. If one considers PBE (which drastically overestimates the atomization energies on its own), the addition of the same dispersion tail makes things much worse.

In recent years, Hobza's S22 and S66 datasets have been quite popular for benchmarking the performance of density functionals on noncovalent interactions. The RMSD of ωB97X-V on the S22 dataset is 0.23 kcal/mol, which is almost 2 times smaller than the next best functional, ωB97X-D (0.41 kcal/mol). It is important to emphasize that the parameters of ωB97X-D were trained on the S22 dataset, while the parameters of ωB97X-V were validated, but not trained, on the S22 dataset. From the Minnesota functionals, M06-L has the best performance for S22, with an RMSD of 0.43 kcal/mol, followed by M06-2X (0.47 kcal/mol) and M11 (0.58 kcal/mol). On the S66 dataset, ωB97X-V has the lowest RMSD at 0.18 kcal/mol, with the next best functional (M06-2X) having an RMSD of 0.29 kcal/mol. The most recent dataset by Hobza is A24, which consists of very accurate CCSD(T)/CBS binding energies for small molecules. ωB97X-V performs the best for this dataset as well, with an RMSD of only 0.09 kcal/mol. The 5 Minnesota functionals have RMSDs larger than 0.20 kcal/mol, and the only functionals that comes close to ωB97X-V are ωB97X-D and LC-VV10, with RMSDs of 0.15 kcal/mol. The last dataset from Hobza that will be discussed is X40, which is comprised of binding energies of halogenated molecules. ωB97X-V has the best performance for this dataset, with an RMSD of 0.21 kcal/mol. M06-2X performs satisfactorily on X40 as well, with an RMSD of 0.28 kcal/mol, while its family members, M06-L (0.48 kcal/mol), M06 (0.57 kcal/mol), M11-L (1.23 kcal/mol), and M11 (0.54 kcal/mol), have RMSDs that are at least twice as large as that of ωB97X-V.

Herbert and coworkers recently reported[152] that density functionals such as LC-VV10 and M06-2X perform poorly for halide-water clusters. Specifically, the systems of interest are $F^-(H_2O)_n$ and $Cl^-(H_2O)_n$, for $n = 1 - 6$. Using the same geometries and reference values, the RMSDs for these 2 datasets were computed for the 16 density functionals in Table 3.6. ωB97X-V has the best performance for the interactions containing the fluoride anion, with an RMSD of 0.36 kcal/mol. In comparison, the

Minnesota functionals perform at least 2.5 times worse. The performance of ωB97X-D is 4 times worse than the performance of its new counterpart, and the only density functional that comes close to ωB97X-V is B97-D2, with an RMSD of 0.59 kcal/mol. However, for the interactions that contain the chloride anion, B97-D2 has the best performance, with an RMSD of only 0.33 kcal/mol. Surprisingly, PBE comes in at second with an RMSD of 0.59 kcal/mol, while ωB97X-V has the third best RMSD at 0.67 kcal/mol. While the Minnesota functionals tend to perform slightly better on this dataset than on HW6F, their RMSDs are still at least 4.5 times larger than that of B97-D2. Both the H2O6Rel and CYCONF datasets in the secondary test set are meant to gauge the performance of density functionals for the relative energies of conformers. ωB97X-V has the best performance for both datasets, with RMSDs of 0.07 kcal/mol and 0.11 kcal/mol, respectively. For H2O6Rel, the next best density functional is ωB97X-D with an RMSD of 0.18 kcal/mol, while the Minnesota functionals have RMSDs larger than 1 kcal/mol. In addition, the RMSD of ωB97X-V (0.11 kcal/mol) for CYCONF is 3 times smaller than that of ωB97X-D (0.41 kcal/mol), but comparable to that of M06 (0.16 kcal/mol). The last dataset in the secondary training set is DS14, which contains binding energies for systems that contain divalent sulfur. Since the molecules in this dataset are small, the reference values were computed at the CCSD(T)/CBS level (with counterpoise corrections), with the following contributions: HF/aQZ + MP2/a(TQ)Z + (CCSD(T)-MP2)/aTZ. The performance of ωB97X-V on DS14 is near perfect, with an RMSD of only 0.05 kcal/mol. The next best density functional, LC-VV10, has an RMSD that is 2 times larger (0.12 kcal/mol). M06-2X has the best performance on DS14 from the Minnesota functionals, but its RMSD of 0.20 kcal/mol is still 4 times larger than that of ωB97X-V.

In Figure 3.6, the potential energy curves (PEC) for the helium dimer (He_2) and argon dimer (Ar_2) are shown. The argon dimer is an important case to consider for ωB97X-V, since the datapoints corresponding to 3 rare-gas dimers (Ne_2, Ar_2, and NeAr) were weighted by 25000 in the calculation of the primary test set RMSD. While the parameters of ωB97X-V were not explicitly optimized on the argon dimer PEC, the functional form that was ultimately selected (2y1n1n) was influenced by its good performance for these 3 dimers. Thus, it is not surprising that the ωB97X-V PEC for the argon dimer in Figure 3.6 is basically superimposed on the "Reference" PEC (Tang-Toennies). Furthermore, it is satisfactory that the good performance of VV10 has been maintained. Besides confirming that the functional selection strategy worked exactly as intended, Figure 3.6 highlights a disadvantage of functionals that were trained without checking for transferability in the selected linear parameters. ωB97X-D predicts an equilibrium bond distance that is 0.5 Å too long and underbinds the dimer. LC-VV10, M06-L, and M06-2X perform comparably to ωB97X-D, while

Dataset	PBE	PBE-D2	B3LYP	B3LYP-D2	B97	B97-D2	B97-D	VV10	LC-VV10	ωB97X-D	ωB97X-V	M06-L	M06	M06-2X	M11-L	M11
HAT707	7.81	7.79	4.31	4.28	3.98	3.98	5.62	7.23	6.81	4.14	4.28	5.86	4.83	3.63	5.43	4.28
BDE99	8.89	8.98	4.07	3.88	3.23	3.18	4.58	6.63	5.39	3.03	3.38	6.14	3.72	2.99	4.61	4.10
TAE_nonMR124	16.72	16.94	5.43	5.28	4.39	4.06	5.18	12.46	5.30	3.65	3.34	5.54	3.94	3.24	6.62	4.37
SN13	5.74	5.36	2.33	1.74	1.28	0.97	4.20	3.67	2.65	0.96	1.01	1.58	2.03	1.01	3.19	2.19
ISOMER20	4.54	4.53	2.35	2.29	2.34	2.30	3.83	4.34	1.85	1.86	1.64	4.11	2.53	1.47	4.46	1.94
DBH24	10.05	10.37	4.83	5.28	4.05	4.36	7.18	9.86	3.02	2.07	1.81	5.38	2.97	1.12	3.54	1.48
EA6	4.71	4.71	3.49	3.49	1.45	1.45	2.07	5.28	2.10	1.89	2.34	2.39	1.91	1.95	6.17	1.33
IP6	5.07	5.07	6.16	6.16	3.19	3.19	2.68	7.64	5.03	3.01	3.76	2.26	3.23	2.86	3.36	5.79
AE8	30.54	30.54	15.89	15.89	2.25	2.25	6.94	64.43	29.82	4.67	1.77	9.22	5.52	1.28	15.27	9.23
SW49Rel345	0.90	1.55	0.39	1.17	0.30	0.79	1.22	1.26	0.36	1.01	0.33	0.53	0.60	0.35	0.22	0.17
SW49Bind345	1.22	2.59	3.07	1.72	3.44	0.69	1.35	2.47	0.58	0.98	0.29	0.60	0.72	0.95	1.54	0.25
NBC10A2	0.91	0.28	1.51	0.20	1.19	0.08	0.08	0.05	0.12	0.13	0.09	0.15	0.28	0.11	0.54	0.16
HBC6A	0.54	1.45	1.27	0.94	1.37	0.39	0.77	1.13	1.25	0.61	0.39	0.36	0.31	0.59	0.68	0.46
BzDC215	1.07	0.69	1.94	0.57	1.54	0.36	0.39	0.22	0.60	0.41	0.27	0.17	0.25	0.60	0.34	0.65
EA7	2.27	2.28	3.52	3.54	2.54	2.55	3.86	3.57	4.19	2.45	2.28	6.07	2.94	1.97	6.44	1.08
IP7	4.28	4.28	5.22	5.22	3.41	3.41	4.03	4.76	5.30	3.26	2.97	3.11	2.93	3.24	2.89	4.56
Gill12	8.24	6.56	5.71	3.27	5.69	3.84	4.84	6.08	2.58	3.24	2.32	6.71	3.82	1.78	5.60	2.58
AlkAtom19	15.83	26.21	14.63	0.64	19.78	9.28	10.48	5.85	19.04	2.90	0.71	8.11	4.63	5.27	29.35	3.94
AlkIsomer11	2.87	0.40	3.96	0.59	3.78	0.57	0.33	0.84	0.13	1.04	0.67	0.95	0.22	0.15	0.74	0.56
AlkIsod14	4.88	0.42	6.57	0.35	6.44	1.93	0.57	2.01	1.09	2.31	1.80	3.86	1.68	1.65	2.06	2.11
HTBH38	9.97	10.57	4.81	5.49	4.53	5.12	7.63	9.22	1.50	2.57	2.28	4.62	2.21	1.26	1.86	1.68
NHTBH38	10.20	10.53	5.41	5.84	3.83	4.15	7.12	10.79	3.38	1.73	1.64	4.79	2.64	1.69	3.72	1.49
SW49Rel6	1.20	2.16	0.29	1.59	0.16	1.01	1.65	1.72	0.52	1.34	0.37	0.80	0.89	0.42	0.21	0.28
SW49Bind6	2.40	4.51	6.31	2.92	7.18	0.90	2.50	4.56	0.91	0.99	0.36	0.57	0.92	1.61	3.38	0.67
NNTT41	0.02	0.09	0.08	0.05	0.03	0.06	0.07	0.03	0.02	0.15	0.02	0.04	0.12	0.05	0.32	0.02
AATT41	0.13	0.05	0.30	0.08	0.19	0.05	0.16	0.02	0.07	0.21	0.02	0.15	0.23	0.08	0.46	0.17
NATT41	0.02	0.09	0.14	0.02	0.06	0.05	0.08	0.04	0.03	0.15	0.02	0.05	0.13	0.03	0.35	0.04
NBC10A1	3.29	0.19	4.77	0.26	3.98	0.67	0.20	0.27	0.15	0.41	0.23	0.26	0.58	0.43	0.47	0.69
NBC10A3	2.57	0.28	3.77	0.29	3.20	0.54	0.19	0.17	0.06	0.27	0.37	0.47	0.69	0.63	0.83	0.80
WATER27	3.46	7.58	3.63	4.63	4.72	1.42	2.56	7.07	3.35	1.48	1.42	1.34	1.70	3.66	4.60	1.62
HW30	0.52	0.71	1.12	0.53	0.83	0.35	0.40	0.43	0.30	0.35	0.20	0.35	0.33	0.46	0.48	0.38
NCCE31	1.97	2.45	1.00	1.05	0.93	0.84	1.27	2.67	0.53	0.38	0.27	1.23	0.56	0.36	0.75	0.51
S22x5	2.91	0.65	3.99	0.61	3.64	0.71	0.40	0.59	0.48	0.43	0.41	0.42	0.80	0.47	0.94	0.55
S66x8	2.19	0.64	3.17	0.65	2.73	0.35	0.43	0.50	0.34	0.44	0.22	0.31	0.48	0.35	0.83	0.42
S22	3.50	0.70	4.74	0.74	4.25	0.60	0.54	0.63	0.51	0.41	0.23	0.43	0.77	0.47	0.91	0.58
S66	2.59	0.63	3.71	0.62	3.16	0.36	0.52	0.52	0.31	0.52	0.18	0.36	0.53	0.29	0.81	0.41
G21EA	3.97	3.96	3.54	3.54	1.83	1.84	3.00	4.60	3.89	1.98	2.27	4.54	2.46	1.99	6.56	2.40
G21IP	4.81	4.81	4.86	4.86	3.48	3.48	4.47	5.43	5.23	3.82	3.57	5.60	3.78	3.49	4.50	4.64
PA8	1.64	1.71	1.20	1.49	2.41	2.67	3.70	1.77	2.74	2.95	2.20	3.36	2.21	1.98	3.57	1.35
A24	0.46	0.59	1.05	0.39	0.75	0.26	0.32	0.41	0.15	0.15	0.09	0.23	0.25	0.28	0.46	0.27
X40	1.63	0.74	2.48	0.47	2.08	0.43	0.59	0.63	0.41	0.49	0.21	0.48	0.57	0.28	1.23	0.54
H2O6Rel	1.77	0.35	2.31	0.59	2.26	0.20	0.41	0.38	0.22	0.18	0.07	1.75	2.10	1.76	1.01	1.45
H2O6Bind	0.90	5.74	3.76	4.10	5.27	0.33	3.82	5.70	2.68	1.55	0.70	0.97	1.01	2.25	4.00	0.67
HW6F	1.72	3.64	4.64	2.78	5.83	0.59	4.34	3.71	2.55	1.46	0.36	2.47	2.11	4.91	1.04	2.95
HW6Cl	0.59	4.74	5.27	2.05	5.49	0.33	3.25	5.26	2.84	0.78	0.67	1.61	2.10	3.65	2.73	2.25
CYCONF	0.95	0.98	0.53	0.32	0.49	0.37	0.56	0.57	0.78	0.41	0.11	0.39	0.16	0.30	0.57	0.51
DS14	1.10	0.57	1.94	0.28	1.55	0.25	0.37	0.52	0.12	0.18	0.05	0.25	0.34	0.20	0.59	0.30
All TC	9.90	10.35	5.11	4.59	4.74	3.97	5.56	9.81	6.79	3.61	3.60	5.63	4.18	3.21	6.68	3.97
All NC	2.00	1.49	2.96	1.01	2.71	0.48	0.82	1.38	0.73	0.54	0.32	0.47	0.60	0.78	1.07	0.56
Training TC	10.27	10.33	4.71	4.66	3.89	3.83	5.43	10.44	6.86	3.82	3.88	5.75	4.45	3.36	5.64	4.23
Training NC	0.89	1.33	1.74	0.91	1.66	0.44	0.76	1.09	0.86	0.61	0.32	0.34	0.38	0.59	0.68	0.48
Primary Test TC	9.76	12.33	7.17	4.43	8.28	5.00	6.79	7.87	7.32	2.42	1.89	5.25	2.86	2.45	11.05	2.32
Primary Test NC	2.29	1.45	3.28	0.99	2.96	0.50	0.68	1.35	0.61	0.50	0.32	0.42	0.59	0.68	1.11	0.51
Secondary Test TC	4.25	4.26	4.12	4.14	2.86	2.89	3.91	4.84	4.54	3.17	3.01	5.01	3.20	2.87	5.26	3.68
Secondary Test NC	1.27	2.12	2.59	1.39	2.74	0.36	1.65	2.15	1.17	0.65	0.29	0.92	0.99	1.60	1.50	1.02

Table 3.6: RMSDs in kcal/mol for all of the datasets in the training, primary test, and secondary test sets for ωB97X-V and 15 existing density functionals. The datasets above the first thick black line are in the training set, the datasets between the first and second thick black lines are in the primary test set, while the datasets below the second thick black line are in the secondary test set. Within the training, primary test, and secondary test sets, datasets above the thin black line contain thermochemistry (TC) datapoints, while datasets below the thin black line contain noncovalent interactions (NC) datapoints. The last section of the table contains overall unweighted statistics. For comparison to B3LYP-D2, the All TC RMSD of B3LYP-D3 is 4.7 kcal/mol, while the All NC RMSD is 0.8 kcal/mol.

Property	#	c_x	Year	Class	Rung	Ref.
PBE	0	0	1996	GGA	2	12
PBE-D2	1	0	2006	GGA w/ DT	2	54
B3LYP	3	20	1993	GH GGA	4	33
B3LYP-D2	4	20	2006	GH GGA w/ DT	4	54
B97	10	19	1997	GH GGA	4	35
B97-D2	11	19	2011	GH GGA w/ DT	4	79
B97-D	10	0	2006	GGA w/ DT	2	54
VV10	2	0	2010	GGA w/ NLC	2	57
LC-VV10	3	0-100	2010	RSH GGA w/ NLC	4	57
ωB97X-D	15	22.2-100	2008	RSH GGA w/ DT	4	58
ωB97X-V	10	16.7-100	2014	RSH GGA w/ NLC	4	
M06-L	34	0	2006	meta-GGA	3	23
M06	33	27	2008	GH meta-GGA	4	45
M06-2X	30	54	2008	GH meta-GGA	4	45
M11-L	44	0	2012	meta-GGA	3	24
M11	40	42.8-100	2011	RSH meta-GGA	4	52

Table 3.7: Details for the density functionals from Table 3.6. GH stands for global hybrid, RSH stands for range-separated hybrid, DT stands for dispersion tail, and NLC stands for nonlocal correlation. The column labeled "#" lists the number of parameters that were optimized on a training set for the specific functional. c_x refers to the percentage of exact exchange.

the biggest surprises are the results produced by the 2 newest Minnesota functionals, M11-L and M11. M11-L has an artificial inflection point at the correct equilibrium distance, but binds the dimer at more than 5 Å. On the other hand, M11 binds the dimer very weakly, with an equilibrium bond length that is too long. Since a (500,974) grid was used for computing the local xc functionals, it is unlikely that the strange behavior of the M11-L functional is related to the integration of the xc functional. Since the argon dimer PEC technically influenced the functional form of ωB97X-V, the PEC for the helium dimer is shown on the left in Figure 3.6 as a completely independent test case. The performance of ωB97X-V is superb for this dimer as well. Neither B3LYP nor B3LYP-D2 bind the helium dimer, while M11 binds the dimer with an equilibrium bond length that is too long. PBE, PBE-D2, M06, and M06-2X all overbind the dimer, but predict reasonable equilibrium bond lengths.

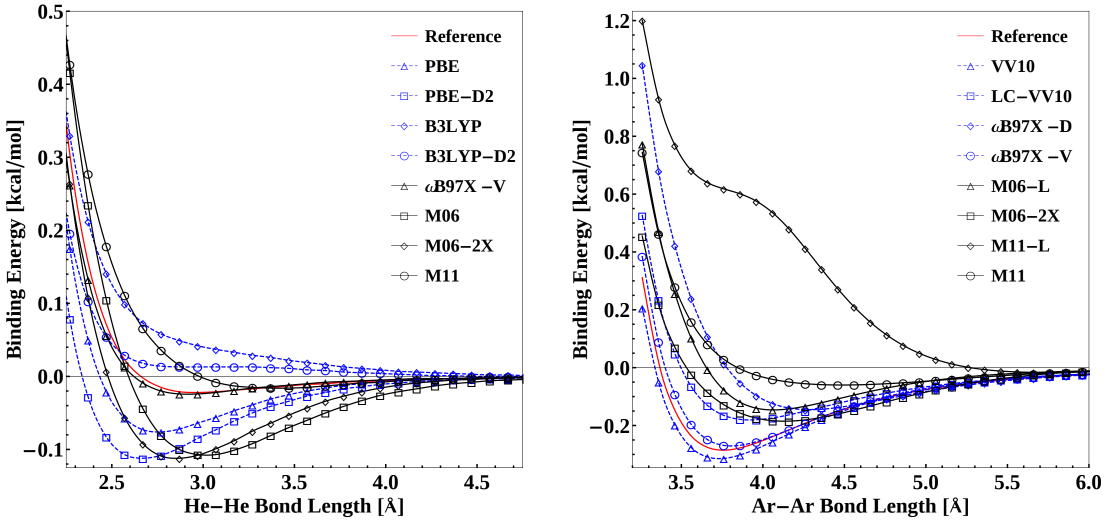


Figure 3.6: Potential energy curves for the helium dimer and the argon dimer. A (500,974) grid was used for computing local xc functionals and a (99,590) grid was used for computing the contribution from the VV10 NLC functional. The calculations were performed with the aug-cc-pVQZ basis set without counterpoise corrections.

3.7.2 Coronene Dimer

The largest intermolecular interactions in the training set for ωB97X-V are the 12 $SO_4^{2-}(H_2O)_5$ isomers from the SW49Rel345 and SW49Bind345 datasets. In the aTZ basis set, these molecules have 694 basis functions. In the primary test set, the largest intermolecular interactions are the 5 adenine-thymine complexes from the S22x5 dataset with 1127 basis functions in the aTZ basis set. In order to assure that ωB97X-V can be successfully applied to even larger interactions, the binding energy of the parallel-displaced coronene dimer was computed in the aDZ and aTZ basis sets (1320 and 2760 basis functions for the dimer, respectively). The binding energy in the aDZ basis set was counterpoise-corrected (CP), while counterpoise corrections were not used (noCP) in the aTZ basis set. The resulting binding energies for the functionals from Table 3.6 are shown in Table 3.8.

While there is no definitive reference value for the binding energy of the parallel-displaced coronene dimer, 2 recent attempts[162, 163] at determining a "CBS" value resulted in binding energies of -20.0 kcal/mol and -24.4 kcal/mol. Furthermore, 2 methods that have been shown to have good performance for dispersion-bound systems (attenuated (aTZ) MP2[164] and MP2.5[165]) give values of -21.3 kcal/mol and -22.8 kcal/mol, respectively. Thus, it is safe to assume that the binding energy of the parallel-displaced coronene dimer is between -20.0 and -25.0 kcal/mol.

The noCP aTZ ωB97X-V binding energy of -22.4 kcal/mol is right in the middle of this range, while the 5 Minnesota functionals underbind the dimer. As expected, the GGA functionals without dispersion corrections fail to bind the dimer (PBE, B3LYP, B97), and even though the DFT-D2 dispersion tail adds -22.4, -31.3, and -22.4 kcal/mol to the binding energy of the parent functionals, respectively, PBE-D2 and B97-D2 still underbind the dimer. However, due to the large s_6 parameter of B3LYP-D2, its noCP aTZ binding energy falls within the aforementioned range. The binding energies of B97-D, VV10, LC-VV10, and ωB97X-D are quite close to that of ωB97X-V, and lie between -20.0 and -25.0 kcal/mol for both the CP aDZ and noCP aTZ calculations. The last column of Table 3.8 lists the difference (kcal/mol) between the CP aDZ and noCP aTZ binding energies for the 16 functionals. It is certainly desirable for this difference to be as small as possible, since the noCP aTZ calculation is at least 10 times more costly than the CP aDZ calculation. For LC-VV10, ωB97X-D, ωB97X-V, M06-L, and M06, this difference is less than 0.60 kcal/mol, while for M11-L and M11, the difference is larger than 2 kcal/mol.

3.7.3 Geometries

To assess the ability of ωB97X-V to calculate accurate geometries, 4 sets of geometries were benchmarked. The first set, A21, includes the 21 equilibrium geometries from the A24 dataset, which were originally optimized (with counterpoise corrections) at a very high level of theory (HF/aQZ + MP2/a(TQ)Z + (CCSD(T)-MP2)/aDZ) with the intention of serving as a benchmark. The DFT benchmark optimizations were performed in the aTZ basis set with a (99,590) grid for local xc functionals and the SG-1 grid for the VV10 NLC functional. The RMSD between each initial (A21) geometry and final (optimized) geometry was calculated with the Kabsch algorithm[166]. Each row in the second column of Table 3.9 contains the RMSD of the 21 RMSDs generated by the Kabsch algorithm. The RMSD of ωB97X-V for these 21 interactions is an outstanding 0.58 pm. The RMSD of the next best functional, B97-D2, is 3 times larger than that of ωB97X-V. Furthermore, the ωB97X-V result is more than 5 times better than that of ωB97X-D, and more than 7 times better than that of the Minnesota functionals. The next dataset includes the equilibrium bond lengths of 6 rare-gas dimers: He_2, HeNe, HeAr, Ne_2, NeAr, and Ar_2. The reference values were taken from the Tang-Toennies potential model[96], and the optimizations were performed in the aTZ basis set with a (500,974) grid for local xc functionals and a (75,302) grid for the VV10 NLC functional. ωB97X-V has

Functional	CP aDZ	noCP aTZ	Δ
PBE	5.26	4.63	-0.64
PBE-D2	-17.12	-17.76	-0.64
B3LYP	11.90	10.52	-1.38
B3LYP-D2	-19.43	-20.81	-1.38
B97	8.22	7.33	-0.89
B97-D2	-14.17	-15.06	-0.89
B97-D	-21.42	-22.36	-0.94
VV10	-20.95	-22.07	-1.12
LC-VV10	-23.75	-23.27	0.48
ωB97X-D	-24.36	-24.36	-0.01
ωB97X-V	-21.97	-22.40	-0.43
M06-L	-17.10	-17.68	-0.58
M06	-14.72	-14.62	0.11
M06-2X	-18.22	-17.47	0.75
M11-L	-14.49	-17.64	-3.15
M11	-17.81	-15.71	2.09

Table 3.8: Binding energies in kcal/mol for the parallel-displaced coronene dimer for the density functionals from Table 3.6. noCP means non-counterpoise-corrected while CP means counterpoise-corrected. In the aDZ basis set, the dimer has 1320 basis functions, while it has 2760 basis functions in the aTZ basis set. The column labeled "Δ" contains the difference between the CP aDZ and noCP aTZ binding energies for each functional.

the best performance, with an RMSD of 7.91 pm, while VV10 follows closely behind at 8.07 pm. From the Minnesota functionals, M06-2X and M06-L perform well, while M11-L has an RMSD of almost 1 Å. From the dispersion-corrected functionals, B97-D2, B97-D, and PBE-D2 perform decently, while ωB97X-D and B3LYP-D2 perform poorly. The third set of geometries, taken from the work of Tentscher and Arey[167], contains 18 bond lengths of 18 small radicals. For this set of geometries, M06-L performs the best, followed by B97-D2 and B3LYP-D2. In general, the bond length RMSDs are very small, ranging from 0.69 pm to 2.32 pm. ωB97X-V performs decently, with an RMSD of 1.15 pm. The last set of geometries, taken from Bak et al.[168], contains 28 bond lengths of 19 small molecules. The reference geometries were computed at the CCSD(T)/cc-pCVQZ level of theory and validated against experimental results. For this set of geometries, M06-L, B3LYP-D2, and B97-D2 perform very well, followed by B97-D and ωB97X-V. For the last 2 datasets, the optimizations were carried out in the aTZ basis set with a (75,302) grid for local xc functionals and SG-1 for the VV10 NLC functional.

RMSD [pm]	A21	Rare-gas	Arey	Bak
PBE-D2	8.33	20.94	1.30	1.20
B3LYP-D2	5.95	84.48	0.82	0.61
B97-D2	1.78	15.19	0.81	0.66
B97-D	5.38	18.55	1.10	0.92
VV10	7.38	8.07	1.43	1.39
LC-VV10	2.93	13.05	1.82	1.74
ωB97X-D	3.34	36.34	1.28	1.11
ωB97X-V	0.58	7.91	1.15	0.98
M06-L	7.02	14.06	0.69	0.60
M06	4.22	34.03	1.46	1.27
M06-2X	4.90	12.65	1.67	1.38
M11-L	4.69	96.02	2.32	2.31
M11	5.00	21.10	1.68	1.36

Table 3.9: RMSDs from the 4 geometry datasets discussed in Section 3.7.3. For the latter 3 datasets, the entries are bond length RMSDs. For the A21 dataset, an RMSD was calculated for each molecule using the Kabsch algorithm. Each row in the second column contains the RMSD of the 21 RMSDs generated by the Kabsch algorithm.

3.8 Using ωB97X-V

Even though ωB97X-V was trained without counterpoise corrections in the aTZ basis set for noncovalent interactions and without counterpoise corrections in the aQZ basis set for thermochemistry, it is inevitable that the functional will be used with different basis sets. As a result, this section explains how the functional should be used and what basis sets are recommended. Ideally, calculations with ωB97X-V should be run as close as possible to the basis set limit. As an example, the binding energy of the water dimer (NCCE31 dataset geometry) was computed with and without counterpoise corrections with the aug-cc-pVXZ (X = D, T, Q, 5) [aXZ] family of basis sets. The resulting counterpoise-corrected binding energies (CP) in kcal/mol are $\{-4.96, -4.98, -5.01, -5.00\}$, while the resulting non-counterpoise-corrected binding energies (noCP) in kcal/mol are $\{-5.18, -5.03, -5.04, -5.01\}$. Since ωB97X-V was trained without counterpoise corrections in the aTZ basis set for noncovalent interactions, the corresponding value of -5.03 kcal/mol is the "best" value. However, the noCP aQZ and noCP a5Z binding energies differ from the noCP aTZ binding energy by 0.02 kcal/mol at most. Thus, using basis sets that are larger and contain higher angular momentum functions than the ones used for training will not degrade the performance of the functional. Regarding the aDZ basis set, it is clear that this basis set should not be used with ωB97X-V without counterpoise corrections, since the basis set superposition error (BSSE) at the aDZ basis set level is larger than 0.20 kcal/mol for the water dimer. However, since the difference between the noCP aTZ and CP aDZ binding energies for the water dimer is only 0.06 kcal/mol, the aDZ basis set can be used with counterpoise corrections if necessary. Furthermore, Table 3.8 indicates that the difference between the noCP aTZ and CP aDZ binding energies for the parallel-displaced coronene dimer is less than 0.50 kcal/mol. Thus, if a noCP aTZ calculation is impractical, a CP aDZ calculation will suffice.

Similar tests were performed on 3 dimers from the NCCE31 dataset (water, hydrogen fluoride, and ammonia) with a variety of basis sets from the EMSL basis set exchange[169, 170], and the results are summarized in Table 3.10. If the MG3S[171] or LP[172, 173] basis sets are desired, we highly recommend that counterpoise corrections be utilized, since ωB97X-V produced poor results when these basis sets were employed without counterpoise corrections. For example, the binding energy for the water dimer in the MG3S basis set is -4.99 kcal/mol with counterpoise corrections and -5.28 kcal/mol without. Compared to the CCSD(T)/CBS estimate of approximately -5.00 kcal/mol, it is clear that the CP MG3S value is a much better result than the noCP MG3S value. Dunning's augmented basis sets were previously analyzed for the water dimer, and the recommendation still holds: noCP aTZ, noCP aQZ, and noCP a5Z are highly recommended, while the aDZ basis set should only be used with counterpoise corrections. If the aug-pc-X (X = 0, 1, 2, 3, 4) [acpX] basis sets[174–176] are desired, we recommend aug-pc-3 and aug-pc-4 without counterpoise corrections. aug-pc-0 should not be used and if aug-pc-1 and aug-pc-2 are desired, they should be used with counterpoise corrections. If the def2-X(Z)VP(P)D (X = S, T, Q) basis sets of Rappoport and Furche[146] are desired, we recommend using the def2-TZVPPD, def2-QZVPD, and def2-QZVPPD basis sets without counterpoise corrections. If calculations with the def2-TZVPD basis set are required, counterpoise corrections should be utilized. In addition, the def2-SVPD basis set should not be used with ωB97X-V. The last family of basis sets that was tested was Truhlar's minimally augmented basis set series[177], maug-cc-pVXZ (X = D, T, Q) [maXZ]. Results with these basis sets indicated that only the maQZ basis set can be recommended with counterpoise corrections, while the maDZ and maTZ cannot be recommended for use with ωB97X-V.

Another looming question in the realm of DFT calculations is the proper choice for the integration grid. While ωB97X-V was trained with the (99,590) grid for the local exchange-correlation functional and the SG-1 grid for the nonlocal correlation functional, several tests were performed in order to quantify the extent to which the binding energies of weakly interacting systems depend on both the local and nonlocal integration grids. The tests were performed by combining 5 local grids {SG-

kcal/mol	$(H_2O)_2$			$(HF)_2$			$(NH_3)_2$		
Basis	BSSE	CP	noCP	BSSE	CP	noCP	BSSE	CP	noCP
MG3S	0.40	-5.03	-5.43	0.39	-4.56	-4.95	0.17	-3.18	-3.35
LP	0.29	-4.99	-5.28	0.38	-4.60	-4.98	0.10	-3.10	-3.21
aDZ	0.22	-4.96	-5.18	0.20	-4.54	-4.74	0.39	-3.06	-3.45
aTZ	0.05	-4.98	-5.03	0.07	-4.58	-4.65	0.03	-3.09	-3.12
aQZ	0.03	-5.01	-5.04	0.03	-4.64	-4.67	0.02	-3.10	-3.12
a5Z	0.01	-5.00	-5.01	0.01	-4.64	-4.64	0.00	-3.10	-3.10
apc0	1.48	-7.76	-9.24	0.85	-6.17	-7.02	1.29	-4.71	-5.99
apc1	0.65	-5.08	-5.73	0.63	-4.67	-5.30	0.53	-3.10	-3.63
apc2	0.06	-5.02	-5.08	0.10	-4.64	-4.74	0.05	-3.11	-3.15
apc3	0.00	-5.01	-5.01	0.00	-4.63	-4.64	0.00	-3.10	-3.10
apc4	0.00	-5.01	-5.01	0.00	-4.64	-4.64	0.00	-3.10	-3.10
SVPD	0.41	-5.13	-5.54	0.41	-4.70	-5.10	0.74	-3.07	-3.81
TZVPD	0.07	-5.00	-5.06	0.05	-4.55	-4.60	0.13	-3.11	-3.23
TZVPPD	0.03	-5.00	-5.03	0.03	-4.60	-4.63	0.09	-3.08	-3.17
QZVPD	0.01	-5.00	-5.01	0.01	-4.63	-4.64	0.02	-3.10	-3.11
QZVPPD	0.01	-5.00	-5.01	0.01	-4.63	-4.64	0.02	-3.10	-3.11
maDZ	0.51	-5.28	-5.79	0.37	-4.54	-4.91	0.35	-3.36	-3.71
maTZ	0.16	-5.05	-5.21	0.09	-4.53	-4.62	0.13	-3.19	-3.32
maQZ	0.07	-5.03	-5.11	0.03	-4.62	-4.66	0.06	-3.13	-3.19

Table 3.10: Assessment of the basis set dependence of ωB97X-V for the binding energies of 3 dimers. Columns labeled BSSE contain the basis set superposition errors for the molecule/basis set combination indicated, columns labeled noCP contain non-counterpoise-corrected binding energies, while columns labeled CP contain counterpoise-corrected binding energies.

1, (75,302), (99,590), (250,590), (500,974)} with 3 nonlocal grids {SG-1, (75,302), (99,590)}, for a total of fifteen distinct local and nonlocal grid combinations. Potential energy curves (PEC) for 2 families of dimers, as well as the water dimer, were computed with these fifteen grid combinations. The first family of dimers was the rare-gas dimers from neon to krypton. The second family of dimers was CH_4—X, where X = F_2, Cl_2, and Br_2 (X40 dataset geometries). The def2-TZVPPD basis set was used for all of the calculations in this test. The PECs each had a total of 51 points, starting from a distance of 1 Å shorter than the equilibrium bond length, and continuing to 4 Å longer than the equilibrium bond length in increments of 0.1 Å. Thus, a total of 105 PECs were computed, requiring 5520 single-point calculations. For the first family of PECs, the ωB97X-V equilibrium binding energies in kcal/mol are $\{-0.13, -0.32, -0.42\}$, compared to the Tang-Toennies estimates of $\{-0.08, -0.28, -0.40\}$. For the second family of PECs, the ωB97X-V equilibrium binding energies in kcal/mol are $\{-0.48, -1.07, -1.31\}$, compared to the X40 estimates of $\{-0.49, -1.08, -1.30\}$. All of these interactions were assessed by calculating the percent error for each point on the fifteen PECs with respect to the (500,974)/(99,590) value, and computing the RMSDs of the 51 percent errors. Percent error RMSDs of approximately 1% or less were deemed acceptable.

Based on the data from Table 3.11, a minimum grid combination of (250,590)/(75,302) must be used for the neon and argon dimers. The krypton dimer calculations appear to be less sensitive to the nonlocal grid, and the (250,590)/SG-1 grid combination gives acceptable results. For the CH_4—X interactions, it appears as if the (99,590)/SG-1 grid combination gives unanimously acceptable results. Finally, for the water dimer (which binds at about -5.00 kcal/mol at equilibrium), using the SG-1 grid for both the local and nonlocal components is sufficient. Based on the results from Table 3.11, the following grids are recommended: (250,590)/(75,302) for interactions weaker than -0.5 kcal/mol at equilibrium, (99,590)/SG-1 for interactions between -0.5 kcal/mol and -2.5 kcal/mol at equilibrium, and (75,302)/SG-1 for interactions stronger than -2.5 kcal/mol at equilibrium.

As a reference value for those interested in implementing this new density functional, the absolute energy (in hartrees) of hydrogen fluoride (HF) with a bond length of 0.9158 Å in the aug-cc-pVTZ basis set with the (75,302)/SG-1 local/nonlocal grid combination is -100.4512112969.

3.9 Conclusions

The primary goal of the development of the ωB97X-V density functional was to create a minimally-parameterized and highly-transferable density functional that could predict accurate energetics for both bonded and non-bonded interactions. With respect to thermochemistry, the performance of ωB97X-V is equivalent to that of ωB97X-D, despite the fact that ωB97X-V has 5 less empirical parameters. Furthermore, the performance of ωB97X-V on noncovalent interactions is considerably better than that of all the functionals tested in this chapter. Table 3.12 ranks the 16 benchmarked density functionals with respect to their overall RMSDs for all of the bonded (1002) and non-bonded (1484) interactions considered in this chapter.

To summarize the main results of this chapter:

(1). We have optimized a new 10-parameter, semi-empirical density functional, ωB97X-V, that is a range-separated hybrid functional based on the B97 GGA model for local exchange and correlation, augmented with nonlocal correlation using the VV10 NLC functional.

(2). A novel feature of the training procedure is that over 16000 candidate functionals were trained and considered. By assessing their performance on both training and test set data, it was determined that 7 linear parameters yielded the most trans-

Ne_2	SG-1	(75,302)	(99,590)	(250,590)	(500,974)
SG-1	1539.82	54.43	34.67	18.67	18.46
(75,302)	1555.95	51.85	21.53	0.82	0.73
(99,590)	1556.16	52.14	21.12	0.50	0.00
Ar_2	SG-1	(75,302)	(99,590)	(250,590)	(500,974)
SG-1	575.68	18.18	8.16	6.53	6.62
(75,302)	581.11	18.22	4.13	0.20	0.23
(99,590)	581.22	18.22	4.01	0.16	0.00
Kr_2	SG-1	(75,302)	(99,590)	(250,590)	(500,974)
SG-1	58.36	10.15	2.90	0.89	0.83
(75,302)	58.75	9.67	2.64	0.33	0.11
(99,590)	58.73	9.70	2.65	0.31	0.00
CH_4F_2	SG-1	(75,302)	(99,590)	(250,590)	(500,974)
SG-1	2.67	0.27	0.14	0.10	0.10
(75,302)	2.67	0.28	0.09	0.05	0.05
(99,590)	2.69	0.28	0.09	0.01	0.00
CH_4Cl_2	SG-1	(75,302)	(99,590)	(250,590)	(500,974)
SG-1	11.52	0.36	0.17	0.15	0.15
(75,302)	11.59	0.37	0.10	0.05	0.05
(99,590)	11.60	0.34	0.09	0.01	0.00
CH_4Br_2	SG-1	(75,302)	(99,590)	(250,590)	(500,974)
SG-1	10.47	1.04	0.18	0.19	0.14
(75,302)	10.54	1.02	0.12	0.13	0.05
(99,590)	10.54	1.01	0.13	0.12	0.00
$(H_2O)_2$	SG-1	(75,302)	(99,590)	(250,590)	(500,974)
SG-1	0.15	0.01	0.01	0.01	0.01
(75,302)	0.15	0.01	0.01	0.01	0.00
(99,590)	0.14	0.01	0.01	0.01	0.00

Table 3.11: Assessment of the grid dependence of ωB97X-V for the potential energy curves of 7 dimers. The numbers shown are RMSDs of 51 percent errors. The percent errors are calculated by assuming that the (500,974)/(99,590) values are exact.

Functional	All TC	Functional	All NC
M06-2X	3.21	ωB97X-V	0.32
ωB97X-V	3.60	M06-L	0.47
ωB97X-D	3.61	B97-D2	0.48
B97-D2	3.97	ωB97X-D	0.54
M11	3.97	M11	0.56
M06	4.18	M06	0.60
B3LYP-D2	4.59	LC-VV10	0.73
B97	4.74	M06-2X	0.78
B3LYP	5.11	B97-D	0.82
B97-D	5.56	B3LYP-D2	1.01
M06-L	5.63	M11-L	1.07
M11-L	6.68	VV10	1.38
LC-VV10	6.79	PBE-D2	1.49
VV10	9.81	PBE	2.00
PBE	9.90	B97	2.71
PBE-D2	10.35	B3LYP	2.96

Table 3.12: Density functionals ranked based on their overall unweighted RMSDs in kcal/mol for all thermochemistry (Columns 1-2) and noncovalent interactions (Columns 3-4) datapoints considered in this chapter.

ferable functional. Increasing the number of linear parameters past 7 degraded test set performance and minimally improved training set performance.

(3). Detailed assessment against 15 existing density functionals on main group thermochemistry and noncovalent interactions suggests that ωB97X-V is the best functional tested for non-bonded interactions by a significant margin. Its performance for thermochemistry is also very good, virtually as good as the best hybrid meta-GGA tested (M06-2X).

(4). ωB97X-V can be recommended for application to a wide range of molecular bonded and non-bonded interactions involving the lighter main group elements. It will be interesting to await further assessment on larger classes of problems in order to characterize its limitations. Such limitations are likely to arise for problems where strong correlation effects are in play.

(5). It is desirable to apply the same training approach used here to develop other semi-empirical density functionals with slightly different physical content (e.g. meta-GGAs), so that the resulting functionals are likewise optimally transferable. We intend to report such developments in due course.

Chapter 4

B97M-V: A meta-GGA density functional

4.1 Introduction

4.1.1 Perspective

Kohn-Sham density functional theory (KS-DFT)[3, 178] is perhaps the most tractable approach to describing the electronic structure of molecules and solids in their ground state. The intractable, exact many-body wave function is replaced by a single determinant of spin orbitals, $\{\phi_i\}$, which describe a reference system of non-interacting electrons, whose density, $\rho(\mathbf{r})$, is to be the same as that of the interacting system of electrons under study. The single determinant of Kohn-Sham molecular orbitals is then used to represent the density, $\rho(\mathbf{r}) = \sum_i |\phi_i|^2$, as well as to evaluate the non-interacting kinetic energy.

The total energy, E_{tot}, is assembled from terms that can be exactly evaluated, and a remaining unknown term, E_{xc}, which is (in its simplest form) taken to be a functional of the electron density:

$$\begin{aligned} E_{tot} = &-\sum_i \frac{1}{2} \left\langle \phi_i \left| \nabla^2 \right| \phi_i \right\rangle + \int \rho(\mathbf{r}) v_{ext}(\mathbf{r}) d\mathbf{r} \\ &+\frac{1}{2} \int \int \frac{\rho(\mathbf{r}_1)\rho(\mathbf{r}_2)}{r_{12}} d\mathbf{r}_1 d\mathbf{r}_2 + E_{nuc} + E_{xc} \end{aligned} \tag{4.1}$$

The task of density functional development is to approximate E_{xc} as accurately as possible, while maintaining the computational tractability that characterizes the preceding 4 terms.

Since density functionals are models, there is no guarantee that a complicated model will be more accurate than a simple one. Nonetheless, if additional, physically-relevant information is incorporated into a density functional with minimal empiricism, it is certainly possible for the resulting density functional to improve upon a non-empirical density functional that excludes such information. This is the basis for systematically exploring semi-empirical density functionals that are increasingly complex functionals of the electron density. Following Perdew[4], who assigned density functionals to various rungs of a ladder ascending from simplest to more complex (with more complex forms offering the possibility of satisfying additional exact conditions), one commonly identifies the following 5 rungs:

1. Local Spin-Density Approximation (LSDA): The LSDA exactly solves the infinite[179] uniform electron gas (UEG) problem and represents a non-empirical density functional that is the starting point for most of the following refinements. The exchange energy and correlation energies are expressed in terms of an integral over the energy density per unit volume, $E_{x/c}^{LSDA} = \int e_{x/c}^{UEG}(\rho(\mathbf{r}))d\mathbf{r}$.

2. Generalized Gradient Approximation (GGA): Density functionals of this type enhance the LSDA energy density with an inhomogeneity correction factor (ICF), $g(s(\mathbf{r}))$, that depends on the reduced spin-density gradient, $s(\mathbf{r}) = \frac{\nabla\rho(\mathbf{r})}{\rho(\mathbf{r})^{4/3}}$, as in $E_{x/c}^{GGA} = \int e_{x/c}^{UEG}(\rho(\mathbf{r}))g_{x/c}(s(\mathbf{r}))d\mathbf{r}$. Leading GGAs include the non-empirical PBE density functional[12], as well as the 9-parameter dispersion-corrected B97-D density functional[54].

3. meta-Generalized Gradient Approximation (meta-GGA): In the early 1970s, while working on modeling nuclear matter forces, Negele and Vautherin[180] proposed the density matrix expansion (DME) in order to derive an approximate expression for the nuclear density matrix in terms of the nuclear density, its gradient, its Laplacian, and the nuclear kinetic energy density. More than 10 years later, Becke, using a similar approach, derived an approximate expression for the exchange charge density in terms of the electron density, its gradient, its Laplacian, as well as the electron kinetic energy density ($\tau = \sum_{i}^{occ.} |\nabla\phi_i|^2$).

   Since then, the term "meta-GGA" or "mGGA"[16] has come to signify density functionals that contain exchange or correlation components that depend on the kinetic energy density. As the central topic of this chapter, existing meta-GGA density functionals will be reviewed in detail in Section 4.1.2.

4. Hybrid density functionals: Originally advocated by Becke[33, 181], hybrid density functionals evaluate a fraction of the exchange energy using the KS determinant as a wave function, leading to an explicit dependence upon the occupied KS orbitals. Such density functionals, beginning with B3PW91[10, 11, 33], significantly improve upon Rung 2 and 3 density functionals for both thermochemical energy differences and reaction barriers, although the cost of evaluating exact exchange is a significant computational burden over local GGAs and meta-GGAs.

5. Double-hybrid density functionals (DHDF): DHDFs include dependence on the unoccupied KS orbitals in addition to the occupied levels, via either a second-order perturbation-like expression, or methods based on the random phase approximation. Such density functionals have demonstrated very high accuracy, albeit with significant additional computational demands beyond hybrids.

Even at Rung 2, and particularly at Rung 3 and higher, there is no unique way of designing non-empirical density functionals. Accordingly, semi-empirically-designed density functionals can compete effectively with those constructed from first principles. Recently, the systematic development of semi-empirical density functionals built upon the ICF of the global hybrid GGA density functional, B97, has been explored[62]. B97 expresses the ICFs for exchange, same-spin correlation, and opposite-spin correlation (g_x, g_{css}, and g_{cos}) as power series in a finite-range dimensionless variable, u, which in turn depends on the semi-infinite-range reduced spin-density gradient, s. Thus, each of the 3 ICFs can be expressed as:

$$g(u) = \sum_{j=0}^{M} c_j u^j \tag{4.2}$$

The aforementioned study first tested the comparative efficacy of different treatments of nonlocal exchange (none vs. global hybrid vs. range-separated hybrid) and nonlocal correlation (none vs. DFT-D2 vs. the VV10 nonlocal correlation (NLC) functional). Within the GGA framework, all of the functional forms from each of the 9 categories were assessed by considering all possible combinations of the approximately 15 parameters. The functional form which trained and tested best across 2301 data points was deemed “optimal” in each category.

The overall winner from the 9 resulting “optimal” functional forms combined range-separation for nonlocal exchange with the VV10 NLC functional for the treatment of long-range dispersion interactions and was subsequently self-consistently trained. The result was a density functional with 7 linear parameters and 3 nonlinear parameters called ωB97X-V. For thermochemistry, ωB97X-V performed similarly to ωB97X-D, which has 13 linear parameters and 2 nonlinear parameters. However, it performed significantly more accurately for non-covalent interactions (NC).

The question naturally arises as to whether a similar approach can be applied to the design of a Rung 3 density functional that uses the kinetic energy density as an additional variable. The primary goal of this chapter is to answer this question. Since meta-GGAs are inherently more flexible than GGAs, the systematic exploration of the space of possible functional forms is considerably more difficult. Consequently, the challenge of systematic design is exponentially greater, because each of the one-dimensional power series expansions of Equation 4.2 must be replaced by two-dimensional expansions, where the new variable, w, depends on the kinetic energy density:

$$g(w,u)=\sum_{i=0}^{M'}\sum_{j=0}^{M}c_{ij}w^{i}u^{j} \tag{4.3}$$

In order to proceed, the values of M and M' must be determined. Historically, most GGAs based on the B97 model have had success with values of M between 2 and 4. For the purposes of this study, M will be set to 4 and M' will be set to 8. With these choices ($M = 4$ and $M' = 2M = 8$), the search problem is of an enormous scale. The number of variables that can be individually included or excluded for each of the 3 components of the exchange-correlation energy is 44, assuming the UEG limit is enforced ($c_{00} = 1$). Under these conditions, the total number of possible functional forms is an astounding $2^{3\times 44} = 2^{132} \sim 10^{40}$. The dimension of this space (10^{16} moles) is so enormous, that it is appropriate to label it the meta-GGA density functional genome.

An exhaustive search of the entire meta-GGA genome for the functional form which demonstrates the highest degree of accuracy on training and test data is clearly impossible. Due to the scale of the problem, this functional search approach has not been previously extensively tried as a means of constructing a meta-GGA density functional (though a step in this direction has been recently reported for a meta-GGA exchange functional[182]). Nonetheless, this approach should be the goal of rational, semi-empirical density functional design, and an intelligent search of an interesting subspace can certainly be attempted. After reviewing the large number of existing meta-GGA density functionals and briefly summarizing dispersion corrections, the approach to attacking the search problem to design a new local meta-GGA density functional will be described.

