Design and analysis of machine learning exchange-correlation functionals via rotationally invariant convolutional descriptors

Xiangyun Lei and Andrew J. Medford

Phys. Rev. Materials **3**, 063801 — Published 12 June 2019

DOI: 10.1103/PhysRevMaterials.3.063801

Design and Analysis of Machine Learning Exchange-Correlation Functionals via Rotationally Invariant Convolutional Descriptors

Xiangyun Lei and Andrew J. Medford*
School of Chemical and Biomolecular Engineering,
Georgia Institute of Technology, Atlanta, GA, 30318 USA
(Dated: May 6, 2019)

In this work we explore the potential of a new data-driven approach to the design of exchange-correlation (XC) functionals. The approach, inspired by convolutional filters in computer vision and surrogate functions from optimization, utilizes convolutions of the electron density to form a feature space to represent local electronic environments and neural networks to map the features to the exchange-correlation energy density. These features are orbital free, and provide a systematic route to including information at various length scales. This work shows that convolutional descriptors are theoretically capable of an exact representation of the electron density, and proposes Maxwell-Cartesian spherical harmonic kernels as a class of rotationally invariant descriptors for the construction of machine-learned functionals. The approach is demonstrated using data from the B3LYP functional on a number of small-molecules containing C, H, O, and N along with a neural network regression model. The machine-learned functionals are compared to standard physical approximations and the accuracy is assessed for the absolute energy of each molecular system as well as formation energies. The results indicate that it is possible to reproduce the exchange-correlation portion of B3LYP formation energies to within chemical accuracy using orbital-free descriptors with a spatial extent of 0.2 Å. The findings provide empirical insight into the spatial range of electron exchange, and suggest that the combination of convolutional descriptors and machine-learning regression models is a promising new framework for XC functional design, although challenges remain in obtaining training data and generating models consistent with pseudopotentials.

I. INTRODUCTION

Since its introduction in the mid-1960s[1,2], density functional theory (DFT) has become a much-used tool in the fields of chemistry, material science, biology and others. The popularity of DFT arises mainly from its simple formalism and low computational cost compared to wavefunction theories (WFTs). The basic formalism of DFT establishes that the exact ground-state energy can be written as functional of the electron density:

$$E_{GS}[\rho(\vec{x})] = T[\rho(\vec{x})] + J[\rho(\vec{x})] + E_{ext}[\rho(\vec{x})] + E_{xc}[\rho(\vec{x})] \quad (1)$$

where $\rho(\vec{x})$ is the ground-state electron density, E_{GS} is the ground-state energy, T is the kinetic energy functional for the non-interacting system, J is the classical Coulomb self-energy (or Hartree energy) functional, E_{ext} is the energy due to external potential (e.g. atomic nuclei), and E_{xc} is the exchange-correlation functional that accounts for the difference between classical and quantum-mechanical electronic repulsion as well as the difference in kinetic energy between the non-interacting and interacting systems[2]. Of these, T, J, and E_{xc} are independent of the external potential and are hence considered "universal" functionals, while E_{ext} depends on the atomic coordinates of a chemical system. Furthermore, T, J and E_{ext} are known exactly, so the challenge is to determine E_{xc}. Although a universal and exact ground-state E_{xc} functional does exist as proved by Hohenberg and Kohn[1], the form of this functional remains unknown. Numerous strategies have been employed to construct density functional approximations (DFAs) for E_{xc}; however, despite over five decades of research and hundreds of trials, no existing functionals are universally comparable to the accuracy of wavefunction theories.

Construction of DFA's relies on two main components: a model space that describes the electronic environment and a functional that connects the model space to the energy density. The concept of improving approximations by increasing the complexity of the model space is captured by Perdew's popular analogy to "Jacob's ladder"[3], which reveals an important trend: the more non-local information that is used to describe the electronic environment, the better the quality of the approximation. In the early days, Kohn and Sham approximated E_{xc} by assuming a uniform electron gas with an electronic density equal to the (spin) density at a local point, referred to as the local density approximation (LDA)[2]. This approximation is surprisingly accurate for delocalized systems such as metals, but its accuracy is considerably lower for molecules. The next major improvement came decades later when several authors proposed using the gradient of electron density as an additional input to the exchange-correlation functional[4–7]. This led to the development of a family of DFA's known as generalized gradient approximation (GGA) functionals that provided an order of magnitude improvement in accuracy and started a rapid increase in the application of DFT. The next logical step would be inclusion of the second derivative in the spirit of a Taylor expansion; however, the kinetic energy density is more commonly used[3]. This "meta-GGA" (mGGA) functional family includes a diverse range of physical and empirical approximations that have generally improved accuracy,

although the improvements are not always systematic[8]. The next class of functionals deviates from the strategy of adding more semi-local information by including a component of fully non-local exact exchange. These "hybrid functionals", introduced by Becke and coworkers in the B3PW91 functional[9,10], exhibit another general (though not necessarily systematic) improvement in accuracy. Hybrid functionals such as the ubiquitous B3LYP functional[9,11] are very popular, particularly in the chemistry community, due to their high accuracy and relatively low cost for molecular systems. However, hybrid functionals have considerably higher computational cost and are difficult to implement for extended systems (e.g. solids and surfaces), leading to screened hybrid functionals such as HSE06[12,13]. Numerous other strategies have also been employed to capture long-range interactions, including fully non-local approaches such as 100% exchange functionals[14,15], approaches that combine multiple approximations[16–20], double-hybrid functionals including wavefunction based correlation[21–23], and functionals including dispersion[24–30]. These diverse options for model spaces indicate that inclusion of additional and increasingly non-local information increase the accuracy of DFA's; however, the improvement of model spaces has been based primarily on chemical intuition. This is advantageous in the derivation of functionals from chemical and physical principles, but also leads to difficulties in systematically improving models or deconvoluting multiple physical effects to avoid double counting.