4.1.2 meta-GGA Density Functionals

The first τ-dependent density functional was a same-spin correlation functional developed by Becke in 1985[183]. Several years later, Becke developed a more advanced, 2-parameter correlation functional (Bc88)[110]. Bc88 had a τ-dependent same-spin correlation component due to a multiplicative factor (later termed the self-correlation correction (SCC) factor) that gave zero correlation energy for one-electron systems. Becke's first τ-dependent exchange functional, BR89, was developed in conjunction with Roussel[109], and was modeled after the exchange hole of the hydrogen atom. Inspired by the SCC factor of the Bc88 correlation functional, Becke proposed a simpler, 2-parameter correlation functional (Bc95)[184] in 1995. Following the success of B97[35], Becke developed the 10-parameter global hybrid meta-GGA density functional, B98[15]. Shortly after, Schmider and Becke[76, 185] parameterized a series of B98-type density functionals on a variety of datasets in order to assess the sensitivity of the resulting empirical parameters to the training set. The inhomogeneity variable of B98 was slightly modified in 2000 with the B00 density functional[111] to give the w variable that would later become the backbone of most of Truhlar's exchange functionals.

Since the early 1990s, Proynov, Salahub, and coworkers have developed a series of τ-dependent correlation functionals, starting with LAP1 and LAP2[186], which were based on the opposite-spin Colle-Salvetti pair-correlation function[187]. Soon thereafter, the LAP3 correlation functional[188] was developed in order to account for same-spin correlation. Its 4 parameters were fit in conjunction with an exchange functional (B88 for BLAP3 and PW86 for PLAP3) to a set of binding energies and bond lengths. The successor to LAP3, the $\tau 1$ correlation functional[189], included higher-order τ-dependent terms, and its 5 parameters were fit in combination with a modified B88 exchange functional to give Bmτ1. Pairings of LAP3 and $\tau 1$ with the OPTX exchange functional[190], OLAP3 and Oτ1, were later tested[191], and Oτ2 was developed by refitting the parameters of both OPTX and $\tau 1$. Additionally[192, 193], TPSSτ1 assessed the pairing of TPSS exchange and $\tau 1$ correlation, while TPSSτ3 involved refitting the parameters of $\tau 1$ with TPSS exchange.

Koehl, Odom, and Scuseria (KOS) used the DME of Negele and Vautherin (NV)[180] to construct an exchange functional[194], employing a more general coordinate system than NV. KOS were able to decrease the mean absolute deviation (MAD) of the exchange energies of 32 molecules more than 100-fold with a 2-parameter expression, giving the VT exchange functional. Van Voorhis and Scuseria (VS) showed that the exchange hole of VT diverged in the asymptotic limit and set out to ameliorate this problem[195], as well as to add explicit dependence on the gradient of the density. The resulting 4-parameter GMVT and KMVT exchange functionals reduced the MAD of VT by another factor of 3 (to around 0.008 E_h). With 2 accurate exchange functionals, VS set out to develop a density functional[27] based on the DME. The resulting local meta-GGA density functional, VS98 (also known as VSXC or GVT4), had a total of 21 empirical parameters that were trained on atomization energies, ionization potentials, and bond lengths.

With only 1 empirical parameter, PKZB[16] was Perdew's first attempt at constructing a local meta-GGA density functional. The exchange ICF of PKZB was an extension of that of the PBE exchange functional, albeit with a more complicated inhomogeneity variable that included kinetic energy density dependence, allowing for further constraint satisfaction. The PKZB correlation functional was similar to the PBE correlation functional, but was modified (with the help of τ) to return zero for one-electron densities (like Becke's SCC factor). PKZB was accurate for surface energies and atomization energies, but less satisfactory for hydrogen-bonded complexes and equilibrium bond lengths[196]. These problems were eventually attributed to shortcomings in the PKZB exchange functional, which was modified (along with minor improvements to the PKZB correlation functional) to yield the well-known, non-empirical local meta-GGA density functional, TPSS[17]. A 1-parameter global hybrid version of TPSS (TPSSh[38]) was developed shortly after, with 10% exact exchange. To improve TPSS atomization energies, modTPSS[197], a 1-parameter version of TPSS which adjusted a previously fixed parameter in the exchange functional, was proposed. In 2009, a revised TPSS (revTPSS[18]) was introduced to overcome shortcomings in predicting accurate lattice constants, as well as to improve additional formal properties. The revTPSS exchange functional was altered to remedy an order of limits problem and paired with a modified PBE correlation functional to define the regTPSS density functional[198]. In 2012, Sun and coworkers[19] investigated the effect of the kinetic energy density on meta-GGA density functionals, which (along with the development of regTPSS) inspired the development of 4 local and 2 global hybrid meta-GGA density functionals[19–21, 199, 200]: MGGA_MS (MGGA_MS0), MGGA_MS1, MGGA_MS2, MGGA_MS2h, MVS, and MVSh.

Other groups were also active in proposing new meta-GGA density functionals in the early 2000s. The EDMGGA exchange functional[201] was based on the DME of the exchange hole and had an ICF that resembled that of B88, while using an inhomogeneity variable which depended (in part) on the kinetic energy density. A global hybrid variant[202] which combined EDMGGA with the Colle-Salvetti meta-GGA correlation functional[14, 187] was also proposed, with 22% exact exchange. KCIS[203] was a meta-GGA correlation functional which depended on τ through its SCC term. KCIS was paired with various exchange functionals, including B88, PBE, and PKZB, and various global hybrid variants were also proposed, such as

B0KCIS (25% exact exchange) and B1KCIS (23.9% exact exchange). Boese and Handy[28] based τ-HCTH on Becke's B97, B98, and B00 density functionals. The same-spin and opposite-spin correlation ICFs used B97 expansions with $m = 3$. The exchange functional had both a "local" exchange component (B97-type with $m = 3$) and a "nonlocal" exchange component (Becke's 3-term B00 ICF[111] multiplying an $m = 3$ B97-type GGA ICF). τ-HCTH had 16 fitted parameters, while its global hybrid counterpart had 17 (with 15% exact exchange). BMK[36] was identical in form to the hybridized τ-HCTH, but since transition states were included in the training set (to better describe barrier heights), its percentage of exact exchange was almost 3 times larger (42%).

Between 2005 and 2011, Truhlar and co-workers have published 10 highly-parameterized local (M06-L[23], M11-L[24]), global hybrid (M05[41], M05-2X[43], M06[45], M06-2X[45], M06-HF[44], M08-HX[46], M08-SO[46]), and range-separated hybrid (M11[52]) meta-GGA density functionals. The exchange component of M05 used the PBE exchange functional as its foundation, enhanced by a 12-term ($m = 11$) power series ICF in Becke's τ-dependent variable[111], w, while the M05 correlation functional employed a 5-term B97-type expansion with Becke's SCC factor multiplying the same-spin component. All 3 UEG limits were satisfied, and a total of 22 fitted parameters (20 linear and 2 nonlinear) were optimized, including 28% exact exchange. The construction of M05-2X was identical to that of M05, except the exact exchange mixing parameter was doubled and fixed (56%) and the 2 nonlinear correlation parameters were borrowed from M05, for a total of 19 fitted parameters. The form of M06-L was an unhybridized M05 or M05-2X, with 1 exception: all 3 components (exchange, same-spin correlation, and opposite spin correlation) had an additional, 5-term, VS98-type ICF. Since the nonlinear correlation parameters were borrowed from M05, M06-L ended up with a total of 34 fitted parameters. The global hybrid M06 density functional was identical to a hybridized M06-L, with 1 exception: the fourth-order terms from the VS98-type exchange ICF were dropped (2 fewer parameters). Accordingly, M06 had a total of 33 fitted parameters, with 27% exact exchange. M06-2X had double the amount of exact exchange (54%) as M06, and the entire VS98-type exchange ICF was dropped, resulting in 29 fitted parameters. M06-HF employed 100% exact exchange, and was otherwise identical to M06, with a total of 32 fitted parameters. M08-HX and M08-SO used exchange functionals based on the summation of the PBE and RPBE[112] ICFs, each enhanced by 12-term power series ICFs in w. For correlation, a third 12-term power series ICF in w enhanced the PW92 correlation energy density per unit volume and a fourth enhanced the PBE correlation functional gradient correction term. M08-HX involved 47 fitted parameters and 52.23% exact exchange, while for M08-SO, 3 additional constraints led to 44 fitted parameters and 56.79% exact exchange. The M11 density functional extended M08-HX and M08-SO with range-separated exact exchange and several minor changes, leading to 40 fitted parameters, 42.8% short-range exact exchange, and $\omega = 0.25$. The local meta-GGA density functional, M11-L, used a novel "dual-range" partitioning of the exchange functional, with a local "long-range" exchange component in addition to the short-range exchange component found in range-separated hybrid exchange functionals. The correlation functional was similar to that of M11, and all 6 power series (2 from short-range exchange, 2 from long-range exchange, and 2 from correlation) were reduced to 9 terms. With 54 initial linear parameters, 1 nonlinear parameter ($\omega = 0.25$), and 11 constraints, M11-L had a total of 44 fitted parameters.

Development of meta-GGAs continues actively to the present day. The meta-VT{8,4} exchange functional[204] was based on revTPSS, but differed in its inhomogeneity variable as well as its ICF, and was paired with the revTPSS correlation functional. ωM05-D was a dispersion-corrected, range-separated hybrid version[59] of Truhlar's M05 density functional with 21 fitted parameters, while ωM06-D3 represented similar modifications to Truhlar's M06-2X density functional and had 25 fitted parameters. These density functionals exhibited the advantages of range-separation and a correct treatment of long-range dispersion interactions. The BLOC exchange functional[205] was constructed by converting a constant in the TPSS and revTPSS exchange functionals to a function of the ratio of the von Weizsäcker kinetic energy density to the exact kinetic energy density. It was paired with the TPSSloc correlation functional[206] to give the BLOC density functional. Two density functionals have recently been developed within a Bayesian error estimation functional framework. BEEF-vdW[207] involved developing a 30-parameter GGA exchange functional and pairing it with 40% PBE correlation and 60% PW92 correlation, along with the vdW-DF-10 NLC functional[74]. This approach was naturally extended[182] to a 64-parameter meta-GGA exchange functional (mBEEF), which added the kinetic energy density as a variable, dropped the nonlocal correlation functional, and replaced the combination of PBE correlation and PW92 correlation with the PBEsol correlation functional[208].

4.1.3 van der Waals Interactions

Since long-range electron correlations that account for the asymptotic $1/r^6$ dependence of van der Waals (vdW) interactions cannot be properly described by conventional density functionals[124, 125], there has been an increased effort in the past decade to remedy this shortcoming. While numerous methods that account for dispersion have been proposed, this brief review is limited to the DFT-D approach of Grimme[53–55] and the nonlocal correlation functionals of Lundqvist and Van

Voorhis[57, 73–75, 136–139]. A comprehensive review of various approaches to extending the applicability of DFT to dispersion interactions can be found in Reference 126.

The simplest and cheapest methods that account for dispersion are Grimme's empirical DFT-D methods[53–55], which are damped interatomic potentials. Grimme's first attempt at an empirical dispersion tail was DFT-D1[53], which was only available for 6 elements (H, C, N, O, F, and Ne). With the atomic C_6 parameters and van der Waals radii predetermined, a single linear optimizable parameter (s_6) was trained onto 3 existing local GGA density functionals (BLYP, BP86, and PBE) and led to a considerable improvement in the description of vdW interactions. Grimme then introduced the DFT-D2 dispersion tail for all elements through Xe, along with a 9-parameter, B97-based, local GGA density functional called B97-D. In addition, independent damping parameters were determined for PBE, BLYP, BP86, TPSS, and B3LYP. This was followed by the DFT-D3 dispersion tail, which used fractional coordination numbers to account for variations in atomic dispersion coefficients in different chemical environments and contained a two-body term and an optional three-body term, as well as an improved damping function. The DFT-D3 dispersion tail was trained onto more than 10 existing density functionals and generally improved upon its predecessors for describing dispersion interactions.

Several density functionals that explicitly account for dispersion have been developed, including vdW-DF-04[73], vdW-DF-10[74], VV09[75], and VV10[57]. These nonlocal correlation functionals rely on a double space integral over the density and a nonlocal correlation kernel, and are computationally more expensive than the DFT-D methods. Langreth, Lundqvist, and coworkers introduced the first NLC scheme (vdW-DF-04) that could be applied to overlapping densities in 2004. Several years later, Vydrov and Van Voorhis implemented the vdW-DF-04 NLC functional for use with Gaussian basis sets[136], and proposed modifications (vdW-DF-09) to improve its compatibility with existing exchange functionals[137]. The VV09[75, 138, 139] NLC functional of Vydrov and Van Voorhis adopted a simple analytic form for the nonlocal correlation kernel, instead of relying on a numerically tabulated kernel. Soon afterwards, Vydrov and Van Voorhis proposed an even simpler NLC functional, VV10, that improved upon its predecessor by employing a less elaborate function for the damping of the $1/r^6$ asymptote. In 2010, vdW-DF-10 was introduced to correct the tendency of vdW-DF-04 to overestimate equilibrium bond lengths and underestimate the binding energies of hydrogen-bonded complexes. While the vdW-DF methods have no optimizable parameters, VV09 and VV10 have 1 and 2 optimizable parameters, respectively.

4.1.4 Outline

This chapter describes the task of designing an accurate and computationally efficient local meta-GGA density functional by attempting a partial search of the functional space defined by the inclusion or exclusion of the parameters in Equation 4.3. To ensure computational efficiency, B97M-V will not include exact exchange; i.e. it is semilocal as far as exchange is concerned. This will impose some limits on the accuracy of B97M-V for properties such as barrier heights that are sensitive to the inclusion of exact exchange. To offer the possibility of high accuracy for non-covalent interactions, particularly in the asymptotic regime, B97M-V will include nonlocal correlation via the VV10 NLC functional, which has already been demonstrated to be highly effective for this purpose[57] and is computationally efficient.

The remainder of the chapter is organized as follows. Details of the methodology used are discussed in Section 4.2, beginning with a full specification of the functional forms that are being trained and tested, followed by details regarding the choice of basis sets and grids, and a brief summary of the datasets that will be employed for training and testing B97M-V. Most importantly, Section 4.2.5 describes the way in which a partial search of the meta-GGA density functional genome (from which B97M-V was drawn) was attempted. Since the inclusion of the kinetic energy density is known to lead to oscillations in potential energy curves (PEC) for weakly interacting systems (i.e. strong integration grid sensitivity), this aspect was heavily emphasized in the development of B97M-V.

The functional form of B97M-V emerged as the leading representative of the "fittest" functional forms encountered in the search. Having selected a functional form, B97M-V was then self-consistently trained and assessed against a broad range of existing density functionals in Section 4.3. These assessments begin with the training and test data from which B97M-V emerged, and then progress to a range of further tests on data which were not used for developing or selecting B97M-V.

4.2 Approach

4.2.1 Theory

The complete functional form for B97M-V is given by Equation 4.4. The components of the exchange functional and correlation functional are described below. The acronyms used in Equation 4.4 (and henceforth) are: exchange-correlation (xc), exchange (x), correlation (c), same-spin (ss), and opposite-spin (os).

$$E_{xc}^{B97M-V} = E_{x}^{B97M-V} + E_{css}^{B97M-V} + E_{cos}^{B97M-V} + E_{NLC}^{VV10} \tag{4.4}$$

The local spin-density approximation (LSDA) for exchange can be expressed in terms of the first-order spinless reduced density matrix for a uniform electron gas (UEG):

$$E_x^{LSDA} = -\frac{1}{2}\sum_{\sigma}^{\alpha,\beta}\int\int\frac{1}{s}|\rho_{\sigma}^{UEG}(\mathbf{r},s)|^2 d\mathbf{r}d\mathbf{s} \tag{4.5}$$

$$\rho_{\sigma}^{UEG}(\mathbf{r},s) = 3\rho_{\sigma}(\mathbf{r})\left(\frac{\sin(k_{F\sigma}s) - k_{F\sigma}s\cos(k_{F\sigma}s)}{(k_{F\sigma}s)^3}\right) \tag{4.6}$$

where $\mathbf{s} = \mathbf{r}_1 - \mathbf{r}_2$, $\mathbf{r} = \frac{1}{2}(\mathbf{r}_1 + \mathbf{r}_2)$, and $k_{F\sigma} = (6\pi^2\rho_{\sigma})^{1/3}$ is the spin-polarized Fermi wave vector. Integration of Equation 4.5 over $\mathbf{s}$ gives the well-known expression for the LSDA exchange energy in terms of the exchange energy density per unit volume of a uniform electron gas:

$$E_x^{LSDA} = \sum_{\sigma}^{\alpha,\beta}\int e_{x,\sigma}^{UEG}(\rho_{\sigma})d\mathbf{r} \tag{4.7}$$

$$e_{x,\sigma}^{UEG}(\rho_{\sigma}) = -\frac{3}{2}\left(\frac{3}{4\pi}\right)^{1/3}\rho_{\sigma}^{4/3} \tag{4.8}$$

Accounting for inhomogeneities in the electron density is achieved by multiplying the integrand of the LSDA exchange functional by a power series inhomogeneity correction factor, $g_x(w_{x,\sigma}, u_{x,\sigma})$, resulting in the B97M-V exchange functional:

$$E_x^{B97M-V} = \sum_{\sigma}^{\alpha,\beta}\int e_{x,\sigma}^{UEG}(\rho_{\sigma})g_x(w_{x,\sigma}, u_{x,\sigma})d\mathbf{r} \tag{4.9}$$

$$g_x(w_{x,\sigma}, u_{x,\sigma}) = \sum_{i=0}\sum_{j=0}c_{x,ij}w_{x,\sigma}^{i}u_{x,\sigma}^{j} \tag{4.10}$$

$$w_{x,\sigma} = \frac{t_{\sigma} - 1}{t_{\sigma} + 1} \tag{4.11}$$

$$u_{x,\sigma} = \frac{\gamma_x s_{\sigma}^2}{1 + \gamma_x s_{\sigma}^2} \tag{4.12}$$

where the dimensionless variable, $w_{x,\sigma} \in [-1,1]$, is a finite domain transformation of the ratio of the UEG kinetic energy density to the exact kinetic energy density, $t_{\sigma} = \frac{\tau_{\sigma}^{UEG}}{\tau_{\sigma}}$, with $\tau_{\sigma}^{UEG} = \frac{3}{5}(6\pi^2)^{2/3}\rho_{\sigma}^{5/3}$, and the dimensionless variable, $u_{x,\sigma} \in [0,1]$, is a finite domain transformation of the reduced spin-density gradient, $s_{\sigma} = \frac{|\nabla\rho_{\sigma}|}{\rho_{\sigma}^{4/3}} \in [0,\infty)$. The linear DFT exchange parameters, $c_{x,ij}$, will be determined by least-squares fitting to a training set in Section 4.2.5, while $\gamma_x = 0.004$ is a nonlinear DFT exchange parameter that was fit to the Hartree–Fock exchange energies of 20 atoms in 1986 by Becke[9].

Closed-form expressions for the correlation energy density per particle of a uniform electron gas, $\epsilon_c^{UEG}(\rho)$, are only known for the low- and high-density limits of the paramagnetic and ferromagnetic cases of the UEG. Using the Monte-Carlo data of Ceperley and Alder[8], Perdew and Wang developed an analytic spin-compensated representation[7], $\epsilon_c^{PW92}(\rho)$, for $\epsilon_c^{UEG}(\rho)$. Combined with the spin-polarization interpolation formula of Vosko, Wilk, and Nusair[6], the spin-polarized PW92 correlation energy density per particle, $\epsilon_c^{PW92}(\rho_\alpha, \rho_\beta)$, is the starting point for the B97M-V correlation functional:

$$E_c^{LSDA} = \int \rho \epsilon_c^{PW92}(\rho_\alpha, \rho_\beta) d\mathbf{r} \tag{4.13}$$

Using the spin decomposition technique of Stoll and coworkers[81], the LSDA correlation energy can be separated into same-spin and opposite-spin components:

$$\begin{aligned} E_{css}^{LSDA} = \sum_{\sigma}^{\alpha,\beta} \int e_{c,\sigma\sigma}^{PW92} d\mathbf{r} = \int \rho_\alpha \epsilon_c^{PW92}(\rho_\alpha, 0) d\mathbf{r} + \\ \int \rho_\beta \epsilon_c^{PW92}(0, \rho_\beta) d\mathbf{r} \end{aligned} \tag{4.14}$$

$$\begin{aligned} E_{cos}^{LSDA} = \int e_{c,\alpha\beta}^{PW92} d\mathbf{r} = \int \rho \epsilon_c^{PW92}(\rho_\alpha, \rho_\beta) d\mathbf{r} - \\ \int \rho_\alpha \epsilon_c^{PW92}(\rho_\alpha, 0) d\mathbf{r} - \int \rho_\beta \epsilon_c^{PW92}(0, \rho_\beta) d\mathbf{r} \end{aligned} \tag{4.15}$$

where $e_{c,\sigma\sigma}^{PW92}$ and $e_{c,\alpha\beta}^{PW92}$ are the PW92 same-spin and opposite-spin correlation energy densities per unit volume. Extending Equations 4.14 and 4.15 to account for inhomogeneities in the electron density is straightforward, since the same approach used for the exchange functional can be utilized:

$$E_{css}^{B97M-V} = \sum_{\sigma}^{\alpha,\beta} \int e_{c,\sigma\sigma}^{PW92} g_{css}(w_{c,\sigma\sigma}, u_{c,\sigma\sigma}) d\mathbf{r} \tag{4.16}$$

$$g_{css}(w_{c,\sigma\sigma}, u_{c,\sigma\sigma}) = \sum_{i=0} \sum_{j=0} c_{css,ij} w_{c,\sigma\sigma}^i u_{c,\sigma\sigma}^j \tag{4.17}$$

$$w_{c,\sigma\sigma} = \frac{t_\sigma - 1}{t_\sigma + 1} \tag{4.18}$$

$$u_{c,\sigma\sigma} = \frac{\gamma_{css} s_\sigma^2}{1 + \gamma_{css} s_\sigma^2} \tag{4.19}$$

$$E_{cos}^{B97M-V} = \int e_{c,\alpha\beta}^{PW92} g_{cos}(w_{c,\alpha\beta}, u_{c,\alpha\beta}) d\mathbf{r} \tag{4.20}$$

$$g_{cos}(w_{c,\alpha\beta}, u_{c,\alpha\beta}) = \sum_{i=0} \sum_{j=0} c_{cos,ij} w_{c,\alpha\beta}^i u_{c,\alpha\beta}^j \tag{4.21}$$

$$w_{c,\alpha\beta} = \frac{t_{\alpha\beta} - 1}{t_{\alpha\beta} + 1} \tag{4.22}$$

$$u_{c,\alpha\beta} = \frac{\gamma_{cos} s_{\alpha\beta}^2}{1 + \gamma_{cos} s_{\alpha\beta}^2} \tag{4.23}$$

where $t_{\alpha\beta} = \frac{1}{2}(t_\alpha + t_\beta)$ and $s_{\alpha\beta}^2 = \frac{1}{2}(s_\alpha^2 + s_\beta^2)$. The linear DFT correlation parameters, $c_{css,ij}$ and $c_{cos,ij}$, will be determined by least-squares fitting to a training set in Section 4.2.5, while $\gamma_{css} = 0.2$ and $\gamma_{cos} = 0.006$ are nonlinear DFT correlation parameters that were fit to the correlation energies of neon and helium in 1997 by Becke[35].

Nonlocal correlation is taken into account via the VV10 NLC functional[57]:

$$E_{NLC}^{VV10} = \int \rho(\mathbf{r}) \left(\frac{1}{32} \left(\frac{3}{b^2} \right)^{3/4} + \frac{1}{2} \int \rho(\mathbf{r}') \Phi(\mathbf{r}, \mathbf{r}', \{b, C\}) d\mathbf{r}' \right) d\mathbf{r} \tag{4.24}$$

where $\Phi(\mathbf{r}, \mathbf{r}', \{b, C\})$ is the nonlocal correlation kernel defined in Reference 57. The VV10 NLC functional introduces 2 nonlinear parameters: b, which controls the short-range damping of the $1/r^6$ asymptote, and C, which controls the accuracy of the asymptotic C_6 coefficients.

4.2.2 Datasets

A semi-empirical density functional is only as applicable as the data used to train and test it. In total, the training, primary test, and secondary test sets contain 2460 data points, requiring 2421 single-point calculations. Of the 2460 data points, 1095 belong to the training set, 1153 belong to the primary test set, and 212 belong to the secondary test set. Furthermore, the training, primary test, and secondary test sets contain both thermochemistry (TC) data as well as non-covalent interactions (NC) data. The training set contains 787 TC data points and 308 NC data points, the primary test set contains 134 TC data points and 1019 NC data points, and the secondary test set contains 81 TC data points and 131 NC data points (for an overall total of 1002 TC data points and 1458 NC data points). Table 4.1 lists the 45 datasets that form the training, primary test, and secondary test sets. Specific details regarding the datasets can be found in Section 4 of Reference 63, with the only difference being that updated reference values for HW6F, HW6Cl, and H2O6Bind8 are used in this work[209].

Name	Description	# of Data Points	References
HAT707	Heavy-atom transfer reaction energies	505	82
BDE99	Bond dissociation reaction energies	83	82
TAE_nonMR124	Total atomization energies (nonmultireference)	124	82
SN13	Nucleophilic substitution reaction energies	13	82
ISOMER20	Isomerization reaction energies	18	82
DBH24	Diverse barrier heights	24	83,84
EA6	Electron affinities of atoms	6	171
IP6	Ionization potentials of atoms	6	171
AE8	Absolute atomic energies	8	86
SW49Rel345	$SO_4^{2-}(H_2O)_n$ ($n = 3 - 5$) relative energies	28	87
SW49Bind345	$SO_4^{2-}(H_2O)_n$ ($n = 3 - 5$) binding energies	30	87
NBC10-2	Methane dimer and benzene-methane dimer PECs	37	88,89
BzDC215	Benzene and first- and second-row hydride PECs	108	91
NBC10-1	Parallel-displaced (3.4 Å), sandwich, and T-shaped benzene dimer PECs	53	88,89
HW30	Hydrocarbon and water dimers	30	100
S22	Equilibrium geometries from S22x5	22	104,89
EA7	Adiabatic electron affinities of small molecules	7	171
IP7	Adiabatic ionization potentials of small molecules	7	171
AlkAtom19	$n = 1 - 8$ alkane atomization energies	19	93
AlkIsomer11	$n = 4 - 8$ alkane isomerization energies	11	93
AlkIsod14	$n = 3 - 8$ alkane isodesmic reaction energies	14	93
HTBH38	Hydrogen transfer barrier heights	38	94
NHTBH38	Non-hydrogen transfer barrier heights	38	95
SW49Rel6	$SO_4^{2-}(H_2O)_n$ ($n = 6$) relative energies	17	87
SW49Bind6	$SO_4^{2-}(H_2O)_n$ ($n = 6$) binding energies	18	87
HBC6	Formic acid, formamide acid, and formamidine acid dimer PECs	118	90,89
NBC10-3	S2 and T3 configuration pyridine dimer PECs	39	97,89
S22x5	Hydrogen-bonded and dispersion-bound complex PECs	110	102
S66x8	Biomolecular structure complex PECs	528	103
S66	Equilibrium geometries from S66x8	66	103,105
NNTT41	Neon-neon PEC	41	96
AATT41	Argon-argon PEC	41	96
NATT41	Neon-argon PEC	41	96
G21EA	Adiabatic electron affinities of atoms and small molecules	25	154,210
G21IP	Adiabatic ionization potentials of atoms and small molecules	36	154,210
PA8	Adiabatic proton affinities of small molecules	8	155,156
Gill12	Neutral, radical, anionic, and cationic isodesmic reaction energies	12	92
A24	Small non-covalent complexes	24	149
X40	Non-covalent interactions of halogenated molecules	40	144
H2O6Bind8	Binding energies of water hexamers	8	152,209
HW6F	Binding energies of $F^-(H_2O)_n$ ($n = 1 - 6$)	6	152,209
HW6Cl	Binding energies of $Cl^-(H_2O)_n$ ($n = 1 - 6$)	6	152,209
CYCONF	Relative energies of cysteine conformers	10	157,210
DS14	Binding energies for complexes containing divalent sulfur	14	153
WATER27	Neutral and charged water interactions	23	98,210

Table 4.1: Summary of the datasets found in the training, primary test, and secondary test sets. The datasets above the first thick black line are in the training set, the datasets between the first and second thick black lines are in the primary test set, while the datasets below the second thick black line are in the secondary test set. Within the training, primary test, and secondary test sets, datasets above the thin black line contain thermochemistry data points, while datasets below the thin black line contain non-covalent interactions data points. PEC stands for potential energy curve.

4.2.3 Computational Details

For the training, primary test, and secondary test sets, the (99,590) grid (99 radial shells with 590 Lebedev points each) was used to evaluate local xc functionals, while the SG-1 grid[61] was used to evaluate the VV10 NLC functional. For the absolute atomic energies in the training set and the rare-gas dimer PECs in the primary test set, the (500,974) grid was used to evaluate local xc functionals, while the (99,590) grid was used to evaluate the VV10 NLC functional. For M06-L and M11-L, calculations in the training, primary test, and secondary test sets were carried out with the (250,590) grid instead of the (99,590) grid. For the sake of simplicity, the following convention will be used henceforth: the grid used to integrate the local xc functional will be appended with a forward slash, followed by the grid used to integrate the VV10 NLC functional (if applicable). Therefore, referring to the second sentence of this paragraph, it will suffice to state that the (500,974)/(99,590) grid was used for the absolute atomic energies and the rare-gas dimer PECs.

The aug-cc-pVQZ (aQZ) basis set[77, 78] was used for all thermochemistry data points in the training, primary test, and secondary test sets except the second-row absolute atomic energies in the training set (aug-cc-pCVQZ)[77, 78], while the aug-cc-pVTZ (aTZ) basis set[77, 78] was used for all non-covalent interactions data points in the training, primary test, and secondary test sets except the rare-gas dimer PECs in the primary test set (aQZ) and the X40 dataset in the secondary test set (def2-TZVPPD[145, 146]). All non-covalent interactions were computed *without* counterpoise corrections (unless otherwise noted).

For B97-D2, Grimme's DFT-D2 dispersion tail was used with an s_6 coefficient of 0.75, as parameterized by Burns and coworkers[79]. Grimme's B97-D density functional[54] uses the DFT-D2 dispersion tail as well, with an s_6 coefficient of 1.25. For B3LYP-D3, Grimme's DFT-D3 dispersion tail was used with the following set of parameters: $\{s_6, s_{r,6}, s_8, s_{r,8}\} = \{1, 1.261, 1.703, 1\}$. For PBE-D3, Grimme's DFT-D3 dispersion tail was used with the following set of parameters: $\{s_6, s_{r,6}, s_8, s_{r,8}\} = \{1, 1.217, 0.722, 1\}$. For TPSS-D3, Grimme's DFT-D3 dispersion tail was used with the following set of parameters: $\{s_6, s_{r,6}, s_8, s_{r,8}\} = \{1, 1.166, 1.105, 1\}$. The density functionals that utilize the DFT-D3 dispersion tail use its zero-damping implementation (commonly referred to as DFT-D3(0)). Electronic energies are exclusively used throughout this chapter and spin-orbit coupling is not taken into consideration. All of the calculations were performed with a development version of Q-Chem 4.0[211].

4.2.4 Least-Squares Procedure

With a total of 2248 data points in the training and primary test sets, it is clear that a comprehensive 2-parameter nonlinear optimization of the parameters associated with VV10 is impractical. As a result, the values of $b = 6$ and $C = 0.01$ that were optimized for ωB97X-V are taken without further adjustment.

For the least-squares fits associated with the development of B97M-V, 1880 molecules comprise the 2248 data points in the training and primary test sets. The molecular orbital (MO) coefficients for these 1880 molecules were initially converged with the LSDA+VV10 density functional and saved to disk (equivalent to $g_x = g_{css} = g_{cos} = 1$). Using these saved MO coefficients, the values (henceforth referred to as contributions) that are enhanced by the exchange ($c_{x,ij}$), same-spin correlation ($c_{css,ij}$), and opposite-spin correlation ($c_{cos,ij}$) coefficients are computed up to 8th order in w and up to 4th order in u (including cross terms) for a total of 135 contributions per molecule, to enable the evaluation of Equation 4.3. These values are computed twice: once with the (99,590)/SG-1 grid and once with the (250,590)/SG-1 grid (the (500,974)/(99,590) grid was always used for AE8, NNTT41, AATT41, and NATT41 throughout this chapter and never altered).

The contributions are used to form a (# of data points) × (# of linear parameters) matrix, A. In addition to the A matrix, a column vector, y, of values corresponding to the errors in the unoptimized density functional ($y = E_{REF} - E_{DFT}$) is computed. Since weights are used during training, a diagonal (# of data points) × (# of data points) training set weight matrix (W_{Train}) is required as well. The diagonal elements corresponding to the training set data contain the appropriate weights, while the remaining diagonal elements corresponding to the primary test set data are set to zero. Thermochemistry data points in the training and primary test sets are given weights of 1 and 2, respectively (except for data points in TAE_nonMR124 (5), EA6 (10), IP6 (10), and DBH24 (25)), non-covalent interactions data points in the training and primary test sets are given weights of 100 and 200, respectively, and data points corresponding to the rare-gas dimer PECs in the primary test set are given a weight of 100,000. The total root-mean-square deviation (RMSD) is defined as a weighted RMSD of all 2248 data points in the training and primary test sets with the aforementioned weights, while the training RMSD is defined as a weighted RMSD of the 1095 data points in the training set with the aforementioned weights.

The change in the linear parameters, Δb, is found by a weighted least-squares fit:

$$\Delta b = (A^T W_{Train} A)^{-1} (A^T W_{Train} y) \tag{4.25}$$

Unless otherwise noted, Δb is computed using the data generated with the (250,590)/SG-1 grid. The training set RMSD is calculated by:

$$RMSD_{Train} = \sqrt{\frac{\text{diag}(W_{Train}) \cdot (y - A\Delta b)^2}{\#_{Train}}} \tag{4.26}$$

while the total RMSD is calculated by:

$$RMSD_{Total} = \sqrt{\frac{\text{diag}(W_{Total}) \cdot (y - A\Delta b)^2}{\#_{Total}}} \tag{4.27}$$

With the contributions calculated with 2 different grids, it is straightforward to analyze the energetic differences between the grids per contribution per data point. Figure 4.1 plots the difference between the (99,590)/SG-1 data and (250,590)/SG-1 data for each of the 135 contributions for all 2248 data points. As a reminder, grid-based errors involve the product of these differences and the corresponding coefficient that appears in the density functional. For instance, the largest grid-based error (-0.0195 kcal/mol for the parallel-displaced benzene dimer from S22x5 at 0.9·R$_e$) comes from $u_{c,\sigma\sigma}$, and a well-behaved density functional like ωB97X-V has $c_{css,01} = -0.257$, meaning that the resulting error is only 0.005 kcal/mol. More attention should be paid to higher-order variables, because while their grid-based errors might seem small, such terms are more likely to have large coefficients. For example, M06-L has $c_{cos,02} = -251.325$, and the largest error for the associated variable is -0.000622 kcal/mol, which leads to a more significant total error of 0.156 kcal/mol. However, since grid-based errors might cancel upon integration, fits are not discarded solely based on the magnitude of the resulting coefficients. The procedure used to ensure that the final density functional is as grid-insensitive as possible will be discussed in Section 4.2.5.

4.2.5 Partial Search of the Functional Space

Following the initial setup described above, a subspace of the full parameter space must be selected to begin the search. The initial parameter space consists of 135 parameters. With the available computing resources (a dedicated 64-core node), the maximum number of fits that can be performed in a single day is around 2 billion. Therefore 135C5 is the largest 1-day calculation that can be performed with all 135 parameters, resulting in only up to 5-parameter fits. A prior search of the much smaller GGA parameter space to design ωB97X-V led to an optimal functional form with a total of 7 linear parameters, but since τ-dependent variables are being included in this fit, larger numbers of parameters are almost certainly required. To reduce the size of the search space, certain higher-order variables are excluded: only contributions from variables up to combined 6th order $(u + w)$ are included in the fits. Furthermore, all 3 uniform electron gas limits are satisfied, further reducing

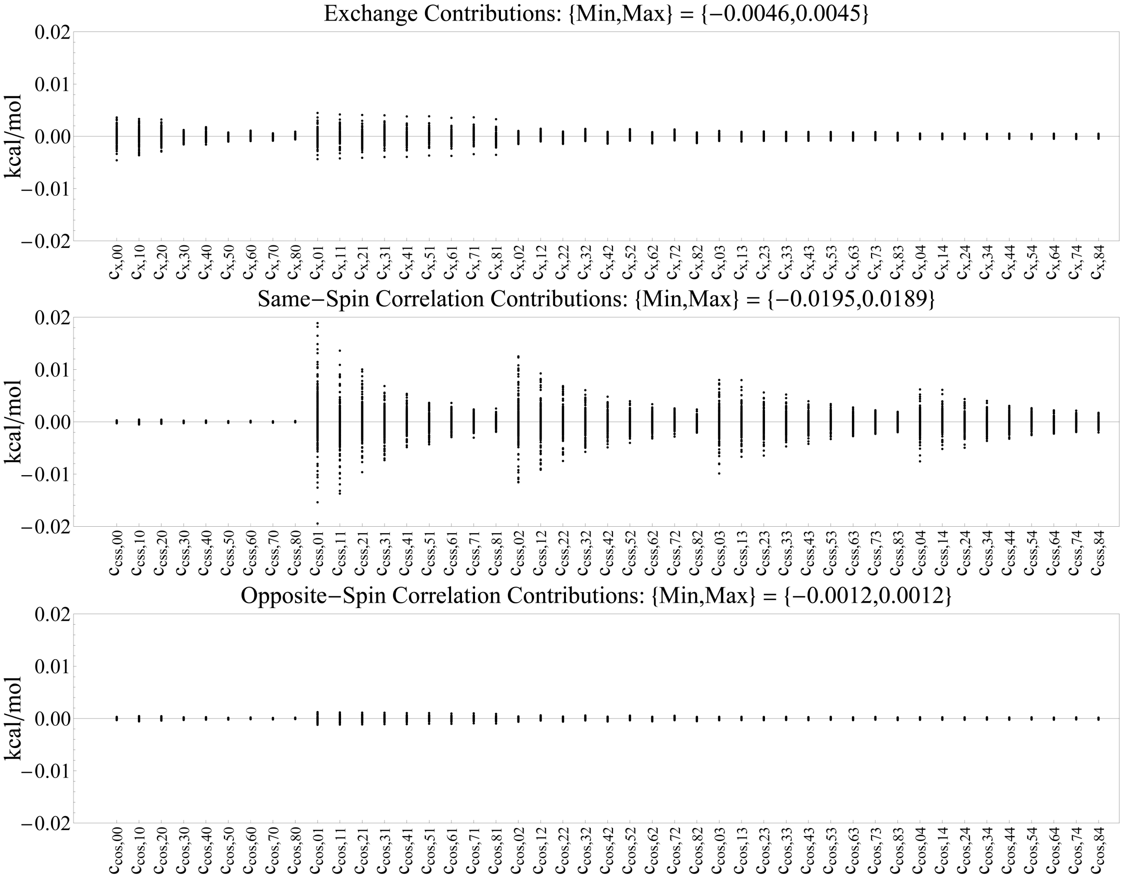


Figure 4.1: Energetic differences in kcal/mol between the (99,590)/SG-1 data and the (250,590)/SG-1 data for each of the 135 contributions for all 2248 data points in the training and primary test sets.

the total number of optional parameters by 3. This truncation reduces the original parameter space from 135 to 72. With 72 selectable parameters, it is possible to expand from 135C5 to 72C7, since 72C7 = 1,473,109,704. Clearly, it is still essential to devise a scheme to allow for the exploration of fits with more than 7 parameters.

From all of the 7-parameter fits resulting from the 72C7 optimization, the top 100,000 (ranked by total RMSD) are analyzed, since the total RMSD is a good indicator of both training set performance (fitting) and primary test set performance (transferability). From these 100,000 fits, those with exchange ICFs that go below zero (resulting in positive exchange energy contributions) are removed on physical grounds. Since the change in the parameters (Δb) from the least-squares fits can be used to predict energies corresponding to all of the data points (via $E_{LSTSQ} = E_{DFT} + A\Delta b$), the (99,590)/SG-1 and (250,590)/SG-1 data is used to compute 2 sets of energies ($E_{LSTSQ}^{(99,590)}$ and $E_{LSTSQ}^{(250,590)}$) for all of the remaining fits. In order to acquire these energies, E_{DFT}, A, and Δb are taken from the respective grid. The 2 sets of energies are compared and only fits that have an absolute maximum grid-based error of 0.01 kcal/mol across all 2248 data points are kept. The remaining fits are filtered once again, such that those with parameters with a magnitude greater than 30 are removed.

The remaining fits are analyzed in order to determine the coefficient that is most commonly used. This coefficient is then compulsorily selected in the next set of least-squares fits in order to allow for the exploration of 8-parameter fits. This procedure was repeated until a minimum in the total RMSD was found. Since the minimum was found at 15 parameters, 9 coefficients ($c_{x,10}$, $c_{x,01}$, $c_{x,11}$, $c_{x,02}$, $c_{css,10}$, $c_{css,02}$, $c_{css,32}$, $c_{css,42}$, $c_{cos,10}$), which will henceforth be referred to as the "Best 9", had to be compulsorily selected in order to reach the 16-parameter mark at which the total RMSD increased. The progression from the 7-parameter fits to the 16-parameter fits can be tracked in Table 4.2. In addition, the total RMSDs of the surviving fits are plotted in Figure 4.2. The 15-parameter fit with the lowest RMSD was self-consistently optimized in order to finalize the parameters. However, after a single self-consistent cycle, it became clear that the least-squares fit RMSDs from the first cycle were inaccurate by more than 0.2 kcal/mol on average, primarily for the thermochemistry datasets.

In order to overcome the issue of inaccurate least-squares fit RMSDs, the contributions were recalculated in the (99,590)/SG-1 grid and (250,590)/SG-1 grid with the "Best 9" parameters fixed. With the updated contributions, all 10-parameter (63C1) through 16-parameter (63C7) fits were recomputed and filtered as before: $\min(g_x(w_{x,\sigma}, u_{x,\sigma}))\geq 0$, $\max(|E_{LSTSQ}^{(99,590)}$ - $E_{LSTSQ}^{(250,590)}|)\leq 0.01$ kcal/mol, and $\max(|b|)\leq 30$. The resulting fits were plotted (on the left in Figure 4.3) in order to identify the "optimal" fit based on the same methodology that was used in Reference 62. However, in order to ensure that the resulting density functional would be accurate for intermolecular geometries as well, the interpolated minima for 19 of the PECs in the training and primary test sets (7 PECs from NBC10, 5 PECs from BzDC215, 4 PECs from HBC6, along with the 3 rare-gas dimer PECs) were computed and compared to reference values in order to determine MADs and maximum absolute deviations (MAX) for each fit. Based on this data, fits with an MAD greater than 0.03 Å and an MAX greater than 0.1 Å were discarded in order to produce the plot on the right in Figure 4.3. The fit that led to the B97M-V density functional is boxed in Figure 4.3.

Frozen	# of Initial Fits	# (Fitted)	g_x	Grid	|b|	Common	# of Remaining Fits
—	1,473,109,704	7	87,506	29,521	29,025	$c_{x,01}$	27,248
$c_{x,01}$	1,329,890,705	8	83,605	45,474	44,828	$c_{x,10}$	44,695
$c_{x,10}$	1,198,774,720	9	55,310	16,572	16,436	$c_{x,02}$	16,155
$c_{x,02}$	1,078,897,248	10	40,900	3,398	3,382	$c_{x,11}$	2,343
$c_{x,11}$	969,443,904	11	17,684	305	290	$c_{css,32}$	224
$c_{css,32}$	869,648,208	12	10,192	90	36	$c_{css,42}$	33
$c_{css,42}$	778,789,440	13	23,664	157	58	$c_{css,10}$	43
$c_{css,10}$	696,190,560	14	3,126	52	42	$c_{css,02}$	41
$c_{css,02}$	621,216,192	15	117	19	16	$c_{cos,10}$	16
$c_{cos,10}$	553,270,671	16	660	62	36	$c_{cos,01}$	33

Table 4.2: Progression from the 7-parameter fits to the 16-parameter fits based on the LSDA+VV10 data. The first column indicates the additional coefficient that was frozen (compulsorily selected) in order to achieve the associated set of fits. The second column contains the total number of least-squares fits that were performed, of which only the top 100,000 (ranked by total RMSD) were analyzed. The fourth column indicates the number of fits (of 100,000) that remained after the first filtering criterion ($\min(g_x(w_{x,\sigma}, u_{x,\sigma}))\geq 0$) was applied. The fifth column indicates the number of fits from the previous column that remained after the second filtering criterion ($\max(|E^{(99,590)}_{LSTSQ}$ - $E^{(250,590)}_{LSTSQ}|)\leq 0.01$ kcal/mol) was applied. The sixth column indicates the number of fits from the previous column that remained after the third filtering criterion ($\max(|b|)\leq 30$) was applied. Finally, the last 2 columns indicate the coefficient that was most commonly utilized in the surviving fits (shown in Column 6) and the number of times that coefficient appeared.

Including the initial cycle (Cycle 1) with the unoptimized B97M-V density functional as well as the "Best 9" cycle, the self-consistent optimization of B97M-V required 7 cycles. For the first 2 cycles, the data points in the training *and* primary test sets were evaluated in order to determine the functional form that would be self-consistently optimized. For the later cycles, only the 1095 data points in the training set were required to finalize the parameters. The parameters from the beginning of all 7 cycles are listed in Table 4.3.

The final parameters of B97M-V can be found in the last column of Table 4.3, and Figure 4.4 shows the exchange, same-spin correlation, and opposite-spin correlation ICF plots for B97M-V. Compared to recent semi-empirical meta-GGA density functionals, the resulting coefficients are very well-behaved. The 2 largest coefficients enhance variables that are fifth and sixth order overall, yet are still around 20 in magnitude. An encouraging but unintended outcome is that all of the exchange coefficients are positive. The resulting ICFs are well-behaved as well, especially the

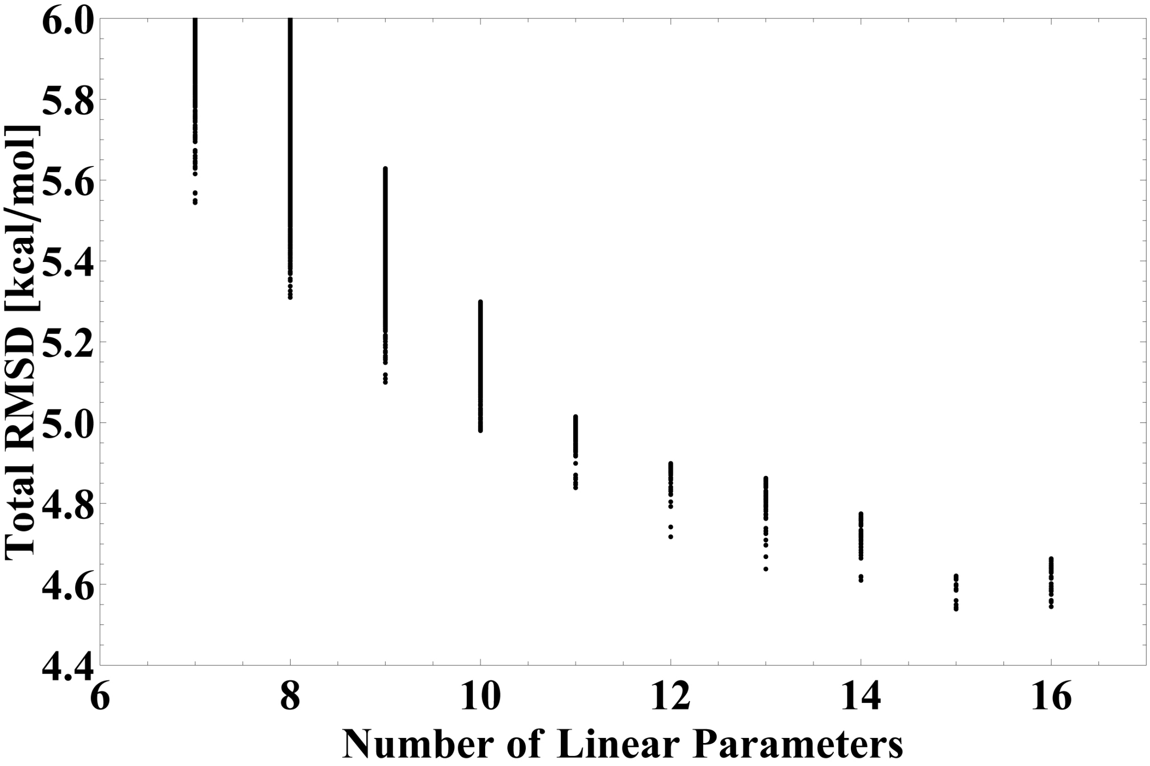
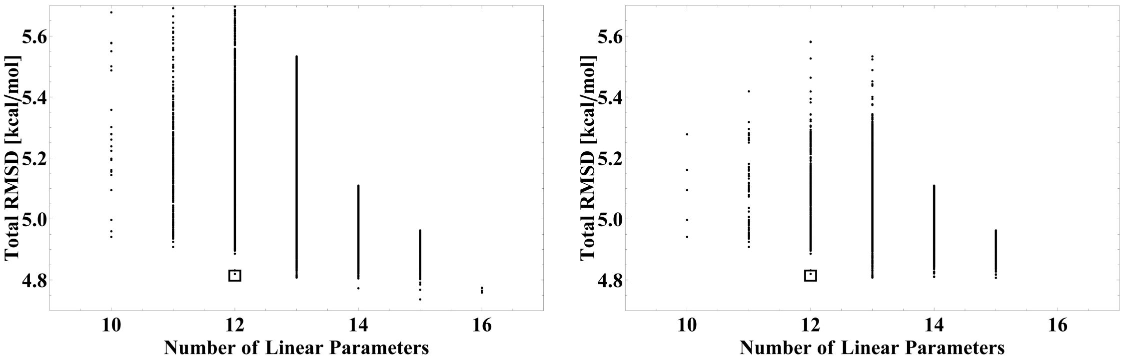


Figure 4.2: Total RMSDs in kcal/mol of the surviving 7-parameter through 16-parameter fits based on the LSDA+VV10 data. The filtering criteria are described in Section 4.2.5 and Table 4.2.

exchange ICF.

Since the resulting functional form of B97M-V is unique and employs terms that are not commonly found in density functionals, it is important to discuss the similarities and differences between B97M-V and existing density functionals. For clarity, the 3 ICFs of B97M-V are given in Equations 4.28-4.30. The motivation for using the particular power series expansion shown in Equation 4.3 came from Becke's B97, B98, and B00 density functionals[15, 35, 111]. With B97, a systematic optimization was carried out specifically with the GGA variable, u, for exchange, same-spin correlation, and opposite-spin correlation, while with B98, the same was done with a meta-GGA variable, w', closely related to the w used in this work. The switch from w' to w occurred with the B00 density functional. For B97M-V, the u-dependent ICF of B97 and the w-dependent ICF of B00 were multiplied in order to give Equation 4.3.



(a) (b)

Figure 4.3: Total RMSDs in kcal/mol of the surviving 10-parameter through 16-parameter fits based on the "Best 9" data. The plot on the left uses the same filtering criteria as Figure 4.2, while the plot on the right uses 2 additional criteria based on the accuracy of 19 interpolated potential energy curve minima. The fit that led to the B97M-V density functional is boxed.

Parameter	1	Best 9	2	3	4	5	6 (Final)
$c_{x,00}$	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$c_{x,10}$	0.000	0.384	0.417	0.417	0.416	0.416	0.416
$c_{x,01}$	0.000	1.344	1.327	1.310	1.309	1.308	1.308
$c_{x,11}$	0.000	3.073	3.071	3.070	3.070	3.070	3.070
$c_{x,02}$	0.000	1.780	1.804	1.895	1.900	1.901	1.901
$c_{css,00}$	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$c_{css,10}$	0.000	-2.543	-5.857	-5.668	-5.670	-5.667	-5.668
$c_{css,02}$	0.000	-1.470	-1.879	-1.855	-1.856	-1.855	-1.855
$c_{css,32}$	0.000	-20.450	-20.406	-20.477	-20.498	-20.495	-20.497
$c_{css,42}$	0.000	-18.847	-19.739	-20.341	-20.361	-20.364	-20.364
$c_{cos,00}$	1.000	1.000	1.000	1.000	1.000	1.000	1.000
$c_{cos,10}$	0.000	0.425	2.659	2.516	2.538	2.534	2.535
$c_{cos,01}$	0.000	0.000	1.475	1.553	1.574	1.573	1.573
$c_{cos,32}$	0.000	0.000	-6.159	-6.371	-6.427	-6.426	-6.427
$c_{cos,03}$	0.000	0.000	-5.723	-6.203	-6.295	-6.297	-6.298

Table 4.3: Linear parameters from the beginning of all 7 cycles of the self-consistent optimization of B97M-V. The "Best 9" column refers to the freezing of 9 commonly occurring parameters. The nonlinear parameters that were taken from previous studies[9, 35, 63] are $\gamma_x = 0.004$, $\gamma_{css} = 0.2$, $\gamma_{cos} = 0.006$, $b = 6$, and $C = 0.01$.

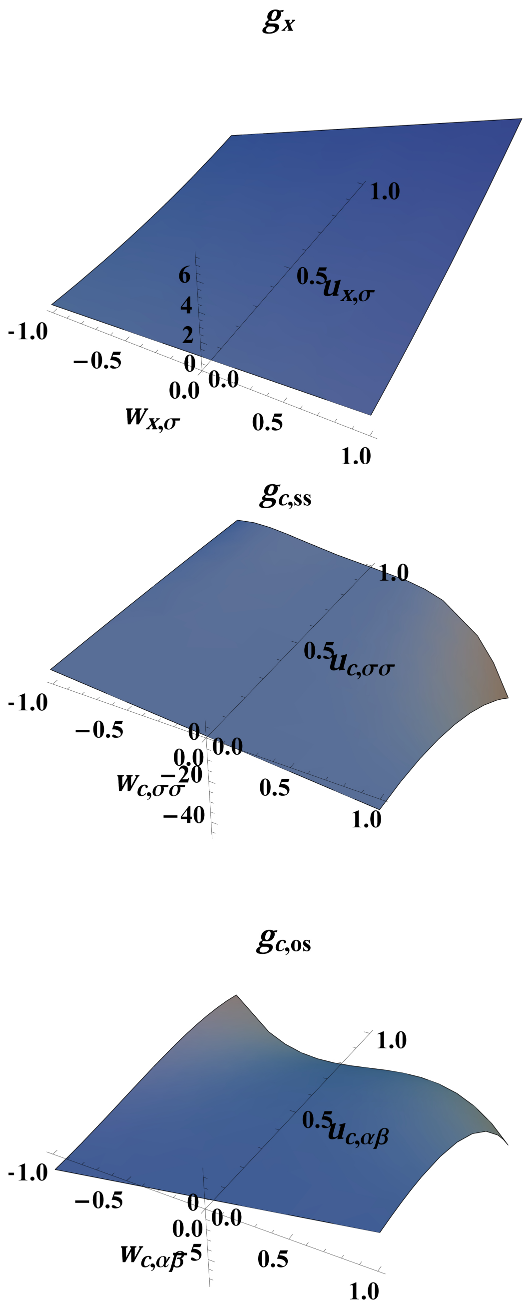


Figure 4.4: Exchange, same-spin correlation, and opposite-spin correlation inhomogeneity correction factor plots for the B97M-V density functional.

$$\begin{aligned} g_x(w_{x,\sigma}, u_{x,\sigma}) = 1 \; + \; 0.416 w_{x,\sigma} \; + \; 1.308 u_{x,\sigma} \; + \\ 3.07 w_{x,\sigma} u_{x,\sigma} \; + \; 1.901 u_{x,\sigma}^2 \end{aligned} \tag{4.28}$$

$$\begin{aligned} g_{css}(w_{c,\sigma\sigma}, u_{c,\sigma\sigma}) = 1 \; - \; 5.668 w_{c,\sigma\sigma} \; - \; 1.855 u_{c,\sigma\sigma}^2 \; - \\ 20.497 w_{c,\sigma\sigma}^3 u_{c,\sigma\sigma}^2 \; - \; 20.364 w_{c,\sigma\sigma}^4 u_{c,\sigma\sigma}^2 \end{aligned} \tag{4.29}$$

$$\begin{aligned} g_{cos}(w_{c,\alpha\beta}, u_{c,\alpha\beta}) = 1 \; + \; 2.535 w_{c,\alpha\beta} \; + \; 1.573 u_{c,\alpha\beta} \; - \\ 6.427 w_{c,\alpha\beta}^3 u_{c,\alpha\beta}^2 \; - \; 6.298 u_{c,\alpha\beta}^3 \end{aligned} \tag{4.30}$$

In the context of empirical density functionals, the power series, as written in Equation 4.3, has never been utilized as the starting point for the systematic parameterization of an empirical density functional, particularly for the same-spin and opposite-spin correlation components. The exchange ICFs of the 2005-2011 Minnesota density functionals can be considered as limiting cases of Equation 4.3, with the u-dependent ICF fixed to the ICFs of either PBE or RPBE (or both), and the w-dependent ICF taking either the B00 form given in Equation 4.3 or the VSXC form (or both). However, the 2012 meta-nonseparable gradient approximation (meta-NGA) density functionals[25, 51], MN12-L and MN12-SX, use an exchange ICF similar in form to Equation 4.3, albeit with an additional third power series that is a function of a density-dependent inhomogeneity variable. While the functional forms for the Minnesota density functionals have always been pre-determined (i.e. the parameters that will be optimized are chosen a priori), B97M-V is unique in that the only restrictions on the functional form were placed during the truncation of the parameter space, and the optimization procedure had the freedom to select the parameters that were most conducive to minimizing the total RMSD as well as satisfying the filtering criteria.

The B97M-V correlation functional distinguishes itself even more from existing correlation functionals, because a two-dimensional optimization of a correlation functional has not yet been reported. As far as the Minnesota meta-GGA and meta-NGA density functionals are concerned, the correlation functionals can be divided into 3 distinct classes. The M05 and M05-2X correlation functionals are practically identical to that of B97, with the exception of the SCC factor enhancing the same-spin correlation component. The 2006 Minnesota density functionals inherit the correlation functional of their 2005 counterparts, as well as that of VSXC. Finally, the post-2006 Minnesota meta-GGAs and meta-NGAs no longer partition the correlation energy into same-spin and opposite-spin components, but instead contain 2 w-dependent ICFs that enhance the PW92 correlation energy density per unit volume and the PBE correlation functional gradient correction term.

Taking a closer look at the B97M-V ICF equations, all of the variables that made their way into the exchange ICF can be rationalized. The importance of the first-order terms in w and u is indisputable, and it is thus plausible that the associated second-order cross-term is also helpful. The appearance of the second-order term in u is not surprising either, since this variable is the foundation of the OPTX exchange functional, which was found to be a sound improvement over the B88 exchange functional. The same-spin correlation ICF also contains 2 variables (w and u^2) that seem intuitive, as does the opposite-spin correlation ICF (w and u). In fact, the variables that seem most out-of-place are the fifth- and sixth-order variables in the same-spin correlation ICF, and the fifth-order variable in the opposite-spin correlation ICF, since these variables are not ones that can be intuitively selected prior to optimization. Ultimately, the optimization procedure must be credited for shedding light on the importance of these variables.

4.3 Results and Discussion

4.3.1 Training Set, Primary Test Set, and Secondary Test Set Performance

To assess the performance of B97M-V, 14 existing density functionals were selected for comparison. These 14 density functionals include 3 local GGA density functionals (PBE-D3, B97-D, VV10), 3 local meta-GGA density functionals (TPSS-D3, M06-L, M11-L), 2 global hybrid GGA density functionals (B3LYP-D3, B97-D2), 2 global hybrid meta-GGA density functionals (M06, M06-2X), 3 range-separated hybrid GGA density functionals (LC-VV10, ωB97X-D, ωB97X-V), and 1 range-separated hybrid meta-GGA density functional (M11). Table 4.4 contains information regarding the benchmarked density functionals. Although there is no unique way to count parameters, each density functional in Table 4.4 is assigned 2 separate numbers regarding its empirical parameter count. The second column lists the total number of linearly independent and nonlinear parameters that were optimized specifically for the given density functional during its development, while the third column lists the total number of linearly independent and nonlinear parameters found in the density functional. As an example, while 12 linear parameters were optimized for B97M-V in this work, 5 additional parameters were borrowed from previous work (γ_x from B86, γ_{css} and γ_{cos} from B97, and b and C from ωB97X-V). Furthermore, parameters that are decided upon prior to optimization, such as s_6 for B97-D and c_x for M06-2X, are only counted in the third column.

Table 4.5 contains the RMSDs for all of the datasets in the training, primary

Functional	# (Fitted)	# (Total)	c_x	Year	Class	Rung	References
PBE-D3	2	2	0	2010	L GGA w/ DT	2	12,55
B97-D	9	13	0	2006	L GGA w/ DT	2	54
VV10	2	2	0	2010	L GGA w/ NLC	2	12,107,57
TPSS-D3	2	2	0	2010	L MGGA w/ DT	3	17,55
M06-L	34	39	0	2006	L MGGA	3	23
M11-L	44	44	0	2012	L MGGA	3	24
B97M-V	12	17	0	2015	L MGGA w/ NLC	3	Present Work
B3LYP-D3	5	6	20	2010	GH GGA w/ DT	4	10,14,33,55
B97-D2	13	14	19.4	2011	GH GGA w/ DT	4	35,54,79
M06	33	38	27	2008	GH MGGA	4	45
M06-2X	29	34	54	2008	GH MGGA	4	45
LC-VV10	3	3	0-100	2010	RSH GGA w/ NLC	4	12,120,57
ωB97X-D	15	18	22.2-100	2008	RSH GGA w/ DT	4	58
ωB97X-V	10	13	16.7-100	2014	RSH GGA w/ NLC	4	63
M11	40	40	42.8-100	2011	RSH MGGA	4	52

Table 4.4: Details for the 15 density functionals from Table 4.5. L stands for local, GH stands for global hybrid, RSH stands for range-separated hybrid, DT stands for dispersion tail, and NLC stands for nonlocal correlation. The column labeled "# (Fitted)" lists the number of parameters that were optimized on a training set for the specified density functional, while the column labeled "# (Total)" lists the total number of empirical parameters. c_x refers to the percentage of exact exchange.

test, and secondary test sets for B97M-V and these 14 existing density functionals. The principal benchmark for success will be how B97M-V compares with other local density functionals: PBE-D3, B97-D, VV10, TPSS-D3, M06-L, and M11-L. However, since hybrid density functionals are usually preferred for higher accuracy DFT calculations, it will be important to see how B97M-V compares with a range of successful hybrids as well.

Before the individual datasets are discussed, the overall performance of the density functionals for TC and NC will be discussed. Of the 15 benchmarked density functionals, M06-2X has the best overall TC performance (3.21 kcal/mol). The next best density functionals for TC are ωB97X-V, ωB97X-D, and B97M-V with RMSDs of 3.6, 3.61, and 3.93 kcal/mol, respectively. It is very encouraging that B97M-V is distinctly the best local density functional for TC, with its closest rivals being B97-D (5.56 kcal/mol) and M06-L (5.63 kcal/mol). This considerable improvement over the best existing local density functionals is a significant gap, and is a validation of the design strategy used to construct B97M-V.

Turning to non-covalent interactions, it is remarkable that B97M-V has the best overall performance (0.22 kcal/mol). After B97M-V, the next best density functionals

kcal/mol	Zero	PBE-D3	B97-D	VV10	TPSS-D3	M06-L	M11-L	B97M-V	B3LYP-D3	B97-D2	M06	M06-2X	LC-VV10	ωB97X-D	ωB97X-V	M11
HAT707	74.79	7.79	5.62	7.23	6.21	5.86	5.43	3.99	4.27	3.98	4.83	3.63	6.81	4.14	4.28	4.28
BDE99	114.98	8.96	4.58	6.63	4.96	6.14	4.61	3.52	3.91	3.18	3.72	2.99	5.39	3.03	3.38	4.10
TAE_nonMR124	381.05	16.85	5.18	12.46	5.92	5.54	6.62	3.79	5.23	4.06	3.94	3.24	5.30	3.65	3.34	4.37
SN13	25.67	5.56	4.20	3.67	3.71	1.58	3.19	1.39	1.89	0.97	2.03	1.01	2.65	0.96	1.01	2.19
ISOMER20	46.43	4.53	3.83	4.34	4.38	4.11	4.46	3.00	2.30	2.30	2.53	1.47	1.85	1.86	1.64	1.94
DBH24	28.34	10.27	7.18	9.86	9.57	5.38	3.54	4.99	5.23	4.36	2.97	1.12	3.02	2.07	1.81	1.48
EA6	46.12	4.71	2.07	5.28	2.27	2.39	6.17	3.86	3.49	1.45	1.91	1.95	2.10	1.89	2.34	1.33
IP6	260.50	5.07	2.68	7.64	4.65	2.26	3.36	3.91	6.16	3.19	3.23	2.86	5.03	3.01	3.76	5.79
AE8	42810.90	30.54	6.94	64.43	17.97	9.22	15.27	8.12	15.89	2.25	5.52	1.28	29.82	4.67	1.77	9.23
SW49Rel345	1.55	1.45	1.22	1.26	1.24	0.53	0.22	0.12	0.76	0.79	0.60	0.35	0.36	1.01	0.33	0.17
SW49Bind345	29.30	2.24	1.35	2.47	1.63	0.60	1.54	0.17	1.34	0.69	0.72	0.95	0.58	0.98	0.29	0.25
NBC10-2	0.75	0.20	0.08	0.05	0.16	0.15	0.54	0.11	0.13	0.08	0.28	0.11	0.12	0.13	0.09	0.16
BzDC215	2.25	0.37	0.39	0.22	0.35	0.17	0.34	0.18	0.34	0.36	0.25	0.60	0.60	0.41	0.27	0.65
NBC10-1	1.85	0.30	0.20	0.27	0.21	0.26	0.47	0.22	0.29	0.67	0.58	0.43	0.15	0.41	0.23	0.69
HW30	2.34	0.48	0.40	0.43	0.34	0.35	0.48	0.19	0.35	0.35	0.33	0.46	0.30	0.35	0.20	0.38
S22	9.65	0.60	0.54	0.63	0.45	0.43	0.91	0.23	0.50	0.60	0.77	0.47	0.51	0.41	0.23	0.58
EA7	39.24	2.28	3.86	3.57	2.78	6.07	6.44	3.11	3.56	2.55	2.94	1.97	4.19	2.45	2.28	1.08
IP7	252.70	4.28	4.03	4.76	3.71	3.11	2.89	3.67	5.22	3.41	2.93	3.24	5.30	3.26	2.97	4.56
AlkAtom19	1829.31	20.93	10.48	5.85	9.46	8.11	29.35	1.36	5.50	9.28	4.63	5.27	19.04	2.90	0.71	3.94
AlkIsomer11	1.81	1.44	0.33	0.84	1.50	0.95	0.74	0.19	1.56	0.57	0.22	0.15	0.13	1.04	0.67	0.56
AlkIsod14	10.35	2.54	0.57	2.01	3.74	3.86	2.06	0.48	2.75	1.93	1.68	1.65	1.09	2.31	1.80	2.11
HTBH38	15.97	10.32	7.63	9.22	8.71	4.62	1.86	4.48	5.31	5.12	2.21	1.26	1.50	2.57	2.28	1.68
NHTBH38	33.30	10.42	7.12	10.79	10.31	4.79	3.72	5.16	5.81	4.15	2.64	1.69	3.38	1.73	1.64	1.49
SW49Rel6	1.26	1.91	1.65	1.72	1.68	0.80	0.21	0.07	0.86	1.01	0.89	0.42	0.52	1.34	0.37	0.28
SW49Bind6	62.11	3.72	2.50	4.56	2.35	0.57	3.38	0.35	2.35	0.90	0.92	1.61	0.91	0.99	0.36	0.67
HBC6	12.69	1.12	0.77	1.13	0.89	0.36	0.68	0.26	0.77	0.39	0.31	0.59	1.25	0.61	0.39	0.46
NBC10-3	2.52	0.37	0.19	0.17	0.30	0.47	0.83	0.34	0.35	0.54	0.69	0.63	0.06	0.27	0.37	0.80
S22x5	7.00	0.82	0.40	0.59	0.59	0.42	0.94	0.28	0.65	0.71	0.80	0.47	0.48	0.43	0.41	0.55
S66x8	5.57	0.52	0.43	0.50	0.38	0.31	0.83	0.17	0.43	0.35	0.48	0.35	0.34	0.44	0.22	0.42
S66	6.91	0.46	0.52	0.52	0.32	0.36	0.81	0.18	0.43	0.36	0.53	0.29	0.31	0.52	0.18	0.41
NNTT41	0.05	0.06	0.07	0.03	0.04	0.04	0.32	0.01	0.02	0.06	0.12	0.05	0.02	0.15	0.02	0.02
AATT41	0.14	0.05	0.16	0.02	0.09	0.15	0.46	0.03	0.07	0.05	0.23	0.08	0.07	0.21	0.02	0.17
NATT41	0.07	0.06	0.08	0.04	0.04	0.05	0.35	0.01	0.02	0.05	0.13	0.03	0.03	0.15	0.02	0.04
G21EA	40.86	3.97	3.00	4.60	2.74	4.54	6.56	3.41	3.56	1.84	2.46	1.99	3.89	1.98	2.27	2.40
G21IP	265.35	4.81	4.47	5.43	4.77	5.60	4.54	3.64	4.86	3.48	3.78	3.49	5.23	3.82	3.57	4.64
PA8	166.17	1.62	3.70	1.77	3.14	3.36	3.57	3.03	1.28	2.67	2.21	1.98	2.74	2.95	2.20	1.35
Gill12	28.47	7.35	4.84	6.08	5.13	6.71	5.60	5.07	4.09	3.84	3.82	1.78	2.58	3.24	2.32	2.58
A24	2.65	0.41	0.32	0.41	0.23	0.23	0.46	0.17	0.23	0.26	0.25	0.28	0.15	0.15	0.09	0.27
X40	4.94	0.59	0.59	0.63	0.47	0.48	1.23	0.20	0.34	0.43	0.57	0.28	0.41	0.49	0.21	0.54
H2O6Bind8	46.96	4.54	3.84	5.68	2.35	0.99	4.02	0.41	3.07	0.32	1.02	2.23	2.66	1.53	0.68	0.66
HW6F	81.42	2.30	4.49	3.55	0.95	2.32	1.18	0.45	1.79	0.73	1.96	4.77	2.40	1.30	0.21	2.81
HW6Cl	57.71	3.54	3.43	5.08	1.75	1.45	2.92	0.25	1.44	0.48	1.93	3.49	2.67	0.62	0.49	2.10
CYCONF	2.10	0.99	0.56	0.57	0.99	0.39	0.57	0.24	0.29	0.37	0.16	0.30	0.78	0.41	0.11	0.51
DS14	3.70	0.47	0.37	0.52	0.25	0.25	0.59	0.09	0.23	0.25	0.34	0.20	0.12	0.18	0.05	0.30
WATER27	67.48	6.53	2.56	7.07	4.44	1.34	4.60	0.74	3.68	1.42	1.70	3.66	3.35	1.48	1.42	1.62
TC	3836.93	10.10	5.56	9.81	6.45	5.63	6.68	3.93	4.66	3.97	4.18	3.21	6.79	3.61	3.60	3.97
TC*	299.75	9.76	5.55	7.98	6.28	5.59	6.57	3.87	4.45	3.98	4.17	3.22	6.27	3.60	3.61	3.90
NC	14.95	1.23	0.82	1.36	0.85	0.42	1.08	0.22	0.77	0.48	0.57	0.77	0.72	0.54	0.32	0.55
TC Train	4319.60	10.30	5.43	10.44	6.31	5.75	5.64	3.96	4.65	3.83	4.45	3.36	6.86	3.82	3.88	4.23
TC Primary Test	691.59	11.19	6.94	8.02	8.19	5.09	11.41	3.84	4.99	5.09	2.76	2.50	7.60	2.33	1.85	2.30
TC Secondary Test	186.16	4.84	4.06	5.04	4.16	5.30	5.33	3.77	4.13	3.05	3.30	2.74	4.31	3.18	2.92	3.54
NC Train	9.67	0.89	0.64	0.90	0.69	0.33	0.65	0.18	0.56	0.51	0.47	0.54	0.45	0.55	0.25	0.53
NC Primary Test	10.56	0.82	0.60	0.87	0.60	0.35	0.89	0.20	0.58	0.43	0.52	0.44	0.54	0.48	0.27	0.44
NC Secondary Test	37.43	3.13	1.92	3.57	2.04	0.90	2.39	0.38	1.80	0.69	1.02	2.07	1.76	0.85	0.64	1.09

Table 4.5: RMSDs in kcal/mol for all of the datasets in the training, primary test, and secondary test sets for B97M-V and 14 existing density functionals. The datasets above the first thick black line are in the training set, the datasets between the first and second thick black lines are in the primary test set, while the datasets between the second and third thick black lines are in the secondary test set. Within the training, primary test, and secondary test sets, datasets above the thin black line contain thermochemistry (TC) data points, while datasets below the thin black line contain non-covalent interactions (NC) data points. The last section of the table contains overall unweighted statistics for TC and NC. The row labeled TC* is TC with AE8 removed. The column labeled "Zero" contains the RMSDs of the energies in each dataset and is meant to give perspective to the magnitudes of the RMSDs in the following columns.

for non-covalent interactions are ωB97X-V, M06-L, and B97-D2, with RMSDs of 0.32, 0.42 and 0.48 kcal/mol, respectively. The fact that the 2 best density functionals both use the VV10 nonlocal correlation functional is a testament to the accuracy of this approach for treating long-range correlation. Comparing against existing local density functionals, B97M-V is almost 2 times better than the best alternative, M06-L. Furthermore, it is a striking indication of the versatility of the meta-GGA form that B97M-V (0.22 kcal/mol) is more than 30% better than ωB97X-V (0.32 kcal/mol), which was designed by a nearly identical procedure and includes range-separation, though not τ dependence.

Since it would be tedious to individually address the performance of the 15 benchmarked density functionals on the 45 datasets in Table 4.5, only a handful of datasets will be discussed. TAE_nonMR124 is comprised of the atomization energies of 124 small molecules computed at the Weizmann-4 (W4) level of theory, and is an indicator of performance for computational thermodynamics. B97M-V has an RMSD of 3.79 kcal/mol on this dataset, performing better than 11 of the benchmarked density functionals, and worse than M06-2X (3.24 kcal/mol), ωB97X-V (3.34 kcal/mol), and ωB97X-D (3.65 kcal/mol). After B97M-V, the next best local density functionals for TAE_nonMR124 are B97-D (5.18 kcal/mol) and M06-L (5.54 kcal/mol).

DBH24 is a training dataset that contains 24 forward and reverse barrier heights computed (at least) at the Weizmann-3.2 (W3.2) level of theory, and is an indicator of performance for computational kinetics. The performance of B97M-V for this dataset (4.99 kcal/mol) is poor relative to the best hybrids (M06-2X with an RMSD of 1.12 kcal/mol) due to the absence of exact exchange. Indeed, most of the hybrid density functionals perform considerably better than B97M-V, and this poor performance is a weakness of B97M-V and all local density functionals. Compared to the 6 existing local density functionals considered, B97M-V ranks second (behind M11-L at 3.54 kcal/mol), is comparable to M06-L (5.38 kcal/mol), and is 2 times more accurate than both PBE-D3 and TPSS-D3. Similar results are seen for the barrier height datasets in the primary test set (HTBH38 and NHTBH38), indicating transferability of these conclusions.

Moving on to the non-covalent interactions in the training set, B97M-V has the best performance for both the relative and binding energies of the $SO_4{}^{2-}(H_2O)_n$ ($n = 3 - 5$) clusters. Compared to density functionals that utilize dispersion tails (DFT-D2 or DFT-D3), B97M-V is 10 times more accurate than PBE-D3 and TPSS-D3 and 5 times more accurate than ωB97X-D for these 2 datasets (SW49Rel345 and SW49Bind345). It even outperforms the range-separated hybrid meta-GGA density functional, M11, which was previously the best performer for both datasets. The S22 dataset by Hobza and coworkers is often used to assess the performance of density functionals for hydrogen-bonded, dispersion-bound, and mixed dimers. The top 2 performers for this dataset are B97M-V and ωB97X-V, both with RMSDs of 0.23 kcal/mol. The next best density functionals are ωB97X-D and M06-L, with RMSDs almost twice as large (0.41 and 0.43 kcal/mol, respectively).

A dataset in the thermochemistry section of the primary test set that is challenging for local and hybrid density functionals alike is AlkAtom19, which contains the atomization energies of 19 alkanes ranging from methane to octane. Surprisingly, B97M-V performs second best (1.36 kcal/mol) out of all 15 density functionals, coming second to ωB97X-V (0.71 kcal/mol). For comparison, M06-2X, which has the best overall thermochemistry performance, has an RMSD of 5.27 kcal/mol for AlkAtom19, while the other Rung 3 density functionals have RMSDs ranging from 8.11 (M06-L) to 29.35 (M11-L) kcal/mol. The performance of B97M-V is also very impressive for 2 datasets derived from the molecules in AlkAtom19: AlkIsomer11 and AlkIsod14. For the isomerization energies, B97M-V (0.19 kcal/mol) performs almost as well as the 2 best density functionals: LC-VV10 (0.13 kcal/mol) and M06-2X (0.15 kcal/mol). In comparison, the RMSD of ωB97X-D is more than 5 times larger, as are the RMSDs of PBE-D3, TPSS-D3, M06-L, and B3LYP-D3. B97M-V has the best performance for the isodesmic reaction energies, with a very small RMSD of 0.48 kcal/mol. The next best density functional is B97-D with an RMSD of 0.57 kcal/mol, followed by LC-VV10 (1.09 kcal/mol). To put the performance of B97M-V in perspective, the range of RMSDs for AlkIsod14 is rather large: 0.48 kcal/mol (B97M-V) to 3.86 kcal/mol (M06-L).

Moving on to the non-covalent interactions in the primary test set, the transferability of the parameters of B97M-V can be checked by considering the relative and binding energies of the $n = 6$ sulfate-water clusters. As intended, B97M-V performs superbly for SW49Rel6 and SW49Bind6, outperforming all of the benchmarked density functionals. Its RMSD for SW49Rel6 (0.07 kcal/mol) is 3 times smaller than that of the next best density functional (M11-L), while its RMSD for SW49Bind6 (0.35 kcal/mol) is on par with ωB97X-V and 3 times smaller than that of ωB97X-D. The S66 dataset was created by Hobza and coworkers to extend the scope of the S22 dataset to non-covalent interactions that are common in biomolecules. As with S22, the top 2 performers for this dataset are B97M-V and ωB97X-V, both with RMSDs of 0.18 kcal/mol. B97M-V has an RMSD that is 40% less than the value of the next best local meta-GGA density functional, which is TPSS-D3.

The A24 dataset consists of very accurate binding energies for small molecules and is a valuable transferability test for B97M-V, since only a few of the interactions that are in A24 were in the training set. Since the binding energies associated with the interactions in A24 are very small, the resulting RMSDs are generally small as well, ranging from 0.09 kcal/mol (ωB97X-V) to 0.46 kcal/mol (M11-L). After ωB97X-V, the next best density functionals are LC-VV10 (0.15 kcal/mol), ωB97X-

D (0.15 kcal/mol), and B97M-V (0.17 kcal/mol). Herbert and coworkers recently reported[152] that density functionals such as LC-VV10 and M06-2X perform poorly for halide-water clusters. Specifically, the systems of interest are $F^-(H_2O)_n$ and $Cl^-(H_2O)_n$, for $n = 1 - 6$. ωB97X-V has the best performance for the interactions containing the fluorine anion (0.21 kcal/mol), followed closely by B97M-V (0.45 kcal/mol). Since these binding energies are large, it is important to consider the RMSD range (0.21 kcal/mol for ωB97X-V to 4.77 kcal/mol for M06-2X) in order to comprehend the superb performance of both ωB97X-V and B97M-V. For HW6F, the next best density functionals are B97-D2 (0.73 kcal/mol) and TPSS-D3 (0.95 kcal/mol). On the other hand, for the interactions that contain the chlorine anion, B97M-V has the best performance, with an RMSD of only 0.25 kcal/mol (the largest is VV10 at 5.08 kcal/mol). B97M-V is followed by B97-D2 (0.48 kcal/mol) and ωB97X-V (0.49 kcal/mol), while the best Minnesota density functional is M06-L, with an RMSD of 1.45 kcal/mol.

Finally, the performance of B97M-V and its fellow density functionals can be assessed for water clusters via the H2O6Bind8 (8 binding energies of water hexamers) and WATER27 (23 binding energies of neutral and charged water clusters from dimers to octamers) datasets. B97M-V has the second smallest RMSD for H2O6Bind8 (0.41 kcal/mol) and the smallest RMSD for WATER27 (0.74 kcal/mol), even though the only water cluster in the entire training set was the water dimer in S22. The performance of B97M-V for WATER27 is almost 2 times better than that of the next best density functional, M06-L (1.34 kcal/mol), and almost 10 times better than that of VV10 (7.07 kcal/mol).

4.3.2 Benzene Dimer and Coronene Dimer

The parallel-displaced benzene dimer is a textbook example of π-π stacking and its binding energy has been recently determined to an extremely high level of accuracy (-2.65 ± 0.02 kcal/mol) by Xantheas and coworkers[212]. Using their CCSD(T)/aug-cc-pVTZ optimized dimer geometry and the (99,590)/SG-1 grid, binding energies were computed for all 15 benchmarked density functionals in the aTZ and aQZ basis sets. In addition, aug-cc-pV5Z (a5Z) binding energies were computed for the Rung 2 and 3 density functionals. With these results, summarized in Table 4.6, the density functionals can be assessed with respect to accuracy, as well as their basis set convergence.

Considering the basis set limit (BSL) values, 8 of the 15 density functionals (PBE-D3, B97-D, VV10, TPSS-D3, B97M-V, B3LYP-D3, LC-VV10, and ωB97X-V) predict the binding energy to within 0.20 kcal/mol of the -2.65 kcal/mol target. All 5 Minnesota density functionals underbind the dimer by between 0.3 kcal/mol (M06-

kcal/mol	aTZ	aQZ	a5Z	Δ
PBE-D3	-2.70	-2.62	-2.55	-0.16
B97-D	-2.71	-2.59	-2.54	-0.17
VV10	-3.08	-2.90	-2.85	-0.23
TPSS-D3	-2.72	-2.62	-2.59	-0.13
M06-L	-2.71	-2.11	-2.23	-0.49
M11-L	-2.91	-1.78	-1.95	-0.96
B97M-V	-2.80	-2.73	-2.73	-0.07
B3LYP-D3	-2.60	-2.48	-	-0.11
B97-D2	-1.90	-1.83	-	-0.08
M06	-2.32	-1.87	-	-0.45
M06-2X	-2.53	-2.37	-	-0.16
LC-VV10	-2.83	-2.69	-	-0.14
ωB97X-D	-3.29	-3.12	-	-0.17
ωB97X-V	-2.80	-2.69	-	-0.12
M11	-2.08	-1.39	-	-0.68

Table 4.6: Equilibrium binding energies in kcal/mol for the parallel-displaced benzene dimer in 3 different basis sets: aug-cc-pVTZ (aTZ), aug-cc-pVQZ (aQZ), and aug-cc-pV5Z (a5Z). Binding energies in the a5Z basis set were only computed for density functionals without exact exchange. The last column contains the difference in binding energy between the aTZ basis set value and the largest basis set value (a5Z for Rungs 2 and 3; aQZ for Rung 4). The (99,590)/SG-1 grid was used for all density functionals. The CCSD(T)/CBS binding energy for the parallel-displaced benzene dimer is -2.65 ± 0.02 kcal/mol, as determined by Xantheas and coworkers[212].

2X) and 1.3 kcal/mol (M11), while ωB97X-D overbinds by almost 0.5 kcal/mol. The basis set convergence of B97M-V is very satisfactory, as the aQZ result is equivalent to the a5Z result, and both differ from the aTZ value by only 0.07 kcal/mol. These results suggest that for weak intermolecular interactions, B97M-V is almost fully converged at the aTZ basis set level, and is fully converged at the aQZ basis set level. Most of the other density functionals also converge satisfactorily with basis set, although consistent with trends reported recently[213], all of the Minnesota density functionals except M06-2X have unusually large differences between the aTZ basis set value and the largest basis set value (reported as Δ values in Table 4.6).

In order to explore whether B97M-V can be successfully applied to larger interactions, the binding energy of the parallel-displaced coronene dimer (Table 4.6)[162, 163, 214], a dispersion-bound system that is nearly 4 times larger than the benzene dimer, was examined. The calculations were performed in the aTZ basis set with the

(99,590)/SG-1 grid. While there is no definitive reference value for the binding energy of the parallel-displaced coronene dimer, 2 recent attempts[162, 163] at determining a complete basis set (CBS) value resulted in counterpoise-corrected binding energies of $E^{CBS}_{bind,1}$ = -19.98 kcal/mol and $E^{CBS}_{bind,2}$ = -24.36 kcal/mol. The first reference value was arrived at via the following equation:

$$E^{CBS}_{bind,1} = E^{aDZ^*}_{QCISD(T)} + ([E^{aTZ^*}_{MP2}] - [E^{aDZ^*}_{MP2}]) \tag{4.31}$$

$$E^{CBS}_{bind,1} = -17.674 + ([-34.610] - [-32.303]) = -19.981 \tag{4.32}$$

with the asterisk indicating that the cc-pVNZ basis was used for hydrogen and the cc-pVNZ and aug-cc-pVNZ basis sets were alternated for the carbon atoms. The second reference value was arrived at via the following equation:

$$E^{CBS}_{bind,2} = E^{CBS^*}_{MP2} + ([E^{aDZ^*}_{QCISD(T)}] - [E^{aDZ^*}_{MP2}]) \tag{4.33}$$

$$E^{CBS}_{bind,2} = -38.984 + ([-17.674] - [-32.303]) = -24.355 \tag{4.34}$$

where $E^{CBS^*}_{MP2} = 1.02(E^{(DT)Z}_{MP2})$ and (DT)Z indicates a two-point extrapolation in the cc-pVDZ and cc-pVTZ basis sets (different extrapolations were used for the Hartree-Fock (HF) energy and the MP2 correlation energy). Since it is straightforward to calculate a more accurate estimate of E^{CBS}_{MP2} via $E^{CBS}_{MP2} = E^{aQZ}_{HF} + E^{a(TQ)Z}_{MP2}$, a better MP2 CBS value, E^{CBS}_{MP2} = -38.08 kcal/mol, can be acquired. Using this updated, counterpoise-corrected MP2 CBS value with the QCISD(T) correction yields an updated reference value of $E^{CBS}_{bind,3}$ = -23.45 kcal/mol:

$$E^{CBS}_{bind,3} = E^{CBS}_{MP2} + ([E^{aDZ^*}_{QCISD(T)}] - [E^{aDZ^*}_{MP2}]) \tag{4.35}$$

$$E^{CBS}_{bind,3} = -38.075 + ([-17.674] - [-32.303]) = -23.446 \tag{4.36}$$

Simply averaging the 3 CBS estimates results in $E^{CBS}_{bind,avg}$ = -22.59 kcal/mol. Thus, it is safe to assume that the binding energy of the parallel-displaced coronene dimer is in the vicinity of -22.59 kcal/mol. The binding energy of B97M-V is -22.46 kcal/mol, which is strikingly close to the ωB97X-V value of -22.4 kcal/mol. All 5 of the Minnesota density functionals underbind the dimer, along with B97-D2, PBE-D3, and to a lesser extent, TPSS-D3 and B3LYP-D3. Along with B97M-V and ωB97X-V, the binding energies of B97-D, VV10, and LC-VV10 lie within 1 kcal/mol of $E^{CBS}_{bind,avg}$, while ωB97X-D is the only density functional that overbinds the dimer. As a note, the binding energies of the hybrid density functionals were computed with the PARI-K algorithm of Manzer and Head-Gordon[215]. In order to test the accuracy of the approximation, B3LYP-D3 binding energies were computed with and without the approximation, and the error due to the approximation was found to be around 0.05 kcal/mol.

kcal/mol	$(C_{24}H_{12})_2$
PBE-D3	-17.33
B97-D	-22.49
VV10	-22.23
TPSS-D3	-19.23
M06-L	-18.04
M11-L	-17.90
B97M-V	-22.46
B3LYP-D3	-20.16
B97-D2	-15.08
M06	-14.62
M06-2X	-17.58
LC-VV10	-23.29
ωB97X-D	-24.35
ωB97X-V	-22.40
M11	-15.75

Table 4.7: Binding energies in kcal/mol for the parallel-displaced coronene dimer. The binding energies were computed in the aug-cc-pVTZ basis set (2760 basis functions for the dimer) with the (99,590)/SG-1 grid.

4.3.3 $(H_2O)_{20}$ Binding Energies and $(H_2O)_{16}$ Relative Energies

While the WATER27 dataset in Grimme's GMTKN30[99, 210, 216] database contains 27 data points, the 4 isomers of $(H_2O)_{20}$ were removed from the secondary test set due to their size. Recently, Anacker and Friedrich[217] have updated the reference values[98, 218] for these 4 clusters: dodecahedron (-198.6 kcal/mol), edge sharing (-209.7 kcal/mol), fused cubes (-208 kcal/mol), and face sharing (-208 kcal/mol). As a reminder, with the exception of the water dimer in the S22, S22x5, S66, S66x8, A24, and DS14 datasets, water clusters have appeared only in the secondary test set via the H2O6Bind8 and WATER27 datasets, with the largest cluster containing 8 water molecules (WATER27).

The binding energies for the 4 isomers were computed in the aTZ basis set with the (99,590)/SG-1 grid and are shown in Table 4.8. B97M-V performs remarkably well with respect to the reference binding energies of the water 20-mers, with an RMSD of 0.77 kcal/mol for 4 interactions that have an average binding energy of more than 200 kcal/mol. The next best density functional (B97-D2) is more than 3 times worse, with an RMSD of 2.51 kcal/mol. After B97M-V, the next best local density functional is M06-L, with an RMSD that is more than 5 times larger than that of B97M-V. The 45-fold improvement of B97M-V over VV10 is surprisingly large, as is its 5-fold improvement over ωB97X-V. Interestingly, while most of the

kcal/mol	dod	es	fc	fs	RMSD
Reference	-198.60	-209.70	-208.00	-208.00	0.00
PBE-D3	-229.70	-238.86	-235.53	-236.61	29.13
B97-D	-189.80	-197.51	-194.71	-195.18	11.91
VV10	-233.93	-244.95	-242.23	-242.69	34.88
TPSS-D3	-218.93	-227.31	-223.54	-224.87	17.67
M06-L	-191.35	-207.63	-210.01	-206.55	3.97
M11-L	-169.20	-185.68	-187.07	-184.76	24.59
B97M-V	-198.46	-210.10	-209.32	-208.66	0.77
B3LYP-D3	-216.15	-227.70	-225.93	-226.08	17.89
B97-D2	-202.88	-211.53	-209.02	-209.61	2.51
M06	-188.27	-206.69	-211.90	-205.54	5.85
M06-2X	-202.47	-216.59	-218.36	-215.74	7.58
LC-VV10	-208.81	-219.07	-216.49	-217.01	9.29
ωB97X-D	-204.23	-215.63	-214.03	-213.60	5.80
ωB97X-V	-203.39	-214.08	-212.10	-212.37	4.42
M11	-193.59	-206.29	-207.43	-205.01	3.39

Table 4.8: Binding energies and RMSDs in kcal/mol for 4 isomers of $(H_2O)_{20}$. The binding energies were computed in the aug-cc-pVTZ basis set (1840 basis functions) with the (99,590)/SG-1 grid. The 4 isomers are dodecahedron (dod), edge sharing (es), fused cubes (fc), and face sharing (fs).

density functionals tend to overbind the isomers, the large errors of B97-D and M11-L are actually due to severe underbinding.

Thus far, B97M-V and the 14 existing density functionals have been thoroughly tested for the accurate prediction of the binding energies of large water clusters. In order to test the performance of these density functionals on the relative energetics of water clusters, the relative energies of 5 isomers of $(H_2O)_{16}$ will be evaluated in the aTZ basis set with the (99,590)/SG-1 grid against reference values[219] computed at the CCSD(T)/aTZ level of theory.

In Table 4.9, the relative energies in the second through sixth columns are computed with respect to the minimum as predicted by the associated method. However, the RMSDs reported in the last column are taken over the relative energies between all 10 unique pairs that can be constructed from the 5 isomers. The RMSD range for the 15 benchmarked density functionals is very large, with values as low as 0.14 kcal/mol (ωB97X-D) and values as high as 4.51 kcal/mol (M06). In perspective, the performance of B97M-V (0.42 kcal/mol) is promising, as it is the best local

kcal/mol	4444-a	4444-b	antiboat	boat-a	boat-b	RMSD
Reference	0.00	0.54	0.51	0.25	0.42	0.00
PBE-D3	1.50	1.85	0.49	0.00	0.25	1.21
B97-D	1.45	1.87	0.77	0.00	0.21	1.17
VV10	0.73	1.07	0.58	0.00	0.24	0.61
TPSS-D3	1.89	2.22	0.36	0.00	0.23	1.54
M06-L	0.00	0.37	3.89	3.52	3.60	2.61
M11-L	0.00	0.49	2.19	2.06	2.17	1.37
B97M-V	0.00	0.58	1.16	0.77	0.90	0.42
B3LYP-D3	0.00	0.51	0.24	0.06	0.26	0.16
B97-D2	1.03	1.55	0.55	0.00	0.18	0.92
M06	0.00	0.31	6.37	5.92	5.97	4.51
M06-2X	0.00	0.25	3.71	2.99	3.11	2.36
LC-VV10	0.60	0.86	0.44	0.00	0.17	0.53
ωB97X-D	0.00	0.35	0.54	0.30	0.45	0.14
ωB97X-V	0.22	0.75	0.33	0.00	0.16	0.35
M11	0.00	0.33	3.13	2.55	2.66	1.94

Table 4.9: Relative energies and RMSDs in kcal/mol for 5 isomers of $(H_2O)_{16}$. The relative energies were computed in the aug-cc-pVTZ basis set (1472 basis functions) with the (99,590)/SG-1 grid and are taken with respect to the minimum as predicted by the associated method. The RMSDs are taken over the relative energies between all 10 unique pairs that can be constructed from the 5 isomers.

density functional. While VV10 drastically overestimated the $(H_2O)_{20}$ binding energies, it performs comparatively well for these relative energies, with an RMSD (0.61 kcal/mol) only slightly worse than that of B97M-V. Other local density functionals, particularly M06-L (RMSD of 2.61 kcal/mol), are significantly worse.

4.3.4 Additional Energetic Tests

Four sets of additional energetic tests were conducted in order to further assess the transferability of the B97M-V density functional: the HB15, HSG, NC15, and Shields38 datasets (Table 4.10). In addition, the potential energy curve of the benzene-argon dimer (a system from the BzDC215 dataset that was not included in the training set) was computed in order to assure that B97M-V maintains the accuracy of VV10 and LC-VV10 in the asymptotic regime. The HB15, HSG, NC15, and Shields38 were carried out in the aTZ basis set with the (99,590)/SG-1 grid. However, due to the delicate nature of the benzene-argon dimer interaction, the (250,590)/SG-1 grid was used instead.

The HB15 dataset[220] is comprised of the equilibrium structures of 15 hydrogen-bonded dimers featuring ionic groups common in biomolecules (acetate, methylammonium, guanidinium, and imidazolium) interacting with neutral donors and acceptors (methanol, water, methylamine, and formaldehyde). The RMSDs of the 15 benchmarked density functionals for this dataset range from 0.23 kcal/mol (B97M-V) to 1.17 kcal/mol (PBE-D3). After B97M-V, the next best performer is ωB97X-V (0.31 kcal/mol), followed by all of the Minnesota density functionals (with RMSDs between 0.41 and 0.54 kcal/mol) besides M11-L (1.03 kcal/mol). The excellent performance of B97M-V further indicates that it can be applied to interactions that were not necessarily included in the training and test sets.

The HSG dataset[221] was developed in order to assess the accuracy with which density functionals and force fields predict the binding affinities of small ligands to protein receptors. A model protein-ligand complex was selected and decomposed into 21 interacting fragment pairs, and the associated binding energies were evaluated at a high level of theory (and later improved by Sherrill and coworkers[89]). Based on these updated (HSG-A) reference values, the density functional with the best performance is B97M-V, with an RMSD of only 0.14 kcal/mol, followed closely by ωB97X-V, with an RMSD of 0.16 kcal/mol. The best Minnesota density functional is M06-L, with an RMSD that is 3 times larger than that of B97M-V, while the density functional with the worst performance, M11-L, has and RMSD that is 7 times larger than that of B97M-V.

The NC15 dataset[222] comes from a recent study on the basis set convergence of the post-CCSD(T) contribution to weakly-interacting systems. The original study included 21 small dimers, but the LiH dimer has been dropped for this assessment, along with the 5 duplicates from A24. The reference values that are used are the CCSD(T)/CBS ones from Table SI in the Supporting Information of Reference 222. Since the systems in NC15 are very small, the RMSDs for the benchmarked density functionals are correspondingly small, ranging from 0.06 kcal/mol (LC-VV10) to 0.47 kcal/mol (M11-L). B97M-V performs very comparably to LC-VV10, with an RMSD of 0.08 kcal/mol, while the next best local density functional is TPSS-D3 with an RMSD more than twice that of B97M-V. The poor performance of B97-D is surprising and mostly due to its severe overbinding of the He-LiH dimer. The same overbinding issue affects B97-D2, indicating that the culprit might be the DFT-D2 dispersion tail.

The Shields38 dataset[223] includes the binding energies of 38 water clusters ranging from dimers to 10-mers. Since the WATER27 dataset in the secondary test set only contains 10 neutral $(H_2O)_n$ $n = 2 - 8$ water clusters and the H2O6Bind8 dataset in the secondary test set only contains 8 $(H_2O)_6$ clusters, it was deemed necessary to further assess B97M-V on medium- to large-sized water clusters, due to

kcal/mol	HB15	HSG	NC15	Shields38
PBE-D3	1.17	0.44	0.29	7.36
B97-D	1.10	0.68	0.35	3.00
VV10	0.99	0.50	0.22	8.75
TPSS-D3	0.87	0.28	0.18	4.52
M06-L	0.43	0.43	0.20	0.78
M11-L	1.03	0.98	0.47	5.69
B97M-V	0.23	0.14	0.08	0.35
B3LYP-D3	0.81	0.33	0.16	4.47
B97-D2	0.64	0.36	0.33	0.90
M06	0.54	0.49	0.19	0.89
M06-2X	0.41	0.43	0.15	2.80
LC-VV10	0.80	0.22	0.06	2.95
ωB97X-D	0.56	0.37	0.13	1.84
ωB97X-V	0.31	0.16	0.07	1.26
M11	0.41	0.69	0.13	0.72

Table 4.10: RMSDs in kcal/mol for 4 datasets. HB15 contains 15 complexes with strong hydrogen bonds, HSG contains 21 interacting fragment pairs from a protein-ligand complex, NC15 contains 15 small non-covalent complexes, and Shields38 contains 38 water clusters with up to 10 water molecules. The binding energies were computed in the aug-cc-pVTZ basis set with the (99,590)/SG-1 grid.

the unimpressive performance of ωB97X-V for the water 20-mers in Section 4.3.3. Consistently, the performance of B97M-V is the best observed for the Shields38 dataset as well, further confirming that it is well-suited for applications involving large water clusters. The RMSD of B97M-V (0.35 kcal/mol) is more than 2 times better than that of the next best density functional, which is M11 (0.72 kcal/mol). VV10 massively overbinds once again, with an RMSD 25 times larger than that of B97M-V, while ωB97X-V is only sixth best. While the best local meta-GGA density functional is B97M-V, the best local GGA density functional (B97-D) has an RMSD that is more than 8.5 times larger than that of B97M-V.