An alternative approach to model space development is to construct model spaces that can be systematically expanded into a theoretically complete description of the system. This is similar to a common approach in the fields of image processing and computer vision where convolutions are used to extract features/information of varying length scales[31,32]. A noteworthy recent triumph in the application of convolutions in image processing are convolutional neural networks (convNets)[33,34] where convolutional kernels are determined through deep learning. ConvNets have achieved unprecedented accuracy in handwriting, object, and facial recognition, and have revolutionized the field of image analysis[34–39]. This approach can be translated to the field of functional development since electron density data can be projected onto a finite-difference grid and treated as a 3D image. With 3D convolutions, any local and semi-local feature of the electronic environment can be extracted, analogous to 2D images. In this work we explore this approach to model space construction, and show that the convolutional descriptor model space is theoretically complete in the limit that the kernel has the same size as the system.

In addition to the model space, a functional requires a mathematical connection between the model space and the exchange-correlation energy. This challenge is at the core of most functional development, and there are two distinct philosophies. The reductionist approach applies physical principles and theoretical constraints to derive "parameter-free" functionals. The PBE functional is a well-known example of this philosophy[40,41]. These derived functionals tend to have less systematic bias toward specific molecular systems and have a predictable accuracy across all systems, although this might not always be true for all kinds of systems[42,43]. The alternative approach is empiricism, with a more practical focus on maximizing the accuracy of DFT in specific applications. Most empirical functionals are based on derived functional forms where some parameters are optimized based on molecular data of the systems of interest. The data is usually obtained from experiments or higher-level calculations. These functionals are usually accurate for systems similar to those used in training, but the accuracy is typically lower for other systems or properties[44]. The B3LYP functional and the Minnesota functionals are well-known examples of empirical functionals[9,11,45–48]. Recently, approaches from machine learning (ML) have taken the empirical approach to functional development to its logical extreme. In a seminal paper by Snyder et al. the idea of using ML to connect density and kinetic energy density of a 1D model system is introduced[49]. The success of this approach inspired substantial interest and subsequent development of employing ML in many different ways related to DFT. An extensive review is beyond the scope of this work, but examples include the use of ML to develop molecular dynamics force-fields[50–52], application of ML models to reproduce DFT results without the use of expensive QM calculations[53–60], application of ML to improve the accuracy or speed of DFT[61–64] and direct inclusion of ML models in the construction of density functionals[44,65–68]. These numerous strategies have illustrated the substantial promise of ML techniques in the field of functional development.

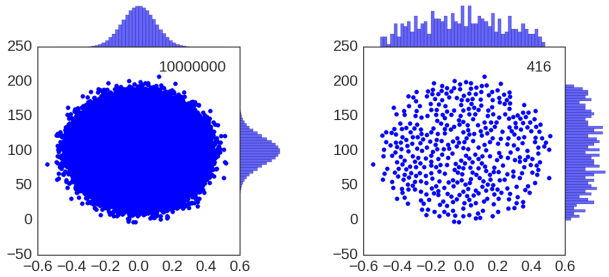
The key concept of ML-based density functionals is that highly-flexible "universal" regression models with thousands or millions of parameters are applied to connect a model space to the exchange-correlation energy. The parameters are optimized using a large amount of known data from experiment or calculations. This strategy does not require any knowledge of the complex underlying physics, but instead the challenge arises from obtaining a sufficient amount of high-quality data and utilizing verification and validation approaches to avoid over-parameterization. Machine learning models also have the advantage of being systematically improvable by addition of training data and/or increase of the regression model complexity (i.e. increase of the number of fitting parameters). Some common choices of regression models are support vector regressors (SVR)[69], kernel ridge regression (KRR)[70], Gaussian process regressors (GPR)[71], and artificial neural networks (NN)[72,73]. Neural networks are a particularly interesting and popular class of non-linear regression models due to their property of being theoretically capable of approximating any function to arbitrary accuracy, as proved by the universal approximation theorem[74,75]. The complexity of a NN can be easily tuned by adding/removing neurons and layers, and prediction is very fast once training is complete.

The quality of the approximation will ultimately depend on the amount of training data available and the heuristics applied during the training process, but in principle NN's provide a route to a systematically-improvable regression model to connect a given model space to the exchange-correlation energy.

In this work, we combine the ideas of convolutional fingerprinting of electronic environments and neural networks to propose a functional design framework with systematically improvable model spaces and regression-based functionals. We show that 3D convolutions can be used to re-formulate finite-difference DFT and are theoretically complete in the trivial limit that the convolutions are equivalent to the input density. We also developed a specific class of convolutional kernels to extract features (or "descriptors") and form model spaces that are complete and rotationally invariant, inspired by the work of Worral et al.[76] and Applequist[77]. This is combined with NN regression models and exchange-correlation (XC) data from the B3LYP hybrid functional for a range of small molecule systems to construct "surrogate" functionals. These surrogate functionals are inspected based on their accuracy as compared to the grid-projected B3LYP XC functional, which is chosen to be the ground truth in this study since it is widely used and there is no semi-local closed form for the exact exchange energy. The resulting convolutional surrogate functionals are orbital-free and have systematically increasing non-locality, providing a route to test the systematic improvement of the resulting functionals and gain insight into the locality of electron exchange. The results indicate that accuracy increases significantly with the size of the model space, and that it is possible to reproduce B3LYP energies to within chemical accuracy using a semi-local orbital-free functional with a range of > 0.1 Å. However, practical challenges remain in the finite-difference representation of all-electron systems, and access to spatially-resolved exchange-correlation data is currently limited. These obstacles are non-trivial, but addressing them represents an alternative strategy for XC functional development.

II. METHODS

The DFT data are generated with the Psi4 package[78]. Single-point spin-paired calculations are performed at the B3LYP/aug-cc-pvtz level with both density and energy convergence set to 10^{-12} Ha. The geometries of molecules are taken from computational chemistry comparison and benchmark database (CCCBDB) maintained by NIST[79] and are static for all calculations. The training set consists of 15 small-molecule systems: C_2H_2, C_2H_4, C_2H_6, CH_3OH, CH_4, CO, CO_2, H_2, H_2O, HCN, HNC, N_2, N_2O, NH_3, O_3. These molecules contain 4 common atom types (C, H, O, N) and a diverse range of single, double, and triple bonds between them. An additional 7 molecular systems that have similar chemistry are used as an independent test set: CH_3CN, CH_3CHO, H_2CCO, H_2CO, H_2O_2, HCOOH, N_2H_4. In addition, 3 extra molecular systems with different chemistry, CH_3NO_2, NH_2CH_2COOH (glycine), NCCN, are used to test the models' ability to extrapolate. This "extrapolation set" is not included in the accuracy analysis, but is used to probe the generality of the model in Sec. III C 3. The converged electronic density (ρ) and exchange-correlation energy density (ϵ_{xc}) of the systems are projected onto a uniform 3D finite-difference grid and stored as 3D arrays. The overall size of each grid is 10 Å × 10 Å × 10 Å with the molecule centered in the cell, and the grid-point spacing is 0.02 Å. This results in a total of $500^3 = 125,000,000$ data points per system. Due to memory limitations, domain decomposition with sub-grids of 2Å × 2Å × 2Å are used for data manipulations including descriptor extraction, sub-sampling, and prediction. The `scipy`[80] implementation of fast Fourier transform convolution (FFT convolution) is used to extract electronic environment descriptors. To ensure correct padding of the convolutions, each sub-grid is combined with the 26 neighboring sub-grids prior to FFT convolutions.



(a) Before sub-sampling

(b) After near-uniform sub-sampling

Figure 1: Example of the near-uniform sub-sampling algorithm applied to a 2D random Gaussian distribution data set with hyper-parameters of $cutoff = 0.2, rate = 0.1$.

The resulting data for each system is sub-sampled to reduce the computational burden of training. A "near-uniform" sub-sampling algorithm is developed to produce roughly uniform sampling density across the high-dimensional space. This improves sampling efficiency by ensuring that rare data points from the tails of the distribution are included in the training sample. An illustration of the procedure is shown in Figure 1 and detailed explanations of the procedure can be found in the Supplementary Information.

The near-uniform sub-sample is supplemented with a random sub-sample to provide information about the relative frequency of points in the original distribution. A fixed random sub-sample size of 10,000,000 in total for all training systems is selected, resulting in total sub-sample size of roughly 680,000 training points per molecule.

The remaining 124,320,000 points in each molecular system are used to test the resulting models, corresponding to roughly 0.55% training data and 99.45% testing data, where <0.05% of the training data is selected non-randomly to ensure that the tails of the distribution are represented. This is supplemented with fully independent validation sets from 7 additional molecular systems that are not used in model training; these systems are considered "test" systems in the remainder of the paper, while molecular systems used in model development are referred to as "training" systems although only 0.55% of their data is actually used to train the NN models. The 3 additional molecular systems in the "extrapolation set" are also used to test the model and no points from these systems are used in training.

The machine-learning models are constructed using a framework similar to Δ machine learning[81] based on residuals from a re-fitted Vosko-Wilk-Nusair (VWN)[82] LDA model. The model is formulated as:

$$\begin{aligned}\tilde{\epsilon}_{xc}(\vec{x}) = E_{r-VWN}(\rho(\vec{x}), C_1, \gamma, \alpha_1, \beta_1, \beta_2, \beta_3, \beta_4) \\ + E_{NN}(\lambda_i[\rho(\vec{x})], W_{jk})\end{aligned} \tag{2}$$

where $\tilde{\epsilon}_{xc}(\vec{x})$ is the predicted XC energy density, E_{r-VWN} is the energy of a VWN LDA model with the numerical values of its parameters (C_1, γ, α_1, β_1, β_2, β_3, β_4) re-fitted to 1,000,000 randomly-selected data points from the B3LYP training systems. This re-parameterization is achieved with the Nelder-Mead algorithm[83] as implemented in `scipy`[80] and the same r-VWN model is used for all surrogate functionals. The E_{NN} term corresponds to a NN with a set of input descriptors, λ_i and weights W_{jk}. The weights are optimized using the Adam algorithm for stochastic gradient descent[84] as implemented in the `Keras` package[85]. A standard NN architecture of 2 hidden layers with 100 neurons and the ReLU activation function[86] is used for consistency. The training data is divided into separate steps with different learning rates and loss functions; details are available in the Supplementary Information.