Finally, the performance of 11 of the 15 density functionals is assessed on the benzene-argon dimer PEC. Figure 4.5 contains the 11 PECs along with the reference PEC and is divided into 3 separate plots: DFT-D containing density functionals (top), Minnesota density functionals (middle), and VV10-containing density functionals (bottom). Considering the 3 DFT-D density functionals, all of them predict bond lengths that are about 0.1 Å too long. Furthermore, TPSS-D3 overbinds the dimer by about 10%, ωB97X-D underbinds it by about 5%, while B97-D underbinds it by about 10%.

Moving on to the Minnesota density functionals, only M06 comes close to predicting an accurate bond length. M06-L predicts a bond length that is about 0.1 Å too long, while the rest predict bond lengths that are about 0.1 Å too short. With respect to equilibrium binding energies, M06 underbinds the dimer by about 30%, while M06-2X, M11-L, and M11 overbind by 10-30%. Only M06-L gives an equilibrium binding energy that is close to the reference value. A feature of the M11-L PEC that is very hard to miss is its strange long-range behavior. From 4.4 to 5 Å, it predicts repulsive binding energies for the dimer. An exaggeration of this feature of M11-L can be seen for the methane dimer in Figure 4.6.

The 3 VV10-containing density functionals have almost perfect bond lengths, but only B97M-V gets the equilibrium binding energy just right. LC-VV10 overbinds by about 10%, while ωB97X-V overbinds by about 20%. In terms of consistency, the VV10-containing density functionals are clearly superior for this system. Their long-range behavior is also noticeably better than the rest of the density functionals.

4.3.5 Geometries

While the energetics of B97M-V have been thoroughly tested (especially for non-covalent interactions), it is time to move on to tests of intramolecular and intermolecular geometries. The first test is basically free to conduct, since it is based on the interpolated minima of 78 PECs from the training and primary test sets, as well as 10 rare-gas dimer PECs that were additionally computed. In the training set, 3 of the datasets (NBC10-2, BzDC215, and NBC10-1) contain potential energy curves, while in the primary test set, 7 of the datasets (HBC6, NBC10-3, S22x5, S66x8, NNTT41, AATT41, and NATT41) contain potential energy curves. Five of these 10 datasets were used to generate equilibrium bond length (and binding energy) statistics based on interpolated PEC minima. The PECs in S22x5 did not have enough points near their minima in order to compute reliable values, while the provided points for several of the HBC6 PECs were all post-equilibrium for some density functionals. NBC10-1, NBC10-2, and NBC10-3 have 3, 2, and 2 PECs, respectively, and were combined to give the NBC10 dataset. For the S66x8 PECs, only the first 5 points (0.9, 0.95, 1, 1.05, and 1.1) were used in the interpolation. Additionally, for each PEC in S66x8, the interpolated minimum was scaled by the closest intermolecular distance of the equilibrium, counterpoise-corrected MP2/cc-pVTZ optimized structure in order to generate the RMSDs in Table 4.11. Finally, instead of using data from NNTT41, AATT41, and NATT41 in the primary test set, PECs (1.5 to 15 Å in increments of 0.05 Å) for all 10 rare-gas dimers containing helium, neon, argon, and krypton were computed in the aQZ basis set with the (250,590)/(75,302) grid.

The resulting equilibrium bond length RMSDs are presented in Table 4.11. The second-to-last column contains the total RMSD for all 88 PECs, while the last column

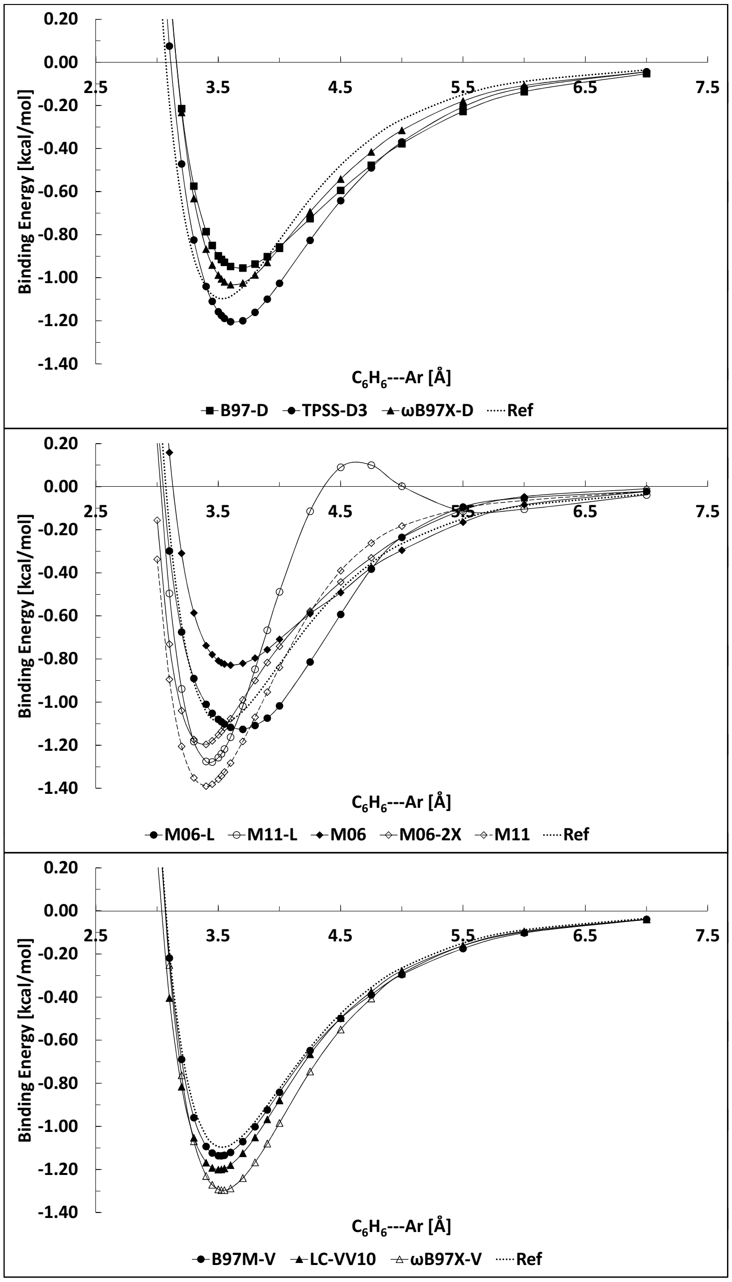


Figure 4.5: Potential energy curves for the benzene-argon dimer from the BzDC215 dataset computed in the aug-cc-pVTZ basis set with the (250,590)/SG-1 grid.

contains the total RMSD for all of the PECs except the rare-gas dimer PECs, for a total of 78. This column was necessary because the rare-gas dimer bond lengths for all of the Minnesota density functionals and ωB97X-D were extremely long and therefore grossly affected the overall RMSD. As a complement to Table 4.11, Table 4.12 contains the same statistics for the associated equilibrium binding energies.

The NBC10 dataset contains weakly-bound dimers such as the methane dimer, the benzene-methane dimer, 3 orientations of the benzene dimer, and 2 orientations of the pyridine dimer. The VV10 density functional predicts incredibly accurate bond lengths for these systems, with an RMSD of 0.012 Å. The next best density functional is B97M-V, with an RMSD (0.03 Å) that is still impressive. On the other hand, Table 4.12 indicates that the equilibrium binding energies of VV10 are 2 times less accurate than those of B97M-V for NBC10. Additional noteworthy performances for the NBC10 bond lengths include B3LYP-D3, LC-VV10, and ωB97X-V, all of which have RMSDs under 0.05 Å. These 3 density functionals have binding energy RMSDs under 0.15 kcal/mol as well, along with B97-D and B97M-V.

The BzDC215 dataset (which is a subset of the true BzDC215 dataset) contains the following PECs: benzene-HF dimer, benzene-water dimer, benzene-ammonia dimer, benzene-methane dimer, and benzene-HCl dimer. For this dataset, the density functionals with the best bond lengths are M06, ωB97X-V, B97M-V, and TPSS-D3, all of which have RMSDs under 0.03 Å. Of these 4 density functionals, B97M-V has the best equilibrium binding energy RMSD (0.21 kcal/mol). The smallest binding energy RMSD is achieved by M06-L (0.17 kcal/mol), but its bond length RMSD of 0.072 Å is 3 times larger than that of B97M-V.

The S66x8 dataset is a diverse collection of 66 dimers, containing electrostatics-dominated interactions, as well as dispersion-dominated interactions. B97M-V gives the most accurate bond lengths for this dataset, with an RMSD of only 0.02 Å, followed by B3LYP-D3 (0.025 Å), VV10 (0.029 Å), and LC-VV10 (0.03 Å). Most of the other density functionals perform reliably, with the exception of B97-D, M11, and M06-2X, all of which have bond length RMSDs in excess of 0.07 Å. With respect to binding energies, the data further confirms that B97M-V and ωB97X-V give the most accurate equilibrium binding energies, with RMSDs of 0.19 and 0.21 kcal/mol, respectively.

As far as the rare-gas dimers are concerned, B97M-V has the smallest bond length RMSD across all 10 dimers, followed by ωB97X-V, VV10, and B3LYP-D3. These 4 density functionals are the only ones with RMSDs smaller than 0.1 Å. Five of the remaining density functionals have RMSDs between 0.1 and 0.2 Å, with the remaining 6 having RMSDs in excess of 0.2 Å: M06-L, M06-2X, ωB97X-D, M06, M11, and M11-L. Given this data, it is straightforward to conclude that systematically-optimized density functionals that are not tested for transferability during fitting (such as ωB97X-D and the Minnesota density functionals) are more likely to perform poorly for the rare-gas dimers. Besides B97M-V and ωB97X-V (both of which were almost guaranteed to perform well for the rare-gas dimers via the transferability tests), only 2 systematically-optimized density functionals (B97-D and B97-D2) perform reasonably well for the rare-gas dimers, most likely due to the conservative $m = 2$ truncation that was employed for their ICFs.

The overall statistics can identify the best density functionals for these types of interactions. Considering the second-to-last column of both Tables 4.11 and 4.12, it is clear that B97M-V has the most accurate bond lengths overall, with an RMSD of 0.028 Å. The next best density functional, VV10, has an RMSD that is 0.01 Å larger, followed by ωB97X-V (0.04 Å) and B3LYP-D3 (0.04 Å). These 4 are the only density functionals that have overall bond length RMSDs under 0.05 Å. With respect to the overall equilibrium binding energies, B97M-V and ωB97X-V are the clear winners, with RMSDs of 0.17 and 0.2 kcal/mol, respectively. The next best density functional (M06-L) has an RMSD of 0.31 kcal/mol, and has an overall bond length RMSD of 0.09 Å (more than 3 times worse than that of B97M-V).

To assess the ability of B97M-V to optimize accurate geometries, 3 sets of geometries were benchmarked in the aug-cc-pVTZ basis set and the results are shown in Table 4.13. The first set, taken from the work of Tentscher and Arey[167], contains 18 bond lengths of 18 small radicals. The *ab Initio* Best Estimate bond lengths from the second-to-last column of Table 1 in Reference 167 are taken as the reference in order to generate the MADs shown in Table 4.13. For this set of geometries, B97M-V and M06-L perform indistinguishably, with bond length MADs of 0.55 pm. The next best density functionals are B3LYP-D3 and B97-D2, with MADs of 0.69 pm, while the worst performers are M11 and M11-L, with MADs of 1.33 and 1.87 pm, respectively. The next set of geometries, taken from Bak et al.[168], contains 28 bond lengths of 19 small molecules. The CCSD(T)/cc-pCVQZ bond lengths from the last column of Table II in Reference 168 are taken as the reference in order to generate the MADs shown in Table 4.13. For this set of geometries, M06-L and B3LYP-D2 (MADs under 0.5 pm) perform very well, followed by B97-D2 and B97M-V (MADs under 0.6 pm). The (99,590)/SG-1 grid was used for both the Arey and Bak datasets.

Moving on to intermolecular geometry optimizations, the A19 dataset includes 19 of the 21 equilibrium geometries from A24. Two dimers (water-methane and methane-ethene) were removed because several density functionals optimized these systems to a different orientation than that of the reference structure. The optimizations for the A19 dataset were carried out with the (150,770)/SG-1 grid and the symmetry of the reference structure was preserved. The metric used to assess the performance of the density functionals is the MAD of 19 RMSDs generated by the Kabsch algorithm[166] (calculated by comparing the reference and final geometries).

Å	NBC10	BzDC215	S66x8	Rare-gas	All	All*
# of Data Points	7	5	66	10	88	78
PBE-D3	0.094	0.035	0.064	0.163	0.083	0.066
B97-D	0.063	0.073	0.070	0.168	0.086	0.069
VV10	0.012	0.032	0.029	0.084	0.038	0.028
TPSS-D3	0.071	0.028	0.057	0.165	0.077	0.057
M06-L	0.072	0.072	0.040	0.233	0.090	0.046
M11-L	0.817	0.042	0.035	1.532	0.566	0.247
B97M-V	0.030	0.024	0.020	0.058	0.028	0.021
B3LYP-D3	0.039	0.043	0.025	0.089	0.040	0.028
B97-D2	0.094	0.047	0.043	0.140	0.067	0.050
M06	0.258	0.020	0.041	0.500	0.187	0.086
M06-2X	0.105	0.116	0.074	0.243	0.112	0.081
LC-VV10	0.040	0.089	0.030	0.152	0.062	0.037
ωB97X-D	0.051	0.042	0.039	0.399	0.140	0.040
ωB97X-V	0.042	0.022	0.036	0.065	0.040	0.036
M11	0.062	0.123	0.071	0.614	0.219	0.075

Table 4.11: Equilibrium bond length RMSDs in Å for interpolated potential energy curve (PEC) minima from 5 of the datasets in the training and primary test sets, as well as an additional dataset containing PECs for all 10 rare-gas dimers (computed in the aug-cc-pVQZ basis set with the (250,590)/(75,302) grid) containing helium, neon, argon, and krypton. NBC10 is a combination of the NBC10-1, NBC10-2, and NBC10-3 datasets from the training and primary test sets. The column labeled "All" contains RMSDs for all 4 datasets, while the column labeled "All*" excludes the rare-gas dimer data.

Of the benchmarked density functionals, ωB97X-V is clearly the best performer, with an MAD of only 1.4 pm, while B97M-V comes in second, with an MAD of 2.35 pm. Following B97M-V, 3 density functionals, B97-D2, B3LYP-D3, and LC-VV10, have MADs between 2.5 pm and 3 pm. The next best local density functional after B97M-V is B97-D, with an MAD (4.92 pm) that is more than twice as large as that of B97M-V. Overall, the performance of B97M-V for both intramolecular and intermolecular geometry optimizations is impressive, since these are features that are not necessarily guaranteed from fitting to energetics.

kcal/mol	NBC10	BzDC215	S66x8	Rare-gas	All	All*
# of Data Points	7	5	66	10	88	78
PBE-D3	0.24	0.49	0.54	0.11	0.49	0.52
B97-D	0.12	0.55	0.52	0.06	0.47	0.50
VV10	0.25	0.30	0.63	0.04	0.55	0.59
TPSS-D3	0.18	0.44	0.40	0.08	0.37	0.39
M06-L	0.38	0.17	0.33	0.13	0.31	0.33
M11-L	0.46	0.19	0.73	0.19	0.65	0.69
B97M-V	0.12	0.21	0.19	0.03	0.17	0.18
B3LYP-D3	0.13	0.49	0.52	0.05	0.47	0.50
B97-D2	0.42	0.51	0.36	0.09	0.36	0.38
M06	0.60	0.28	0.52	0.12	0.48	0.51
M06-2X	0.52	0.78	0.31	0.11	0.36	0.38
LC-VV10	0.11	0.64	0.40	0.07	0.38	0.41
ωB97X-D	0.38	0.57	0.59	0.08	0.54	0.57
ωB97X-V	0.11	0.36	0.21	0.03	0.20	0.21
M11	0.75	0.80	0.39	0.16	0.45	0.47

Table 4.12: Equilibrium binding energy RMSDs in kcal/mol for interpolated potential energy curve (PEC) minima from 5 of the datasets in the training and primary test sets, as well as an additional dataset containing PECs for all 10 rare-gas dimers (computed in the aug-cc-pVQZ basis set with the (250,590)/(75,302) grid) containing helium, neon, argon, and krypton. NBC10 is a combination of the NBC10-1, NBC10-2, and NBC10-3 datasets from the training and primary test sets. The column labeled "All" contains RMSDs for all 4 datasets, while the column labeled "All*" excludes the rare-gas dimer data.

4.4 Using B97M-V

4.4.1 Basis Sets

Even though B97M-V was trained in the aQZ basis set for thermochemistry and the aTZ basis set (without counterpoise corrections) for non-covalent interactions, it is inevitable that it will be used with different basis sets. As a result, this section explains how B97M-V should be used and what basis sets are recommended.

In order to assess the basis set dependence of thermochemical quantities, the data points in the W4-11 dataset were computed in a variety of basis sets with the (99,590)/SG-1 grid: cc-pVDZ (DZ), cc-pVTZ (TZ), cc-pVQZ (QZ), 6-311++G(3df,3pd) (LP), pc-1 (pc1), pc-2 (pc2), pc-3 (pc3), def2-SVP (SVP), def2-TZVP (TZVP), def2-QZVP (QZVP), and aug-cc-pVQZ (aQZ). The goal was to identify basis sets substantially smaller than aQZ that can provide results similar in quality to the aQZ

pm	Arey	Bak	A19
# of Data Points	18	28	19
PBE-D3	1.13	1.01	5.57
B97-D	0.90	0.75	4.92
VV10	1.21	1.06	5.87
M06-L	0.55	0.42	5.58
M11-L	1.87	1.78	5.08
B97M-V	0.55	0.57	2.35
B3LYP-D3	0.69	0.48	2.64
B97-D2	0.69	0.52	2.62
M06	1.10	0.87	5.93
M06-2X	1.20	0.87	5.23
LC-VV10	1.31	1.08	2.93
ωB97X-D	0.90	0.71	3.80
ωB97X-V	0.91	0.69	1.40
M11	1.33	0.96	4.25

Table 4.13: MADs in picometers (pm) for the 3 geometry datasets discussed in Section 4.3.5. For the Arey and Bak datasets, the entries are bond length MADs and the optimizations were carried out in the aug-cc-pVTZ basis set with the (99,590)/SG-1 grid. For the A19 dataset, the optimizations were carried out with the (150,770)/SG-1 grid in the aug-cc-pVTZ basis set and the errors are calculated by using the Kabsch algorithm. The A19 column contains MADs of 19 RMSDs calculated by the Kabsch algorithm.

results. The corresponding data is shown in Table 4.14. The values in the second row correspond to the number of basis functions (BF) the given basis set has for propane, while the values in the following row are the total atomization energies (TAE) of propane computed with the given basis set. The W4-11 reference value for the TAE of propane is -1007.91 kcal/mol, while the B97M-V aQZ value is -1007.28 kcal/mol. Considering just the values from the third row of Table 4.14, it is clear that the DZ, pc1, SVP, and TZVP basis sets are insufficient for the calculation of thermochemical quantities. However, in order to arrive at a more solid conclusion, the RMSDs for the 5 subdatasets in the W4-11 dataset are shown in the fourth through eighth rows of Table 4.14. In addition, the last 6 rows contain overall statistics for all of W4-11 with respect to both the aQZ values as well as the reference values.

Considering both performance with respect to the aQZ basis set as well as the reference, the following basis sets can be potentially recommended as alternatives to

kcal/mol	DZ	TZ	QZ	LP	pc1	pc2	pc3	SVP	TZVP	QZVP	aQZ
BF (C_3H_8)	82	202	405	261	82	202	464	82	141	411	608
TAE (C_3H_8)	-1004.83	-1008.49	-1007.78	-1007.82	-1000.09	-1008.61	-1008.09	-1020.79	-1005.43	-1007.73	-1007.28
HAT707	8.75	4.73	4.08	4.08	7.28	4.19	3.85	8.08	4.19	3.94	3.99
BDE99	5.74	3.76	3.57	3.64	5.29	3.56	3.64	5.70	3.70	3.59	3.52
TAE_nonMR124	11.72	4.17	3.74	3.92	10.84	3.42	3.85	7.78	4.11	3.88	3.79
SN13	4.65	1.66	1.47	1.35	3.51	1.41	1.37	3.16	1.56	1.41	1.39
ISOMER20	3.87	3.10	3.07	3.01	3.80	3.07	3.01	4.06	3.26	3.03	3.00
RMSD vs. aQZ	6.86	1.36	0.45	0.80	5.98	0.92	1.07	5.47	1.07	0.96	0.00
RMSD vs. Reference	8.91	4.47	3.92	3.95	7.70	3.94	3.78	7.66	4.07	3.84	3.85
MAD vs. aQZ	4.51	0.98	0.34	0.58	4.26	0.69	0.55	3.87	0.72	0.52	0.00
MAD vs. Reference	6.25	3.38	2.95	2.96	5.77	2.94	2.84	5.72	3.12	2.89	2.90
MSE vs. aQZ	1.33	-0.04	-0.13	-0.32	1.02	-0.27	-0.27	-0.91	-0.05	-0.20	0.00
MSE vs. Reference	0.30	-1.07	-1.16	-1.34	-0.01	-1.30	-1.30	-1.94	-1.08	-1.23	-1.03

Table 4.14: RMSDs in kcal/mol computed with the (99,590)/SG-1 grid for the 5 subdatasets comprising the W4-11 dataset in a variety of basis sets: cc-pVDZ (DZ), cc-pVTZ (TZ), cc-pVQZ (QZ), 6-311++G(3df,3pd) (LP), pc-1 (pc1), pc-2 (pc2), pc-3 (pc3), def2-SVP (SVP), def2-TZVP (TZVP), def2-QZVP (QZVP), and aug-cc-pVQZ (aQZ). The last 6 rows contain statistics for the entire W4-11 dataset with respect to both the basis set limit values (aQZ) as well as the reference values. The second row indicates the number of basis functions (BF) the given basis set has for propane, while the third row contains the total atomization energy (TAE) of propane in kcal/mol. The W4-11 reference value for the TAE of propane is -1007.91 kcal/mol.

the aQZ basis set for thermochemistry (pending further tests): QZ, LP, pc2, pc3, and QZVP. From these recommended basis sets, the smallest one is pc2, which only has 202 basis functions for propane, while aQZ has 608. Thus, even though B97M-V was trained in the aQZ basis set for thermochemistry, it might be sufficient to use a basis set that is 3 times smaller to arrive at results that are of aQZ quality.

Since the molecules in the W4-11 dataset are relatively small, the 5 promising basis sets from the W4-11 study were applied to larger interactions via the AlkAtom19, AlkIsomer11, and AlkIsod14 datasets. The results (computed with the (75,302)/SG-0 grid) are shown in Table 4.15. The second row of Table 4.15 lists the number of basis functions the given basis set has for octane, while the third row contains the TAE of octane (the reference is -2482.64 kcal/mol and the B97M-V aQZ result is -2480.94 kcal/mol). At first glance, all of the basis sets appear to predict a reasonable value for the TAE of octane, with the largest deviation being the aQZ result itself (0.07% error). In fact, in the aQZ basis set, B97M-V slightly underestimates the atomization energies of the larger alkanes in AlkAtom19, with an RMSD of 1.35 kcal/mol and an MSE (mean signed error) of 0.73 kcal/mol. Decreasing the size of the basis set increases the effect of basis set superposition error (BSSE) and leads to slightly more binding, resulting in a 7-fold improvement in the AlkAtom19 RMSD with the pc3 basis set. The isomerization energies are relatively insensitive to the

kcal/mol	QZ	LP	pc2	pc3	QZVP	aQZ
BF (C_8H_{18})	980	636	492	1124	996	1468
TAE (C_8H_{18})	-2481.93	-2481.88	-2483.40	-2482.73	-2481.37	-2480.94
AlkAtom19	0.60	0.61	0.67	0.19	0.99	1.35
AlkIsomer11	0.21	0.20	0.11	0.30	0.27	0.21
AlkIsod14	0.54	0.43	0.39	0.64	0.62	0.51
RMSD vs. aQZ	0.50	0.50	1.31	0.86	0.28	0.00
RMSD vs. Reference	0.51	0.48	0.49	0.41	0.75	0.94
MAD vs. aQZ	0.33	0.35	0.88	0.60	0.22	0.00
MAD vs. Reference	0.42	0.38	0.42	0.30	0.60	0.73
MSE vs. aQZ	-0.31	-0.35	-0.88	-0.48	-0.13	0.00
MSE vs. Reference	0.42	0.38	-0.16	0.25	0.60	0.73

Table 4.15: RMSDs in kcal/mol computed with the (75,302)/SG-0 grid for 3 datasets from the primary test set in a variety of basis sets: cc-pVQZ (QZ), 6-311++G(3df,3pd) (LP), pc-2 (pc2), pc-3 (pc3), def2-QZVP (QZVP), and aug-cc-pVQZ (aQZ). The last 6 rows contain statistics for the 3 combined datasets with respect to both the basis set limit values (aQZ) as well as the reference values. The second row indicates the number of basis functions (BF) the given basis set has for octane, while the third row contains the total atomization energy (TAE) of octane in kcal/mol. The AlkAtom19 reference value for the TAE of octane is -2482.64 kcal/mol.

basis set, as are the isodesmic reaction energies. **Based on these results, the cc-pVQZ, 6-311++G(3df,3pd), pc-2, pc-3, and def2-QZVP basis sets can be recommended as alternatives to the aug-cc-pVQZ basis set for thermochemistry.**

Moving on to non-covalent interactions, the A24 dataset was used as an initial test of the basis set convergence of small molecules. The calculations were performed with the (99,590)/SG-1 grid and a total of 25 basis sets were assessed: cc-pVNZ for N={D,T,Q,5} (NZ), aug-cc-pVNZ for N={D,T,Q,5} (aNZ), pc-N for N={0,1,2,3,4} (pcN), aug-pc-N for N={0,1,2,3,4} (apcN), def2-NVP for N={S,TZ,QZ} (NVP), def2-NVPD for N={S,TZ,QZ} (NVPD), and 6-311++G(3df,3pd) (LP). The results are summarized in Table 4.16. For each basis set, 2 types of statistics with respect to the reference values are reported: RMSDs and MSEs. Furthermore, the results are presented both with counterpoise corrections (CP), without counterpoise corrections (noCP), as well as their average (AVG=(CP+noCP)/2). Since B97M-V was trained at the noCP aTZ basis set level, the corresponding RMSD of 0.17 kcal/mol is a useful guide for assessing the performance of B97M-V in the remaining 24 basis sets. Immediately, it is clear that the pc0 and apc0 basis sets are incompatible with B97M-V, both with and without counterpoise corrections. In addition, the DZ, pc1, and SVP basis sets have RMSDs greater than 0.25 kcal/mol with counterpoise corrections, and are unacceptable without counterpoise corrections. Basis sets that perform well both with and without counterpoise corrections include 5Z, aTZ, aQZ, a5Z, pc3, pc4, apc2, apc3, apc4, TZVPD, QZVP, QZVPD, and LP, while basis sets that perform well only with counterpoise corrections include TZ, QZ, aDZ, pc2, apc1, SVPD, and TZVP.

For comparison to B97M-V, the last row of Table 4.16 contains data corresponding to the M06-L density functional computed in the aTZ basis set. While the RMSD of B97M-V improves very slightly from 0.17 kcal/mol to 0.15 kcal/mol when going from noCP aTZ to CP aTZ, the reverse happens for M06-L: the RMSD increases from 0.23 kcal/mol to 0.41 kcal/mol. The fact that the CP and noCP a5Z RMSDs of B97M-V closely match the CP and noCP aTZ RMSDs indicates that for intermolecular interactions involving small molecules, the aTZ basis set can be considered close to the basis set limit for B97M-V. However, since the largest dimer in the A24 dataset is the ethene dimer (the number of basis functions for each basis set is listed in the second column of Table 4.16), the promising basis sets that were smaller than or equivalent to aTZ were applied to larger intermolecular interactions via the S22 and S66 datasets.

The S22 data (Table 4.17) looks significantly different from the A24 data. As a reminder, the S22 dataset was included in the training set of B97M-V with a weight of 100. The noCP aTZ RMSD of 0.23 kcal/mol is the smallest RMSD possible, while the pc2, TZVPD, and aQZ basis sets look promising without counterpoise corrections, presenting a degradation of 20% at most compared with the noCP aTZ RMSD. However, the CP aTZ RMSD of 0.33 kcal/mol presents a degradation of 0.1 kcal/mol (40%) from the noCP aTZ result. Considering the aQZ basis set results, it is clear that the noCP and CP results have not converged, indicating that it is much more difficult to converge the binding energies of larger interactions. Additionally, since the interactions in S22 were heavily emphasized in training the parameters of B97M-V, it is possible that going to a larger basis set is leading to underbinding. Considering the CP MSE and noCP MSE values for aTZ and aQZ provides confirmation of this effect. Based on the S22 data, only the TZVPD basis set (to be used without counterpoise corrections) can be recommended as an alternative to noCP aTZ for non-covalent interactions.

For comparison to B97M-V, the last row of Table 4.17 contains data corresponding to the M06-L density functional computed in the aTZ basis set. While the RMSD of B97M-V degrades by 40% when going from noCP aTZ to CP aTZ, the RMSD of M06-L worsens by more than a factor of 2.5.

kcal/mol	BF	CP RMSD	noCP RMSD	AVG RMSD	CP MSE	noCP MSE	AVG MSE
DZ	96	0.29	1.27	0.63	0.16	-0.86	-0.35
TZ	232	0.19	0.65	0.36	0.03	-0.42	-0.19
QZ	460	0.17	0.34	0.22	0.02	-0.20	-0.09
5Z	804	0.15	0.17	0.16	0.01	-0.04	-0.02
aDZ	164	0.16	0.29	0.19	0.03	-0.21	-0.09
aTZ	368	0.15	0.17	0.16	-0.01	-0.05	-0.03
aQZ	688	0.15	0.15	0.15	0.00	-0.03	-0.01
a5Z	1148	0.15	0.15	0.15	0.00	-0.01	0.00
pc0	52	0.95	3.75	2.30	-0.32	-2.45	-1.38
pc1	96	0.25	0.87	0.51	0.00	-0.56	-0.28
pc2	232	0.17	0.28	0.22	0.00	-0.13	-0.06
pc3	528	0.15	0.15	0.15	0.00	-0.01	0.00
pc4	940	0.15	0.15	0.15	0.01	0.00	0.00
apc0	76	1.00	1.96	1.44	-0.43	-1.39	-0.91
apc1	164	0.17	0.49	0.26	0.06	-0.41	-0.18
apc2	368	0.15	0.18	0.16	0.00	-0.09	-0.04
apc3	756	0.15	0.15	0.15	0.00	-0.02	-0.01
apc4	1284	0.15	0.15	0.15	0.00	0.00	0.00
SVP	96	0.26	1.26	0.67	0.10	-0.90	-0.40
TZVP	172	0.21	0.41	0.30	-0.03	-0.22	-0.13
QZVP	468	0.15	0.18	0.16	0.01	-0.05	-0.02
SVPD	144	0.16	0.52	0.28	0.02	-0.47	-0.23
TZVPD	220	0.15	0.15	0.15	0.00	-0.06	-0.03
QZVPD	516	0.15	0.15	0.15	0.00	-0.01	0.00
LP	300	0.15	0.24	0.19	0.01	-0.11	-0.05
aTZ*	368	0.41	0.23	0.31	0.37	0.15	0.26

Table 4.16: RMSDs and MSEs in kcal/mol for the A24 dataset computed in 25 basis sets. The abbreviations are explained in Section 4.4.1. The second column contains the number of basis functions (BF) for the largest interaction in A24: the ethene dimer. The interactions were computed with the (99,590)/SG-1 grid with counterpoise corrections (CP), without counterpoise corrections (noCP), as well as their average (AVG). The last row contains data for the M06-L density functional in the aTZ basis set for comparison to the B97M-V results.

kcal/mol	BF	CP RMSD	noCP RMSD	AVG RMSD	CP MSE	noCP MSE	AVG MSE
DZ	321	0.75	1.91	0.67	0.64	-1.45	-0.41
TZ	724	0.40	0.75	0.34	0.30	-0.54	-0.12
aDZ	536	0.44	0.55	0.31	0.34	-0.35	-0.01
aTZ	1127	0.33	0.23	0.26	0.24	0.04	0.14
pc1	321	0.53	1.81	0.95	0.15	-1.46	-0.66
pc2	724	0.36	0.27	0.25	0.27	-0.05	0.11
apc1	536	0.55	2.15	0.95	0.43	-1.68	-0.62
apc2	1127	0.34	0.40	0.25	0.25	-0.18	0.04
SVP	321	0.53	2.05	0.89	0.44	-1.56	-0.56
SVPD	474	0.34	1.57	0.70	0.26	-1.21	-0.48
TZVP	655	0.37	0.41	0.29	0.24	-0.21	0.01
TZVPD	808	0.35	0.25	0.28	0.28	0.07	0.17
LP	939	0.38	0.35	0.29	0.28	-0.09	0.09
aQZ	2026	0.34	0.27	0.30	0.25	0.15	0.20
QZVPD	1566	0.34	0.31	0.32	0.26	0.20	0.23
aTZ*	1127	1.08	0.43	0.73	0.96	0.30	0.63

Table 4.17: RMSDs and MSEs in kcal/mol for the S22 dataset computed in 15 basis sets. The abbreviations are explained in Section 4.4.1. The second column contains the number of basis functions (BF) for the largest interaction in S22: the adenine-thymine complex. The interactions were computed with the (75,302)/SG-0 grid with counterpoise corrections (CP), without counterpoise corrections (noCP), as well as their average (AVG). The last row contains data for the M06-L density functional in the aTZ basis set for comparison to the B97M-V results.

Since the S22 dataset was in the training set, it is important to test a dataset of larger interactions from one of the test sets in order to assure that the effect seen with the S22 data is less pronounced for datasets that were not included in the training set. Accordingly, the results for the S66 dataset are provided in Table 4.18. The S66 data is much more reasonable, as the CP and noCP RMSDs for the aTZ basis set differ by no more than 0.02 kcal/mol. While the noCP aTZ RMSD is 0.19 kcal/mol, the lowest possible RMSD is achieved in the TZVPD basis set without counterpoise corrections (0.18 kcal/mol). According to the S66 data, 2 basis sets that are substantially smaller than aTZ can be employed without counterpoise corrections to match the quality of the noCP aTZ result: pc2 and TZVPD. In addition, it appears that a variety of basis sets can be employed with counterpoise corrections: pc2, apc2, SVPD, and TZVPD. **Based on the results from the A24, S22, and S66 datasets, the def2-SVPD basis set (to be used with counterpoise corrections) and the def2-TZVPD basis set (to be used without counterpoise corrections) can be recommended as alternatives to the non-counterpoise-corrected aug-**

kcal/mol	BF	CP RMSD	noCP RMSD	AVG RMSD	CP MSE	noCP MSE	AVG MSE
DZ	260	0.67	1.69	0.66	0.52	-1.39	-0.43
TZ	636	0.32	0.73	0.31	0.21	-0.60	-0.19
aDZ	446	0.28	0.54	0.28	0.19	-0.43	-0.12
aTZ	1012	0.21	0.19	0.17	0.13	-0.04	0.05
pc1	260	0.37	1.34	0.66	0.10	-1.21	-0.56
pc2	636	0.22	0.21	0.15	0.15	-0.13	0.01
apc1	446	0.35	1.58	0.69	0.28	-1.35	-0.54
apc2	1012	0.21	0.32	0.19	0.13	-0.20	-0.03
SVP	260	0.49	1.75	0.78	0.36	-1.46	-0.55
SVPD	392	0.21	1.21	0.55	0.14	-1.10	-0.48
TZVP	454	0.24	0.39	0.22	0.13	-0.28	-0.08
TZVPD	586	0.22	0.18	0.18	0.16	-0.02	0.07
LP	822	0.24	0.25	0.18	0.15	-0.14	0.00
aQZ	1904	0.22	0.18	0.19	0.14	0.05	0.09
QZVPD	1422	0.22	0.20	0.21	0.14	0.10	0.12
aTZ*	1012	0.81	0.36	0.55	0.74	0.17	0.46

Table 4.18: RMSDs and MSEs in kcal/mol for the S66 dataset computed in 15 basis sets. The abbreviations are explained in Section 4.4.1. The second column contains the number of basis functions (BF) for the largest interaction in S66: the pentane dimer. The interactions were computed with the (75,302)/SG-0 grid with counterpoise corrections (CP), without counterpoise corrections (noCP), as well as their average (AVG). The last row contains data for the M06-L density functional in the aTZ basis set for comparison to the B97M-V results.

cc-pVTZ basis set for non-covalent interactions. Furthermore, for systems where the use of diffuse functions may lead to issues with linear dependence, the pc-2 basis set is recommended for use without counterpoise corrections.

4.4.2 Grids

Different density functionals converge to their “infinite grid” limit at different rates. To ensure that B97M-V does not converge too slowly, it was mentioned in Section 4.2.5 that the billions of fits were filtered such that the least-squares fit energies generated in the (99,590)/SG-1 and (250,590)/SG-1 grids differed by a maximum of 0.01 kcal/mol. The effectiveness of this decision is tested by analyzing the grid sensitivity of B97M-V relative to other density functionals on the methane dimer PEC from NBC10-2, calculated in the aTZ basis set. The results are shown in Figure 4.6, with the associated maximum absolute deviations (with respect to the (250,590) grid) in cal/mol in Table 4.19. For the VV10 NLC functional, the SG-1 grid is used and not varied.

The B97M-V PEC looks nearly acceptable in the SG-1 grid, and is fully converged in the (75,302) grid. The same can be said for all of the density functionals, except M06-L, M06, and M11. The density functional with the smallest MAX in the SG-1 grid (11.36 cal/mol) is ωB97X-V, while B97M-V and the non-empirical PBE density functional have maximum absolute deviations of approximately 27 cal/mol. For comparison, the least grid-sensitive Minnesota density functional (M06-2X) has a MAX of around 45 cal/mol in the SG-1 grid, with M11, M06-L, and M06 having deviations between 146 and 190 cal/mol. In the (75,302) grid, the MAX for most of the density functionals drops below 10 cal/mol. The outliers include M06-L, M06, and M11, and this result is repeated in the (99,590) grid. In fact, the M06 potential energy curve computed with the finest grid does not exhibit a proper well shape, indicating that it is not at its “infinite grid” limit even with the (250,590) grid. Furthermore, the strange behavior that was seen for M11-L with the benzene-argon dimer is much more pronounced in the case of the methane dimer, and it is clear that the feature cannot be eliminated by using a finer grid.

B97M-V is the only semi-empirical meta-GGA density functional out of those considered that is as grid-insensitive as its GGA counterparts. In fact, in the (75,302) grid, its MAX (4.57 cal/mol) is considerably smaller than that of 3 of the tested GGA density functionals: B97-D (10.79 cal/mol), B97-D2 (6.77 cal/mol), and ωB97X-D (6.81 cal/mol). Furthermore, out of the 7 meta-GGA density functionals considered, B97M-V is the least grid-sensitive density functional (even when compared to the non-empirical TPSS density functional).

Based on the data from Figure 4.6 and Table 4.19, it appears as if the (75,302)/SG-1 grid can be recommended for B97M-V. However, the methane dimer test only covers a very small fraction of the types of interactions that B97M-V can be applied to. In order to be absolutely certain of this recommendation, all of the data points in the training, primary test, and secondary test sets (with the exception of the absolute atomic energies and the rare-gas dimer PECs) were computed with the following grids: SG-1/SG-1, (75,302)/SG-1, (99,590)/SG-1, and (250,590)/SG-1.

Table 4.20 summarizes the results of this comprehensive test, which are categorized with respect to 3 metrics: absolute percent error (APE), absolute error (AE), as well as their product (APE·AE). With the assumption that the (250,590)/SG-1 results are fully converged with respect to the grid, Table 4.20 was populated with data from the aforementioned 2329 data points from the training, primary test, and secondary test sets. Starting with the (99,590)/SG-1 grid, it is clear that the filtering applied during the training stage has completely transferred to the final functional form, since all 2329 data points have absolute errors less than 0.01 kcal/mol. The SG-

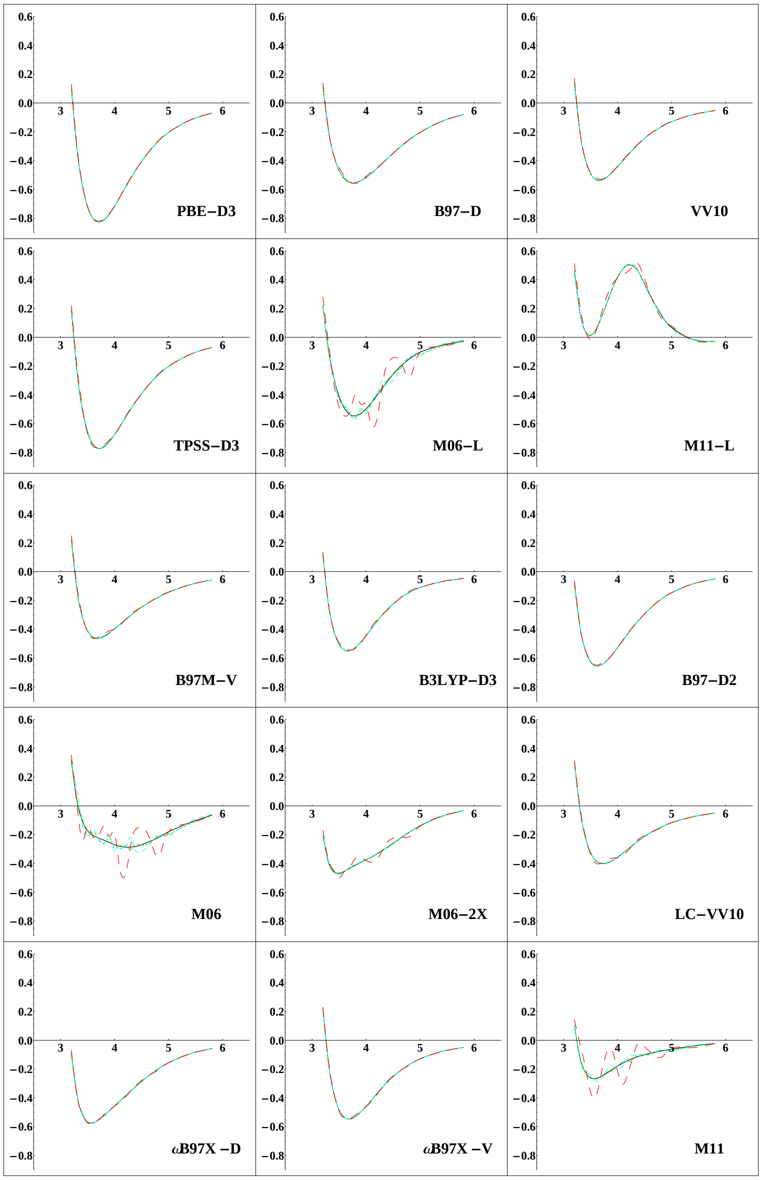


Figure 4.6: Methane dimer potential energy curves for 15 density functionals with 4 different local exchange-correlation grids (SG-1, (75,302), (99,590), and (250,590)), calculated in the aug-cc-pVTZ basis set without counterpoise corrections. The x-axis shows bond length in Å and the y-axis shows binding energy in kcal/mol. The color scheme is as follows: SG-1 (Red with Large Dashes), (75,302) (Cyan; Medium Dashes), (99,590) (Green; Small Dashes), (250,590) (Black; No Dashes). For the VV10 NLC functional, the SG-1 grid is used and not varied.

cal/mol	SG-1	(75,302)	(99,590)
PBE-D3	27.06	4.39	0.34
B97-D	22.62	10.79	1.97
VV10	27.94	4.44	0.68
TPSS-D3	34.77	7.91	2.12
M06-L	165.54	56.24	23.32
M11-L	54.86	8.82	5.66
B97M-V	27.79	4.57	1.52
B3LYP-D3	19.61	3.75	0.40
B97-D2	18.90	6.77	1.10
M06	189.29	68.93	29.78
M06-2X	44.84	11.49	4.87
LC-VV10	20.17	3.25	1.01
ωB97X-D	18.89	6.81	2.11
ωB97X-V	11.36	1.91	0.57
M11	146.59	24.55	10.61

Table 4.19: Maximum absolute deviations in cal/mol for the 18 points on the methane dimer potential energy curve from NBC10-2. The deviations are taken with respect to the (250,590) grid results. For the VV10 NLC functional, the SG-1 grid is used and not varied.

1/SG-1 results are clearly unacceptable, with 9 data points having APEs larger than 100% and 507 data points having an AE of more than 0.1 kcal/mol. The (75,302)/SG-1 grid strikes a reasonable balance between the SG-1/SG-1 and (99,590)/SG-1 grids, and the largest absolute errors are no greater than 0.1 kcal/mol. However, an outlier appears that has an APE of more than 50%. This specific data point is in the S22x5 dataset and corresponds to the parallel-displaced benzene dimer at a separation of 2.0·R$_e$. This data point was investigated further in order to identify grids that can reduce this outlying APE.

Table 4.21 contains the APEs for this investigation in a variety of radial (columns) and angular (rows) grids. The APEs are calculated with respect to the (500,974)/SG-1 grid and it is clear that the accuracy for this specific data point relies entirely on the number of angular grid points. Accordingly, the coarsest grid that can substantially reduce the 55.6% error of the (75,302)/SG-1 grid is the (75,590)/SG-1 grid, which reduces the APE to less than 3%. As a consequence of this test, the (75,590)/SG-1 grid was added to Table 4.20 in order to assess its performance for the remaining 2328 data points. As predicted, the (75,590)/SG-1 grid substantially improves upon the coarser (75,302)/SG-1 grid, with a maximum APE of only 5.5% for all 2329 data

APE (%)	[0,1)	[1,2)	[2,5)	[5,10)	[10,50)	[50,100)	[100,∞)
SG-1/SG-1	1699	233	213	67	82	26	9
(75,302)/SG-0	2184	87	40	9	8	1	0
(75,302)/SG-1	2213	72	25	12	6	1	0
(75,590)/SG-1	2298	22	7	2	0	0	0
(99,590)/SG-1	2328	1	0	0	0	0	0
AE (kcal/mol)	[0,0.01)	[0.01,0.02)	[0.02,0.05)	[0.05,0.10)	[0.1,0.5)	[0.5,1)	[1,∞)
SG-1/SG-1	623	320	517	362	441	52	14
(75,302)/SG-0	1950	246	119	14	0	0	0
(75,302)/SG-1	1996	238	87	8	0	0	0
(75,590)/SG-1	2270	47	12	0	0	0	0
(99,590)/SG-1	2329	0	0	0	0	0	0
APE·AE (%·kcal/mol)	[0,0.1)	[0.1,0.2)	[0.2,0.5)	[0.5,1)	[1,5)	[5,10)	[10,∞)
SG-1/SG-1	1900	98	109	66	96	16	44
(75,302)/SG-0	2303	13	7	4	2	0	0
(75,302)/SG-1	2309	9	7	3	1	0	0
(75,590)/SG-1	2327	2	0	0	0	0	0
(99,590)/SG-1	2329	0	0	0	0	0	0

Table 4.20: Error ranges for 2329 data points from the training, primary test, and secondary tests sets. From the original 2460 data points, the 8 data points from AE8 as well as the 123 data points corresponding to the rare-gas dimer PECs were dropped. The errors are taken with respect to the (250,590)/SG-1 grid. The grids are assessed with respect to 3 metrics: absolute percent error (APE), absolute error (AE), as well as their product (APE·AE).

points.

Finally, in order to test the sensitivity of the VV10 NLC functional to the integration grid, the SG-0 grid was used for the nonlocal integration, along with the (75,302) grid for the integration of the local component of B97M-V. A comparison of the (75,302)/SG-0 and (75,302)/SG-1 grids in Table 4.20 indicates that the integration of the VV10 NLC functional is not very sensitive to the grid, and the SG-0 grid can be applied if necessary. **Based on these results, the (75,302)/SG-0 and (75,302)/SG-1 grids are recommended as the bare minimum for use with B97M-V (particularly for quick calculations), while the (99,590)/SG-1 grid is recommended if results near the "infinite grid" limit are required. Finally, the (75,590)/SG-1 grid can serve as a compromise between these 2 limits.**

APE (%)	75	87	99	150	250	500
302	55.6	56.2	52.7	53.9	54.0	54.0
350	19.7	19.8	20.3	19.3	19.3	19.3
434	20.1	20.6	16.9	20.0	19.9	19.9
590	2.9	3.3	3.4	3.2	3.2	3.2
770	1.0	0.4	1.0	1.1	1.0	1.0
974	0.2	0.2	0.1	0.0	0.0	0.0

Table 4.21: Absolute percent errors (APE) for the S22x5 parallel-displaced benzene dimer at a separation of $2.0{\cdot}R_e$ for a variety of radial (columns) and angular (rows) grids. The APE is calculated with respect to the binding energy in the (500,974) grid. The SG-1 grid is used throughout for integrating the VV10 NLC functional. For reference, the APE for the SG-1 grid is 243.9%.