All energies are evaluated non-self-consistently by directly evaluating the integral of the predicted XC energy density with the self-consistent B3LYP electron density is used as an input. The accuracy of the resulting models is assessed with three different error metrics at the chemical system level: sum of local absolute error ($\varepsilon_{absolute}$), energy prediction error ($\varepsilon_{predict}$) and formation energy prediction error ($\varepsilon_{formation}$), each probing different aspects of the models. The sum of local absolute error corresponds to the integral of absolute difference between the predicted and actual XC energy density:

$$\varepsilon_{absolute} = \sum_i |\epsilon_{xc}(\vec{x}_i) - \tilde{\epsilon}_{xc}(\vec{x}_i)| \times h^3 \tag{3}$$

This is a straightforward definition of the error from the perspective of model training, and is proportional to the mean absolute error. It directly probes the absolute accuracy of the model for a system and gives an upper bound for energy prediction error. The energy prediction error corresponds to the error of the integral of the XC energy density over the system, as approximated by the sum of the predicted energy at all grid points:

$$\begin{aligned}\varepsilon_{predict} &= E_{xc} - \sum_i \tilde{\epsilon}_{xc}(\vec{x}_i) \times h^3 \\ &= \sum_i (\epsilon_{xc}(\vec{x}_i) - \tilde{\epsilon}_{xc}(\vec{x}_i)) \times h^3\end{aligned} \tag{4}$$

This is proportional to the mean signed error, and cancellation of error will result in errors lower than the sum of local absolute error. The final metric of formation energy prediction error is the most practical since these relative quantities are most relevant in chemistry, and cancellation of systematic errors is a common feature of DFT functionals. The predicted formation energy error is obtained by computing the formation energy of each species relative to the following atomic reference states that are commonly employed in DFT studies:

$$\varepsilon_{form} = E_{xc} - \sum_i n_i \mu_i \tag{5}$$

where n_i is the number of atoms i in the molecule, and μ_i is the reference energy of each atomic species. In this case, the following molecular references are used for each atomic species: C = CH_4, N = NH_3, O = H_2O, and H = H_2. Formation energy errors enable the most cancellation of error, but anti-cancellation is also possible in the case of over-fitted models. Hence, formation energy errors compared to local and system level errors provide a convenient and practical measure of model accuracy and over-fitting.

The code for constructing descriptors, training, and evaluating models is available via the supporting information.

III. RESULTS AND DISCUSSION

The results are presented in three parts. In Sec. III A the theoretical motivation for using convolutions to construct model spaces is presented by re-formulating the XC energy in terms of convolution kernels, and some advantages and limitations are discussed. In Sec. III B a specific class of "convolutional descriptors" based on spherical harmonics are applied to the dataset of small molecules. In Sec. III C the machine learning framework for XC energy is introduced through discussion of both the re-parameterized VWN model and the NN models that are used to fit the residuals. The accuracy of models based on various model spaces are presented and discussed.

A. Convolutional reformulation of exchange-correlation functional

The theoretical motivation for using convolutions to form a basis set for XC model spaces is based on two properties: systematic increase of spatial range, and theoretical completeness in the limit that the range is equal to the system size. The spatial range can be increased by increasing the maximum size of the convolution kernels. To show completeness we re-formulate the XC functional, $E_{xc}[\rho(\vec{x})]$, as a function of convolutions between a set of kernels and the electron density. We show that the functional is equivalent to the function with the proper choice of kernels in the case where (i) the density is discretized onto a finite-difference (FD) grid, (ii) the maximum kernel size is equal to the system size, and (iii) the number of convolutions is equal to the number of points in the finite difference grid. This equivalence between functionals and functions for numerical representations in DFT has been noted before[87]; here we briefly examine the specific case of finite difference representations and convolutions. We restrict the discussion to the spin-paired case for simplicity, but the same arguments hold in the case of spin-polarized systems. Similar arguments are expected to hold for any energy functional including the kinetic energy or full system energy, though we focus only on XC energy here. First, we consider spatially-resolved XC energy densities:

$$E_{xc}[\rho(\vec{x})] = \int_{\mathbb{R}^3} \epsilon_{xc}[\rho(\vec{x})](\vec{x}) \mathrm{d}^3\vec{x} \tag{6}$$

where $\epsilon_{xc}[\rho(\vec{x})](\vec{x})$ is the exchange-correlation energy density defined at each point $\vec{x}$ in space. The existence of a locally-resolved XC energy density and methods to extract it have been examined previously[88], although extracting this quantity presents a practical challenge for wave-function theories. Next, consider a finite-difference representation of the electron density:

$$\rho(\vec{x}) = \rho^{xyz} \tag{7}$$

where the density is represented on a 3D grid and x, y, z are indices of each voxel along the (x, y, z) Cartesian axes. We note that these indices can be "unraveled" such that ρ^{xyz} can be considered as a 1-dimensional vector of N^3 points with a single index, although it is conceptually simpler to consider ρ^{xyz} as a 3-dimensional array of voxels. Each voxel has dimensions of h_x, h_y, h_z Å and a corresponding volume of $v = h_x h_y h_z$ Å^3; for simplicity we consider the isotropic case where $h_x = h_y = h_z = h$. The finite difference representation is chosen because it is intuitive, convenient for convolutions, commonly used in solid-state codes[89–91], and systematically converges to the exact density in the limit of $h_i \to 0$. The XC energy can also be written in terms of a finite difference basis:

$$\epsilon_{xc}[\rho(\vec{x})](\vec{x}) = \epsilon^{lmn}(\rho^{xyz}) \tag{8}$$

where l, m, n are also indices of each voxel. Here we have exploited the fact that a functional becomes a function when its argument is represented in a numerical basis (i.e. the XC energy density at each grid point is a function of the value of the electron density at every grid point). Finally, we introduce the concept of density convolutions:

$$\lambda_q^{xyz} = (C_q \circledast \rho)^{xyz} \tag{9}$$

where λ_q^{xyz} represents a vector of "convolutional descriptors" (indexed by q) spatially-resolved at each grid point (indexed by xyz), and C_q is an arbitrary convolution kernel of size $n_x \times n_y \times n_z$. For convenience we restrict discussion to the case where $n_x = n_y = n_z = n$, and n is odd, such that we can define the range r_q of a convolution kernel C_q as $r_q = h(n-1)/2$. Furthermore, we consider only the case of periodic boundary conditions to avoid issues of padding. In this case λ_q^{xyz} is always the same dimension as ρ^{xyz}, corresponding to a feature vector indexed by q at each grid point xyz. The restriction to periodic boundary conditions is a minor limitation, considering that any finite system can be represented as a periodic system with sufficient vacuum padding; this is commonly exploited in plane-wave codes. Considering a set of convolution kernels produces a set of N_d ($q \leq N_d$) local descriptors for a point xyz in space. These descriptors capture information out to a distance of a total range $R = max(r_q)$. If the largest dimension of the unit cell of the system is given by L_{max} then in the limit of $R \to L_{max}/2$ and $N_d \to n^3$ the full non-local density can be recovered by using n^3 delta-function kernels:

$$\tilde{\lambda}_{xyz}^{lmn} = (\delta_{xyz} \circledast \rho)^{lmn} = \rho^{xyz} \tag{10}$$

where $\delta_{xyz} = 1$ if $xyz = lmn$, 0 otherwise and $\tilde{\lambda}_{xyz}^{lmn}$ is the fully non-local descriptor set, equivalent to "unraveling" the entire density grid as a vector (indexed by xyz) at each spatially-resolved grid point (indexed by lmn). Substitution into Eq. 8 yields:

$$\epsilon^{lmn}(\rho^{xyz}) = \epsilon^{lmn}(\tilde{\lambda}_{xyz}^{lmn}) \tag{11}$$

This expression is a trivial re-statement of Eq. 8, but it has the advantage of being a system-independent mapping between a locally-centered electronic environment (as characterized by its convolutional descriptors) and a corresponding local XC energy density. However, in practice Eq. 11 is no more efficient or practical than Equation 8 since both ultimately require a fully non-local 6-dimensional evaluation of the energy functional. However, Eq. 11 provides a natural starting point for establishing controlled orbital-free approximations to the XC energy density based on sets of descriptors λ_q^{lmn} where $R << L_{max}/2$ and $N_d << n^3$.

$$\epsilon^{lmn}(\rho^{xyz}) = \epsilon^{lmn}(\tilde{\lambda}_{xyz}^{lmn}) \approx \epsilon^{lmn}(\lambda_q^{lmn}) \tag{12}$$

The two most common classes of orbital-free XC functionals, LDA and GGA, are easily reformulated in terms of convolutional descriptors. For example, the LDA functional approximates the exchange-correlation energy at a point $\vec{x}$ with the XC energy of the homogeneous electron gas with density equivalent to that point:

$$\epsilon_{LDA}[\rho(\vec{x})](\vec{x}) = \epsilon_{HEG}(\rho(\vec{x})) \tag{13}$$

or, in convolutional notation:

$$\epsilon^{lmn}_{LDA}(\rho^{xyz}) = \epsilon^{lmn}_{HEG}(\rho^{lmn}) = \epsilon^{lmn}_{HEG}(\lambda^{lmn}_0) \tag{14}$$

where $\lambda^{lmn}_0 = (\delta_{lmn} \circledast \rho)^{lmn} = \rho^{lmn}$. In the case of GGA functionals the XC energy density depends on the density and its gradient:

$$\epsilon_{GGA}[\rho(\vec{x})](\vec{x}) = \epsilon_{GGA}(\rho(\vec{x}), \nabla\rho(\vec{x})) \tag{15}$$

or, in convolutional notation:

$$\epsilon^{lmn}_{GGA}(\rho^{xyz}) = \epsilon^{lmn}_{GGA}(\rho^{lmn}, \nabla\rho^{lmn}) = \epsilon^{lmn}_{GGA}(\lambda^{lmn}_0, \lambda^{lmn}_1) \tag{16}$$

where $\lambda^{lmn}_0 = \rho^{lmn}$ as before, and $\lambda^{lmn}_1 = (\nabla_1 \circledast \rho)^{lmn}$ where ∇_1 is a finite difference stencil corresponding to the gradient. This pattern can be generalized to higher-order derivatives to produce a class of convolutional XC functionals based on differential stencils:

$$\begin{aligned}\epsilon^{lmn}_{\nabla N}(\rho^{xyz}) &\approx \epsilon^{lmn}(\nabla^{lmn}_0, \nabla^{lmn}_1, \nabla^{lmn}_2, ...\nabla^{lmn}_N) \\ &= \epsilon^{lmn}(\nabla^{lmn}_q)\end{aligned} \tag{17}$$