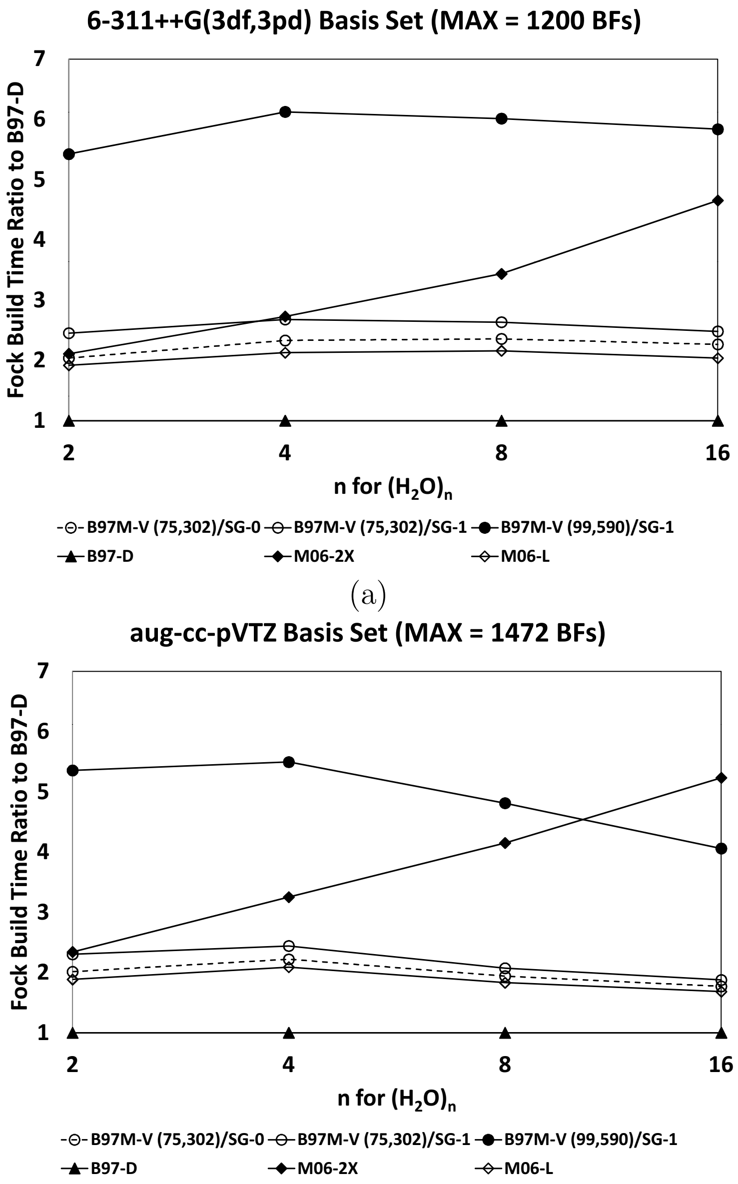
4.4.3 Timings

Since B97M-V does not contain any exact exchange, it has the potential of being much faster than its hybrid counterparts. However, the VV10 NLC functional requires additional grid point evaluations and as a result, B97M-V will be slightly slower than a local meta-GGA density functional like M06-L (for a single Fock build). In order to quantify the additional time required to evaluate the VV10 NLC functional, as a well as to compare the effect of using different grids, timings for a single Fock build were carried out with B97-D, B97M-V, M06-L, and M06-2X. For B97M-V, 3 different grids were used ((75,302)/SG-0, (75,302)/SG-1, and (99,590)/SG-1), while the rest of the density functionals employed the (75,302) grid. The resulting timings, shown in Figure 4.7, are taken as a ratio to the B97-D result, which is the cheapest density functional of the 4 tested. The molecules considered were $(H_2O)_n$ for $n = \{2, 4, 8, 16\}$ in 2 basis sets: LP and aTZ. The results indicate that the additional cost of evaluating the VV10 NLC functional is negligible, when either the SG-0 or SG-1 grids are used in conjunction with the (75,302) grid for the local component. Using the (99,590) grid for the local component of B97M-V is very costly and mostly unnecessary according to the results from the previous section.

As a reference value for those interested in implementing this new density functional, the absolute energy (in hartrees) of hydrogen fluoride with a bond length of 0.9158 Å in the aug-cc-pVTZ basis set with the (75,302)/SG-1 grid is -100.4472797104.

4.5 Conclusions

The primary goal of the development of the B97M-V density functional was to create a minimally-parameterized and highly-transferable local meta-GGA density



(a)

(b)

Figure 4.7: Single Fock build timings for B97-D, B97M-V, M06-L, and M06-2X for water clusters of varying size in 2 different basis sets: 6-311++G(3df,3pd) and aug-cc-pVTZ. The timings are displayed as a ratio to the B97-D value. For B97-D, M06-2X, and M06-L, the (75,302) grid was used, while for B97M-V, timings for 3 different grid combinations are shown. The number of basis functions (BF) for the largest water cluster within each basis set is shown in the title of the associated plot.

Functional	TC RMSD	Functional	NC RMSD
M06-2X	3.21	B97M-V	0.22
ωB97X-V	3.60	ωB97X-V	0.32
ωB97X-D	3.61	M06-L	0.42
B97M-V	3.93	B97-D2	0.48
B97-D2	3.97	ωB97X-D	0.54
M11	3.97	M11	0.55
M06	4.18	M06	0.57
B3LYP-D3	4.66	LC-VV10	0.72
B97-D	5.56	M06-2X	0.77
M06-L	5.63	B3LYP-D3	0.77
TPSS-D3	6.45	B97-D	0.82
M11-L	6.68	TPSS-D3	0.85
LC-VV10	6.79	M11-L	1.08
VV10	9.81	PBE-D3	1.23
PBE-D3	10.10	VV10	1.36

Table 4.22: Density functionals ranked based on their overall unweighted RMSDs in kcal/mol for all thermochemistry (Columns 1-2) and non-covalent interactions (Columns 3-4) data points from the training, primary test, and secondary test sets.

functional that could predict accurate energetics for both bonded and non-bonded interactions. The most important aspects of the development process, and the resulting density functional, may be summarized as follows:

1. Unlike the development of most meta-GGA density functionals, where only one or a small number of functional forms are considered, a combinatorial screening approach was applied to the density functional design problem. Out of a space of almost 10^{40} possible functional forms, over 10 billion were screened for optimal accuracy, transferability, numerical stability, and desired physical properties. The results are necessarily sensitive to the data used for training and testing (as well as the associated weights), which include a wide range of main group thermochemistry and non-covalent interactions, but no transition metal systems or strong correlation cases.

2. From the partial search of the space of possible meta-GGA density functionals, many strong candidates emerged, and the best one (as measured by the defined criteria) has been self-consistently optimized. The resulting semi-empirical density functional, B97M-V, is a 12-parameter local meta-GGA density functional based on the B97 GGA and B00 meta-GGA models for local exchange and correlation, augmented with nonlocal correlation using the VV10 nonlocal correlation functional. Since it does not contain exact exchange, B97M-V complements the range-separated hybrid GGA ωB97X-V density functional that was

recently developed from a complete search of a much smaller GGA functional space.

3. Detailed assessment against 14 existing density functionals on main group thermochemistry and non-covalent interactions suggests that B97M-V is the best density functional tested for non-bonded interactions by a very significant margin, even outperforming ωB97X-V. The RMSD of B97M-V for non-covalent interactions is almost 2 times smaller than that of the next best local density functional. Its performance for thermochemistry is also very good, 30% better than the next best local density functional tested. Table 4.22 ranks the 15 benchmarked density functionals with respect to their overall RMSDs for all of the bonded and non-bonded interactions in the training, primary test, and secondary test sets. These results suggest that B97M-V should be a useful improvement over existing local meta-GGA density functionals for a wide range of chemical applications, a conclusion which is supported by a large variety of additional tests conducted in this work.

4. As a semi-empirical density functional developed using a computationally efficient but physically inexact form, B97M-V also necessarily has significant weaknesses in some applications, despite the successes discussed above. The main weakness arises in predicting reaction barrier heights, where due to the lack of exact exchange, the RMSDs of B97M-V are 3 to 4 times larger than the best hybrid density functionals (though comparable to those of existing local meta-GGAs). It is also likely that B97M-V will perform poorly for relative energies that are sensitive to self-interaction (delocalization) error or for systems that exhibit strong correlations.

5. B97M-V was trained in the aug-cc-pVQZ basis set for thermochemistry and the aug-cc-pVTZ basis set (without counterpoise corrections) for non-covalent interactions. Its basis set dependence has been thoroughly tested in order to identify smaller basis sets that can provide results similar in quality to those acquired with the basis sets used for training the parameters. For thermochemistry, the cc-pVQZ, 6-311++G(3df,3pd), pc-2, pc-3, and def2-QZVP basis sets can be recommended as smaller alternatives to the aug-cc-pVQZ basis set. For non-covalent interactions, the def2-SVPD basis set used with counterpoise corrections and the def2-TZVPD basis set used without counterpoise corrections can be recommended as smaller alternatives to the non-counterpoise-corrected aug-cc-pVTZ basis set. Furthermore, for systems where the use of diffuse functions may lead to issues with linear dependence, the pc-2 basis set is recommended for use without counterpoise corrections. It is important to note

that while B97M-V was trained very close to the basis set limit for thermochemistry, it was not necessarily trained at the basis set limit for non-covalent interactions.

6. Since existing semi-empirical meta-GGA density functionals are very sensitive to the integration grid, B97M-V was trained with the goal of making the (99,590)/SG-1 grid the "infinite grid" limit. The (75,302)/SG-0 and (75,302)/SG-1 grids are recommended as the bare minimum for use with B97M-V (particularly for quick calculations), while the (99,590)/SG-1 grid is recommended if results near the "infinite grid" limit are required. Finally, the (75,590)/SG-1 grid can serve as a compromise between these 2 limits.

7. It is desirable to apply the same training approach used here to develop other semi-empirical density functionals with improved physical content, so that the resulting density functionals are likewise minimally parameterized and optimally transferable. Perhaps the most obvious candidate is a meta-GGA that includes nonlocal exchange through range-separation. We hope to report such a development in due course.

Chapter 5

ωB97M-V: An RSH meta-GGA density functional

5.1 Introduction

Density functional theory (DFT) is built on exact foundations[3], but the exact functional, even if it were accessible, would likely be so complicated that it would give little practical advantage relative to the best wave function theories. The great achievement of functional development to date is the fact that very approximate functionals can provide useful levels of accuracy for many electronic structure problems in chemistry and condensed matter physics. The quest to obtain improved functionals that are computationally tractable continues in many research groups today, and this chapter describes a promising effort in that direction.

The parameterization of empirical density functionals via linear least-squares fitting is perhaps the most widely-used method for functional development in the quantum chemistry community. Introduced by Axel Becke with the B97 density functional[35], it relies on expanding inhomogeneity variables based on physically-relevant ingredients, such as the spin-density (ρ_σ), its gradient ($|\nabla \rho_\sigma|$), or the kinetic energy density ($\tau_\sigma = \sum\limits_i^{n_\sigma} |\nabla \psi_{i\sigma}|^2$), in one or more power series, whose coefficients are determined with the use of a training set of high-quality reference values.

Since 1997, at least 40 semi-empirical density functionals have been developed based on the concept introduced by B97. These functionals range from generalized gradient approximation (GGA) and nonseparable gradient approximation (NGA) functionals to meta-GGA and meta-NGA functionals. GGAs, representing Rung 2 of Perdew's Jacob's Ladder[4], usually depend on a single inhomogeneity variable that is a function of both ρ_σ and $|\nabla\rho_\sigma|$, while NGAs additionally depend on an inhomogeneity variable that is solely a function of ρ_σ. Meta-GGAs (Rung 3) expand upon GGAs by including an additional dependence on an inhomogeneity variable that depends on ρ_σ and τ_σ, while meta-NGAs expand upon meta-GGAs in the same way that NGAs expand upon GGAs.

The inclusion of exact exchange, popularized in 1993 with the B3PW91 density functional[33], has conventionally been of the global variety, meaning that the fraction of exact exchange is uniform for all interelectronic distances. More recently, these global hybrids have been replaced by range-separated hybrid (RSH) functionals[47], which have a fraction of short-range exact exchange that typically either smoothly increases to 1 (long-range-corrected) or smoothly decreases to 0 (long-range-screened).

Finally, since density functionals are unable to properly account for long-range correlation, most modern parameterizations simultaneously train a dispersion correction onto the exchange-correlation functional. The simplest form for a dispersion correction is a damped, atom-atom potential (DFT-D) such as Grimme's DFT-D2 or DFT-D3 models[54–56]. A more rigorous treatment of dispersion is provided by nonlocal correlation (NLC) functionals such as VV10[57] and vdW-DF-2[74]. However, the most elaborate and computationally demanding choice for a dispersion correction is a post-Hartree–Fock correlation (post-HFC) method such as MP2, RPA, or CCSD.

In light of the above considerations, Figure 5.1 presents an alternate view of elements that can be combined to define most existing density functionals. The first element (**Local Exchange-Correlation**) pertains to the ingredients that constitute the local exchange-correlation functional, with the available choices mimicking the first three rungs of Jacob's Ladder. The second element (**Exact Exchange**) pertains to the optional use of exact exchange contributions to define hybrid functionals. Finally, the third element (**Dispersion Correction**) generally accounts for the optional treatment of long-range correlation by one of the aforementioned approaches.

A selection of semi-empirical density functionals based on the B97 concept is listed below (dispersion-corrected functionals are underlined):

- **GGA Functionals**
  - **Local**: HCTH/93, HCTH/120, HCTH/147, HCTH/407, <u>B97-D</u>, SOGGA11[54, 67–69, 224]
  - **Global Hybrid**: B97-1, B97-2, B97-K, B97-3, SOGGA11-X[36, 37, 67, 70, 71]

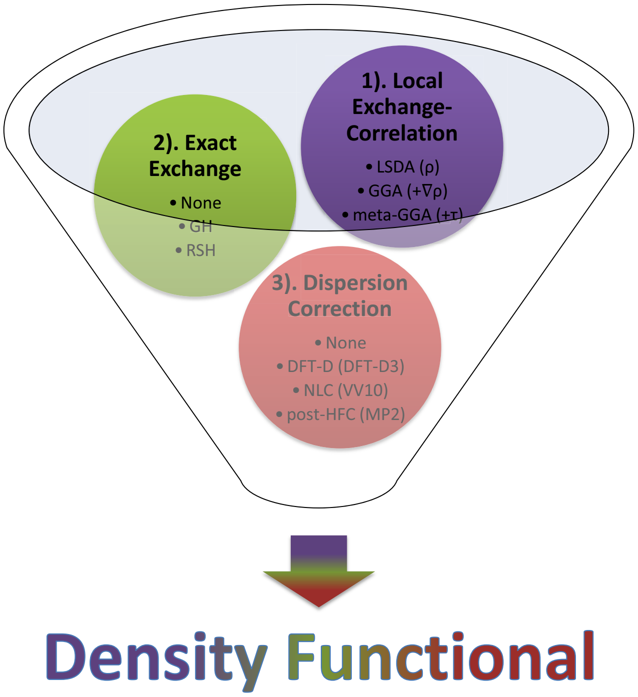


Figure 5.1: An alternate view of elements that can be combined to define most existing density functionals.

  - **Range-Separated Hybrid**: ωB97, ωB97X, ωB97X-D, ωB97X-D3, ωB97X-V[48, 58, 60, 63]

- **NGA Functionals**
  - **Local**: N12, GAM[225, 226]
  - **Range-Separated Hybrid**: N12-SX[51]

- **meta-GGA Functionals**
  - **Local**: τ-HCTH, M06-L, M11-L, B97M-V[23, 24, 28, 64]
  - **Global Hybrid**: τ-HCTHh, BMK, M05, M05-2X, M06, M06-2X, M06-HF, M08-HX, M08-SO[28, 36, 41, 43–46]
  - **Range-Separated Hybrid**: M11, ωM05-D, ωM06-D3[52, 59, 60]

- **meta-NGA Functionals**
  - **Local**: MN12-L, MN15-L[25, 26]

  - **Global Hybrid**: MN15[227]
  - **Range-Separated Hybrid**: MN12-SX[51]

The simplest form for the power series utilized in GGA functionals is

$$f(y) = \sum_{j=0}^{N} c_j y^j, \tag{5.1}$$

with y representing an inhomogeneity variable based on one of the aforementioned physically-relevant ingredients, and N representing the maximum truncation order for the summation. Conventionally, the value of N has either been chosen a priori or determined based on a "goodness-of-fit" to the training set. Smaller values of N can yield smoother and perhaps more transferable inhomogeneity correction factors, while larger values necessarily provide better fits to training data, whose transferability must subsequently be assessed.

The most general approach, however, is to choose a value for N and consider all possible combinations of the $N+1$ coefficients. This approach was explored several years ago[62], resulting in the development of the ωB97X-V functional[63]. Using this combinatorial approach leads to a total of $2^{N+1}-1$ fits, instead of just a handful. With a large number of candidate fits, the transferability of the fits can be assessed on a test set, allowing them to be ranked based on both their training set and test set performance. Furthermore, fits can be discarded based on undesirable physical characteristics or other relevant criteria.

In contrast to the one-dimensional power series that characterizes a GGA density functional, the most general power series that can accommodate a meta-GGA density functional is two-dimensional,

$$f(x, y) = \sum_{i=0}^{N'} \sum_{j=0}^{N} c_{ij} x^i y^j. \tag{5.2}$$

In the spirit of the original B97 density functional, three components of the local exchange-correlation functional require parameterization: exchange, same-spin correlation, and opposite-spin correlation. With each component contributing $(N'+1)(N+1)$ coefficients, the total number of possible fits is $2^{3(N'+1)(N+1)}-1$. Setting N', the meta-GGA maximum truncation order, to 8, and N, the GGA maximum truncation order, to 4, brings the total number of possible combinations to an astounding $2^{135}-1 \approx 10^{41}$, a "functional genome" whose rank is approaching the square of Avogadro's number.

The development of the B97M-V density functional[64] was a first attempt at a partial search of the meta-GGA functional genome within the combinatorial optimization approach. Apart from the difficult issue of choosing appropriate weights for different sets of training and testing data, the main compromise made in the design of B97M-V was the choice to exclude exact exchange.

The goal of this chapter is to improve upon B97M-V by revisiting the meta-GGA combinatorial search problem with the inclusion of exact exchange. The objective is to combinatorially design a range-separated hybrid, meta-GGA density functional which includes VV10 nonlocal correlation. It must be stressed that the combinatorial search performed to define B97M-V should not and will not be used in any direct way for this purpose. The addition of exact exchange means that different coefficients in Equation 5.2 will emerge as significant, perhaps apart from the few lowest-order terms. Unfortunately, the whole reason for adopting a combinatorial design approach is that it is impossible to anticipate *which* other terms will emerge as significant. Accordingly, it is a brand new search problem.

After describing the search process and its outcome, the functional which emerges as most transferable (ωB97M-V) can then be compared to existing functionals. ωB97M-V will be compared against two functionals that were designed in a similar fashion (ωB97X-V and B97M-V), as well as some of the best alternative functionals from other groups, particularly those that include similar functional ingredients. However, no other existing functional yet combines precisely the same components as ωB97M-V.

A few of the issues that will be particularly interesting to examine are the following:

1. Comparing ωB97M-V against ωB97X-V will demonstrate the value of adding kinetic energy density dependence to a range-separated hybrid GGA functional containing VV10 nonlocal correlation.

2. Comparing ωB97M-V against B97M-V will demonstrate the value of adding exact exchange to a local meta-GGA functional containing VV10 nonlocal correlation.

3. Comparing ωB97M-V against the best available semi-empirical hybrid meta-GGAs including M06-2X, M08-HX, M11, ωM05-D, and MN15 will test the value of the combinatorial optimization strategy.

The hope is that within the limits of the functional form that has been chosen for optimization, the combinatorial design approach will permit the discovery of the most broadly accurate Rung 1-4 density functional to date.

5.2 Outline

This chapter describes the task of designing, self-consistently training, and extensively testing an accurate, transferable, and well-behaved range-separated hybrid meta-GGA density functional with VV10 nonlocal correlation. The use of highly accurate datasets is integral to the functional development process, and the 84 datasets utilized in this work are described in Section 5.4 and Table 5.1. This is followed by a description of the equations that characterize the chosen functional form (Section 5.5). The design process, described in detail in Section 5.6, involves searching through over 10 billion potential fits (Table 5.2 and Figure 5.2), whose parameters are individually trained on a training set and tested for transferability on a primary test set. The fit with the best performance across the training and primary test sets is self-consistently parameterized (Table 5.3 and Figure 5.3) and named ωB97M-V. In Section 5.7, the new functional is compared to the 11 existing functionals shown in Table 5.4, across the training set (Figure 5.5), the primary test set (Figure 5.6), and a secondary test set (Figure 5.7). Since the test set performance is the most significant measure of the merit of a semi-empirical density functional, Figure 5.4 shows the performance of all 12 functionals for the 8 datatypes defined in Section 5.4 across 3547 test data points. The interpolated equilibrium bond lengths and equilibrium binding energies of 90 potential energy curves are analyzed for all 12 functionals in Figure 5.8, and the benzene-argon potential energy curve is displayed in Figure 5.9 as an example of the performance of these functionals for weak interactions. In Section 5.8, the basis set dependence of ωB97M-V is thoroughly assessed by applying 21 basis sets to 4 datasets, and the results are summarized in Figure 5.10. In Section 5.9, the integration grid sensitivity of ωB97M-V is thoroughly assessed (Table 5.5) by calculating 3247 of the 3834 data points in the training and primary test sets with a variety of integration grids. Finally, the chapter is concluded in Section 5.10 with a summary (Figure 5.11) of the performance of all 12 functionals for the 8 aforementioned datatypes across 4399 data points.

5.3 Computational Details

Since several of the density functionals that appear in this chapter contain both a local exchange-correlation functional and a nonlocal correlation functional, the integration grids used to evaluate these two components will be reported together, separated by a forward slash (local/nonlocal). Regarding the integration grid, the notation (x,y) indicates x radial shells with y angular grid points per shell.

The (99,590)/SG-1 grid was used for all of the datasets in the training, primary test, and secondary test sets, except AE18 and RG10. The (500,974)/(75,302) grid was used for both of these datasets. The def2-QZVPPD basis set was used for all of the datasets in the training, primary test, and secondary test sets, without counterpoise corrections.

For the basis set limit tests, the (99,590)/SG-1 grid was used, while 21 basis sets from 4 different families were tested: the cc-pVXZ and aug-cc-pVXZ (X=D,T,Q) Dunning basis sets[77, 78, 147], the pc-X and aug-pc-X (X=0,1,2,3) Jensen basis sets[174–176], the def2-SVP, def2-SVPD, and def2-XZVPP and def2-XZVPPD (X=T,Q) Karlsruhe basis sets[146, 228–231], and the 6-311++G(3df,3pd) Pople basis set. For the integration grid limit tests, the def2-QZVPPD basis set was used, while seven different grid combinations were tested: (250,974)/SG-1, (99,590)/SG-1, (99,302)/SG-1, (75,590)/SG-1, (75,302)/SG-1, (75,302)/SG-0, and SG-1/SG-0[61, 232].

All of the calculations were performed with a development version of Q-Chem 4.0[211].

5.4 Datasets

A total of 84 existing datasets are employed in this work, containing 4986 data points (and requiring 5931 single-point calculations). 82 of these 84 datasets (AE18 and RG10 are excluded) are classified according to 8 categories (or datatypes): NCED (non-covalent dimers (easy)), NCEC (non-covalent clusters (easy)), NCD (non-covalent dimers (difficult)), IE (isomerization energies (easy)), ID (isomerization energies (difficult)), TCE (thermochemistry (easy)), TCD (thermochemistry (difficult)), and BH (barrier heights). The number of data points (and datasets) that are classified according to NCED, NCEC, NCD, IE, ID, TCE, TCD, and BH are 1744 (18), 243 (12), 91 (5), 755 (12), 155 (5), 947 (15), 258 (7), and 206 (8), respectively.

For the purposes of functional development and testing, the datasets are divided into three categories. A training set is used to fit the parameters of each candidate functional in the combinatorial search, and then again to self-consistently train the best candidate. A primary test set is used in conjunction with the training set to perform the combinatorial search, and identify the best candidate (details of this procedure are given in Section 5.6). Finally, a secondary test set is used to assess the final optimized functional. Detailed information about the training, primary test, and secondary test sets can be found in Table 5.1. The training set contains 870 data points overall, the primary test set contains 2964 data points overall, and the secondary test set contains 1152 data points overall. Thus, the training set constitutes only 17.5% of the entire database used to develop and assess ωB97M-V.

5.5 Theory

The complete functional form for ωB97M-V is given by Equations 5.3-5.5. The acronyms used in this section are: exchange-correlation (xc), exchange (x), correlation (c), short-range (sr), long-range (lr), meta-GGA (mGGA), same-spin (ss), opposite-spin (os), and nonlocal (nl).

$$E_{xc}^{\omega B97M-V} = E_{x}^{\omega B97M-V} + E_{c}^{\omega B97M-V} \tag{5.3}$$

$$E_{x}^{\omega B97M-V} = E_{x,sr}^{mGGA} + c_x E_{x,sr}^{exact} + E_{x,lr}^{exact} \tag{5.4}$$

$$E_{c}^{\omega B97M-V} = E_{c,ss}^{mGGA} + E_{c,os}^{mGGA} + E_{c,nl}^{VV10} \tag{5.5}$$

The local spin-density approximation (LSDA) for exchange can be expressed in terms of the first-order spinless reduced density matrix for an infinite uniform electron gas (UEG),

$$E_{x}^{LSDA} = -\frac{1}{2}\sum_{\sigma}^{\alpha,\beta}\int\int\frac{1}{s}\left|\rho_{\sigma}^{UEG}\right|^{2} d\mathbf{r}d\mathbf{s}, \tag{5.6}$$

$$\rho_{\sigma}^{UEG} = 3\rho_{\sigma}\left(\frac{\sin\left(k_{F\sigma}s\right) - k_{F\sigma}s\cos\left(k_{F\sigma}s\right)}{\left(k_{F\sigma}s\right)^{3}}\right), \tag{5.7}$$

where $\mathbf{s} = \mathbf{r}_1 - \mathbf{r}_2$, $\mathbf{r} = \frac{1}{2}\left(\mathbf{r}_1 + \mathbf{r}_2\right)$, and $k_{F\sigma} = \left(6\pi^2\rho_\sigma\right)^{1/3}$ is the spin-polarized Fermi wave vector. Integration of Equation 5.6 over $\mathbf{s}$ gives the well-known expression for the LSDA exchange energy in terms of the exchange energy density per unit volume,

$$E_{x}^{LSDA} = \sum_{\sigma}^{\alpha,\beta}\int e_{x,\sigma}^{UEG} d\mathbf{r}, \tag{5.8}$$

$$e_{x,\sigma}^{UEG} = -\frac{3}{2}\left(\frac{3}{4\pi}\right)^{1/3}\rho_{\sigma}^{4/3}, \tag{5.9}$$

Transforming E_x^{LSDA} to its short-range counterpart, $E_{x,sr}^{LSDA}$, is accomplished by replacing $\frac{1}{s}$ in Equation 5.6 with $\frac{\mathrm{erfc}(\omega s)}{s}$ and carrying out the same integration. The resulting SR-LSDA exchange functional,

$$E_{x,sr}^{LSDA} = \sum_{\sigma}^{\alpha,\beta}\int e_{x,\sigma}^{UEG} F_{x,\sigma} d\mathbf{r}, \tag{5.10}$$

Name	Set	Datatype	#	Description	ΔE (kcal/mol)	Ref.
A24	Train	NCED	24	Binding energies of small non-covalent complexes	2.65	149
DS14	Train	NCED	14	Binding energies of complexes containing divalent sulfur	3.70	153
HB15	Train	NCED	15	Binding energies of hydrogen-bonded dimers featuring ionic groups common in biomolecules	19.91	220
HSG	Train	NCED	21	Binding energies of small ligands interacting with protein receptors	6.63	221,89
NBC10	Train	NCED	184	PECs for BzBz (5), BzMe (1), MeMe (1), BzH$_2$S (1), and PyPy (2)	1.91	97,88,233,89
S22	Train	NCED	22	Binding energies of hydrogen-bonded and dispersion-bound non-covalent complexes	9.65	104,89
X40	Train	NCED	31	Binding energies of non-covalent interactions involving halogenated molecules	5.26	144
A21x12	PriTest	NCED	252	PECs for the 21 equilibrium complexes from A24	1.43	234
BzDC215	PriTest	NCED	215	PECs for benzene interacting with two rare-gas atoms and eight first- and second-row hydrides	1.81	91
HW30	PriTest	NCED	30	Binding energies of hydrocarbon-water dimers	2.34	100
NC15	PriTest	NCED	15	Binding energies of very small non-covalent complexes	0.95	222
S66	PriTest	NCED	66	Binding energies of non-covalent interactions found in organic molecules and biomolecules	6.88	103,105
S66x8	PriTest	NCED	528	PECs for the 66 complexes from S66x8	5.57	103
3B-69-DIM	SecTest	NCED	207	Binding energies of all relevant pairs of monomers from 3B-69-TRIM	5.87	235
AlkBind12	SecTest	NCED	12	Binding energies of saturated and unsaturated hydrocarbon dimers	3.14	236
CO2Nitrogen16	SecTest	NCED	16	Binding energies of CO$_2$ to molecular models of pyridinic N-doped graphene	3.84	237
HB49	SecTest	NCED	49	Binding energies of small- and medium-sized hydrogen-bonded systems	14.12	238,239,240
Ionic43	SecTest	NCED	43	Binding energies of anion-neutral, cation-neutral, and anion-cation dimers	69.94	241
H2O6Bind8	Train	NCEC	8	Binding energies of isomers of $(H_2O)_6$	46.96	152,209
HW6Cl	Train	NCEC	6	Binding energies of $Cl^-(H_2O)_n$ $(n = 1-6)$	57.71	152,209
HW6F	Train	NCEC	6	Binding energies of $F^-(H_2O)_n$ $(n = 1-6)$	81.42	152,209
FmH2O10	PriTest	NCEC	10	Binding energies of isomers of $F^-(H_2O)_{10}$	168.50	152,209
Shields38	PriTest	NCEC	38	Binding energies of $(H_2O)_n$ $(n = 2-10)$	51.54	223
SW49Bind345	PriTest	NCEC	31	Binding energies of isomers of $SO_4^{2-}(H_2O)_n$ $(n = 3-5)$	28.83	87
SW49Bind6	PriTest	NCEC	18	Binding energies of isomers of $SO_4^{2-}(H_2O)_6$	62.11	87
WATER27	PriTest	NCEC	23	Binding energies of neutral and charged water clusters	67.48	98,216
3B-69-TRIM	SecTest	NCEC	69	Binding energies of trimers, with three different orientations of 23 distinct molecular crystals	14.36	235
CE20	SecTest	NCEC	20	Binding energies of water, ammonia, and hydrogen fluoride clusters	30.21	242,243
H2O20Bind10	SecTest	NCEC	10	Binding energies of isomers of $(H_2O)_{20}$ (low-energy structures)	198.16	209
H2O20Bind4	SecTest	NCEC	4	Binding energies of isomers of $(H_2O)_{20}$ (dod, fc, fs, and es)	206.12	218,98,216,217
TA13	Train	NCD	13	Binding energies of dimers involving radicals	22.00	244
XB18	Train	NCD	8	Binding energies of small halogen-bonded dimers	5.23	245
Bauza30	PriTest	NCD	30	Binding energies of halogen-, chalcogen-, and pnicogen-bonded dimers	23.65	246,247
CT20	PriTest	NCD	20	Binding energies of charge-transfer complexes	1.07	248
XB51	PriTest	NCD	20	Binding energies of large halogen-bonded dimers	6.06	245
AlkIsomer11	Train	IE	11	Isomerization energies of $n = 4-8$ alkanes	1.81	93
Butanediol65	Train	IE	65	Isomerization energies of butane-1,4-diol	2.89	249
ACONF	PriTest	IE	15	Isomerization energies of alkane conformers	2.23	250,216
CYCONF	PriTest	IE	11	Isomerization energies of cysteine conformers	2.00	157,216
Pentane14	PriTest	IE	14	Isomerization energies of stationary points on the n-pentane torsional surface	6.53	251
SW49Rel345	PriTest	IE	31	Isomerization energies of $SO_4^{2-}(H_2O)_n$ $(n = 3-5)$	1.47	87
SW49Rel6	PriTest	IE	18	Isomerization energies of $SO_4^{2-}(H_2O)_6$	1.22	87
H2O16Rel5	SecTest	IE	5	Isomerization energies of $(H_2O)_{16}$ (boat and fused cube structures)	0.40	219
H2O20Rel10	SecTest	IE	10	Isomerization energies of $(H_2O)_{20}$ (low-energy structures)	2.62	209
H2O20Rel4	SecTest	IE	4	Isomerization energies of $(H_2O)_{20}$ (dod, fc, fs, and es)	5.68	218,98,216,217
Melatonin52	SecTest	IE	52	Isomerization energies of melatonin	5.54	252
YMPJ519	SecTest	IE	519	Isomerization energies of the proteinogenic amino acids	8.33	253
EIE22	Train	ID	22	Isomerization energies of enecarbonyls	4.97	254
Styrene45	Train	ID	45	Isomerization energies of C_8H_8	68.69	255
DIE60	PriTest	ID	60	Isomerization energies of reactions involving double-bond migration in conjugated dienes	5.06	256
ISOMERIZATION20	PriTest	ID	20	Isomerization energies	44.05	82
C20C24	SecTest	ID	8	Isomerization energies of the ground state structures of C_{20} and C_{24}	36.12	257
AlkAtom19	Train	TCE	19	$n = 1-8$ alkane atomization energies	1829.31	93
BDE99nonMR	Train	TCE	83	Bond dissociation energies (SR)	114.98	82
G21EA	Train	TCE	25	Adiabatic electron affinities of atoms and small molecules	40.86	154,216
G21IP	Train	TCE	36	Adiabatic ionization potentials of atoms and small molecules	265.35	154,216
TAE140nonMR	Train	TCE	124	Total atomization energies (SR)	381.05	82
AlkIsod14	PriTest	TCE	14	$n = 3-8$ alkane isodesmic reaction energies	10.35	93
BH76RC	PriTest	TCE	30	Reaction energies from HTBH38 and NHTBH38	30.44	95,94,216
EA13	PriTest	TCE	13	Adiabatic electron affinities	42.51	171
HAT707nonMR	PriTest	TCE	505	Heavy-atom transfer energies (SR)	74.79	82
IP13	PriTest	TCE	13	Adiabatic ionization potentials	256.24	171
NBPRC	PriTest	TCE	12	Reactions involving NH_3/BH_3 and PH_3/BH_3	30.52	216,258,210
SN13	PriTest	TCE	13	Nucleophilic substitution energies	25.67	82
BSR36	SecTest	TCE	36	Hydrocarbon bond separation reaction energies	20.06	259,210
HNBrBDE18	SecTest	TCE	18	Homolytic N–Br bond dissociation energies	56.95	260
WCPT6	SecTest	TCE	6	Tautomerization energies for water-catalyzed proton-transfer reactions	7.53	261
BDE99MR	PriTest	TCD	16	Bond dissociation energies (MR)	54.51	82
HAT707MR	PriTest	TCD	202	Heavy-atom transfer energies (MR)	83.41	82
TAE140MR	PriTest	TCD	16	Total atomization energies (MR)	147.20	82
PlatonicHD6	SecTest	TCD	6	Homodesmotic reactions involving platonic hydrocarbon cages, C_nH_n $(n = 4, 6, 8, 10, 12, 20)$	136.71	262
PlatonicID6	SecTest	TCD	6	Isodesmic reactions involving platonic hydrocarbon cages, C_nH_n $(n = 4, 6, 8, 10, 12, 20)$	96.19	262
PlatonicIG6	SecTest	TCD	6	Isogyric reactions involving platonic hydrocarbon cages, C_nH_n $(n = 4, 6, 8, 10, 12, 20)$	356.33	262
PlatonicTAE6	SecTest	TCD	6	Total atomization energies of platonic hydrocarbon cages, C_nH_n $(n = 4, 6, 8, 10, 12, 20)$	2539.27	262
BHPERI26	Train	BH	26	Barrier heights of pericyclic reactions	23.15	216,263
CRBH20	Train	BH	20	Barrier heights for cycloreversion of heterocyclic rings	46.40	264
DBH24	Train	BH	24	Diverse barrier heights	28.34	83,84
CR20	PriTest	BH	20	Cycloreversion reaction energies	22.31	265
HTBH38	PriTest	BH	38	Hydrogen transfer barrier heights	16.05	94
NHTBH38	PriTest	BH	38	Non-hydrogen transfer barrier heights	33.48	95
PX13	SecTest	BH	13	Barrier heights for proton exchange in water, ammonia, and hydrogen fluoride clusters	28.83	242,243
WCPT27	SecTest	BH	27	Barrier heights of water-catalyzed proton-transfer reactions	38.73	261
AE18	Train	–	18	Absolute atomic energies of hydrogen through argon	148,739.00	86
RG10	PriTest	–	569	PECs for the 10 rare-gas dimers involving helium through krypton	1.21	96

Table 5.1: Summary of the 84 datasets that comprise the training, primary test, and secondary test sets. The datatypes are explained in Section 5.4. The sixth column contains the root-mean-squares of the dataset reaction energies. PEC stands for potential energy curve, SR stands for single-reference, MR stands for multi-reference, Bz stands for benzene, Me stands for methane, and Py stands for pyridine.

is conveniently identical to its unattenuated counterpart, with the exception of a multiplicative factor,

$$F_{x,\sigma} = 1 - \frac{2}{3}a_\sigma\left(2\sqrt{\pi}\text{erf}\left(\frac{1}{a_\sigma}\right) - 3a_\sigma + a_\sigma^3 + \left(2a_\sigma - a_\sigma^3\right)\exp\left(-\frac{1}{a_\sigma^2}\right)\right), \tag{5.11}$$

where $a_\sigma = \frac{\omega}{k_{F\sigma}}$ and ω is a nonlinear parameter that controls the transition from local DFT exchange to nonlocal exact exchange with respect to interelectronic distance.

Accounting for inhomogeneities in the electron density is achieved by multiplying the integrand of the SR-LSDA exchange functional (Equation 5.10) by a power series inhomogeneity correction factor (ICF), g_x, resulting in the SR-mGGA exchange functional,

$$E_{x,sr}^{mGGA} = \sum_\sigma^{\alpha,\beta}\int e_{x,\sigma}^{UEG}F_{x,\sigma}g_x d\mathbf{r}, \tag{5.12}$$

$$g_x = \sum_i^{N'}\sum_j^{N} c_{x,ij}w_{x,\sigma}^i u_{x,\sigma}^j, \tag{5.13}$$

$$w_{x,\sigma} = \frac{t_\sigma - 1}{t_\sigma + 1}, \tag{5.14}$$

$$u_{x,\sigma} = \frac{\gamma_x s_\sigma^2}{1 + \gamma_x s_\sigma^2}, \tag{5.15}$$

where the variable, $w_{x,\sigma} \in [-1, 1]$, is a finite domain transformation of the dimensionless ratio of the UEG kinetic energy density to the exact kinetic energy density, $t_\sigma = \frac{\tau_\sigma^{UEG}}{\tau_\sigma}$, with $\tau_\sigma^{UEG} = \frac{3}{5}\left(6\pi^2\right)^{2/3}\rho_\sigma^{5/3}$, and the variable, $u_{x,\sigma} \in [0, 1]$, is a finite domain transformation of the dimensionless spin-density gradient, $s_\sigma = \frac{|\nabla\rho_\sigma|}{\rho_\sigma^{4/3}}$. The linear DFT exchange parameters, $c_{x,ij}$, will be determined via least-squares fitting to a training set in Section 5.6, while $\gamma_x = 0.004$ is a fixed nonlinear DFT exchange parameter that was fit to the Hartree–Fock exchange energies of 20 atoms in 1986 by Becke[9].

Nonlocal exact exchange is introduced by splitting the conventional $\frac{1}{r}$ Coulomb operator into a short-range component ($E_{x,sr}^{exact}$) and a long-range component ($E_{x,lr}^{exact}$) with the $\frac{\text{erfc}(\omega \text{r})}{r}$ and $\frac{\text{erf}(\omega \text{r})}{r}$ Coulomb functions, respectively,

$$E_{x,sr}^{exact} = -\frac{1}{2}\sum_\sigma^{\alpha,\beta}\sum_{i,j}^{n_\sigma}\int\int \psi_{i\sigma}^*\left(\mathbf{r}_1\right)\psi_{j\sigma}^*\left(\mathbf{r}_2\right)\frac{\text{erfc}\left(\omega r\right)}{r} \psi_{j\sigma}\left(\mathbf{r}_1\right)\psi_{i\sigma}\left(\mathbf{r}_2\right)d\mathbf{r}_1 d\mathbf{r}_2, \tag{5.16}$$

$$E_{x,lr}^{exact} = -\frac{1}{2}\sum_{\sigma}^{\alpha,\beta}\sum_{i,j}^{n_\sigma}\int\int \psi_{i\sigma}^{*}(\mathbf{r}_1)\,\psi_{j\sigma}^{*}(\mathbf{r}_2)\,\frac{\mathrm{erf}(\omega r)}{r}\,\psi_{j\sigma}(\mathbf{r}_1)\,\psi_{i\sigma}(\mathbf{r}_2)\,d\mathbf{r}_1 d\mathbf{r}_2, \tag{5.17}$$

where $\psi_{i\sigma}$ and $\psi_{j\sigma}$ are occupied Kohn–Sham spatial orbitals and $r = |\mathbf{r}_1 - \mathbf{r}_2|$. Instead of setting the percentage of exact-exchange at $r = 0$ to zero, an optimizable parameter, c_x, controls the amount of short-range exact exchange.

Closed-form expressions for the correlation energy density per particle of an infinite uniform electron gas, ϵ_c^{UEG}, are only known for the low- and high-density limits of the paramagnetic and ferromagnetic cases. Using the Monte-Carlo data of Ceperley and Alder[8], Perdew and Wang developed an analytic spin-compensated representation[7] for ϵ_c^{UEG}. Combined with the spin-polarization interpolation formula of Vosko, Wilk, and Nusair[6], the spin-polarized PW92 correlation energy density per electron, ϵ_c^{PW92}, is the starting point for the correlation functional,

$$E_c^{LSDA} = \int \rho\epsilon_c^{PW92} d\mathbf{r}. \tag{5.18}$$

Using the spin decomposition technique of Stoll and coworkers[81], the LSDA correlation energy is separated into same-spin and opposite-spin components,

$$E_{c,ss}^{LSDA} = \sum_{\sigma}^{\alpha,\beta}\int e_{c,\sigma\sigma}^{PW92} d\mathbf{r} = \int \rho_\alpha \epsilon_c^{PW92}(\rho_\alpha, 0)\, d\mathbf{r} + \int \rho_\beta \epsilon_c^{PW92}(0, \rho_\beta)\, d\mathbf{r}, \tag{5.19}$$

$$E_{c,os}^{LSDA} = \int e_{c,\alpha\beta}^{PW92} d\mathbf{r} = \int \rho\epsilon_c^{PW92} d\mathbf{r} - \int \rho_\alpha \epsilon_c^{PW92}(\rho_\alpha, 0)\, d\mathbf{r} - \int \rho_\beta \epsilon_c^{PW92}(0, \rho_\beta)\, d\mathbf{r}, \tag{5.20}$$

where $e_{c,\sigma\sigma}^{PW92}$ and $e_{c,\alpha\beta}^{PW92}$ are the PW92 same-spin and opposite-spin correlation energy densities per unit volume, respectively. Extending Equations 5.19 and 5.20 to account for inhomogeneities in the electron density is straightforward, since the approach used for the exchange functional can be utilized,

$$E_{c,ss}^{mGGA} = \sum_{\sigma}^{\alpha,\beta}\int e_{c,\sigma\sigma}^{PW92} g_{c,ss} d\mathbf{r}, \tag{5.21}$$

$$g_{c,ss} = \sum_{i}^{N'}\sum_{j}^{N} c_{css,ij} w_{c,\sigma\sigma}^{i} u_{c,\sigma\sigma}^{j}, \tag{5.22}$$

$$w_{c,\sigma\sigma} = \frac{t_\sigma - 1}{t_\sigma + 1}, \tag{5.23}$$

$$u_{c,\sigma\sigma} = \frac{\gamma_{c,ss} s_\sigma^2}{1 + \gamma_{c,ss} s_\sigma^2}, \tag{5.24}$$

$$E_{c,os}^{mGGA} = \int e_{c,\alpha\beta}^{PW92} g_{c,os} d\mathbf{r}, \tag{5.25}$$

$$g_{c,os} = \sum_i^{N'} \sum_j^{N} c_{cos,ij} w_{c,\alpha\beta}^i u_{c,\alpha\beta}^j, \tag{5.26}$$

$$w_{c,\alpha\beta} = \frac{t_{\alpha\beta} - 1}{t_{\alpha\beta} + 1}, \tag{5.27}$$

$$u_{c,\alpha\beta} = \frac{\gamma_{c,os} s_{\alpha\beta}^2}{1 + \gamma_{c,os} s_{\alpha\beta}^2}, \tag{5.28}$$

where $t_{\alpha\beta} = \frac{1}{2}\left(t_\alpha + t_\beta\right)$ and $s_{\alpha\beta}^2 = \frac{1}{2}\left(s_\alpha^2 + s_\beta^2\right)$. The linear DFT correlation parameters, $c_{css,ij}$ and $c_{cos,ij}$, will be determined via least-squares fitting to a training set in Section 5.6, while $\gamma_{c,ss} = 0.2$ and $\gamma_{c,os} = 0.006$ are nonlinear DFT correlation parameters that were fit to the correlation energies of neon and helium in 1997 by Becke[35].

Nonlocal correlation is taken into account via the VV10 NLC functional[57],

$$E_{c,nl}^{VV10} = \int \rho\left(\mathbf{r}_1\right) \left[\frac{1}{32} \left[\frac{3}{b^2} \right]^{3/4} + \frac{1}{2} \int \rho\left(\mathbf{r}_2\right) \Phi\left(\mathbf{r}_1, \mathbf{r}_2; \{b, C\}\right) d\mathbf{r}_2 \right] d\mathbf{r}_1, \tag{5.29}$$

where $\Phi\left(\mathbf{r}_1, \mathbf{r}_2; \{b, C\}\right)$ is the nonlocal correlation kernel defined in Reference 57. The VV10 NLC functional contains two nonlinear parameters: b, which controls the short-range damping of the $1/r^6$ asymptote, and C, which controls the accuracy of the asymptotic C_6 coefficients.

5.6 Training

With a total of 3834 data points in the training and primary test sets, a three-dimensional nonlinear optimization of the parameters associated with range-separation (ω) and VV10 (b and C) is impractical. As a result, the values of $\omega = 0.3$, $b = 6$, and $C = 0.01$ that were optimized for ωB97X-V are taken without further investigation. Any inaccuracies in these parameters will be accounted for during the optimization of the linear parameters.

In order to generate the data that is needed to carry out the least-squares fits, it is necessary to choose an initial guess for the linear coefficients. As explained in Section 5.1, the value of N' (the maximum truncation order for w in Equations 5.13, 5.22, and 5.26) is set to 8 and the value of N (the maximum truncation order for u

in Equations 5.13, 5.22, and 5.26) is set to 4. This results in 135 coefficients that arise from the local exchange-correlation functional. Additionally, the fraction of short-range exact exchange (c_x) from Equation 5.4 is the 136th coefficient.

For the 135 coefficients that arise due to the three power series ICFs, the most flexible initial guess is that of the SR-LSDA, where $c_{x,00} = c_{css,00} = c_{cos,00} = 1$, and the remaining 132 coefficients are zero. For c_x, the most straightforward initial guess is zero. However, since the contribution from exact exchange is bound to constitute a large fraction of the total exchange-correlation energy, it is beneficial to pick a value for c_x that is as close to the final value as possible, in order to minimize discrepancies between the root-mean-square deviations that are generated using the initial guess, and those of the final, self-consistently optimized functional. With the value of $c_x = 0.167$ from ωB97X-V serving as a guide, the initial guess for the fraction of short-range exact exchange is set to $c_x = 0.15$. Finally, the only constraint that is explicitly enforced is the UEG limit for exchange ($c_{x,00} + c_x = 1$).

The fundamental equations that will be used throughout the least-squares fitting procedure are

$$W_{Tr}^{1/2} A \Delta x = W_{Tr}^{1/2} b \tag{5.30}$$

and

$$\Delta x = \left(A^T W_{Tr} A\right)^{-1} \left(A^T W_{Tr} b\right), \tag{5.31}$$

where $\Delta x = x_{i+1} - x_i$ is the change in the linear coefficients (length: 136), W_{Tr} is a diagonal matrix of training weights (dimensions: 3834 x 3834), $b = E_{ref} - E_i$ is the difference between the reference and initial guess energies (length: 3834), and $A = A\left(x_i\right)$ is the matrix of ICF contributions (dimensions: 3834 x 136). While the first three quantities are conceptually straightforward, it is worthwhile to further explain how to generate the A matrix.

A is most generally a (# of data points) × (# of linear parameters) matrix. As an example, the element $A_{7,23}$ corresponds to the contribution to the 7th data point from the 23rd linear parameter ($c_{x,42}$). Assuming that the seventh data point corresponds to the binding energy of the water dimer from the S22 dataset, computing $A_{7,23}$ requires calculating the following quantity,

$$\sum_{\sigma}^{\alpha,\beta} \int e_{x,\sigma}^{UEG} F_{x,\sigma} w_{x,\sigma}^4 u_{x,\sigma}^2 d\mathbf{r}, \tag{5.32}$$

for the water dimer and both of its monomers, and computing the associated contribution to the binding energy. In this work, the A matrix is computed twice: once with the (99,590)/SG-1 grid and once with the (250,974)/SG-1 grid (the (500,974)/(75,302) grid is always used for AE18 and RG10). The rationale for computing A in two different grids will be explained when the filtering procedure is discussed. Unless otherwise specified, terms from Equations 5.30 and 5.31 that appear henceforth are computed with the (250,974)/SG-1 grid. To refer to a quantity, Z, computed in the (99,590)/SG-1 grid, the following notation will be used: Z'.

Since the optimization procedure involves both a training set (for determining the values of the coefficients) and a primary test set (for assessing the transferability of the resulting coefficients), a measure of the overall performance of the fits is necessary for ranking them. For this purpose, the total weighted root-mean-square-deviation (WRMSD),

$$WRMSD_{Total} = \sqrt{\frac{\text{diag}\left(W\right) \cdot \left(A\Delta x - b\right)^2}{\#_{Total}}}, \tag{5.33}$$

is used, where W is a diagonal matrix of training and primary test set weights (dimensions: 3834 x 3834). The diagonals of W and W_{Tr} are identical for the elements that belong to the training set, while W_{Tr} contains zeros for the elements that belong to the primary test set.