where $\nabla^{xyz}_q = (\nabla_q \circledast \rho)^{xyz}$ and ∇_q is the q^{th} unmixed partial derivative stencil, with $\nabla_0 \equiv \delta_{xyz}$. Thus, $\epsilon^{lmn}_{\nabla 0}$ corresponds to any fully local functional (e.g. LDA) and $\epsilon^{lmn}_{\nabla 1}$ corresponds to GGA functionals, while $\epsilon^{lmn}_{\nabla 2}$ corresponds to functionals that include the Laplacian[92], etc. The idea is analogous to that of Taylor series expansion in that it uses linear combinations of different orders of derivatives to approximate a function. Hence the functional should become more accurate as higher order derivatives are included and longer-range information is taken into account. However, this approach suffers from a few issues, in theory and in practice: it's hard, if not impossible, to construct isotropic or rotation-invariant stencils for higher order derivatives, gradient expansions only improve accuracy when the density varies slowly, and higher order derivatives tend to become numerically unstable with practical grid spacing, or even not integrable at all[93,94]. It is clear that the descriptors must stay constant as the system rotates and translates. In other words, the stencils need to be rotation- and translation invariant. The magnitude of gradient operator and the Laplacian operator (trace of the Hessian) that are effectively adapted in GGA and mGGA, respectively, are known to be isotropic. However, the invariant norms of higher-order derivative tensors are not well known. Furthermore, issues with numerical stability are encountered when computing the Laplacian and higher order derivatives, even with analytical basis sets. For these reasons most functionals based on ∇^{xyz}_2 have been abandoned for the "meta-GGA" approach, which substitutes the kinetic energy density, τ^{xyz}, for ∇^{xyz}_2[95,96]. The kinetic energy density is not orbital-free, and this transition deviates from the formalism of the Taylor expansion, making it unclear how to systematically improve model spaces beyond mGGA.

B. Maxwell-Cartesian spherical harmonic descriptors

Prior to selecting a convolutional descriptor set to fingerprint electronic environments, it is important to first define the necessary properties of the descriptor set. First the descriptor set needs to be complete. This means, at the limit of taking all the descriptors from the set, they should form a complete basis that can describe all possible variations in the electronic environment, or that of any 3D function in general. This complete set would be infinite, therefore the set should have a clear entry point and a systematic route toward convergence. Furthermore, the descriptors need to be invariant under translation and rotation, consistent with the symmetries of the Hamiltonian.

To find such descriptor set, the variation of 3D functions is decomposed into two parts: variations in angular coordinate (rotational variation) and variations in the radial coordinate (radial variation). The Maxwell-Cartesian spherical harmonics (MCSH) capture the rotational variations, and the radial variation is captured by varying a cutoff distance. The MCSH descriptors are selected over standard spherical harmonics because they posess straightforward rotationally-invariant norms. This is inspired by the circular-harmonic-based 2D rotationally equivariant features developed by Worral et al.[76] that has been generalized to 3D by the work of Thomas et al.[97], as well as Applequist's work[77] on MCSHs. Specifically, the descriptor set is defined as follows:

$$\left\{ M^{(n)}_{r,ijk} = \sqrt{\sum\nolimits_{P(i,j,k)} \mu^2_{r,ijk}} \mid i,j,k \in \mathbb{N}, r \in \mathbb{R}^+ \right\} \tag{18}$$

where M denotes the descriptors, $P(i,j,k)$ denotes the permutation group of i,j,k, $n = i+j+k$ is the order of the descriptor, and $\mu_{r,ijk}$ is the convolution result using spherical harmonic S_{ijk} with cutoff distance r as the stencil:

$$\mu_{r,ijk} = input \circledast \overbrace{\left[f_r(x,y,z) \times S^{(n)}_{ijk}(x,y,z) \right]}^{\text{stencil}} \tag{19}$$

f_r is a step function that controls the cutoff distance, and $S^{(n)}_{ijk}$ is the Maxwell-Cartesian spherical harmonic:

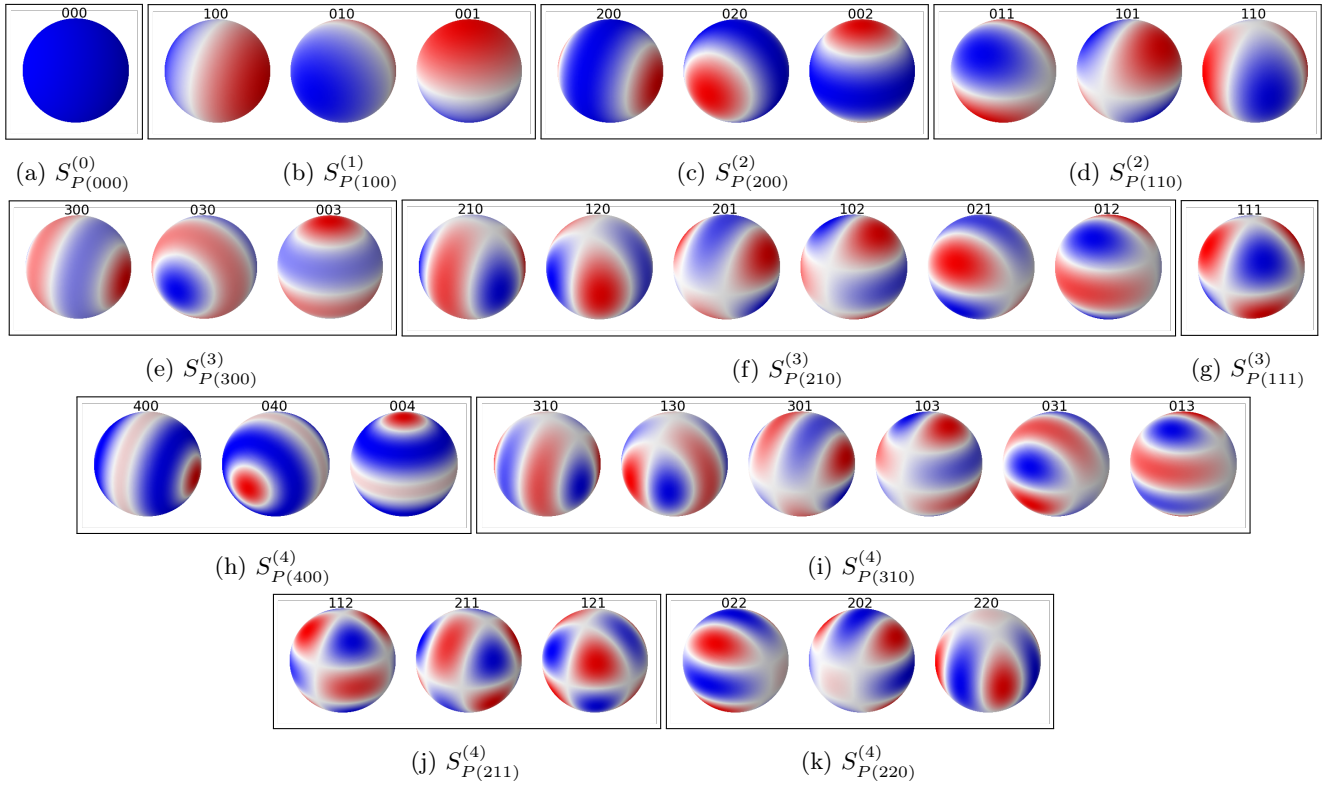


Figure 2: Graphical illustrations of the first 4 orders of Maxwell-Cartesian spherical harmonics (MCSH) descriptor kernels denoted by $S^{(n)}_{P(ijk)}$. n is the order and $P(ijk)$ denotes the permutation group of the index ijk. The Euclidean norm of MCSH stencils in each group provides the 3D rotation-invariant descriptors that are used as inputs to the neural networks.

$$f_r(x,y,z) = \begin{cases} 1 & \text{if } \sqrt{x^2+y^2+z^2} \le r \\ 0 & \text{if } x < 0 \end{cases} \tag{20}$$

$$S^{(n)}_{ijk}(x,y,z) = \sum_{m_1=0}^{i/2}\sum_{m_2=0}^{j/2}\sum_{m_3=0}^{k/2} (-1)^m (2n-2m-1)!! \times \begin{bmatrix} i \\ m_1 \end{bmatrix} \begin{bmatrix} j \\ m_2 \end{bmatrix} \begin{bmatrix} k \\ m_3 \end{bmatrix} r^{2m} x^{i-2m_1} y^{j-2m_2} z^{k-2m_3} \tag{21}$$

where $m = m_1 + m_2 + m_3$, and

$$\begin{bmatrix} a \\ b \end{bmatrix} = \frac{a!}{2^b b!(a-2b)!} \tag{22}$$

Thus, each descriptor is the Euclidean norm of $\mu_{r,ijk}$ with all possible combination of i, j, k, and the whole set can be written as:

$$\left\{ \begin{array}{l} \sqrt{\mu^2_{r_1,000}}, \sqrt{\mu^2_{r_1,100}+\mu^2_{r_1,010}+\mu^2_{r_1,001}}, \sqrt{\mu^2_{r_1,200}+\mu^2_{r_1,020}+\mu^2_{r_1,002}}, \sqrt{\mu^2_{r_1,110}+\mu^2_{r_1,101}+\mu^2_{r_1,011}}, \dots \\ \sqrt{\mu^2_{r_2,000}}, \sqrt{\mu^2_{r_2,100}+\mu^2_{r_2,010}+\mu^2_{r_2,001}}, \sqrt{\mu^2_{r_2,200}+\mu^2_{r_2,020}+\mu^2_{r_2,002}}, \sqrt{\mu^2_{r_2,110}+\mu^2_{r_2,101}+\mu^2_{r_2,011}}, \dots \\ \dots \end{array} \right\} \tag{23}$$

The first four orders of the MCSHs are listed in Table I and illustrated in Figure 2. The detailed properties of

n	{ijk}	$S_{ijk}^{(n)}$	n	{ijk}	$S_{ijk}^{(n)}$
0	000	1	4	400	$105\hat{x}^4 - 90\hat{x}^2 + 9$
1	100	$\hat{x}$		040	$105\hat{y}^4 - 90\hat{y}^2 + 9$
	010	$\hat{y}$		004	$105\hat{z}^4 - 90\hat{z}^2 + 9$
	001	$\hat{z}$		310	$105\hat{x}^3\hat{y} - 45\hat{x}\hat{y}$
2	200	$3\hat{x}^2 - 1$		301	$105\hat{x}^3\hat{z} - 45\hat{x}\hat{z}$
	020	$3\hat{y}^2 - 1$		031	$105\hat{y}^3\hat{z} - 45\hat{y}\hat{z}$
	002	$3\hat{z}^2 - 1$		130	$105\hat{x}\hat{y}^3 - 45\hat{x}\hat{y}$
	110	$3\hat{x}\hat{y}$		103	$105\hat{x}\hat{z}^3 - 45\hat{x}\hat{z}$
	101	$3\hat{x}\hat{z}$		013	$105\hat{y}\hat{z}^3 - 45\hat{y}\hat{z}$
	011	$3\hat{y}\hat{z}$		220	$105\hat{x}^2\hat{y}^2 - 15\hat{x}^2 - 15\hat{y}^2 + 3$
3	300	$15\hat{x}^3 - 9\hat{x}$		202	$105\hat{x}^2\hat{z}^2 - 15\hat{x}^2 - 15\hat{z}^2 + 3$
	030	$15\hat{y}^3 - 9\hat{y}$		022	$105\hat{y}^2\hat{z}^2 - 15\hat{y}^2 - 15\hat{z}^2 + 3$
	003	$15\hat{z}^3 - 9\hat{z}$		211	$105\hat{x}^2\hat{y}\hat{z} - 3\hat{y}\hat{z}$
	210	$15\hat{x}^2\hat{y} - 3\hat{y}$		121	$105\hat{x}\hat{y}^2\hat{z} - 3\hat{x}\hat{z}$
	201	$15\hat{x}^2\hat{z} - 3\hat{z}$		112	$105\hat{x}\hat{y}\hat{z}^2 - 3\hat{x}\hat{y}$
	021	$15\hat{y}^2\hat{z} - 3\hat{z}$			
	120	$15\hat{x}\hat{y}^2 - 3\hat{x}$			
	102	$15\hat{x}\hat{z}^2 - 3\hat{x}$			
	012	$15\hat{y}\hat{z}^2 - 3\hat{y}$			
	111	$15\hat{x}\hat{y}\hat{z}$			