The weighting scheme used for ωB97M-V is considerably different than that used for B97M-V. Initially, each data point is given a weight that corresponds to the inverse of the product of the number of data points in the associated dataset and the root-mean-square of the reaction energies in the associated dataset. These values can be found in the fourth and sixth columns of Table 5.1, respectively. Within each of the datatypes, the weights are normalized by dividing by the smallest weight, and then exponentiated such that they lie between 1 and 2. For the determination of the weights only, the NCED and NCEC datatypes are consolidated into a single datatype, NCE, giving a total of seven datatypes. Furthermore, AE18 is included in the TCE datatype in order to receive a weight. Finally, the seven datatypes get a multiplicative weight based on intuition: 0.1 for TCD, 1 for TCE, 10 for NCD, ID, and BH, and 100 for NCE and IE. As RG10 does not belong to a datatype, it gets a weight of 10000. At this point, all of the data points (besides those in RG10) have a weight between 0.1 and 200. In order to promote transferability, the datasets in the primary test set get another multiplicative weight of 2.

Following the initial setup described above, the search for the optimal least-squares fit can proceed. Applying the single UEG constraint to the initial parameter space of 136 brings the total number of linearly independent parameters to 135. With the available computing resources (a 64-core node), the maximum number of fits that can be performed in a single day is around two billion. Therefore 135C5=346,700,277, which returns only 5-parameter fits, is the largest calculation that can be performed in a single day using all 135 parameters (since 135C6=7,511,839,335). In order to explore fits with more parameters, it is necessary to either consider different truncations of the parameter space or compulsorily-select commonly-occurring parameters. The first option reduces the value of n (from nCk), allowing for larger values of k. For example, the binomial coefficients 135C5 and 82C6 are both similar in value, yet the latter allows the exploration of 6-parameter fits. The second option requires F commonly-occurring parameters to be compulsorily-selected (or frozen), allowing for the exploration of (n-F)Ck (k+F)-parameter fits. For example, if the results from 135C5 indicate that $c_{x,01}$ is the most important parameter, a possible next step would be to freeze $c_{x,01}$ (not its value but simply its inclusion in all successive fits) and explore the results of 134C5 with $c_{x,01}$ frozen, giving 6-parameter fits.

For the optimization of the functional at hand, a combination of these two options is utilized. While the meta-GGA parameter space described thus far contains 135 linearly independent parameters, its GGA subset contains only 15, which amounts to $2^{15} - 1 = 32767$ total possible fits. The GGA parameter space is fully searchable and the recent ωB97X-V density functional was developed within this subspace. Since ωB97X-V has seven linear parameters, it is plausible to assume that the minimum number of linear parameters necessary to parameterize a functional within the meta-GGA parameter space is seven. Thus, the initial parameter space of 135 should be truncated such that 7-parameter fits are possible without freezing. After considering a multitude of different truncations, a parameter space designated as 2I6I6I was selected. This includes variables up to second order individually in w and u for exchange and up to sixth order individually in w and u for same-spin and opposite-spin correlation. With 9 coefficients from exchange, 35 coefficients from same-spin correlation, 35 coefficients from opposite-spin correlation, an additional coefficient from short-range exact exchange, and a single constraint, the 2I6I6I designation reduces the parameter space from 135 to 79.

Although the 2I6I6I designation reduces the parameter space substantially, the largest calculation that is possible is still only 79C7=2,898,753,715. Thus, only up to 7-parameter fits can be explored. In order to advance past seven, it is necessary to compulsorily-select commonly-occurring parameters. The process of selecting the parameter that will be frozen involves a series of steps. First, from all of the 7-parameter fits resulting from the 79C7 optimization, the top 100,000 (ranked by total WRMSD) are saved and filtered twice. The first set of filters deals with the physical characteristics of the fits:

- $|x_{i+1}| \leq 25$
- $0 \leq g_x \leq 2.273$
- $-10 \leq g_{c,ss} \leq 10$

- $-10 \leq g_{c,os} \leq 10$
- $|E_i - E_i' + (A - A')\,\Delta x| \leq 0.015$ kcal/mol

Of the conditions listed above, the first ensures that the coefficients are small, the second ensures that the exchange functional obeys the local Lieb-Oxford bound *and* that all contributions to exchange are negative, the third and fourth ensure that the correlation functionals are well-behaved, and the last ensures that the absolute interaction energy errors attributed to the integration grid are no larger than 0.015 kcal/mol.

Following the first set of filters, at most 1000 of the remaining fits are passed through a second filter, which deals with the accuracy of the fits for predicting equilibrium bond lengths for non-covalent interactions. The BzDC215, NBC10, RG10, and S66x8 datasets are utilized to this effect. In total, these four datasets contain 96 potential energy curves (PEC), with BzDC215, NBC10, and RG10 each having 10, and S66x8 having 66. However, the benzene-neon dimer PEC from BzDC215 is removed because of severe integration grid issues. The remaining 95 PECs are interpolated and the equilibrium bond lengths are evaluated for all of the fits that pass through the first filter. Only fits that have an RMSD of less than 0.03 Å across the 95 PECs are allowed through the second filter.

Finally, the surviving fits are analyzed in order to determine the coefficient that is most commonly used. This coefficient is then compulsorily-selected in the next set of least-squares fits in order to allow for the exploration of 8-parameter fits. This procedure was repeated until a total of eight parameters were frozen ($c_{x,01}$, $c_{x,10}$, $c_{css,10}$, $c_{css,20}$, $c_{cos,20}$, $c_{css,00}$, $c_{cos,10}$, $c_{cos,21}$). The progression from the 7-parameter fits to the 15-parameter fits can be tracked in Table 5.2.

Due to the nonlinear nature of the self-consistent field method, the A matrix changes with every update to the parameters (since $A = A\left(x_i\right)$), and it is very likely that the A matrix that is constructed from the initial guess will be vastly different from the last A matrix that will be used to finalize the parameters. Furthermore, the larger the difference between the starting point and the ending point, the higher the chance that the initial RMSDs will differ significantly from the RMSDs of the final functional. In fact, this phenomenon was first encountered during the development of the B97M-V density functional and was bypassed by updating the initial guess with a better guess formed by the first nine parameters that were frozen. Since this methodology worked well during the development of B97M-V, it is utilized in the present work. Thus, the eight parameters shown in the first column of Table 5.2 ("Best 8") are used to update the SR-LSDA+VV10 initial guess.

Frozen	# of Initial Fits	# (Fitted)	Filter 1	Filter 2	Common
–	2,898,753,715	7	3342	413	$c_{x,01}$
$c_{x,01}$	2,641,902,120	8	1476	454	$c_{x,10}$
$c_{x,10}$	2,404,808,340	9	2806	623	$c_{css,10}$
$c_{css,10}$	2,186,189,400	10	7144	764	$c_{css,20}$
$c_{css,20}$	1,984,829,850	11	18021	591	$c_{cos,20}$
$c_{cos,20}$	1,799,579,064	12	6531	469	$c_{css,00}$
$c_{css,00}$	1,629,348,612	13	2985	696	$c_{cos,10}$
$c_{cos,10}$	1,473,109,704	14	1868	726	$c_{cos,21}$
$c_{cos,21}$	1,329,890,705	15	120	101	$c_{cos,60}$

Table 5.2: Progression from the 7-parameter fits to the 15-parameter fits based on the SR-LSDA+VV10 initial guess. The first column indicates the additional coefficient that is frozen (compulsorily-selected) in order to achieve the associated set of fits. The second column contains the total number of least-squares fits that are performed, of which only the top 100,000 (ranked by total WRMSD) are analyzed. The fourth column indicates the number of fits (of 100,000) that remain after the first filter is applied. The fifth column indicates the number of fits (of at most 1000) from the previous column that remain after the second filter is applied. Finally, the last column indicates the coefficient that is most commonly utilized in the surviving fits (shown in Column 5).

With the updated guess, the A matrix is computed with both the (99,590)/SG-1 and (250,974)/SG-1 grids and the same procedure outlined above is followed. However, no additional parameters need to be frozen, since with eight parameters already frozen, it is trivial to explore 9- through 15-parameter fits (71C1 through 71C7). However, both sets of filters are still applied. Beginning with the 9-parameter fits, an additional parameter is accepted only if it improves the total WRMSD by more than 0.05 kcal/mol (a protocol which was successfully utilized during the exploration of the GGA subspace[62]). Consequently, a 12-parameter fit emerges as the optimal fit and is self-consistently optimized. The progression of the minimum total WRMSDs from the 9- to 15-parameters fits is: {–, 5.04, 3.52, **3.46**, 3.44, 3.42, 3.39}. The WRMSDs corresponding to the least-squares fits from the SR-LSDA+VV10 initial guess and the "Best 8" updated guess are shown in Figure 5.2 in red and blue, respectively. The smallest WRMSD for a given number of linear parameters is displayed only for the "Best 8" data. The WRMSD of the least-squares fit that corresponds to ωB97M-V is boxed.

Including the initial cycle (Cycle 1) with the unoptimized ωB97M-V density functional as well as the "Best 8" cycle, the self-consistent optimization of ωB97M-V required six cycles. The final parameters of ωB97M-V can be found in the last col-

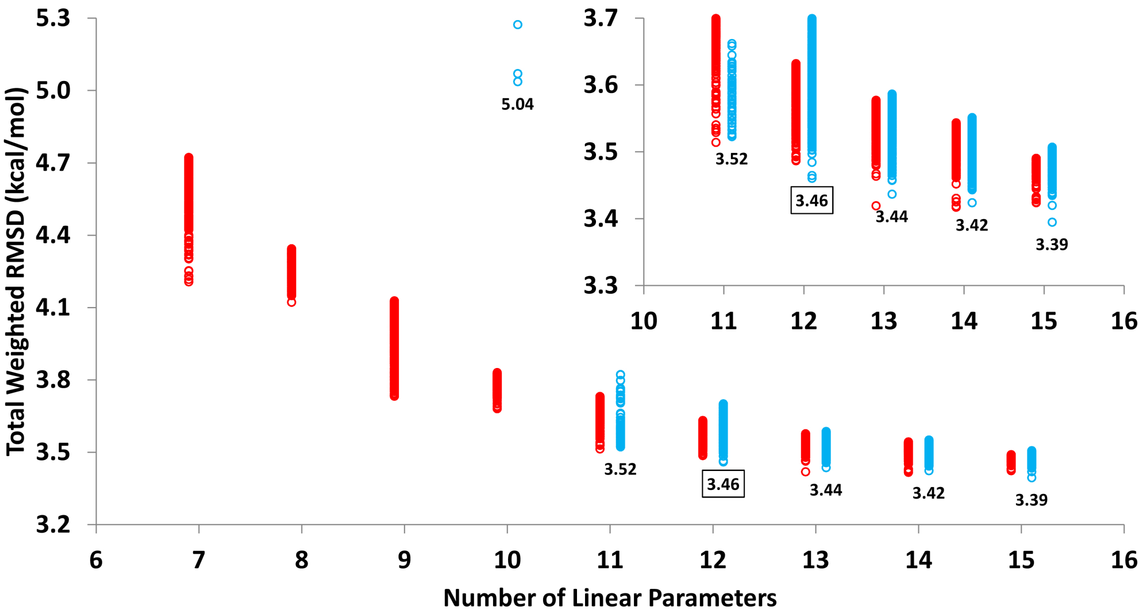


Figure 5.2: WRMSDs corresponding to the least-squares fits from the SR-LSDA+VV10 initial guess (red) and the "Best 8" updated guess (blue). The smallest WRMSD for a given number of linear parameters is displayed only for the "Best 8" data. The WRMSD of the least-squares fit that corresponds to ωB97M-V is boxed.

umn of Table 5.3, and Figure 5.3 shows the exchange, same-spin correlation, and opposite-spin correlation ICF plots for ωB97M-V (bottom). In addition, it displays the ICFs for the SR-LSDA+VV10 initial guess (top) and the "Best 8" updated guess (middle). The final ICFs of ωB97M-V are smooth and well-behaved, with the maximum value of the exchange ICF (2.116) lying well under the Lieb-Oxford bound (2.273). The lower and upper bounds of all three ICFs are:

- $0.591 \leq g_x \leq 2.116$
- $-7.482 \leq g_{c,ss} \leq 4.429$
- $-1.957 \leq g_{c,os} \leq 4.222$

The four parameters that are chosen in addition to the "Best 8" are $c_{css,43}$, $c_{css,04}$, $c_{cos,60}$, and $c_{cos,61}$. Interestingly, the chosen fit does not optimize the value of short-range exact exchange away from $c_x = 0.15$.

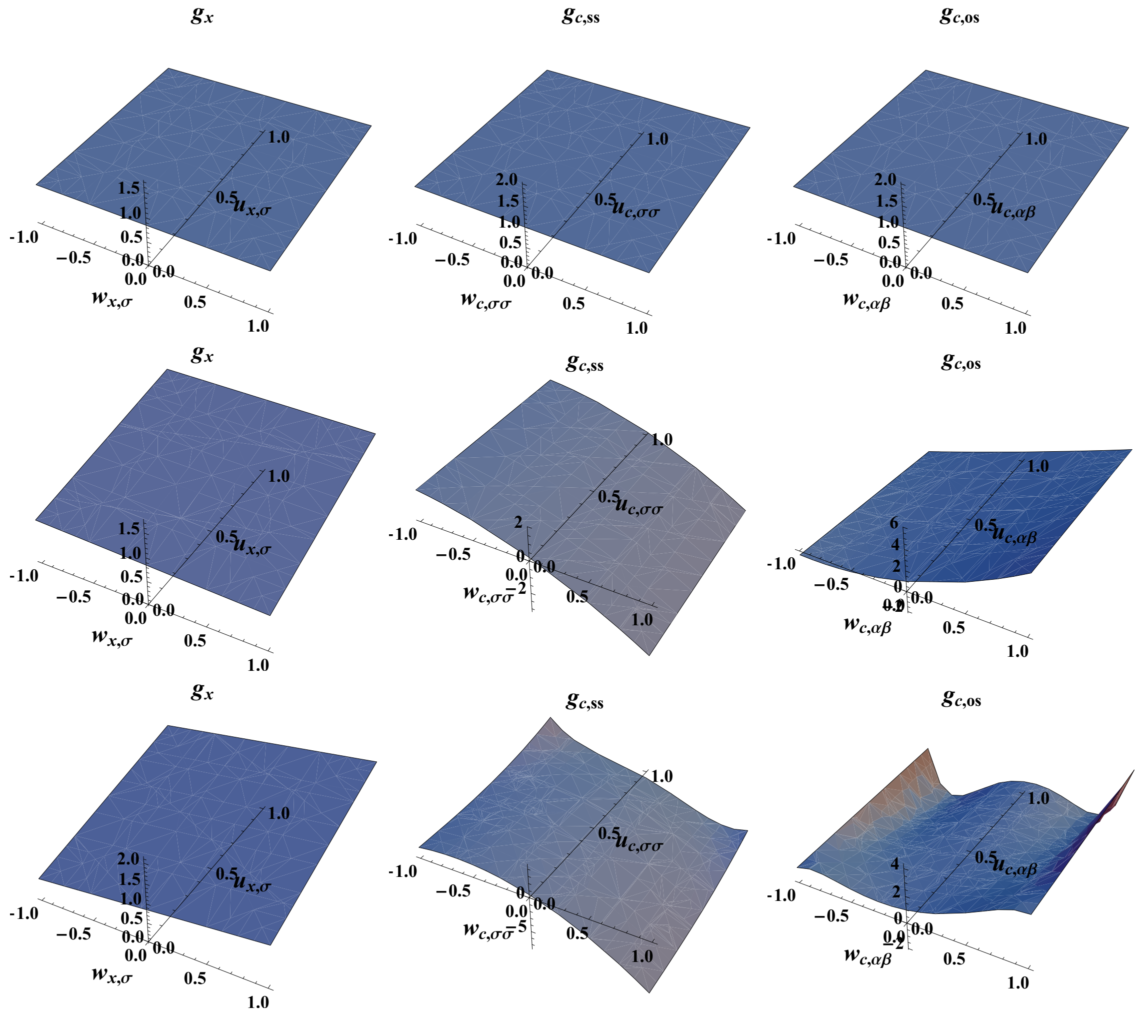


Figure 5.3: Exchange, same-spin correlation, and opposite-spin correlation inhomogeneity correction factor plots for the ωB97M-V density functional (bottom). In addition, the inhomogeneity correction factor plots for the SR-LSDA+VV10 initial guess (top) and the "Best 8" updated guess (middle) are included for comparison.

Parameter	1	2 (Best 8)	3	4	5	6 (Final)
$c_{x,00}$	0.85	0.85	0.85	0.85	0.85	0.85
$c_{x,10}$	0	-0.097	0.265	0.259	0.259	0.259
$c_{x,01}$	0	0.789	1.014	1.007	1.007	1.007
$c_{css,00}$	1	0.216	0.457	0.443	0.443	0.443
$c_{css,10}$	0	-2.496	-4.868	-4.55	-4.536	-4.535
$c_{css,20}$	0	-0.816	-3.401	-3.391	-3.39	-3.39
$c_{css,43}$	0	0	4.107	4.267	4.278	4.278
$c_{css,04}$	0	0	-1.533	-1.438	-1.437	-1.437
$c_{cos,00}$	1	1	1	1	1	1
$c_{cos,10}$	0	3.13	1.573	1.372	1.359	1.358
$c_{cos,20}$	0	1.736	3.002	2.931	2.925	2.924
$c_{cos,60}$	0	0	-1.736	-1.419	-1.392	-1.39
$c_{cos,21}$	0	-1.591	-8.241	-8.776	-8.81	-8.812
$c_{cos,61}$	0	0	8.582	9.113	9.141	9.142
c_x	0.15	0.15	0.15	0.15	0.15	0.15

Table 5.3: Linear parameters from the beginning of all six cycles of the self-consistent optimization of ωB97M-V. The "Best 8" column refers to the freezing of eight commonly-occurring parameters. The nonlinear parameters that are taken from previous studies[9, 35, 63] are γ_x=0.004, $\gamma_{c,ss}$=0.2, $\gamma_{c,os}$=0.006, ω=0.3, b=6, and C=0.01.

5.7 Results

In order to verify that ωB97M-V is a broadly accurate density functional, it is necessary to compare it to existing density functionals. Thus, 11 exemplary density functionals are selected for comparison to ωB97M-V across the nearly 5000 data points in the training, primary test, and secondary test sets. Furthermore, all 12 functionals are benchmarked on 90 potential energy curves in order to assess equilibrium binding energies and equilibrium bond lengths. Details for the 11 density functionals selected for comparison to ωB97M-V are given in Table 5.4.

5.7.1 Overall Test Set Performance

Since the training, primary test, and secondary test sets contain 84 datasets overall, it is easier to obtain a general idea of the performance of ωB97M-V by first considering the eight datatypes defined in Section 5.4. However, in order to make the comparison as unbiased as possible, only the data points from the primary and secondary test sets (henceforth referred to as the test set) are considered. Further-

Functional	# (Fitted)	c_x (ω)	Class	Rung	UEG	Ref.
B97-D3(BJ)	12	0	L GGA DFT-D3(BJ)	2	No	54,55,56
B97M-V	12	0	L meta-GGA VV10	3	Yes	64
ωB97X-V	10	16.7-100 (0.3)	RSH GGA VV10	4	No	63
ωB97M-V	12	15-100 (0.3)	RSH meta-GGA VV10	4	No	–
ωB97X-D	15	22.2-100 (0.2)	RSH GGA CHG	4	Yes	58
ωM05-D	21	37.0-100 (0.2)	RSH meta-GGA CHG	4	Yes	59
M06-2X	29	54	GH meta-GGA	4	Yes	45
M08-HX	47	52.23	GH meta-GGA	4	Yes	46
M11	40	42.8-100 (0.25)	RSH meta-GGA	4	Yes	52
MN15	59	44	GH meta-NGA	4	No	227
M06-L	34	0	L meta-GGA	3	Yes	23
MN15-L	58	0	L meta-NGA	3	No	26

Table 5.4: Details for the 11 exemplary functionals chosen for comparison to ωB97M-V. L stands for local, GH stands for global hybrid, and RSH stands for range-separated hybrid. The second column lists the number of parameters that were optimized on a training set for the given functional. The third column lists the percentage of exact exchange, c_x, as well as the value for ω in parentheses, if applicable. The column labeled UEG indicates whether or not the uniform electron gas limits were satisfied.

more, since RG10 does not belong to any of the datatypes, it is excluded from this analysis. The total number of data points and datasets considered in Figure 5.4 are 3547 and 58, respectively. The density functionals are compared using root-mean-square deviations (RMSD) because this convention provides a better summary of the distribution of residuals than the mean absolute deviation (MAD)[266].

Beginning with the NCED datatype, the functional with the best performance is ωB97M-V, with an RMSD of 0.18 kcal/mol. The next best functionals are ωB97X-V and B97M-V, both with RMSDs of 0.23 kcal/mol. Thus, the test set performance of ωB97M-V for over 1400 non-covalent dimer binding energies is at least 20% better than that of both of its predecessors. Besides the VV10-containing functionals, ωM05-D and ωB97X-D perform similarly, both with RMSDs of 0.38 kcal/mol, followed by M06-2X, B97-D3(BJ), and MN15, which have RMSDs of 0.42, 0.44, and 0.48 kcal/mol, respectively. The remaining functionals (M06-L, M08-HX, M11, and MN15-L) have RMSDs in excess of 0.5 kcal/mol, with MN15-L having by far the largest (1.39 kcal/mol).

Moving on to the 223 non-covalent cluster binding energies in the test set with the NCEC datatype, the best performance is once again reserved for ωB97M-V, which has an RMSD of 0.50 kcal/mol. The next best functional (ωB97X-V) is about 30%

worse, with an RMSD of 0.66 kcal/mol. Only two other functionals have RMSDs under or around 1 kcal/mol: B97M-V and ωB97X-D. The rest of the functionals have RMSDs that range from 1.73 kcal/mol (M08-HX) to 13.20 kcal/mol (MN15-L).

The NCD datatype contains non-covalent dimer binding energies that are susceptible to self-interaction error. Thus, local functionals should perform significantly worse than hybrid functionals, while hybrid functionals with a large fraction of exact exchange should perform best. Accordingly, the density functionals with the largest RMSDs across the 70 data points in the test set are the local ones, with RMSDs between 1.35 and 1.75 kcal/mol. Surprisingly, the best density functional is ωB97M-V, with an RMSD of 0.49 kcal/mol, followed by MN15, ωB97X-V, and ωM05-D, which have RMSDs of 0.61, 0.63, and 0.78 kcal/mol, respectively. The rest of the hybrid functionals have RMSDs around 1 kcal/mol.

The test set contains a total of 679 standard isomerization energies, ranging from alkane conformers to amino acid conformers. The three functionals with the best performance for the IE data points in the test set are ωB97X-V, B97M-V, and ωB97M-V, with RMSDs slightly under 0.30 kcal/mol. The next best functionals are M06-2X (0.52 kcal/mol) and ωM05-D (0.54 kcal/mol), while six of the remaining seven functionals (M08-HX, ωB97X-D, MN15, M06-L, M11, and B97-D3(BJ)) have RMSDs between 0.6 and 0.8 kcal/mol. MN15-L, on the other hand, has an RMSD (1.57 kcal/mol) that is more than 2 times larger than that of M06-L.

The ID datatype contains isomerization energies that are sensitive to self-interaction error. As with the NCD category, the local functionals exhibit the worst performance across the 88 data points in the test set, with RMSDs larger than 7.9 kcal/mol. From the hybrid functionals, ωB97X-V and ωB97M-V perform almost indistinguishably, with RMSDs of around 2.3 kcal/mol, while the next best functional, ωB97X-D, performs about 65% worse, with an RMSD of 3.84 kcal/mol. The other hybrids have RMSDs that are 2 to 3 times larger than that of ωB97M-V.

The 660 TCE data points in the test set include heavy-atom transfer energies, homolytic bond dissociation energies, as well as isodesmic reaction energies. ωB97M-V is the best-performing functional by a comfortable margin, with an RMSD of 2.45 kcal/mol. The next best functionals, M06-2X, MN15, and ωM05-D perform more than 30% worse, with RMSDs of 3.17, 3.19, and 3.28 kcal/mol, respectively. The remaining four hybrids (M08-HX, M11, ωB97X-V, and ωB97X-D) perform comparably and have RMSDs around 3.6 kcal/mol. The best local functionals are B97M-V (3.59 kcal/mol) and MN15-L (3.95 kcal/mol), both performing significantly better than M06-L (5.26 kcal/mol).

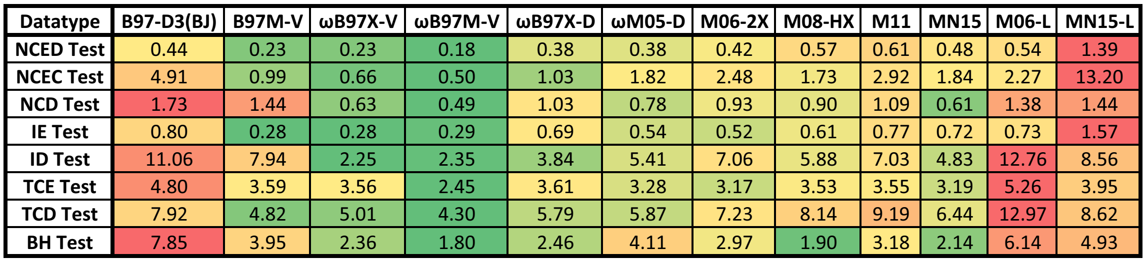
Of the 258 TCD data points in the test set, 234 are multi-reference atomization energies, bond dissociation energies, and heavy-atom transfer energies from the W4-11 database, while 24 are atomization energies and homodesmotic, isodesmic, and isogyric reactions from the Platonic24 dataset. While the first grouping should be a major challenge for hybrid functionals, the second grouping should present difficulties for local functionals. The best-performing density functional is a hybrid (ωB97M-V with an RMSD of 4.30 kcal/mol), followed by a local functional (B97M-V with an RMSD of 4.82 kcal/mol). While ωB97X-V comes in third with an RMSD of 5.01 kcal/mol, the next best functionals are ωB97X-D and ωM05-D, with RMSDs of 5.79 and 5.87 kcal/mol, respectively. MN15 and M06-2X perform about 50% and 65% worse than ωB97M-V, respectively, while M08-HX and M11 perform about 2 times worse. The second-best local functional is B97-D3(BJ), with an RMSD (7.92 kcal/mol) that is significantly smaller than that of M06-L (12.97 kcal/mol).

Finally, the BH datatype contains five test datasets (136 data points), two of which (HTBH38 and NHTBH38) are found in the Minnesota density functional training sets. Nevertheless, ωB97M-V, with an RMSD of 1.80 kcal/mol, has the smallest RMSD out of the 12 benchmarked functionals, followed by M08-HX, which performs only slightly worse, and MN15, which performs approximately 20% worse. Surprisingly, M06-2X is only sixth best, with an RMSD of 2.97 kcal/mol, followed closely by M11 (3.18 kcal/mol) and distantly by ωM05-D (4.11 kcal/mol). From the local functionals, B97M-V performs the best, with an RMSD of 3.95 kcal/mol, followed by MN15-L, which has an RMSD of 4.93 kcal/mol. M06-L performs about 25% worse than MN15-L, while the worst functional overall is B97-D3(BJ), with an RMSD of 7.85 kcal/mol.

Overall, the performance of ωB97M-V across 3547 test data points is very encouraging. Across the eight datatypes, ωB97M-V performs significantly better than the next best functional for five of the datatypes (NCED, NCEC, NCD, TCE, and TCD), is indistinguishable from the next best functional for one of the datatypes (BH), and is indistinguishable from the best functional for the two remaining datatypes (IE and ID). It is worth noting that the size of the test set (3547 data points) used to validate ωB97M-V is more than 7 times larger than the entire 2015A Minnesota database used to train and test the MN15-L functional. Furthermore, of the 4986 total data points in the training, primary test, and secondary test sets, only 870 data points are used for training, while the other 82.5% are used for testing. Thus, the transferability of ωB97M-V is satisfactorily demonstrated with the results documented thus far.

5.7.2 Results for Individual Datasets

Figures 5.5, 5.6, and 5.7 contain RMSDs for datasets in the training, primary test, and secondary test sets, respectively. Although there are 84 datasets in total, the AE18 and RG10 datasets from the training and primary test sets are excluded both because they are not associated with any of the datatypes and because the



Datatype	B97-D3(BJ)	B97M-V	ωB97X-V	ωB97M-V	ωB97X-D	ωM05-D	M06-2X	M08-HX	M11	MN15	M06-L	MN15-L
NCED Test	0.44	0.23	0.23	0.18	0.38	0.38	0.42	0.57	0.61	0.48	0.54	1.39
NCEC Test	4.91	0.99	0.66	0.50	1.03	1.82	2.48	1.73	2.92	1.84	2.27	13.20
NCD Test	1.73	1.44	0.63	0.49	1.03	0.78	0.93	0.90	1.09	0.61	1.38	1.44
IE Test	0.80	0.28	0.28	0.29	0.69	0.54	0.52	0.61	0.77	0.72	0.73	1.57
ID Test	11.06	7.94	2.25	2.35	3.84	5.41	7.06	5.88	7.03	4.83	12.76	8.56
TCE Test	4.80	3.59	3.56	2.45	3.61	3.28	3.17	3.53	3.55	3.19	5.26	3.95
TCD Test	7.92	4.82	5.01	4.30	5.79	5.87	7.23	8.14	9.19	6.44	12.97	8.62
BH Test	7.85	3.95	2.36	1.80	2.46	4.11	2.97	1.90	3.18	2.14	6.14	4.93

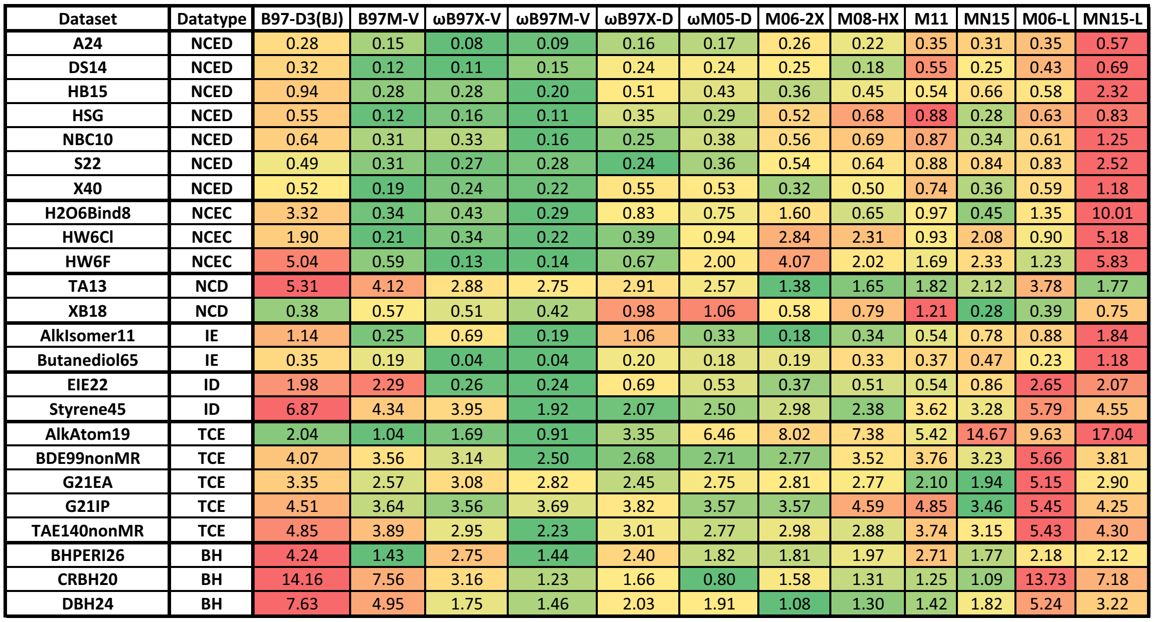
Figure 5.4: RMSDs in kcal/mol for 8 datatypes for 12 density functionals. These datatype RMSDs include data points from the primary and secondary test sets only. NCED stands for non-covalent dimers (easy), NCEC stands for non-covalent clusters (easy), NCD stands for non-covalent dimers (difficult), IE stands for isomerization energies (easy), ID stands for isomerization energies (difficult), TCE stands for thermochemistry (easy), TCD stands for thermochemistry (difficult), and BH stands for barrier heights. The partitioning of the 3547 data points contained in this figure into the 8 datatypes is: 1433, 223, 70, 679, 88, 660, 258, and 136.

corresponding RMSDs are not very meaningful. The performance of ωB97M-V on the training datasets will be discussed very briefly, since it is bad scientific practice to compare the performance of a semi-empirical density functional to that of existing functionals using its own training set.

One training result worth mentioning is that across the 124 atomization energies in the TAE140nonMR dataset[82], ωB97M-V affords an RMSD of 2.23 kcal/mol, which significantly improves over ωB97X-V (2.95 kcal/mol) and B97M-V (3.89 kcal/mol). This shows the improvement that is possible by including exact exchange, local meta-GGA contributions, as well as nonlocal correlation, since TAE140nonMR was included in the training set of all three functionals.

Moving away from the relatively unimportant training datasets toward the more meaningful primary test datasets (Figure 5.6), the performance of ωB97M-V is generally satisfactory, and the new functional has the smallest RMSD for 14 of the 34 primary test datasets considered. In order to circumvent the laborious process of documenting the performance of ωB97M-V for all 34 primary test datasets in Figure 5.6, only a handful of the datasets will be analyzed.

The S66 dataset[103, 105] is certainly the most popular of the NCED datasets in the primary test set, and the performance of ωB97M-V is very satisfactory. Its RMSD of 0.15 kcal/mol is slightly larger than that of ωB97X-V (0.13 kcal/mol), which is the best performer. ωB97M-V performs more than 2 times better than M06-2X, M08-HX, and ωB97X-D, more than 3 times better than ωM05-D, and more than 4



Dataset	Datatype	B97-D3(BJ)	B97M-V	ωB97X-V	ωB97M-V	ωB97X-D	ωM05-D	M06-2X	M08-HX	M11	MN15	M06-L	MN15-L
A24	NCED	0.28	0.15	0.08	0.09	0.16	0.17	0.26	0.22	0.35	0.31	0.35	0.57
DS14	NCED	0.32	0.12	0.11	0.15	0.24	0.24	0.25	0.18	0.55	0.25	0.43	0.69
HB15	NCED	0.94	0.28	0.28	0.20	0.51	0.43	0.36	0.45	0.54	0.66	0.58	2.32
HSG	NCED	0.55	0.12	0.16	0.11	0.35	0.29	0.52	0.68	0.88	0.28	0.63	0.83
NBC10	NCED	0.64	0.31	0.33	0.16	0.25	0.38	0.56	0.69	0.87	0.34	0.61	1.25
S22	NCED	0.49	0.31	0.27	0.28	0.24	0.36	0.54	0.64	0.88	0.84	0.83	2.52
X40	NCED	0.52	0.19	0.24	0.22	0.55	0.53	0.32	0.50	0.74	0.36	0.59	1.18
H2O6Bind8	NCEC	3.32	0.34	0.43	0.29	0.83	0.75	1.60	0.65	0.97	0.45	1.35	10.01
HW6Cl	NCEC	1.90	0.21	0.34	0.22	0.39	0.94	2.84	2.31	0.93	2.08	0.90	5.18
HW6F	NCEC	5.04	0.59	0.13	0.14	0.67	2.00	4.07	2.02	1.69	2.33	1.23	5.83
TA13	NCD	5.31	4.12	2.88	2.75	2.91	2.57	1.38	1.65	1.82	2.12	3.78	1.77
XB18	NCD	0.38	0.57	0.51	0.42	0.98	1.06	0.58	0.79	1.21	0.28	0.39	0.75
AlkIsomer11	IE	1.14	0.25	0.69	0.19	1.06	0.33	0.18	0.34	0.54	0.78	0.88	1.84
Butanediol65	IE	0.35	0.19	0.04	0.04	0.20	0.18	0.19	0.33	0.37	0.47	0.23	1.18
EIE22	ID	1.98	2.29	0.26	0.24	0.69	0.53	0.37	0.51	0.54	0.86	2.65	2.07
Styrene45	ID	6.87	4.34	3.95	1.92	2.07	2.50	2.98	2.38	3.62	3.28	5.79	4.55
AlkAtom19	TCE	2.04	1.04	1.69	0.91	3.35	6.46	8.02	7.38	5.42	14.67	9.63	17.04
BDE99nonMR	TCE	4.07	3.56	3.14	2.50	2.68	2.71	2.77	3.52	3.76	3.23	5.66	3.81
G21EA	TCE	3.35	2.57	3.08	2.82	2.45	2.75	2.81	2.77	2.10	1.94	5.15	2.90
G21IP	TCE	4.51	3.64	3.56	3.69	3.82	3.57	3.57	4.59	4.85	3.46	5.45	4.25
TAE140nonMR	TCE	4.85	3.89	2.95	2.23	3.01	2.77	2.98	2.88	3.74	3.15	5.43	4.30
BHPERI26	BH	4.24	1.43	2.75	1.44	2.40	1.82	1.81	1.97	2.71	1.77	2.18	2.12
CRBH20	BH	14.16	7.56	3.16	1.23	1.66	0.80	1.58	1.31	1.25	1.09	13.73	7.18
DBH24	BH	7.63	4.95	1.75	1.46	2.03	1.91	1.08	1.30	1.42	1.82	5.24	3.22

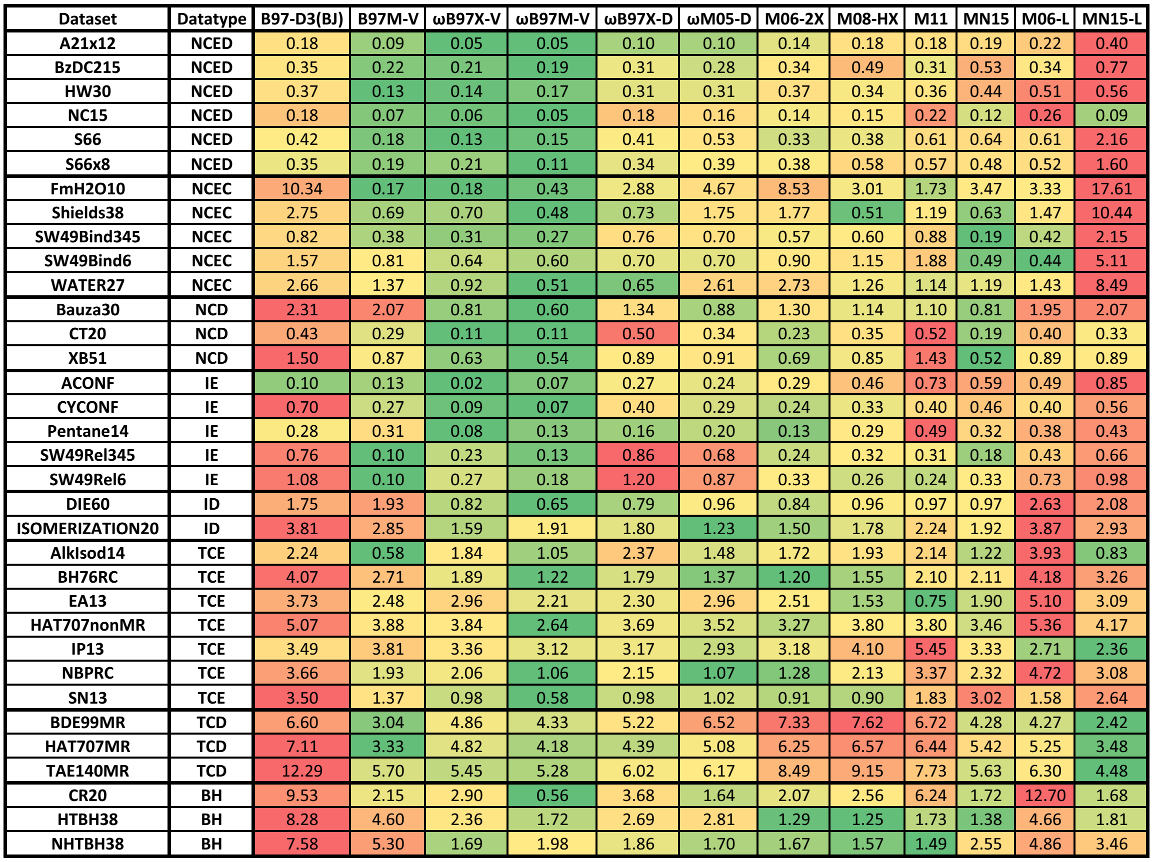
Figure 5.5: RMSDs in kcal/mol for 24 of the 25 training datasets (AE18 is excluded) for 12 density functionals. Table 5.1 contains information regarding the datasets, and the datatypes are explained in Section 5.4.

times better than M11 and MN15. At 0.18 kcal/mol, B97M-V has the third smallest RMSD, and is the best local functional tested.

The Shields38 dataset[223] contains 38 water clusters, ranging from dimers to decamers. ωB97M-V performs the best, with an RMSD of 0.48 kcal/mol, followed by M08-HX (0.51 kcal/mol), MN15 (0.63 kcal/mol), and B97M-V, ωB97X-V, and ωB97X-D, which have RMSDs around 0.7 kcal/mol. The rest of the functionals have RMSDs between 1 and 3 kcal/mol, with the exception of MN15-L, which performs very poorly (10.44 kcal/mol). Binding energies for larger clusters are evaluated later, with the 14 water 20-mers in the secondary test set.

The CYCONF dataset is taken from the GMTKN30 database[157, 216] and contains the isomerization energies of 10 cysteine conformers. Again, ωB97M-V performs the best for this dataset, with a small RMSD of 0.07 kcal/mol that is virtually indistinguishable from that of ωB97X-V, and nearly 3.5 times better than that of the next best functional, M06-2X.

The HAT707nonMR dataset from the W4-11 database[82] contains 505 heavy-atom transfer energies, and is one of the largest datasets in the primary test set. ωB97M-V affords an impressive RMSD of 2.64 kcal/mol on this dataset, performing



Dataset	Datatype	B97-D3(BJ)	B97M-V	ωB97X-V	ωB97M-V	ωB97X-D	ωM05-D	M06-2X	M08-HX	M11	MN15	M06-L	MN15-L
A21x12	NCED	0.18	0.09	0.05	0.05	0.10	0.10	0.14	0.18	0.18	0.19	0.22	0.40
BzDC215	NCED	0.35	0.22	0.21	0.19	0.31	0.28	0.34	0.49	0.31	0.53	0.34	0.77
HW30	NCED	0.37	0.13	0.14	0.17	0.31	0.31	0.37	0.34	0.36	0.44	0.51	0.56
NC15	NCED	0.18	0.07	0.06	0.05	0.18	0.16	0.14	0.15	0.22	0.12	0.26	0.09
S66	NCED	0.42	0.18	0.13	0.15	0.41	0.53	0.33	0.38	0.61	0.64	0.61	2.16
S66x8	NCED	0.35	0.19	0.21	0.11	0.34	0.39	0.38	0.58	0.57	0.48	0.52	1.60
FmH2O10	NCEC	10.34	0.17	0.18	0.43	2.88	4.67	8.53	3.01	1.73	3.47	3.33	17.61
Shields38	NCEC	2.75	0.69	0.70	0.48	0.73	1.75	1.77	0.51	1.19	0.63	1.47	10.44
SW49Bind345	NCEC	0.82	0.38	0.31	0.27	0.76	0.70	0.57	0.60	0.88	0.19	0.42	2.15
SW49Bind6	NCEC	1.57	0.81	0.64	0.60	0.70	0.70	0.90	1.15	1.88	0.49	0.44	5.11
WATER27	NCEC	2.66	1.37	0.92	0.51	0.65	2.61	2.73	1.26	1.14	1.19	1.43	8.49
Bauza30	NCD	2.31	2.07	0.81	0.60	1.34	0.88	1.30	1.14	1.10	0.81	1.95	2.07
CT20	NCD	0.43	0.29	0.11	0.11	0.50	0.34	0.23	0.35	0.52	0.19	0.40	0.33
XB51	NCD	1.50	0.87	0.63	0.54	0.89	0.91	0.69	0.85	1.43	0.52	0.89	0.89
ACONF	IE	0.10	0.13	0.02	0.07	0.27	0.24	0.29	0.46	0.73	0.59	0.49	0.85
CYCONF	IE	0.70	0.27	0.09	0.07	0.40	0.29	0.24	0.33	0.40	0.46	0.40	0.56
Pentane14	IE	0.28	0.31	0.08	0.13	0.16	0.20	0.13	0.29	0.49	0.32	0.38	0.43
SW49Rel345	IE	0.76	0.10	0.23	0.13	0.86	0.68	0.24	0.32	0.31	0.18	0.43	0.66
SW49Rel6	IE	1.08	0.10	0.27	0.18	1.20	0.87	0.33	0.26	0.24	0.33	0.73	0.98
DIE60	ID	1.75	1.93	0.82	0.65	0.79	0.96	0.84	0.96	0.97	0.97	2.63	2.08
ISOMERIZATION20	ID	3.81	2.85	1.59	1.91	1.80	1.23	1.50	1.78	2.24	1.92	3.87	2.93
AlkIsod14	TCE	2.24	0.58	1.84	1.05	2.37	1.48	1.72	1.93	2.14	1.22	3.93	0.83
BH76RC	TCE	4.07	2.71	1.89	1.22	1.79	1.37	1.20	1.55	2.10	2.11	4.18	3.26
EA13	TCE	3.73	2.48	2.96	2.21	2.30	2.96	2.51	1.53	0.75	1.90	5.10	3.09
HAT707nonMR	TCE	5.07	3.88	3.84	2.64	3.69	3.52	3.27	3.80	3.80	3.46	5.36	4.17
IP13	TCE	3.49	3.81	3.36	3.12	3.17	2.93	3.18	4.10	5.45	3.33	2.71	2.36
NBPRC	TCE	3.66	1.93	2.06	1.06	2.15	1.07	1.28	2.13	3.37	2.32	4.72	3.08
SN13	TCE	3.50	1.37	0.98	0.58	0.98	1.02	0.91	0.90	1.83	3.02	1.58	2.64
BDE99MR	TCD	6.60	3.04	4.86	4.33	5.22	6.52	7.33	7.62	6.72	4.28	4.27	2.42
HAT707MR	TCD	7.11	3.33	4.82	4.18	4.39	5.08	6.25	6.57	6.44	5.42	5.25	3.48
TAE140MR	TCD	12.29	5.70	5.45	5.28	6.02	6.17	8.49	9.15	7.73	5.63	6.30	4.48
CR20	BH	9.53	2.15	2.90	0.56	3.68	1.64	2.07	2.56	6.24	1.72	12.70	1.68
HTBH38	BH	8.28	4.60	2.36	1.72	2.69	2.81	1.29	1.25	1.73	1.38	4.66	1.81
NHTBH38	BH	7.58	5.30	1.69	1.98	1.86	1.70	1.67	1.57	1.49	2.55	4.86	3.46

Figure 5.6: RMSDs in kcal/mol for 34 of the 35 primary test datasets (RG10 is excluded) for 12 density functionals. Table 5.1 contains information regarding the datasets, and the datatypes are explained in Section 5.4.

nearly 20% better than the next best functional, M06-2X (3.27 kcal/mol), and 30% better than ωB97X-V (3.84 kcal/mol).

The most interesting datasets contained in this chapter are found in the secondary test set, as most of them are taken from papers that were very recently published. Furthermore, the secondary test set is the truest form of transferability testing, as it was compiled and evaluated after ωB97M-V was fully self-consistently trained. Thus, the bulk of the remaining discussion in this section will be focused on the datasets from the secondary test set (Figure 5.7).

3B-69-DIM is a dataset created from the 3B-69 dataset of Beran and coworkers[235], and contains all relevant pairs of monomers that can be constructed from the 69 trimers. This results in a total of 207 dimer binding energies and serves as a stringent test of transferability for the new functional. ωB97M-V performs outstandingly for this dataset, with an RMSD of 0.16 kcal/mol, followed by ωB97X-V and B97M-V, with RMSDs of 0.20 and 0.21 kcal/mol, respectively. The next best functional, ωM05-D, performs nearly 2 times worse than ωB97M-V, while the best Minnesota functional, MN15, performs 3 times worse.

Since the two NCED datasets discussed thus far (S66 and 3B-69-DIM) contain a mixture of hydrogen-bonded and dispersion-bound systems, it is useful to consider a dataset that deals specifically with the latter. For this purpose, the AlkBind12 dataset[236], which contains saturated and unsaturated hydrocarbon dimers, will be analyzed. The RMS of the 12 reference energies comprising AlkBind12 is only 3.14 kcal/mol, indicating that the systems are fairly weakly bound. The two best functionals are ωB97X-V and ωB97M-V, with almost identical RMSDs of 0.12 and 0.13 kcal/mol, respectively. Only three other functionals manage RMSDs under 0.3 kcal/mol (an error of roughly 10%): B97M-V, M08-HX, and M06-2X. MN15 overbinds the dimers with an RMSD (1.18 kcal/mol) that is almost 10 times larger than that of ωB97M-V, while the worst-performing functional is MN15-L, which completely overbinds the alkanes with an RMSD of 3.49 kcal/mol.

The HB49 dataset[238–240] is a very interesting dataset constructed by Boese, and contains the binding energies of 49 hydrogen-bonded dimers. In fact, a recent benchmark of density functionals on the HB49 dataset found that MP2 at the basis set limit, with an RMSD of approximately 0.3 kcal/mol, performed better than all of the tested Rung 1-4 density functionals. Therefore, it is of interest to assess the performance of ωB97M-V on the HB49 dataset. The results are very encouraging: with a low RMSD of 0.23 kcal/mol, ωB97M-V is the only functional tested which significantly outperforms MP2. In addition, ωB97X-V performs very comparably to MP2, with an RMSD of 0.29 kcal/mol. From the local functionals, B97M-V performs best, with an RMSD of 0.47 kcal/mol.

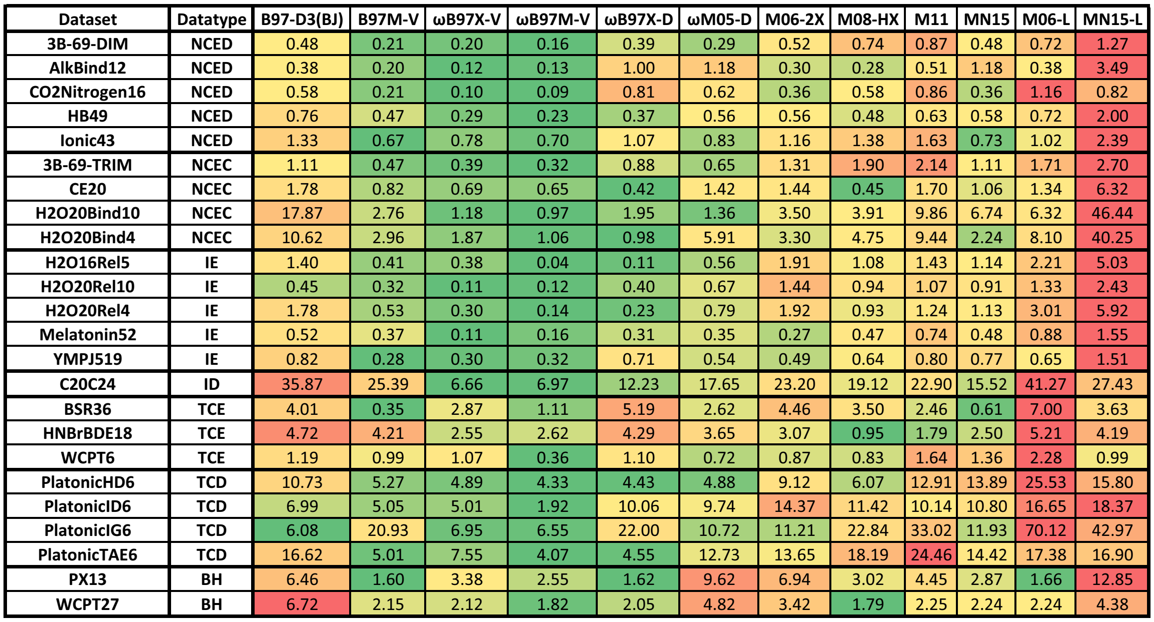
While the 3B-69 dataset was originally intended as a benchmark for three-body intermolecular interaction energies, it can also be used as a benchmark for trimer binding energies (3B-69-TRIM). This is a good transferability test for ωB97M-V, since very few trimers are found in its training set. ωB97M-V performs very well for this dataset, with the smallest RMSD of 0.32 kcal/mol. The next best functionals are ωB97X-V and B97M-V, with RMSDs of 0.39 and 0.47 kcal/mol, respectively. Only two other functionals manage RMSDs under 1 kcal/mol: ωM05-D and ωB97X-D, with RMSDs of 0.65 and 0.88 kcal/mol, respectively. The best Minnesota functional, MN15 (1.11 kcal/mol), performs more than 3 times worse than ωB97M-V.

Having tested the performance of ωB97M-V for small- to medium-sized water clusters with the Shields38 dataset in the primary test set, it is time to consider the H2O20Bind10 and H2O20Bind4 datasets in the secondary test set, as they contain a total of 14 water 20-mer binding energies. In order to address both datasets simultaneously, the geometric mean of the two RMSDs (GMRMSD) will be considered. ωB97M-V performs best overall, with a GMRMSD of 1.01 kcal/mol, while ωB97X-D and ωB97X-V perform second and third best, with GMRMSDs of 1.38 and 1.48 kcal/mol. ωM05-D and B97M-V perform very similarly, with GMRMSDs around 2.85 kcal/mol, while M06-2X is the best of the tested Minnesota functionals, with a GMRMSD of 3.40 kcal/mol. After B97M-V, the next best local functional is M06-L, with a GMRMSD (7.16 kcal/mol) that is 6 times smaller than that of MN15-L (43.24 kcal/mol).

While the binding energies of small, medium, and large water clusters have been thoroughly addressed thus far, it is important to assess the performance of ωB97M-V for the relative energies of water clusters. This is done with the help of three datasets from the secondary test set: H2O16Rel5, H2O20Rel10, and H2O20Rel4. Once again, the geometric mean of these three datasets will be considered for brevity. The GMRMSD of ωB97M-V across these three datasets is remarkably small, at only 0.09 kcal/mol. The next best functionals are ωB97X-D and ωB97X-V, with very similar GMRMSDs of 0.22 and 0.24 kcal/mol, respectively. B97M-V is the best local functional, with a GMRMSD of 0.41 kcal/mol, while the two remaining non-Minnesota functionals (ωM05-D and B97-D3(BJ)) have GMRMSDs of 0.67 and 1.04 kcal/mol, respectively. All of the tested Minnesota functionals perform poorly for these isomerization energies, with GMRMSDs ranging from 0.98 kcal/mol (M08-HX) to 4.17 kcal/mol (MN15-L).

A recent benchmark by Karton and coworkers on the YMPJ519 dataset of amino acid isomerization energies[253] found ωB97X-V to be the best Rung 1-4 density functional. Thus, it is important to verify that ωB97M-V performs as well as its GGA counterpart. Accordingly, both ωB97X-V and ωB97M-V have impressive RMSDs of 0.30 and 0.32 kcal/mol, respectively, while the smallest RMSD is surprisingly reserved for B97M-V, at 0.28 kcal/mol. The rest of the functionals have RMSDs that range from 0.49 kcal/mol (M06-2X) to 0.77 kcal/mol (MN15) to 1.51 kcal/mol (MN15-L).

Another recent benchmark by Martin and coworkers assessed the performance of various density functionals for the relative energies of a handful of C_{20} and C_{24} structures[257]. The study found that only double hybrid functionals were able to afford RMSDs under 10 kcal/mol (with the smallest RMSD being around 8.3 kcal/mol). Therefore, it is interesting to assess the performance of ωB97M-V on this dataset to see if it can break the 10 kcal/mol barrier. Both ωB97X-V and ωB97M-V manage RMSDs under 7 kcal/mol, with the former slightly outperforming the latter. By contrast, the best local functional is B97M-V with an RMSD of 25.39 kcal/mol, followed closely by MN15-L (27.43 kcal/mol). From the remaining hybrid functionals, ωB97X-D, MN15, ωM05-D, and M08-HX manage RMSDs under 20 kcal/mol, while

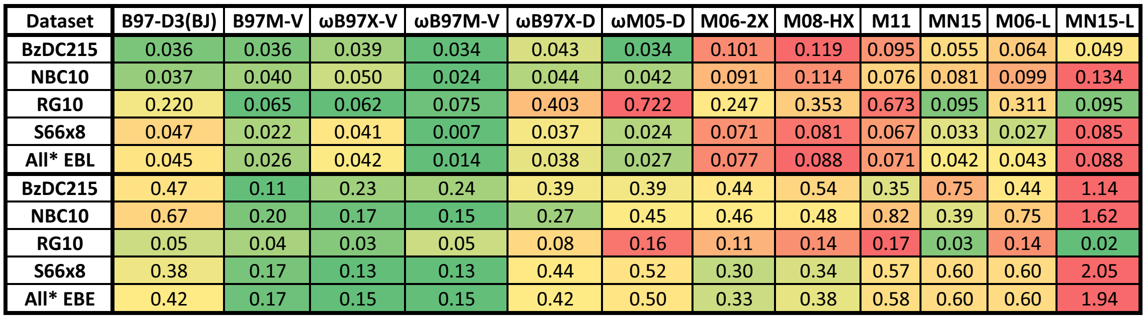


Dataset	Datatype	B97-D3(BJ)	B97M-V	ωB97X-V	ωB97M-V	ωB97X-D	ωM05-D	M06-2X	M08-HX	M11	MN15	M06-L	MN15-L
3B-69-DIM	NCED	0.48	0.21	0.20	0.16	0.39	0.29	0.52	0.74	0.87	0.48	0.72	1.27
AlkBind12	NCED	0.38	0.20	0.12	0.13	1.00	1.18	0.30	0.28	0.51	1.18	0.38	3.49
CO2Nitrogen16	NCED	0.58	0.21	0.10	0.09	0.81	0.62	0.36	0.58	0.86	0.36	1.16	0.82
HB49	NCED	0.76	0.47	0.29	0.23	0.37	0.56	0.56	0.48	0.63	0.58	0.72	2.00
Ionic43	NCED	1.33	0.67	0.78	0.70	1.07	0.83	1.16	1.38	1.63	0.73	1.02	2.39
3B-69-TRIM	NCEC	1.11	0.47	0.39	0.32	0.88	0.65	1.31	1.90	2.14	1.11	1.71	2.70
CE20	NCEC	1.78	0.82	0.69	0.65	0.42	1.42	1.44	0.45	1.70	1.06	1.34	6.32
H2O20Bind10	NCEC	17.87	2.76	1.18	0.97	1.95	1.36	3.50	3.91	9.86	6.74	6.32	46.44
H2O20Bind4	NCEC	10.62	2.96	1.87	1.06	0.98	5.91	3.30	4.75	9.44	2.24	8.10	40.25
H2O16Rel5	IE	1.40	0.41	0.38	0.04	0.11	0.56	1.91	1.08	1.43	1.14	2.21	5.03
H2O20Rel10	IE	0.45	0.32	0.11	0.12	0.40	0.67	1.44	0.94	1.07	0.91	1.33	2.43
H2O20Rel4	IE	1.78	0.53	0.30	0.14	0.23	0.79	1.92	0.93	1.24	1.13	3.01	5.92
Melatonin52	IE	0.52	0.37	0.11	0.16	0.31	0.35	0.27	0.47	0.74	0.48	0.88	1.55
YMPJ519	IE	0.82	0.28	0.30	0.32	0.71	0.54	0.49	0.64	0.80	0.77	0.65	1.51
C20C24	ID	35.87	25.39	6.66	6.97	12.23	17.65	23.20	19.12	22.90	15.52	41.27	27.43
BSR36	TCE	4.01	0.35	2.87	1.11	5.19	2.62	4.46	3.50	2.46	0.61	7.00	3.63
HNBrBDE18	TCE	4.72	4.21	2.55	2.62	4.29	3.65	3.07	0.95	1.79	2.50	5.21	4.19
WCPT6	TCE	1.19	0.99	1.07	0.36	1.10	0.72	0.87	0.83	1.64	1.36	2.28	0.99
PlatonicHD6	TCD	10.73	5.27	4.89	4.33	4.43	4.88	9.12	6.07	12.91	13.89	25.53	15.80
PlatonicID6	TCD	6.99	5.05	5.01	1.92	10.06	9.74	14.37	11.42	10.14	10.80	16.65	18.37
PlatonicIG6	TCD	6.08	20.93	6.95	6.55	22.00	10.72	11.21	22.84	33.02	11.93	70.12	42.97
PlatonicTAE6	TCD	16.62	5.01	7.55	4.07	4.55	12.73	13.65	18.19	24.46	14.42	17.38	16.90
PX13	BH	6.46	1.60	3.38	2.55	1.62	9.62	6.94	3.02	4.45	2.87	1.66	12.85
WCPT27	BH	6.72	2.15	2.12	1.82	2.05	4.82	3.42	1.79	2.25	2.24	2.24	4.38

Figure 5.7: RMSDs in kcal/mol for the 24 secondary test datasets for 12 density functionals. Table 5.1 contains information regarding the datasets, and the datatypes are explained in Section 5.4.

M11 and M06-2X perform more than 3 times worse than ωB97M-V.

A very challenging benchmark set by Martin and coworkers[262] assessed the performance of density functionals for various reactions (homodesmotic, isodesmic, and isogyric) involving platonic hydrocarbon cages, in addition to their atomization energies. While the individual RMSDs for these four datasets are given in Figure 5.7, the functionals will be assessed based on the geometric mean of the four RMSDs. Although the original study found ωB97X-V to be the most promising Rung 1-4 functional overall, ωB97M-V manages a GMRMSD of only 3.86 kcal/mol, compared to ωB97X-V's GMRMSD of 5.99 kcal/mol. Thus, ωB97M-V improves over ωB97X-V by over 35%. The next best functional (B97M-V) is surprisingly a local one, and has a GMRMSD of 7.27 kcal/mol. The performance of B97M-V is certainly noteworthy, since the two local Minnesota functionals have GMRMSDs larger than 20 kcal/mol. While M06-2X, MN15, and M08-HX perform at least 3 times worse than ωB97M-V, M11 performs more than 4.5 times worse than ωB97M-V.



Dataset	B97-D3(BJ)	B97M-V	ωB97X-V	ωB97M-V	ωB97X-D	ωM05-D	M06-2X	M08-HX	M11	MN15	M06-L	MN15-L
BzDC215	0.036	0.036	0.039	0.034	0.043	0.034	0.101	0.119	0.095	0.055	0.064	0.049
NBC10	0.037	0.040	0.050	0.024	0.044	0.042	0.091	0.114	0.076	0.081	0.099	0.134
RG10	0.220	0.065	0.062	0.075	0.403	0.722	0.247	0.353	0.673	0.095	0.311	0.095
S66x8	0.047	0.022	0.041	0.007	0.037	0.024	0.071	0.081	0.067	0.033	0.027	0.085
All* EBL	0.045	0.026	0.042	0.014	0.038	0.027	0.077	0.088	0.071	0.042	0.043	0.088
BzDC215	0.47	0.11	0.23	0.24	0.39	0.39	0.44	0.54	0.35	0.75	0.44	1.14
NBC10	0.67	0.20	0.17	0.15	0.27	0.45	0.46	0.48	0.82	0.39	0.75	1.62
RG10	0.05	0.04	0.03	0.05	0.08	0.16	0.11	0.14	0.17	0.03	0.14	0.02
S66x8	0.38	0.17	0.13	0.13	0.44	0.52	0.30	0.34	0.57	0.60	0.60	2.05
All* EBE	0.42	0.17	0.15	0.15	0.42	0.50	0.33	0.38	0.58	0.60	0.60	1.94

Figure 5.8: Equilibrium bond length (EBL) RMSDs in Å and equilibrium binding energy (EBE) RMSDs in kcal/mol for 12 density functionals. The first section contains the EBL RMSDs while the second section contains the EBE RMSDs. The All* category contains 81 data points and is a combination of BzDC215, NBC10, and S66x8. More information regarding the datasets and excluded potential energy curves can be found in Table 5.1 and Section 5.7.3, respectively.

5.7.3 Potential Energy Curves

Within the NCED category, the BzDC215, NBC10, and S66x8 datasets contain potential energies curves that can be used to assess the accuracy of density functionals for predicting equilibrium properties of dimers. Furthermore, the RG10 dataset contains all 10 PECs that can be constructed between the rare-gas dimers from helium to krypton. In total, these four datasets contain 96 PECs, with BzDC215, NBC10, and RG10 each having 10, and S66x8 having 66. Unfortunately, even with the (99,590)/SG-1 grid, some of the resulting potential energy curves are too oscillatory to be accurately interpolated[141–143], primarily for the Minnesota density functionals. Consequently, the benzene-neon dimer and the benzene-argon dimer PECs from BzDC215 were removed, the sandwich benzene dimer, the methane dimer, and the sandwich (S2) pyridine dimer PECs from NBC10 were removed, and the helium dimer PEC from RG10 was removed, leaving a total of 90 potential energy curves. Figure 5.8 contains the equilibrium bond length (EBL) and equilibrium binding energy (EBE) RMSDs for these four datasets, along with the corresponding total RMSDs with RG10 excluded (All*). In order to keep the discussion succinct, only the RG10 and All* results will be discussed.

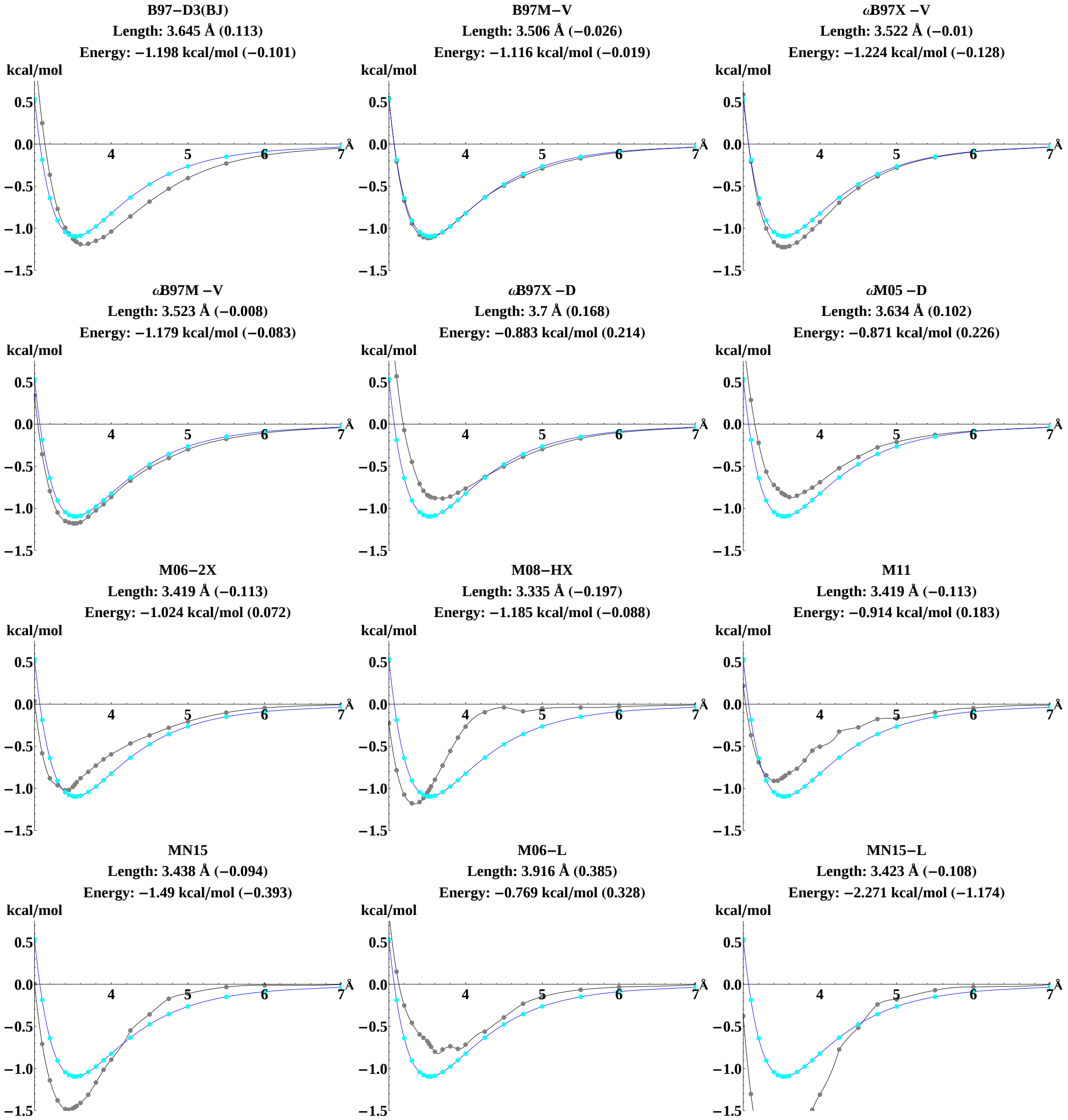
For the nine rare-gas dimers, the three VV10-containing functionals predict reasonably accurate equilibrium bond lengths, with RMSDs around 0.07 Å. Only two other functionals manage an EBL RMSD of under 0.1 Å for the rare-gas dimers: MN15-L and MN15. However, it is important to mention that both of these functionals were fit to at least a single point from the PEC of six of the nine rare-gas dimers considered here. The rest of the Minnesota functionals perform very poorly, with RMSDs between 0.2 and 0.7 Å. ωB97X-D also performs poorly, with an RMSD of 0.403, while the worst overall performer is ωM05-D, with an RMSD of 0.722 Å.

Moving on to the 81 PECs in the All* category, the best performance is exhibited by ωB97M-V, with a very impressive equilibrium bond length RMSD of only 0.014 Å. In fact, ωB97M-V performs almost 2 times better than the next best functional, B97M-V, and 3 times better than ωB97X-V. The six Minnesota functionals have RMSDs that range from 0.042 Å (MN15) to 0.088 Å (M08-HX and MN15-L), while ωM05-D, a range-separated hybrid functional based on the M05 functional form[41], performs almost as well as B97M-V, with an RMSD of 0.027 Å.

As for the All* equilibrium binding energies, ωB97X-V, ωB97M-V, and B97M-V perform very well, with RMSDs between 0.15 and 0.17 kcal/mol, while the rest of the functionals (except MN15-L) have RMSDs that range from 0.33 kcal/mol (M06-2X) to 0.60 kcal/mol (M06-L and MN15). MN15-L, on the other hand, has an All* EBE RMSD of nearly 2 kcal/mol, which is more than 3 times larger than that of M06-L.

Although the benzene-argon dimer was removed from the BzDC215 dataset in order to generate the RMSDs discussed thus far, it is nevertheless an interesting example of a system bound primarily by dispersion. Furthermore, due to the inherent weakness of the interaction, it is a case that can be used to assess the sensitivity of density functionals (especially meta-GGAs) to the integration grid. Figure 5.9 displays the PEC for the benzene-argon dimer as calculated by the 12 benchmarked density functionals with the (99,590)/SG-1 grid. It is evident that the grid filtering that was applied in Section 5.6 worked successfully, since the PEC of ωB97M-V is nearly as smooth as that of ωB97X-V for this system. By contrast, the Minnesota functionals are far harder to converge with respect to the grid, with M06-2X, M08-HX, and MN15 appearing to behave better than M06-L, M11, and MN15-L.

Considering the accuracy of the PECs themselves, ωB97M-V, ωB97X-V, and B97M-V are very accurate, with equilibrium bond length errors of -0.008, -0.01, and -0.026 Å, respectively. Of the remaining nine functionals, ωM05-D, B97-D3(BJ), ωB97X-D, and M06-L predict equilibrium bond lengths that are at least 0.1 Å too long, while MN15, MN15-L, M06-2X, M11, and M08-HX predict equilibrium bond lengths that are at least 0.1 Å too short. The three VV10-containing functionals manage to reproduce the equilibrium binding energy rather well, with the largest error (-0.128 kcal/mol) attributed to ωB97X-V, an error of -0.083 kcal/mol attributed to ωB97M-V, and the smallest error associated with B97M-V (-0.019 kcal/mol). Despite predicting a bond length that is more than 0.1 Å too short, M06-2X underestimates the EBE of the benzene-argon dimer by only 0.072 kcal/mol. By contrast, MN15-L overbinds the system by more than 1.15 kcal/mol.



B97−D3(BJ)
Length: 3.645 Å (0.113)
Energy: −1.198 kcal/mol (−0.101)

B97M−V
Length: 3.506 Å (−0.026)
Energy: −1.116 kcal/mol (−0.019)

ωB97X −V
Length: 3.522 Å (−0.01)
Energy: −1.224 kcal/mol (−0.128)

ωB97M −V
Length: 3.523 Å (−0.008)
Energy: −1.179 kcal/mol (−0.083)

ωB97X −D
Length: 3.7 Å (0.168)
Energy: −0.883 kcal/mol (0.214)

ωM05 −D
Length: 3.634 Å (0.102)
Energy: −0.871 kcal/mol (0.226)

M06−2X
Length: 3.419 Å (−0.113)
Energy: −1.024 kcal/mol (0.072)

M08−HX
Length: 3.335 Å (−0.197)
Energy: −1.185 kcal/mol (−0.088)

M11
Length: 3.419 Å (−0.113)
Energy: −0.914 kcal/mol (0.183)

MN15
Length: 3.438 Å (−0.094)
Energy: −1.49 kcal/mol (−0.393)

M06−L
Length: 3.916 Å (0.385)
Energy: −0.769 kcal/mol (0.328)

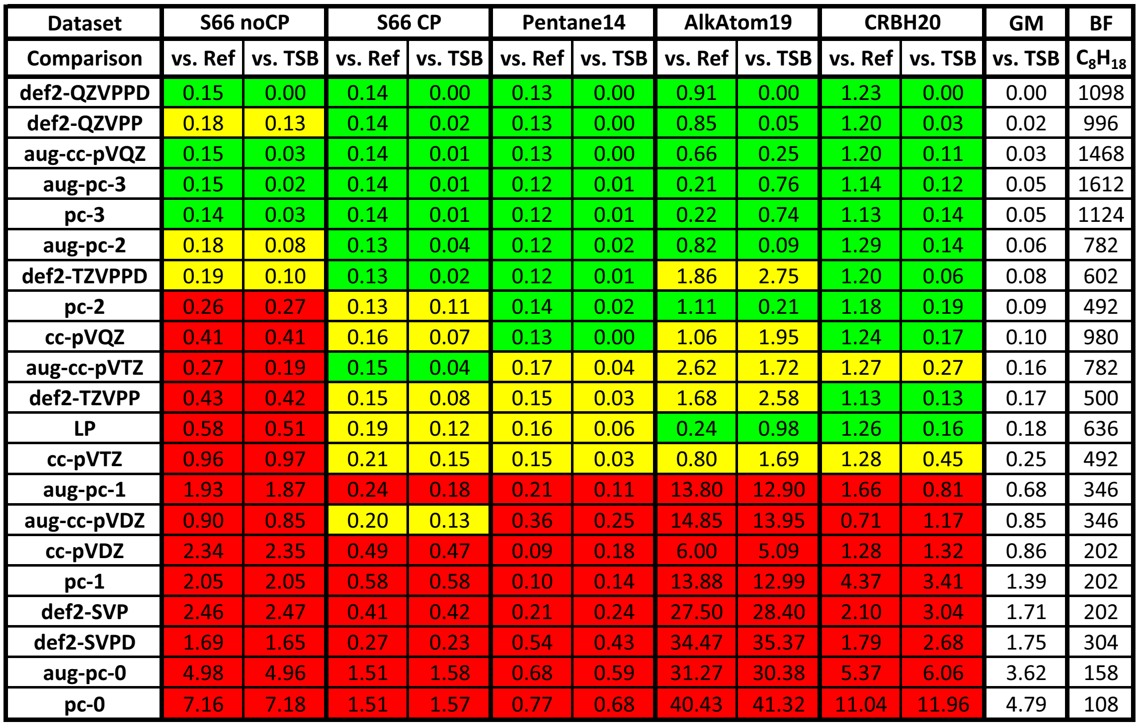
MN15−L
Length: 3.423 Å (−0.108)
Energy: −2.271 kcal/mol (−1.174)

Figure 5.9: Potential energy curves (computed with the def2-QZVPPD basis set and the (99,590)/SG-1 grid) for the benzene-argon dimer from BzDC215 as computed by the 12 benchmarked density functionals. The gray curve represents the DFT method, while the blue curve represents the reference method. The line immediately following the functional name contains the equilibrium bond length in Å and the error (with respect to the reference) in parentheses. The following line contains the same information for the equilibrium binding energy (in kcal/mol).

5.8 Reaching the Basis Set Limit

Although ωB97M-V was consistently trained and tested in the def2-QZVPPD basis set (without counterpoise corrections), it is inevitable that it will be used with different basis sets. As a result, this section explores the use of ωB97M-V with 21 basis sets from 4 different families, and makes recommendations based on how closely these basis sets can mimic the results of the training set basis (TSB), def2-QZVPPD. For this purpose, four datasets are selected and tested: S66 representing non-covalent interactions, Pentane14 representing isomerization energies, AlkAtom19 representing thermochemistry, and CRBH20 representing barrier heights. For the S66 dataset, the calculations are performed both with and without counterpoise corrections (designated CP and noCP, respectively), because it is very unlikely that a double- or triple-zeta basis set without counterpoise corrections will be able to reproduce the quadruple-zeta, def2-QZVPPD basis set binding energies. The results, summarized in Figure 5.10, are analyzed using two sets of RMSDs (the first relative to the reference values and the second relative to the def2-QZVPPD values) for each of the five categories of interest: S66 CP, S66 noCP, Pentane14, AlkAtom19, and CRBH20. In order to facilitate the use of Figure 5.10, the basis sets are sorted based on the geometric mean (GM) of the S66 CP, Pentane14, AlkAtom19, and CRBH20 RMSDs relative to the TSB. The S66 noCP RMSD is excluded from the GM because it unfairly disadvantages triple-zeta basis sets. Furthermore, the RMSDs within each dataset are color-coded, with green indicating that the use of the corresponding basis set with the type of interaction represented by the corresponding dataset is recommended, yellow indicating that the pairing should be used with caution, and red indicating that the pairing should not be used. Finally, the number of basis functions that each basis set contains for octane is shown in the last column of Figure 5.10.

From the outset, it is clear that a handful of basis sets are entirely incompatible with ωB97M-V, namely def2-SVP, def2-SVPD, pc-0, aug-pc-0, pc-1, aug-pc-1, and cc-pVDZ. This result is expected, since the functional is trained as close to the basis set limit as possible. On the other hand, it is clear that certain basis sets are very compatible with ωB97M-V, namely pc-3, aug-pc-3, aug-cc-pVQZ, and of course, def2-QZVPPD. These basis sets work superbly well for isomerization energies, thermochemistry, and barrier heights, and provide accurate binding energies for non-covalent interactions with and even without counterpoise corrections. With counterpoise corrections, aug-pc-2 and def2-QZVPP additionally provide satisfactory results for all four types of interactions. While the smallest basis set that can successfully handle all four categories is aug-pc-2 with 782 basis functions for octane, two smaller basis sets, pc-2 and def2-TZVPPD, are almost always satisfactory,



Dataset	S66 noCP		S66 CP		Pentane14		AlkAtom19		CRBH20		GM	BF
Comparison	vs. Ref	vs. TSB	vs. Ref	vs. TSB	vs. Ref	vs. TSB	vs. Ref	vs. TSB	vs. Ref	vs. TSB	vs. TSB	C_8H_{18}
def2-QZVPPD	0.15	0.00	0.14	0.00	0.13	0.00	0.91	0.00	1.23	0.00	0.00	1098
def2-QZVPP	0.18	0.13	0.14	0.02	0.13	0.00	0.85	0.05	1.20	0.03	0.02	996
aug-cc-pVQZ	0.15	0.03	0.14	0.01	0.13	0.00	0.66	0.25	1.20	0.11	0.03	1468
aug-pc-3	0.15	0.02	0.14	0.01	0.12	0.01	0.21	0.76	1.14	0.12	0.05	1612
pc-3	0.14	0.03	0.14	0.01	0.12	0.01	0.22	0.74	1.13	0.14	0.05	1124
aug-pc-2	0.18	0.08	0.13	0.04	0.12	0.02	0.82	0.09	1.29	0.14	0.06	782
def2-TZVPPD	0.19	0.10	0.13	0.02	0.12	0.01	1.86	2.75	1.20	0.06	0.08	602
pc-2	0.26	0.27	0.13	0.11	0.14	0.02	1.11	0.21	1.18	0.19	0.09	492
cc-pVQZ	0.41	0.41	0.16	0.07	0.13	0.00	1.06	1.95	1.24	0.17	0.10	980
aug-cc-pVTZ	0.27	0.19	0.15	0.04	0.17	0.04	2.62	1.72	1.27	0.27	0.16	782
def2-TZVPP	0.43	0.42	0.15	0.08	0.15	0.03	1.68	2.58	1.13	0.13	0.17	500
LP	0.58	0.51	0.19	0.12	0.16	0.06	0.24	0.98	1.26	0.16	0.18	636
cc-pVTZ	0.96	0.97	0.21	0.15	0.15	0.03	0.80	1.69	1.28	0.45	0.25	492
aug-pc-1	1.93	1.87	0.24	0.18	0.21	0.11	13.80	12.90	1.66	0.81	0.68	346
aug-cc-pVDZ	0.90	0.85	0.20	0.13	0.36	0.25	14.85	13.95	0.71	1.17	0.85	346
cc-pVDZ	2.34	2.35	0.49	0.47	0.09	0.18	6.00	5.09	1.28	1.32	0.86	202
pc-1	2.05	2.05	0.58	0.58	0.10	0.14	13.88	12.99	4.37	3.41	1.39	202
def2-SVP	2.46	2.47	0.41	0.42	0.21	0.24	27.50	28.40	2.10	3.04	1.71	202
def2-SVPD	1.69	1.65	0.27	0.23	0.54	0.43	34.47	35.37	1.79	2.68	1.75	304
aug-pc-0	4.98	4.96	1.51	1.58	0.68	0.59	31.27	30.38	5.37	6.06	3.62	158
pc-0	7.16	7.18	1.51	1.57	0.77	0.68	40.43	41.32	11.04	11.96	4.79	108

Figure 5.10: RMSDs in kcal/mol for 4 datasets computed with 21 different basis sets. S66 represents non-covalent interactions, Pentane14 represents isomerization energies, AlkAtom19 represents thermochemistry, and CRBH20 represents barrier heights. The S66 dataset is computed both with and without counterpoise corrections (designated CP and noCP, respectively). The RMSDs are taken with respect to both the reference values (vs. Ref) as well as the training set basis (vs. TSB), which is def2-QZVPPD. The basis sets are sorted based on the geometric mean (GM) of the S66 CP, Pentane14, AlkAtom19, and CRBH20 RMSDs relative to the TSB. The S66 noCP RMSD is excluded from the GM because it unfairly disadvantages triple-zeta basis sets. The number of basis functions (BF) that each basis set contains for octane is shown in the last column.

with the former being considerably smaller than aug-pc-2. In fact, the only result that makes pc-2 not fully satisfactory is the S66 CP RMSD relative to the TSB (0.11 kcal/mol). However, the RMSD relative to the reference values is actually very impressive (0.13 kcal/mol). Thus, pc-2, with only 492 basis functions for octane, is the most economical option that can be recommended for use with ωB97M-V. def2-TZVPPD, on the other hand, has an RMSD of 2.75 kcal/mol for AlkAtom19, relative to the TSB. However, the RMSD of 1.86 kcal/mol relative to the reference values is still acceptable, making def2-TZVPPD, with only 602 basis functions for octane, another economical basis set choice. The rest of the basis sets that have not been mentioned explicitly must be used very cautiously.

5.9 Reaching the Integration Grid Limit

Different density functionals converge to the integration grid limit at different rates. During the training process of ωB97M-V, the billions of candidate functionals were filtered such that the least-squares fit energies generated in the (99,590)/SG-1 and (250,974)/SG-1 grids differed by an absolute maximum of 0.015 kcal/mol. The effectiveness of this decision is tested by analyzing the grid sensitivity of ωB97M-V on all of the data points in the training and primary test sets (with the exception of those in AE18 and RG10) with the following grids: (250,974)/SG-1, (99,590)/SG-1, (99,302)/SG-1, (75,590)/SG-1, (75,302)/SG-1, (75,302)/SG-0, and SG-1/SG-0.

Table 5.5 summarizes the results of this comprehensive test involving 3247 data points, which are binned with respect to the absolute error (AE) in kcal/mol. The table is populated with the assumption that the (250,974)/SG-1 results are fully converged with respect to the grid. Starting with the (99,590)/SG-1 grid, it is clear that the filtering applied during the training stage has completely transferred to the final functional form, since all 3247 data points have absolute errors less than 0.015 kcal/mol. Furthermore, changing the number of radial shells from 99 to 75 (while keeping the number of angular grid points constant) seems to have a negligible effect on the results, yet accelerates the integration of the local exchange-correlation functional by 25%. On the other hand, changing the number of angular grid points from 590 to 302 (while keeping the number of radial shells constant) seems to have a much more profound effect. Based on the negligible effect of transitioning from (99,590)/SG-1 to (75,590)/SG-1, it is reasonable to assume that transitioning from (99,302)/SG-1 to (75,302)/SG-1 should have a negligible effect on the (99,302)/SG-1 results. The (75,302)/SG-1 results indicate that this is indeed true.

The goal of this grid analysis is to recommend three tiers of integration grids for use with ωB97M-V: fine, medium, and coarse. So far, it is clear that the (99,590)/SG-

AE (kcal/mol)	[0,0.015)	[0.015,0.03)	[0.03,0.045)	[0.045,0.06)	[0.06,0.075)	[0.075,0.09)	[0.09,∞)
(99,590)/SG-1	3247	0	0	0	0	0	0
(75,590)/SG-1	3244	3	0	0	0	0	0
(99,302)/SG-1	3200	39	3	3	2	0	0
(75,302)/SG-1	3189	50	3	3	2	0	0
(75,302)/SG-0	3122	107	11	5	2	0	0
SG-1/SG-0	1851	587	303	160	101	83	162

Table 5.5: Grid error ranges for 3247 data points from the training and primary test sets. From the original 3834 data points, the 18 data points from AE18 and the 569 data points from RG10 are excluded. The errors are taken with respect to the (250,974)/SG-1 grid. The grids are assessed with respect to the absolute error (AE) in kcal/mol.

1 grid is certainly the finest grid that is necessary to obtain fully converged results. Thus, the (99,590)/SG-1 grid is deemed to be the "fine" option for ωB97M-V. Furthermore, the only other combination that is computationally more efficient than the (99,590)/SG-1 grid yet maintains its accuracy is the (75,590)/SG-1 grid, which receives the "medium" certification. While (75,302)/SG-1 appears to be a viable "coarse" option, it is useful to see if the nonlocal grid can be reduced without incurring substantial additional errors. Modifying the nonlocal grid from SG-1 to SG-0 (while maintaining the (75,302) local grid) only negligibly affects the results. On the other hand, modifying the local grid from (75,302) to SG-1 (with SG-0 as the nonlocal grid) has a devastating effect on the quality of the results and is absolutely not recommended. Therefore, the "coarse" specification is deemed to be the (75,302)/SG-0 grid.

Based on these results, the following three grids are recommended for use with ωB97M-V:

- fine: (99,590)/SG-1
- medium: (75,590)/SG-1
- coarse: (75,302)/SG-0

5.10 Conclusions

For semi-empirical density functionals, universality (or transferability) is impossible to fully guarantee, because such functionals are necessarily approximate. In other words, for a given system, a new density functional *cannot necessarily* improve upon existing ones, even though it may *often* do so. Nonetheless, within a class of functionals, transferability can be enhanced by minimizing the number of empirical parameters (i.e. avoiding overfitting), while increasing the size of the training and test sets. Even then, the use of a new density functional should only be advocated if it *statistically* improves upon a wide variety of existing competitors in and below its class, across a very diverse set of benchmark systems.

ωB97M-V was developed upon these foundations. A combinatorial, "survival-of-the-most-transferable" approach was utilized to screen over 10 billion candidate least-squares fits based on accuracy, transferability, and desired physical properties, an immense database of nearly 5000 data points was used to train and test the most promising fit, and the final, self-consistently-optimized density functional was assessed against 11 well-respected semi-empirical functionals across all 4986 data points. The results are very encouraging, beginning with a large reduction in the number of trained parameters versus other meta-GGA functionals from 29 (M06-2X), 40 (M11), 47 (M08-HX), or 59 (MN15), to 12 in ωB97M-V. *The use of additional parameters does not yield significantly better transferability in the screening and gradually leads to potential problems with overfitting.*

The combined training, primary test, and secondary test set results (summarized in Figure 5.11) indicate that ωB97M-V is remarkably accurate for non-covalent interactions, isomerization energies, thermochemistry, and barrier heights across the main-group elements. For both NCED and NCEC, ωB97M-V is at least 25% more accurate than the next best tested functional, which is ωB97X-V. ωB97M-V is equivalent to ωB97X-V and B97M-V for IE, but outperforms all tested functionals by at least 25% for ID. Additionally, ωB97M-V is almost 30% more accurate than ωB97X-V for TCE, and 25% more accurate than any tested functional. For TCD, ωB97M-V significantly outperforms the next best functionals, which are B97M-V and ωB97X-V. Finally, despite only having 15% short-range exact exchange, ωB97M-V is the best tested density functional for BH.

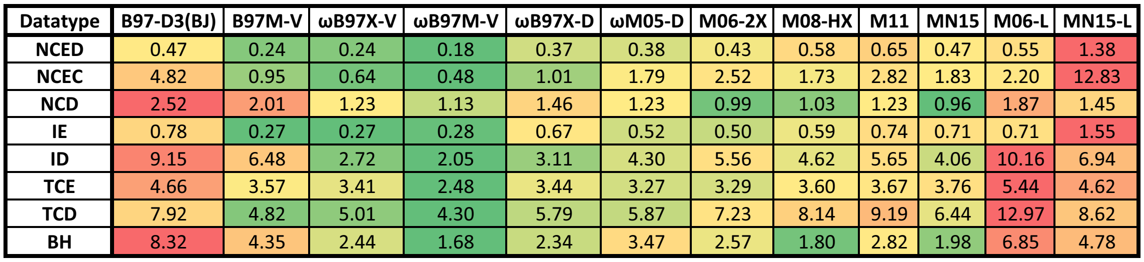
ωB97M-V was consistently trained and tested in the def2-QZVPPD basis set (without counterpoise corrections). Thus, it is meant to be used as close as practically possible to the basis set limit. Its basis set dependence has been thoroughly tested across four types of interactions (non-covalent interactions, isomerization energies, thermochemistry, and barrier heights) in order to identify basis sets that can provide results similar in quality to those acquired with the basis set used for training the parameters. The def2-QZVPPD, pc-3, aug-pc-3, and aug-cc-pVQZ basis sets are recommended for use, both with and without counterpoise corrections (when applicable). Additionally, the def2-QZVPP and aug-pc-2 basis sets are recommended for use with counterpoise corrections (when applicable). Finally, the pc-2 and def2-

TZVPPD basis sets (to be used with counterpoise corrections, when applicable) should serve as economical choices under most circumstances.

Since the evaluation of the kinetic energy density is very sensitive to the integration grid, ωB97M-V was trained with the intention of making the (99,590)/SG-1 grid the integration grid limit. This goal was met by filtering fits during the training stage based on their absolute energetic deviation from the (250,974)/SG-1 grid. Based on tests spanning 3247 data points, the (75,302)/SG-0 grid is recommended as a viable coarse option for use with ωB97M-V (particularly for quick calculations), while the (99,590)/SG-1 grid is recommended as the fine option if results near the integration grid limit are required. For most applications, the medium-sized (75,590)/SG-1 grid can serve as a compromise between these two limits.

It is important to discuss the remaining limitations of ωB97M-V. Like most Kohn–Sham density functionals, it is not appropriate for use when strong correlation effects are significant (e.g. see the TCD results in Figure 5.11). It contains some self-interaction error, which causes larger errors in problems involving odd electrons or holes (e.g. see the NCD results in Figure 5.11). Additionally, it is trained and tested on main-group elements only, so its performance on transition metal-containing systems remains to be tested as suitable reference values become available. However, the minimal empiricism of ωB97M-V gives reason for cautious optimism in cases such as organometallic systems, where strong correlation is not important.

Finally, it is desirable to apply the same approach used here to develop other semi-empirical density functionals with improved physical content, so that the resulting density functionals are likewise minimally parameterized and optimally transferable. Perhaps the most obvious next step is a range-separated hybrid, meta-GGA density functional that includes nonlocal correlation through virtual orbitals. A functional of this type should have significantly lower errors due to self-interaction. We hope to report such a development in due course.



Datatype	B97-D3(BJ)	B97M-V	ωB97X-V	ωB97M-V	ωB97X-D	ωM05-D	M06-2X	M08-HX	M11	MN15	M06-L	MN15-L
NCED	0.47	0.24	0.24	0.18	0.37	0.38	0.43	0.58	0.65	0.47	0.55	1.38
NCEC	4.82	0.95	0.64	0.48	1.01	1.79	2.52	1.73	2.82	1.83	2.20	12.83
NCD	2.52	2.01	1.23	1.13	1.46	1.23	0.99	1.03	1.23	0.96	1.87	1.45
IE	0.78	0.27	0.27	0.28	0.67	0.52	0.50	0.59	0.74	0.71	0.71	1.55
ID	9.15	6.48	2.72	2.05	3.11	4.30	5.56	4.62	5.65	4.06	10.16	6.94
TCE	4.66	3.57	3.41	2.48	3.44	3.27	3.29	3.60	3.67	3.76	5.44	4.62
TCD	7.92	4.82	5.01	4.30	5.79	5.87	7.23	8.14	9.19	6.44	12.97	8.62
BH	8.32	4.35	2.44	1.68	2.34	3.47	2.57	1.80	2.82	1.98	6.85	4.78

Figure 5.11: RMSDs in kcal/mol for 8 datatypes for 12 density functionals. These datatype RMSDs include data points from the training, primary test, and secondary test sets. NCED stands for non-covalent dimers (easy), NCEC stands for non-covalent clusters (easy), NCD stands for non-covalent dimers (difficult), IE stands for isomerization energies (easy), ID stands for isomerization energies (difficult), TCE stands for thermochemistry (easy), TCD stands for thermochemistry (difficult), and BH stands for barrier heights. The partitioning of the 4399 data points contained in this figure into the 8 datatypes is: 1744, 243, 91, 755, 155, 947, 258, and 206.

References

(1) Hohenberg, P.; Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev.* **1964**, *136*, B864–B871.

(2) Levy, M. Universal variational functionals of electron densities, first-order density matrices, and natural spin-orbitals and solution of the v-representability problem. *Proceedings of the National Academy of Sciences* **1979**, *76*, 6062–6065.

(3) Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, *140*, A1133–A1138.

(4) 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. *The Journal of Chemical Physics* **2005**, *123*, 062201.

(5) Perdew, J. P.; Zunger, A. Self-interaction correction to density-functional approximations for many-electron systems. *Phys. Rev. B* **1981**, *23*, 5048–5079.

(6) Vosko, S. H.; Wilk, L.; Nusair, M. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Can. J. Phys.* **1980**, *58*, 1200–1211.

(7) Perdew, J. P.; Wang, Y. Accurate and simple analytic representation of the electron-gas correlation energy. *Phys. Rev. B* **1992**, *45*, 13244–13249.

(8) Ceperley, D. M.; Alder, B. J. Ground State of the Electron Gas by a Stochastic Method. *Phys. Rev. Lett.* **1980**, *45*, 566–569.

(9) Becke, A. D. Density functional calculations of molecular bond energies. *The Journal of Chemical Physics* **1986**, *84*, 4524–4529.

(10) Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* **1988**, *38*, 3098–3100.

(11) Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. Atoms, molecules, solids, and surfaces: Applications of the generalized gradient approximation for exchange and correlation. *Phys. Rev. B* **1992**, *46*, 6671–6687.

(12) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

(13) Perdew, J. P. Density-functional approximation for the correlation energy of the inhomogeneous electron gas. *Phys. Rev. B* **1986**, *33*, 8822–8824.

(14) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **1988**, *37*, 785–789.

(15) Becke, A. D. A new inhomogeneity parameter in density-functional theory. *The Journal of Chemical Physics* **1998**, *109*, 2092–2098.

(16) Perdew, J. P.; Kurth, S.; Zupan, A.; Blaha, P. Accurate Density Functional with Correct Formal Properties: A Step Beyond the Generalized Gradient Approximation. *Phys. Rev. Lett.* **1999**, *82*, 2544–2547.

(17) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. Climbing the Density Functional Ladder: Nonempirical Meta–Generalized Gradient Approximation Designed for Molecules and Solids. *Phys. Rev. Lett.* **2003**, *91*, 146401.

(18) Perdew, J. P.; Ruzsinszky, A.; Csonka, G. I.; Constantin, L. A.; Sun, J. Workhorse Semilocal Density Functional for Condensed Matter Physics and Quantum Chemistry. *Phys. Rev. Lett.* **2009**, *103*, 026403.

(19) Sun, J.; Xiao, B.; Ruzsinszky, A. Communication: Effect of the orbital-overlap dependence in the meta generalized gradient approximation. *The Journal of Chemical Physics* **2012**, *137*, 051101.

(20) Sun, J.; Haunschild, R.; Xiao, B.; Bulik, I. W.; Scuseria, G. E.; Perdew, J. P. Semilocal and hybrid meta-generalized gradient approximations based on the understanding of the kinetic-energy-density dependence. *The Journal of Chemical Physics* **2013**, *138*, 044113.

(21) Sun, J.; Perdew, J. P.; Ruzsinszky, A. Semilocal density functional obeying a strongly tightened bound for exchange. *Proceedings of the National Academy of Sciences* **2015**, *112*, 685–689.

(22) Sun, J.; Ruzsinszky, A.; Perdew, J. P. Strongly Constrained and Appropriately Normed Semilocal Density Functional. *Phys. Rev. Lett.* **2015**, *115*, 036402.

(23) Zhao, Y.; Truhlar, D. G. A new local density functional for main-group thermochemistry, transition metal bonding, thermochemical kinetics, and noncovalent interactions. *The Journal of Chemical Physics* **2006**, *125*, 194101.

(24) Peverati, R.; Truhlar, D. G. M11-L: A Local Density Functional That Provides Improved Accuracy for Electronic Structure Calculations in Chemistry and Physics. *The Journal of Physical Chemistry Letters* **2012**, *3*, 117–124.

(25) Peverati, R.; Truhlar, D. G. An improved and broadly accurate local approximation to the exchange-correlation density functional: The MN12-L functional for electronic structure calculations in chemistry and physics. *Phys. Chem. Chem. Phys.* **2012**, *14*, 13171–13174.

(26) Yu, H. S.; He, X.; Truhlar, D. G. MN15-L: A New Local Exchange-Correlation Functional for Kohn–Sham Density Functional Theory with Broad Accuracy for Atoms, Molecules, and Solids. *Journal of Chemical Theory and Computation* **2016**, *12*, 1280–1293.

(27) Voorhis, T. V.; Scuseria, G. E. A novel form for the exchange-correlation energy functional. *The Journal of Chemical Physics* **1998**, *109*, 400–410.

(28) Boese, A. D.; Handy, N. C. New exchange-correlation density functionals: The role of the kinetic-energy density. *The Journal of Chemical Physics* **2002**, *116*, 9559–9569.

(29) Harris, J; Jones, R. O. The surface energy of a bounded electron gas. *Journal of Physics F: Metal Physics* **1974**, *4*, 1170.

(30) Gunnarsson, O.; Lundqvist, B. I. Exchange and correlation in atoms, molecules, and solids by the spin-density-functional formalism. *Phys. Rev. B* **1976**, *13*, 4274–4298.

(31) Langreth, D. C.; Perdew, J. P. Exchange-correlation energy of a metallic surface: Wave-vector analysis. *Phys. Rev. B* **1977**, *15*, 2884–2901.

(32) Harris, J. Adiabatic-connection approach to Kohn-Sham theory. *Phys. Rev. A* **1984**, *29*, 1648–1659.

(33) Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *The Journal of Chemical Physics* **1993**, *98*, 5648–5652.

(34) Adamo, C.; Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *The Journal of Chemical Physics* **1999**, *110*, 6158–6170.

(35) Becke, A. D. Density-functional thermochemistry. V. Systematic optimization of exchange-correlation functionals. *The Journal of Chemical Physics* **1997**, *107*, 8554–8560.

(36) Boese, A. D.; Martin, J. M. L. Development of density functionals for thermochemical kinetics. *The Journal of Chemical Physics* **2004**, *121*, 3405–3416.

(37) 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. *The Journal of Chemical Physics* **2011**, *135*, 191102.

(38) Staroverov, V. N.; Scuseria, G. E.; Tao, J.; Perdew, J. P. Comparative assessment of a new nonempirical density functional: Molecules and hydrogen-bonded complexes. *The Journal of Chemical Physics* **2003**, *119*, 12129–12137.

(39) Csonka, G. I.; Perdew, J. P.; Ruzsinszky, A. Global Hybrid Functionals: A Look at the Engine under the Hood. *Journal of Chemical Theory and Computation* **2010**, *6*, 3688–3703.

(40) Zhao, Y.; Truhlar, D. G. Hybrid Meta Density Functional Theory Methods for Thermochemistry, Thermochemical Kinetics, and Noncovalent Interactions: The MPW1B95 and MPWB1K Models and Comparative Assessments for Hydrogen Bonding and van der Waals Interactions. *The Journal of Physical Chemistry A* **2004**, *108*, 6908–6918.

(41) Zhao, Y.; Schultz, N. E.; Truhlar, D. G. Exchange-correlation functional with broad accuracy for metallic and nonmetallic compounds, kinetics, and noncovalent interactions. *The Journal of Chemical Physics* **2005**, *123*, 161103.

(42) Zhao, Y.; Truhlar, D. G. Design of Density Functionals That Are Broadly Accurate for Thermochemistry, Thermochemical Kinetics, and Nonbonded Interactions. *The Journal of Physical Chemistry A* **2005**, *109*, 5656–5667.

(43) 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. *Journal of Chemical Theory and Computation* **2006**, *2*, 364–382.

(44) Zhao, Y.; Truhlar, D. G. Density Functional for Spectroscopy: No Long-Range Self-Interaction Error, Good Performance for Rydberg and Charge-Transfer States, and Better Performance on Average than B3LYP for Ground States. *The Journal of Physical Chemistry A* **2006**, *110*, 13126–13130.

(45) Zhao, Y.; Truhlar, D. 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. *Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta)* **2008**, *120*, 215–241.

(46) Zhao, Y.; Truhlar, D. G. Exploring the Limit of Accuracy of the Global Hybrid Meta Density Functional for Main-Group Thermochemistry, Kinetics, and Noncovalent Interactions. *Journal of Chemical Theory and Computation* **2008**, *4*, 1849–1868.

(47) Gill, P. M. W.; Adamson, R. D.; Pople, J. A. Coulomb-attenuated exchange energy density functionals. *Molecular Physics* **1996**, *88*, 1005–1009.

(48) Chai, J.-D.; Head-Gordon, M. Systematic optimization of long-range corrected hybrid density functionals. *The Journal of Chemical Physics* **2008**, *128*, 084106.

(49) Yanai, T.; Tew, D. P.; Handy, N. C. A new hybrid exchange–correlation functional using the Coulomb-attenuating method (CAM-B3LYP). *Chemical Physics Letters* **2004**, *393*, 51 –57.

(50) Krukau, A. V.; Vydrov, O. A.; Izmaylov, A. F.; Scuseria, G. E. Influence of the exchange screening parameter on the performance of screened hybrid functionals. *The Journal of Chemical Physics* **2006**, *125*, 224106.

(51) Peverati, R.; Truhlar, D. G. Screened-exchange density functionals with broad accuracy for chemistry and solid-state physics. *Phys. Chem. Chem. Phys.* **2012**, *14*, 16187–16191.

(52) Peverati, R.; Truhlar, D. G. Improving the Accuracy of Hybrid Meta-GGA Density Functionals by Range Separation. *The Journal of Physical Chemistry Letters* **2011**, *2*, 2810–2817.

(53) Grimme, S. Accurate description of van der Waals complexes by density functional theory including empirical corrections. *J. Comput. Chem.* **2004**, *25*, 1463–1473.

(54) Grimme, S. Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *Journal of Computational Chemistry* **2006**, *27*, 1787–1799.

(55) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *The Journal of Chemical Physics* **2010**, *132*, 154104.

(56) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **2011**, *32*, 1456–1465.

(57) Vydrov, O. A.; Voorhis, T. V. Nonlocal van der Waals density functional: The simpler the better. *The Journal of Chemical Physics* **2010**, *133*, 244103.

(58) Chai, J.-D.; Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom-atom dispersion corrections. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615–6620.

(59) Lin, Y.-S.; Tsai, C.-W.; Li, G.-D.; Chai, J.-D. Long-range corrected hybrid meta-generalized-gradient approximations with dispersion corrections. *The Journal of Chemical Physics* **2012**, *136*, 154109.

(60) Lin, Y.-S.; Li, G.-D.; Mao, S.-P.; Chai, J.-D. Long-Range Corrected Hybrid Density Functionals with Improved Dispersion Corrections. *Journal of Chemical Theory and Computation* **2013**, *9*, 263–272.

(61) Gill, P. M.; Johnson, B. G.; Pople, J. A. A standard grid for density functional calculations. *Chemical Physics Letters* **1993**, *209*, 506 –512.

(62) Mardirossian, N.; Head-Gordon, M. Exploring the limit of accuracy for density functionals based on the generalized gradient approximation: Local, global hybrid, and range-separated hybrid functionals with and without dispersion corrections. *The Journal of Chemical Physics* **2014**, *140*, 18A527.

(63) 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.

(64) Mardirossian, N.; Head-Gordon, M. Mapping the genome of meta-generalized gradient approximation density functionals: The search for B97M-V. *The Journal of Chemical Physics* **2015**, *142*, 074111.

(65) Herman, F.; Van Dyke, J. P.; Ortenburger, I. B. Improved Statistical Exchange Approximation for Inhomogeneous Many-Electron Systems. *Phys. Rev. Lett.* **1969**, *22*, 807–811.

(66) Slater, J. C. A Simplification of the Hartree-Fock Method. *Phys. Rev.* **1951**, *81*, 385–390.

(67) Hamprecht, F. A.; Cohen, A. J.; Tozer, D. J.; Handy, N. C. Development and assessment of new exchange-correlation functionals. *The Journal of Chemical Physics* **1998**, *109*, 6264–6271.

(68) Boese, A. D.; Doltsinis, N. L.; Handy, N. C.; Sprik, M. New generalized gradient approximation functionals. *The Journal of Chemical Physics* **2000**, *112*, 1670–1678.

(69) Boese, A. D.; Handy, N. C. A new parametrization of exchange–correlation generalized gradient approximation functionals. *The Journal of Chemical Physics* **2001**, *114*, 5497–5503.

(70) Wilson, P. J.; Bradley, T. J.; Tozer, D. J. Hybrid exchange-correlation functional determined from thermochemical data and ab initio potentials. *The Journal of Chemical Physics* **2001**, *115*, 9233–9242.

(71) Keal, T. W.; Tozer, D. J. Semiempirical hybrid functional with improved performance in an extensive chemical assessment. *The Journal of Chemical Physics* **2005**, *123*, 121103.

(72) Chai, J.-D.; Head-Gordon, M. Long-range corrected double-hybrid density functionals. *The Journal of Chemical Physics* **2009**, *131*, 174105.

(73) Dion, M.; Rydberg, H.; Schröder, E.; Langreth, D. C.; Lundqvist, B. I. Van der Waals Density Functional for General Geometries. *Phys. Rev. Lett.* **2004**, *92*, 246401.

(74) Lee, K.; Murray, E. D.; Kong, L.; Lundqvist, B. I.; Langreth, D. C. Higher-accuracy van der Waals density functional. *Phys. Rev. B* **2010**, *82*, 081101.

(75) Vydrov, O. A.; Van Voorhis, T. Nonlocal van der Waals Density Functional Made Simple. *Phys. Rev. Lett.* **2009**, *103*, 063004.

(76) Becke, A. D. Exploring the limits of gradient corrections in density functional theory. *J. Comput. Chem.* **1999**, *20*, 63–69.

(77) Thom H. Dunning, J. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *The Journal of Chemical Physics* **1989**, *90*, 1007–1023.

(78) Woon, D. E.; Thom H. Dunning, J. Gaussian basis sets for use in correlated molecular calculations. III. The atoms aluminum through argon. *The Journal of Chemical Physics* **1993**, *98*, 1358–1371.

(79) Burns, L. A.; Álvaro Vázquez-Mayagoitia; Sumpter, B. G.; Sherrill, C. D. Density-functional approaches to noncovalent interactions: A comparison of dispersion corrections (DFT-D), exchange-hole dipole moment (XDM) theory, and specialized functionals. *The Journal of Chemical Physics* **2011**, *134*, 084107.

(80) Shao, Y. et al. Advances in methods and algorithms in a modern quantum chemistry program package. *Phys. Chem. Chem. Phys.* **2006**, *8*, 3172–3191.

(81) Stoll, H.; Golka, E.; Preuß, H. Correlation energies in the spin-density functional formalism. *Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta)* **1980**, *55*, 29–41.

(82) Karton, A.; Daon, S.; Martin, J. M. W4-11: A high-confidence benchmark dataset for computational thermochemistry derived from first-principles W4 data. *Chemical Physics Letters* **2011**, *510*, 165 –178.

(83) Zheng, J.; Zhao, Y.; Truhlar, D. G. Representative Benchmark Suites for Barrier Heights of Diverse Reaction Types and Assessment of Electronic Structure Methods for Thermochemical Kinetics. *Journal of Chemical Theory and Computation* **2007**, *3*, 569–582.

(84) Karton, A.; Tarnopolsky, A.; Lamère, J.-F.; Schatz, G. C.; Martin, J. M. L. Highly Accurate First-Principles Benchmark Data Sets for the Parametrization and Validation of Density Functional and Other Approximate Methods. Derivation of a Robust, Generally Applicable, Double-Hybrid Functional for Thermochemistry and Thermochemical Kinetics. *The Journal of Physical Chemistry A* **2008**, *112*, 12868–12886.

(85) Lynch, B. J.; Truhlar, D. G. Robust and Affordable Multicoefficient Methods for Thermochemistry and Thermochemical Kinetics: The MCCM/3 Suite and SAC/3. *The Journal of Physical Chemistry A* **2003**, *107*, 3898–3906.

(86) 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.

(87) Mardirossian, N.; Lambrecht, D. S.; McCaslin, L.; Xantheas, S. S.; Head-Gordon, M. The Performance of Density Functionals for Sulfate–Water Clusters. *Journal of Chemical Theory and Computation* **2013**, *9*, 1368–1380.

(88) Sherrill, C. D.; Takatani, T.; Hohenstein, E. G. An Assessment of Theoretical Methods for Nonbonded Interactions: Comparison to Complete Basis Set Limit Coupled-Cluster Potential Energy Curves for the Benzene Dimer, the Methane Dimer, Benzene–Methane, and Benzene–H_2S. *The Journal of Physical Chemistry A* **2009**, *113*, 10146–10159.

(89) Marshall, M. S.; Burns, L. A.; Sherrill, C. D. Basis set convergence of the coupled-cluster correction, $\delta_{MP2}^{CCSD(T)}$: Best practices for benchmarking non-covalent interactions and the attendant revision of the S22, NBC10, HBC6, and HSG databases. *The Journal of Chemical Physics* **2011**, *135*, 194102.

(90) Thanthiriwatte, K. S.; Hohenstein, E. G.; Burns, L. A.; Sherrill, C. D. Assessment of the Performance of DFT and DFT-D Methods for Describing Distance Dependence of Hydrogen-Bonded Interactions. *Journal of Chemical Theory and Computation* **2011**, *7*, 88–96.

(91) Crittenden, D. L. A Systematic CCSD(T) Study of Long-Range and Noncovalent Interactions between Benzene and a Series of First- and Second-Row Hydrides and Rare Gas Atoms. *The Journal of Physical Chemistry A* **2009**, *113*, 1663–1669.

(92) Brittain, D. R. B.; Lin, C. Y.; Gilbert, A. T. B.; Izgorodina, E. I.; Gill, P. M. W.; Coote, M. L. The role of exchange in systematic DFT errors for some organic reactions. *Phys. Chem. Chem. Phys.* **2009**, *11*, 1138–1142.

(93) Karton, A.; Gruzman, D.; Martin, J. M. L. Benchmark Thermochemistry of the C_nH_{2n+2} Alkane Isomers (n = 2–8) and Performance of DFT and Composite Ab Initio Methods for Dispersion-Driven Isomeric Equilibria. *The Journal of Physical Chemistry A* **2009**, *113*, 8434–8447.

(94) Zhao, Y.; Lynch, B. J.; Truhlar, D. G. Multi-coefficient extrapolated density functional theory for thermochemistry and thermochemical kinetics. *Phys. Chem. Chem. Phys.* **2005**, *7*, 43–52.

(95) 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. *The Journal of Physical Chemistry A* **2005**, *109*, 2012–2018.

(96) Tang, K. T.; Toennies, J. P. The van der Waals potentials between all the rare gas atoms from He to Rn. *The Journal of Chemical Physics* **2003**, *118*, 4976–4983.

(97) Hohenstein, E. G.; Sherrill, C. D. Effects of Heteroatoms on Aromatic π–π Interactions: Benzene–Pyridine and Pyridine Dimer. *The Journal of Physical Chemistry A* **2009**, *113*, 878–886.

(98) Bryantsev, V. S.; Diallo, M. S.; van Duin, A. C. T.; Goddard, W. A. Evaluation of B3LYP, X3LYP, and M06-Class Density Functionals for Predicting the Binding Energies of Neutral, Protonated, and Deprotonated Water Clusters. *Journal of Chemical Theory and Computation* **2009**, *5*, 1016–1026.

(99) Goerigk, L.; Grimme, S. A thorough benchmark of density functional methods for general main group thermochemistry, kinetics, and noncovalent interactions. *Phys. Chem. Chem. Phys.* **2011**, *13*, 6670–6688.

(100) Copeland, K. L.; Tschumper, G. S. Hydrocarbon/Water Interactions: Encouraging Energetics and Structures from DFT but Disconcerting Discrepancies for Hessian Indices. *Journal of Chemical Theory and Computation* **2012**, *8*, 1646–1656.

(101) Zhao, Y.; Truhlar, D. G. Benchmark Databases for Nonbonded Interactions and Their Use To Test Density Functional Theory. *Journal of Chemical Theory and Computation* **2005**, *1*, 415–432.

(102) Gráfová, L.; Pitoňák, M.; Řezáč, J.; Hobza, P. Comparative Study of Selected Wave Function and Density Functional Methods for Noncovalent Interaction Energy Calculations Using the Extended S22 Data Set. *Journal of Chemical Theory and Computation* **2010**, *6*, 2365–2376.

(103) Řezáč, J.; Riley, K. E.; Hobza, P. S66: A Well-balanced Database of Benchmark Interaction Energies Relevant to Biomolecular Structures. *Journal of Chemical Theory and Computation* **2011**, *7*, 2427–2438.

(104) Jurečka, P.; Šponer, J.; Černý, J.; Hobza, P. Benchmark database of accurate (MP2 and CCSD(T) complete basis set limit) interaction energies of small model complexes, DNA base pairs, and amino acid pairs. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985–1993.

(105) Řezáč, J.; Riley, K. E.; Hobza, P. Extensions of the S66 Data Set: More Accurate Interaction Energies and Angular-Displaced Nonequilibrium Geometries. *Journal of Chemical Theory and Computation* **2011**, *7*, 3466–3470.

(106) Hujo, W.; Grimme, S. Performance of the van der Waals Density Functional VV10 and (hybrid)GGA Variants for Thermochemistry and Noncovalent Interactions. *Journal of Chemical Theory and Computation* **2011**, *7*, 3866–3871.

(107) Murray, E. D.; Lee, K.; Langreth, D. C. Investigation of Exchange Energy Density Functional Accuracy for Interacting Molecules. *J. Chem. Theory Comput.* **2009**, *5*, 2754–2762.

(108) Zhao, Q.; Morrison, R. C.; Parr, R. G. From electron densities to Kohn-Sham kinetic energies, orbital energies, exchange-correlation potentials, and exchange-correlation energies. *Phys. Rev. A* **1994**, *50*, 2138–2142.

(109) Becke, A. D.; Roussel, M. R. Exchange holes in inhomogeneous systems: A coordinate-space model. *Phys. Rev. A* **1989**, *39*, 3761–3767.

(110) Becke, A. D. Correlation energy of an inhomogeneous electron gas: A coordinate-space model. *The Journal of Chemical Physics* **1988**, *88*, 1053–1062.

(111) Becke, A. D. Simulation of delocalized exchange by local density functionals. *The Journal of Chemical Physics* **2000**, *112*, 4020–4026.

(112) Hammer, B.; Hansen, L. B.; Nørskov, J. K. Improved adsorption energetics within density-functional theory using revised Perdew-Burke-Ernzerhof functionals. *Phys. Rev. B* **1999**, *59*, 7413–7421.

(113) Leininger, T.; Stoll, H.; Werner, H.-J.; Savin, A. Combining long-range configuration interaction with short-range density functionals. *Chemical Physics Letters* **1997**, *275*, 151 –160.

(114) Iikura, H.; Tsuneda, T.; Yanai, T.; Hirao, K. A long-range correction scheme for generalized-gradient-approximation exchange functionals. *The Journal of Chemical Physics* **2001**, *115*, 3540–3544.

(115) Tsuneda, T.; Hirao, K. A new spin-polarized Colle-Salvetti-type correlation energy functional. *Chemical Physics Letters* **1997**, *268*, 510 –520.

(116) Tsuneda, T.; Suzumura, T.; Hirao, K. A new one-parameter progressive Colle–Salvetti-type correlation functional. *The Journal of Chemical Physics* **1999**, *110*, 10664–10678.

(117) Tawada, Y.; Tsuneda, T.; Yanagisawa, S.; Yanai, T.; Hirao, K. A long-range-corrected time-dependent density functional theory. *The Journal of Chemical Physics* **2004**, *120*, 8425–8433.

(118) Vydrov, O. A.; Scuseria, G. E. Assessment of a long-range corrected hybrid functional. *The Journal of Chemical Physics* **2006**, *125*, 234109.

(119) Vydrov, O. A.; Heyd, J.; Krukau, A. V.; Scuseria, G. E. Importance of short-range versus long-range Hartree-Fock exchange for the performance of hybrid density functionals. *The Journal of Chemical Physics* **2006**, *125*, 074106.

(120) Henderson, T. M.; Janesko, B. G.; Scuseria, G. E. Generalized gradient approximation model exchange holes for range-separated hybrids. *The Journal of Chemical Physics* **2008**, *128*, 194105.

(121) Weintraub, E.; Henderson, T. M.; Scuseria, G. E. Long-Range-Corrected Hybrids Based on a New Model Exchange Hole. *Journal of Chemical Theory and Computation* **2009**, *5*, 754–762.

(122) Rohrdanz, M. A.; Herbert, J. M. Simultaneous benchmarking of ground- and excited-state properties with long-range-corrected density functional theory. *The Journal of Chemical Physics* **2008**, *129*, 034107.

(123) Rohrdanz, M. A.; Martins, K. M.; Herbert, J. M. A long-range-corrected density functional that performs well for both ground-state properties and time-dependent density functional theory excitation energies, including charge-transfer excited states. *The Journal of Chemical Physics* **2009**, *130*, 054112.

(124) Kristyan, S.; Pulay, P. Can (semi)local density functional theory account for the London dispersion forces? *Chemical Physics Letters* **1994**, *229*, 175 –180.

(125) Hobza, P.; Šponer, J.; Reschel, T. Density functional theory and molecular clusters. *J. Comput. Chem.* **1995**, *16*, 1315–1325.

(126) Klimeš, J.; Michaelides, A. Perspective: Advances and challenges in treating van der Waals dispersion forces in density functional theory. *The Journal of Chemical Physics* **2012**, *137*, 120901.

(127) Jurečka, P.; Černý, J.; Hobza, P.; Salahub, D. R. Density functional theory augmented with an empirical dispersion term. Interaction energies and geometries of 80 noncovalent complexes compared with ab initio quantum mechanics calculations. *J. Comput. Chem.* **2007**, *28*, 555–569.

(128) Becke, A. D.; Johnson, E. R. A density-functional model of the dispersion interaction. *The Journal of Chemical Physics* **2005**, *123*, 154101.

(129) Becke, A. D.; Johnson, E. R. Exchange-hole dipole moment and the dispersion interaction. *The Journal of Chemical Physics* **2005**, *122*, 154104.

(130) Johnson, E. R.; Becke, A. D. A post-Hartree–Fock model of intermolecular interactions. *The Journal of Chemical Physics* **2005**, *123*, 024101.

(131) Becke, A. D.; Johnson, E. R. Exchange-hole dipole moment and the dispersion interaction: High-order dispersion coefficients. *The Journal of Chemical Physics* **2006**, *124*, 014104.

(132) Johnson, E. R.; Becke, A. D. A post-Hartree-Fock model of intermolecular interactions: Inclusion of higher-order corrections. *The Journal of Chemical Physics* **2006**, *124*, 174104.

(133) Johnson, E. R.; Becke, A. D. Van der Waals interactions from the exchange hole dipole moment: Application to bio-organic benchmark systems. *Chemical Physics Letters* **2006**, *432*, 600 –603.

(134) Becke, A. D.; Johnson, E. R. Exchange-hole dipole moment and the dispersion interaction revisited. *The Journal of Chemical Physics* **2007**, *127*, 154108.

(135) Kong, J.; Gan, Z.; Proynov, E.; Freindorf, M.; Furlani, T. R. Efficient computation of the dispersion interaction with density-functional theory. *Phys. Rev. A* **2009**, *79*, 042510.

(136) Vydrov, O. A.; Wu, Q.; Voorhis, T. V. Self-consistent implementation of a nonlocal van der Waals density functional with a Gaussian basis set. *The Journal of Chemical Physics* **2008**, *129*, 014106.

(137) Vydrov, O. A.; Voorhis, T. V. Improving the accuracy of the nonlocal van der Waals density functional with minimal empiricism. *The Journal of Chemical Physics* **2009**, *130*, 104105.

(138) Vydrov, O. A.; Voorhis, T. V. Implementation and assessment of a simple nonlocal van der Waals density functional. *The Journal of Chemical Physics* **2010**, *132*, 164113.

(139) Vydrov, O. A.; Van Voorhis, T. Dispersion interactions from a local polarizability model. *Phys. Rev. A* **2010**, *81*, 062708.

(140) Johnson, E. R.; Wolkow, R. A.; DiLabio, G. A. Application of 25 density functionals to dispersion-bound homomolecular dimers. *Chemical Physics Letters* **2004**, *394*, 334 –338.

(141) Grafenstein, J.; Izotov, D.; Cremer, D. Avoiding singularity problems associated with meta-GGA (generalized gradient approximation) exchange and correlation functionals containing the kinetic energy density. *The Journal of Chemical Physics* **2007**, *127*, 214103.

(142) Johnson, E. R.; Becke, A. D.; Sherrill, C. D.; DiLabio, G. A. Oscillations in meta-generalized-gradient approximation potential energy surfaces for dispersion-bound complexes. *The Journal of Chemical Physics* **2009**, *131*, 034111.

(143) 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. *Journal of Chemical Theory and Computation* **2010**, *6*, 395–404.

(144) Řezáč, J.; Riley, K. E.; Hobza, P. Benchmark Calculations of Noncovalent Interactions of Halogenated Molecules. *Journal of Chemical Theory and Computation* **2012**, *8*, 4285–4292.

(145) Peterson, K. A.; Figgen, D.; Goll, E.; Stoll, H.; Dolg, M. Systematically convergent basis sets with relativistic pseudopotentials. II. Small-core pseudopotentials and correlation consistent basis sets for the post-d group 16–18 elements. *The Journal of Chemical Physics* **2003**, *119*, 11113–11123.

(146) Rappoport, D.; Furche, F. Property-optimized Gaussian basis sets for molecular response calculations. *The Journal of Chemical Physics* **2010**, *133*, 134105.

(147) Kendall, R. A.; Thom H. Dunning, J.; Harrison, R. J. Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions. *The Journal of Chemical Physics* **1992**, *96*, 6796–6806.

(148) Pekeris, C. L. 1^1S and 2^3S States of Helium. *Phys. Rev.* **1959**, *115*, 1216–1221.

(149) 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? *Journal of Chemical Theory and Computation* **2013**, *9*, 2151–2155.

(150) Dahlke, E. E.; Olson, R. M.; Leverentz, H. R.; Truhlar, D. G. Assessment of the Accuracy of Density Functionals for Prediction of Relative Energies and Geometries of Low-Lying Isomers of Water Hexamers. *The Journal of Physical Chemistry A* **2008**, *112*, 3976–3984.

(151) Bates, D. M.; Tschumper, G. S. CCSD(T) Complete Basis Set Limit Relative Energies for Low-Lying Water Hexamer Structures. *The Journal of Physical Chemistry A* **2009**, *113*, 3555–3559.

(152) Lao, K. U.; Herbert, J. M. An improved treatment of empirical dispersion and a many-body energy decomposition scheme for the explicit polarization plus symmetry-adapted perturbation theory (XSAPT) method. *The Journal of Chemical Physics* **2013**, *139*, 034107.

(153) Mintz, B. J.; Parks, J. M. Benchmark Interaction Energies for Biologically Relevant Noncovalent Complexes Containing Divalent Sulfur. *The Journal of Physical Chemistry A* **2012**, *116*, 1086–1092.

(154) Curtiss, L. A.; Raghavachari, K.; Trucks, G. W.; Pople, J. A. Gaussian-2 theory for molecular energies of first- and second-row compounds. *The Journal of Chemical Physics* **1991**, *94*, 7221–7230.

(155) 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. *The Journal of Chemical Physics* **2001**, *114*, 6014–6029.

(156) 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. *The Journal of Physical Chemistry A* **2006**, *110*, 10478–10486.

(157) Wilke, J. J.; Lind, M. C.; Schaefer, H. F.; Csaszar, A. G.; Allen, W. D. Conformers of Gaseous Cysteine. *Journal of Chemical Theory and Computation* **2009**, *5*, 1511–1523.

(158) Lieb, E. H.; Oxford, S. Improved lower bound on the indirect Coulomb energy. *Int. J. Quantum Chem.* **1981**, *19*, 427–439.

(159) Zupan, A.; Perdew, J. P.; Burke, K.; Causa, M. Density-gradient analysis for density functional theory: Application to atoms. *Int. J. Quantum Chem.* **1997**, *61*, 835–845.

(160) Amovilli, C.; March, N.; Bogár, F.; Gál, T. Use of ab initio methods to classify four existing energy density functionals according to their possible variational validity. *Physics Letters A* **2009**, *373*, 3158 –3160.

(161) Malli, G. L.; Da Silva, A. B. F.; Ishikawa, Y. Universal Gaussian basis set for accurate ab initio relativistic Dirac-Fock calculations. *Phys. Rev. A* **1993**, *47*, 143–146.

(162) Janowski, T.; Ford, A. R.; Pulay, P. Accurate correlated calculation of the intermolecular potential surface in the coronene dimer. *Molecular Physics* **2010**, *108*, 249–257.

(163) Sedlak, R.; Janowski, T.; Pitoňák, M.; Řezáč, J.; Pulay, P.; Hobza, P. Accuracy of Quantum Chemical Methods for Large Noncovalent Complexes. *Journal of Chemical Theory and Computation* **2013**, *9*, 3364–3374.

(164) Goldey, M.; Dutoi, A.; Head-Gordon, M. Attenuated second-order Møller-Plesset perturbation theory: performance within the aug-cc-pVTZ basis. *Phys. Chem. Chem. Phys.* **2013**, *15*, 15869–15875.

(165) Pitoňák, M.; Neogrády, P.; Černý, J.; Grimme, S.; Hobza, P. Scaled MP3 Non-Covalent Interaction Energies Agree Closely with Accurate CCSD(T) Benchmark Data. *ChemPhysChem* **2009**, *10*, 282–289.

(166) Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A* **1976**, *32*, 922–923.

(167) Tentscher, P. R.; Arey, J. S. Geometries and Vibrational Frequencies of Small Radicals: Performance of Coupled Cluster and More Approximate Methods. *Journal of Chemical Theory and Computation* **2012**, *8*, 2165–2179.

(168) Bak, K. L.; Gauss, J.; Jørgensen, P.; Olsen, J.; Helgaker, T.; Stanton, J. F. The accurate determination of molecular equilibrium structures. *The Journal of Chemical Physics* **2001**, *114*, 6548–6556.

(169) Feller, D. The role of databases in support of computational chemistry calculations. *J. Comput. Chem.* **1996**, *17*, 1571–1586.

(170) 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. *Journal of Chemical Information and Modeling* **2007**, *47*, 1045–1052.

(171) Lynch, B. J.; Zhao, Y.; Truhlar, D. G. Effectiveness of Diffuse Basis Functions for Calculating Relative Energies by Density Functional Theory. *The Journal of Physical Chemistry A* **2003**, *107*, 1384–1388.

(172) Frisch, M. J.; Pople, J. A.; Binkley, J. S. Self-consistent molecular orbital methods 25. Supplementary functions for Gaussian basis sets. *The Journal of Chemical Physics* **1984**, *80*, 3265–3269.

(173) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. Self-consistent molecular orbital methods. XX. A basis set for correlated wave functions. *The Journal of Chemical Physics* **1980**, *72*, 650–654.

(174) Jensen, F. Polarization consistent basis sets: Principles. *The Journal of Chemical Physics* **2001**, *115*, 9113–9125.

(175) Jensen, F. Polarization consistent basis sets. II. Estimating the Kohn–Sham basis set limit. *The Journal of Chemical Physics* **2002**, *116*, 7372–7379.

(176) Jensen, F. Polarization consistent basis sets. III. The importance of diffuse functions. *The Journal of Chemical Physics* **2002**, *117*, 9234–9240.

(177) Papajak, E.; Leverentz, H. R.; Zheng, J.; Truhlar, D. G. Efficient Diffuse Basis Sets: cc-pVxZ+ and maug-cc-pVxZ. *Journal of Chemical Theory and Computation* **2009**, *5*, 3330–3330.

(178) Kohn, W.; Becke, A. D.; Parr, R. G. Density Functional Theory of Electronic Structure. *The Journal of Physical Chemistry* **1996**, *100*, 12974–12980.

(179) Gill, P. M. W.; Loos, P.-F. Uniform electron gases., English *Theoretical Chemistry Accounts* **2011**, *131*, 1–9.

(180) Negele, J. W.; Vautherin, D. Density-Matrix Expansion for an Effective Nuclear Hamiltonian. *Phys. Rev. C* **1972**, *5*, 1472–1493.

(181) Becke, A. D. A new mixing of Hartree–Fock and local density-functional theories. *The Journal of Chemical Physics* **1993**, *98*, 1372–1377.

(182) Wellendorff, J.; Lundgaard, K. T.; Jacobsen, K. W.; Bligaard, T. mBEEF: An accurate semi-local Bayesian error estimation density functional. *The Journal of Chemical Physics* **2014**, *140*, 144107.

(183) Becke, A. D. Local exchange-correlation approximations and first-row molecular dissociation energies. *International Journal of Quantum Chemistry* **1985**, *27*, 585–594.

(184) Becke, A. D. Density-functional thermochemistry. IV. A new dynamical correlation functional and implications for exact-exchange mixing. *The Journal of Chemical Physics* **1996**, *104*, 1040–1046.

(185) Schmider, H. L.; Becke, A. D. Density functionals from the extended G2 test set: Second-order gradient corrections. *The Journal of Chemical Physics* **1998**, *109*, 8188–8199.

(186) Proynov, E.; Vela, A.; Salahub, D. Nonlocal correlation functional involving the Laplacian of the density. *Chemical Physics Letters* **1994**, *230*, 419 –428.

(187) Colle, R.; Salvetti, O. Approximate calculation of the correlation energy for the closed shells. *Theoretical Chemistry Accounts: Theory, Computation, and Modeling (Theoretica Chimica Acta)* **1975**, *37*, 329–334.

(188) Proynov, E. I.; Sirois, S.; Salahub, D. R. Extension of the LAP functional to include parallel spin correlation. *Int. J. Quantum Chem.* **1997**, *64*, 427–446.

(189) Proynov, E.; Chermette, H.; Salahub, D. R. New tau-dependent correlation functional combined with a modified Becke exchange. *The Journal of Chemical Physics* **2000**, *113*, 10013–10027.

(190) Handy, N. C.; Cohen, A. J. Left-right correlation energy. *Molecular Physics* **2001**, *99*, 403–412.

(191) Proynov, E. I.; Thakkar, A. J. Is combining meta-GGA correlation functionals with the OPTX exchange functional useful? *Int. J. Quantum Chem.* **2006**, *106*, 436–446.

(192) Zhang, Y.; Salahub, D. R. A reparametrization of a meta-GGA exchange-correlation functional with improved descriptions of van der Waals interactions. *Chemical Physics Letters* **2007**, *436*, 394 –399.

(193) Zhang, Y.; Vela, A.; Salahub, D. Reparameterization of a meta-generalized gradient approximation functional by combining TPSS exchange with $\tau 1$ correlation., English *Theoretical Chemistry Accounts* **2007**, *118*, 693–707.

(194) Koehl, R. M.; Odom, G. K.; Scuseria, G. E. The use of density matrix expansions for calculating molecular exchange energies. *Molecular Physics* **1996**, *87*, 835–843.

(195) Voorhis, T. V.; Scuseria, G. E. Exchange energy functionals based on the density matrix expansion of the Hartree-Fock exchange term. *Molecular Physics* **1997**, *92*, 601–608.

(196) 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.

(197) Perdew, J. P.; Ruzsinszky, A.; Tao, J.; Csonka, G. I.; Scuseria, G. E. One-parameter optimization of a nonempirical meta-generalized-gradient-approximation for the exchange-correlation energy. *Phys. Rev. A* **2007**, *76*, 042506.

(198) Ruzsinszky, A.; Sun, J.; Xiao, B.; Csonka, G. I. A meta-GGA Made Free of the Order of Limits Anomaly. *Journal of Chemical Theory and Computation* **2012**, *8*, 2078–2087.

(199) Hao, P.; Sun, J.; Xiao, B.; Ruzsinszky, A.; Csonka, G. I.; Tao, J.; Glindmeyer, S.; Perdew, J. P. Performance of meta-GGA Functionals on General Main Group Thermochemistry, Kinetics, and Noncovalent Interactions. *Journal of Chemical Theory and Computation* **2013**, *9*, 355–363.

(200) Sun, J.; Xiao, B.; Fang, Y.; Haunschild, R.; Hao, P.; Ruzsinszky, A.; Csonka, G. I.; Scuseria, G. E.; Perdew, J. P. Density Functionals that Recognize Covalent, Metallic, and Weak Bonds. *Phys. Rev. Lett.* **2013**, *111*, 106401.

(201) Tao, J. Exchange energy density of an atom as a functional of the electron density. *The Journal of Chemical Physics* **2001**, *115*, 3519–3530.

(202) Tao, J. An accurate MGGA-based hybrid exchange-correlation functional. *The Journal of Chemical Physics* **2002**, *116*, 2335–2337.

(203) Toulouse, J.; Savin, A.; Adamo, C. Validation and assessment of an accurate approach to the correlation problem in density functional theory: The Kriger–Chen–Iafrate–Savin model. *The Journal of Chemical Physics* **2002**, *117*, 10465–10473.

(204) Del Campo, J. M.; Gázquez, J. L.; Trickey, S.; Vela, A. A new meta-GGA exchange functional based on an improved constraint-based GGA. *Chemical Physics Letters* **2012**, *543*, 179 –183.

(205) Constantin, L. A.; Fabiano, E.; Della Sala, F. Meta-GGA Exchange-Correlation Functional with a Balanced Treatment of Nonlocality. *Journal of Chemical Theory and Computation* **2013**, *9*, 2256–2263.

(206) Constantin, L. A.; Fabiano, E.; Sala, F. D. Semilocal dynamical correlation with increased localization. *Phys. Rev. B* **2012**, *86*, 035130.

(207) 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.

(208) Perdew, J. P.; Ruzsinszky, A.; Csonka, G. I.; Vydrov, O. A.; Scuseria, G. E.; Constantin, L. A.; Zhou, X.; Burke, K. Restoring the Density-Gradient Expansion for Exchange in Solids and Surfaces. *Phys. Rev. Lett.* **2008**, *100*, 136406.

(209) Lao, K. U.; Herbert, J. M. Accurate and Efficient Quantum Chemistry Calculations for Noncovalent Interactions in Many-Body Systems: The XSAPT Family of Methods. *The Journal of Physical Chemistry A* **2015**, *119*, 235–252.

(210) Goerigk, L.; Grimme, S. Efficient and Accurate Double-Hybrid-Meta-GGA Density Functionals–Evaluation with the Extended GMTKN30 Database for General Main Group Thermochemistry, Kinetics, and Noncovalent Interactions. *Journal of Chemical Theory and Computation* **2011**, *7*, 291–309.

(211) Shao, Y. et al. Advances in molecular quantum chemistry contained in the Q-Chem 4 program package. *Molecular Physics* **2015**, *113*, 184–215.

(212) Miliordos, E.; Aprà, E.; Xantheas, S. S. Benchmark Theoretical Study of the π–π Binding Energy in the Benzene Dimer. *The Journal of Physical Chemistry A* **2014**, *118*, 7568–7578.

(213) Mardirossian, N.; Head-Gordon, M. Characterizing and Understanding the Remarkably Slow Basis Set Convergence of Several Minnesota Density Functionals for Intermolecular Interaction Energies. *Journal of Chemical Theory and Computation* **2013**, *9*, 4453–4461.

(214) Janowski, T.; Pulay, P. A Benchmark Comparison of σ/σ and π/π Dispersion: the Dimers of Naphthalene and Decalin, and Coronene and Perhydrocoronene. *Journal of the American Chemical Society* **2012**, *134*, 17520–17525.

(215) Manzer, S. F.; Epifanovsky, E.; Head-Gordon, M. Efficient Implementation of the Pair Atomic Resolution of the Identity Approximation for Exact Exchange for Hybrid and Range-Separated Density Functionals. *Journal of Chemical Theory and Computation* **2015**, *11*, 518–527.

(216) Goerigk, L.; Grimme, S. A General Database for Main Group Thermochemistry, Kinetics, and Noncovalent Interactions – Assessment of Common and Reparameterized (meta-)GGA Density Functionals. *Journal of Chemical Theory and Computation* **2010**, *6*, 107–126.

(217) Anacker, T.; Friedrich, J. New accurate benchmark energies for large water clusters: DFT is better than expected. *J. Comput. Chem.* **2014**, *35*, 634–643.

(218) Fanourgakis, G. S.; Aprà, E.; Xantheas, S. S. High-level ab initio calculations for the four low-lying families of minima of $(H_2O)_{20}$. I. Estimates of MP2/CBS binding energies and comparison with empirical potentials. *The Journal of Chemical Physics* **2004**, *121*, 2655–2663.

(219) Yoo, S.; Aprà, E.; Zeng, X. C.; Xantheas, S. S. High-Level Ab Initio Electronic Structure Calculations of Water Clusters $(H_2O)_{16}$ and $(H_2O)_{17}$: A New Global Minimum for $(H_2O)_{16}$. *The Journal of Physical Chemistry Letters* **2010**, *1*, 3122–3127.

(220) Řezáč, J.; Hobza, P. Advanced Corrections of Hydrogen Bonding and Dispersion for Semiempirical Quantum Mechanical Methods. *Journal of Chemical Theory and Computation* **2012**, *8*, 141–151.

(221) Faver, J. C.; Benson, M. L.; He, X.; Roberts, B. P.; Wang, B.; Marshall, M. S.; Kennedy, M. R.; Sherrill, C. D.; Merz, K. M. Formal Estimation of Errors in Computed Absolute Interaction Energies of Protein–Ligand Complexes. *Journal of Chemical Theory and Computation* **2011**, *7*, 790–797.

(222) 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. *Journal of Chemical Theory and Computation* **2014**, *10*, 3140–3150.

(223) Temelso, B.; Archer, K. A.; Shields, G. C. Benchmark Structures and Binding Energies of Small Water Clusters with Anharmonicity Corrections. *The Journal of Physical Chemistry A* **2011**, *115*, 12034–12046.

(224) Peverati, R.; Zhao, Y.; Truhlar, D. G. Generalized Gradient Approximation That Recovers the Second-Order Density-Gradient Expansion with Optimized Across-the-Board Performance. *The Journal of Physical Chemistry Letters* **2011**, *2*, 1991–1997.

(225) Peverati, R.; Truhlar, D. G. Exchange-Correlation Functional with Good Accuracy for Both Structural and Energetic Properties while Depending Only on the Density and Its Gradient. *Journal of Chemical Theory and Computation* **2012**, *8*, 2310–2319.

(226) 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.

(227) Yu, H. S.; He, X.; Li, S. L.; Truhlar, D. G. MN15: A Kohn-Sham Global-Hybrid Exchange-Correlation Density Functional with Broad Accuracy for Multi-Reference and Single-Reference Systems and Noncovalent Interactions. *Chem. Sci.* **2016**, *0*, 0.

(228) Schäfer, A.; Horn, H.; Ahlrichs, R. Fully optimized contracted Gaussian basis sets for atoms Li to Kr. *The Journal of Chemical Physics* **1992**, *97*, 2571–2577.

(229) Schäfer, A.; Huber, C.; Ahlrichs, R. Fully optimized contracted Gaussian basis sets of triple zeta valence quality for atoms Li to Kr. *The Journal of Chemical Physics* **1994**, *100*, 5829–5835.

(230) Weigend, F.; Furche, F.; Ahlrichs, R. Gaussian basis sets of quadruple zeta valence quality for atoms H–Kr. *The Journal of Chemical Physics* **2003**, *119*, 12753–12762.

(231) Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.

(232) Chien, S.-H.; Gill, P. M. W. SG-0: A small standard grid for DFT quadrature on large systems. *Journal of Computational Chemistry* **2006**, *27*, 730–739.

(233) Takatani, T.; David Sherrill, C. Performance of spin-component-scaled Møller-Plesset theory (SCS-MP2) for potential energy curves of noncovalent interactions. *Phys. Chem. Chem. Phys.* **2007**, *9*, 6106–6114.

(234) Witte, J.; Goldey, M.; Neaton, J. B.; Head-Gordon, M. Beyond Energies: Geometries of Nonbonded Molecular Complexes as Metrics for Assessing Electronic Structure Approaches. *Journal of Chemical Theory and Computation* **2015**, *11*, 1481–1492.

(235) Řezáč, J.; Huang, Y.; Hobza, P.; Beran, G. J. O. Benchmark Calculations of Three-Body Intermolecular Interactions and the Performance of Low-Cost Electronic Structure Methods. *Journal of Chemical Theory and Computation* **2015**, *11*, 3065–3079.

(236) Granatier, J.; Pitoňák, M.; Hobza, P. Accuracy of Several Wave Function and Density Functional Theory Methods for Description of Noncovalent Interaction of Saturated and Unsaturated Hydrocarbon Dimers. *Journal of Chemical Theory and Computation* **2012**, *8*, 2282–2292.

(237) Li, S.; Smith, D. G. A.; Patkowski, K. An accurate benchmark description of the interactions between carbon dioxide and polyheterocyclic aromatic compounds containing nitrogen. *Phys. Chem. Chem. Phys.* **2015**, *17*, 16560–16574.

(238) Boese, A. D. Assessment of Coupled Cluster Theory and more Approximate Methods for Hydrogen Bonded Systems. *Journal of Chemical Theory and Computation* **2013**, *9*, 4403–4413.

(239) Boese, A. D. Basis set limit coupled-cluster studies of hydrogen-bonded systems. *Molecular Physics* **2015**, *113*, 1618–1629.

(240) Boese, A. D. Density Functional Theory and Hydrogen Bonds: Are We There Yet? *ChemPhysChem* **2015**, *16*, 978–985.

(241) Lao, K. U.; Schäffer, R.; Jansen, G.; Herbert, J. M. Accurate Description of Intermolecular Interactions Involving Ions Using Symmetry-Adapted Perturbation Theory. *Journal of Chemical Theory and Computation* **2015**, *11*, 2473–2486.

(242) Karton, A.; O'Reilly, R. J.; Chan, B.; Radom, L. Determination of Barrier Heights for Proton Exchange in Small Water, Ammonia, and Hydrogen Fluoride Clusters with G4(MP2)-Type, MPn, and SCS-MPn Procedures–A Caveat. *Journal of Chemical Theory and Computation* **2012**, *8*, 3128–3136.

(243) Chan, B.; Gilbert, A. T. B.; Gill, P. M. W.; Radom, L. Performance of Density Functional Theory Procedures for the Calculation of Proton-Exchange Barriers: Unusual Behavior of M06-Type Functionals. *Journal of Chemical Theory and Computation* **2014**, *10*, 3777–3783.

(244) Tentscher, P. R.; Arey, J. S. Binding in Radical-Solvent Binary Complexes: Benchmark Energies and Performance of Approximate Methods. *Journal of Chemical Theory and Computation* **2013**, *9*, 1568–1579.

(245) Kozuch, S.; Martin, J. M. L. Halogen Bonds: Benchmarks and Theoretical Analysis. *Journal of Chemical Theory and Computation* **2013**, *9*, 1918–1931.

(246) Bauzá, A.; Alkorta, I.; Frontera, A.; Elguero, J. On the Reliability of Pure and Hybrid DFT Methods for the Evaluation of Halogen, Chalcogen, and Pnicogen Bonds Involving Anionic and Neutral Electron Donors. *Journal of Chemical Theory and Computation* **2013**, *9*, 5201–5210.

(247) De-la Roza, A. O.; Johnson, E. R.; DiLabio, G. A. Halogen Bonding from Dispersion-Corrected Density-Functional Theory: The Role of Delocalization Error. *Journal of Chemical Theory and Computation* **2014**, *10*, 5436–5447.

(248) Steinmann, S. N.; Piemontesi, C.; Delachat, A.; Corminboeuf, C. Why are the Interaction Energies of Charge-Transfer Complexes Challenging for DFT? *Journal of Chemical Theory and Computation* **2012**, *8*, 1629–1640.

(249) Kozuch, S.; Bachrach, S. M.; Martin, J. M. Conformational Equilibria in Butane-1,4-diol: A Benchmark of a Prototypical System with Strong Intramolecular H-bonds. *The Journal of Physical Chemistry A* **2014**, *118*, 293–303.

(250) 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). *The Journal of Physical Chemistry A* **2009**, *113*, 11974–11983.

(251) Martin, J. M. L. What Can We Learn about Dispersion from the Conformer Surface of n-Pentane? *The Journal of Physical Chemistry A* **2013**, *117*, 3118–3132.

(252) Fogueri, U. R.; Kozuch, S.; Karton, A.; Martin, J. M. The Melatonin Conformer Space: Benchmark and Assessment of Wave Function and DFT Methods for a Paradigmatic Biological and Pharmacological Molecule. *The Journal of Physical Chemistry A* **2013**, *117*, 2269–2277.

(253) 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. *Journal of Chemical Theory and Computation* **2016**, *12*, 444–454.

(254) Yu, L.-J.; Sarrami, F.; Karton, A.; O'Reilly, R. J. An assessment of theoretical procedures for π-conjugation stabilisation energies in enones. *Molecular Physics* **2015**, *113*, 1284–1296.

(255) Karton, A.; Martin, J. M. Explicitly correlated benchmark calculations on C_8H_8 isomer energy separations: how accurate are DFT, double-hybrid, and composite ab initio procedures? *Molecular Physics* **2012**, *110*, 2477–2491.

(256) Yu, L.-J.; Karton, A. Assessment of theoretical procedures for a diverse set of isomerization reactions involving double-bond migration in conjugated dienes. *Chemical Physics* **2014**, *441*, 166 –177.

(257) Manna, D.; Martin, J. M. L. What Are the Ground State Structures of C_{20} and C_{24}? An Explicitly Correlated Ab Initio Approach. *The Journal of Physical Chemistry A* **2016**, *120*, 153–160.

(258) Grimme, S.; Kruse, H.; Goerigk, L.; Erker, G. The Mechanism of Dihydrogen Activation by Frustrated Lewis Pairs Revisited. *Angewandte Chemie International Edition* **2010**, *49*, 1402–1405.

(259) Krieg, H.; Grimme, S. Thermochemical benchmarking of hydrocarbon bond separation reaction energies: Jacob's ladder is not reversed! *Molecular Physics* **2010**, *108*, 2655–2666.

(260) O'Reilly, R. J.; Karton, A. A dataset of highly accurate homolytic N–Br bond dissociation energies obtained by Means of W2 theory. *International Journal of Quantum Chemistry* **2016**, *116*, 52–60.

(261) 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. *The Journal of Physical Chemistry A* **2012**, *116*, 4211–4221.

(262) Karton, A.; Schreiner, P. R.; Martin, J. M. L. Heats of formation of platonic hydrocarbon cages by means of high-level thermochemical procedures. *Journal of Computational Chemistry* **2016**, *37*, 49–58.

(263) Karton, A.; Goerigk, L. Accurate reaction barrier heights of pericyclic reactions: Surprisingly large deviations for the CBS-QB3 composite method and their consequences in DFT benchmark studies. *Journal of Computational Chemistry* **2015**, *36*, 622–632.

(264) Yu, L.-J.; Sarrami, F.; O'Reilly, R. J.; Karton, A. Reaction barrier heights for cycloreversion of heterocyclic rings: An Achilles' heel for DFT and standard ab initio procedures. *Chemical Physics* **2015**, *458*, 1 –8.

(265) Yu, L.-J.; Sarrami, F.; O'Reilly, R. J.; Karton, A. Can DFT and ab initio methods describe all aspects of the potential energy surface of cycloreversion reactions? *Molecular Physics* **2016**, *114*, 21–33.

(266) Ruscic, B. Uncertainty quantification in thermochemistry, benchmarking electronic structure computations, and Active Thermochemical Tables. *Int. J. Quantum Chem.* **2014**, *114*, 1097–1101.