Table I: The analytical expressions of first 4 orders of MCSH denoted by $S_{ijk}^{(n)}$, where $\hat{x} = x/r$, $\hat{y} = y/r$ and $\hat{z} = z/r$

MCSH are introduced in the work of Applequist[77]. The MCSHs are used to construct the descriptors because it is known that spherical harmonics form a complete basis for functions defined on the 3D unit sphere. This idea is analogous to that of multi-pole expansion, where the original 3D function is expressed as a linear combination of terms with progressively finer angular features[98]. Examining the MCSHs in Figure 2 reveals that the order 0 MCSH corresponds to the monopole, and captures features that are constant and independent of angle (order 0 angular feature); the order 1 MCSH corresponds to the dipole, and captures features that vary once, from positive to negative, with angle (order 1 angular features); order 2 MCSH corresponds to the quadrupole, and captures features that varies more quickly with angle (order 2 angular features), and so on. Any rotational variations can be approximated by this linear combinations of angular features of different order, and will be exact in the limit of the entire series. The descriptors are empirically verified as rotation-invariant (see Supporting Information) and the mathematical proof is in progress but is beyond the scope of this work. In addition to the rotational variations it is necessary to capture radial variations. This is achieved by taking the rotation-invariant descriptors with different cutoff distances through the cutoff function f_r in Eq. 20, where cutoff radii are discretized based on the underlying finite-difference grid. We conjecture that this descriptor set provides a complete basis on the rotation-invariant subspace of the 3D finite difference grid. Moreover, $\mu_{0,000}^2$, which is equivalent to fully local information (i.e. ρ), is the clear entry point of this descriptor set.

Descriptor Set	Number of Descriptors	Descriptor Set	Number of Descriptors
$\bar{\lambda}_{(0.00)}^{(0)}$	1	$\bar{\lambda}_{(0.08)}^{(1)}$	9
$\bar{\lambda}_{(0.02)}^{(0)}$	2	$\bar{\lambda}_{(0.2)}^{(1)}$	21
$\bar{\lambda}_{(0.04)}^{(0)}$	3	$\bar{\lambda}_{(0.02)}^{(2)}$	5
$\bar{\lambda}_{(0.08)}^{(0)}$	5	$\bar{\lambda}_{(0.04)}^{(2)}$	9
$\bar{\lambda}_{(0.2)}^{(0)}$	11	$\bar{\lambda}_{(0.08)}^{(2)}$	17
$\bar{\lambda}_{(0.02)}^{(1)}$	3	$\bar{\lambda}_{(0.2)}^{(2)}$	41
$\bar{\lambda}_{(0.04)}^{(1)}$	5		

Table II: List of number of features for each of the descriptor sets

There are 3 directions for systematic expansion: higher orders of MCSH, longer cutoff distances, and a finer grid for discretization. In this work we fix the grid spacing and explore the impact of higher orders of MCSH and longer cutoff distances.

The MCSH descriptors provide a route to include semi-local and non-local information about a local electronic environment. MCSH descriptor sets are defined by their maximum range (R) and the maximum order of spherical harmonics (n) that is the same as the maximum order of angular features captured, and are denoted as $\bar{\lambda}_R^{(n)}$. Here, we consider 13 MCSH descriptor sets with ranges of 0 Å , 0.02 Å, 0.04 Å , 0.08 Å and 0.2 Å, and orders of 0, 1 and 2. The descriptor sets are designed such that information of longer range and higher order of angular feature are gradually added:

$$
\begin{aligned}
\bar{\lambda}_{(0.0)}^{(0)} &= \{M_{(0.0,000)}^{(0)}\} = \{\rho\} \\
\bar{\lambda}_{(0.02)}^{(0)} &= \{\rho, M_{(0.02,000)}^{(0)}\} \\
\bar{\lambda}_{(0.02)}^{(1)} &= \{\rho, M_{(0.02,000)}^{(0)}, M_{(0.02,100)}^{(1)}\} \\
\bar{\lambda}_{(0.02)}^{(2)} &= \{\rho, M_{(0.02,000)}^{(0)}, M_{(0.02,100)}^{(1)}, M_{(0.02,200)}^{(2)}, M_{(0.02,110)}^{(2)}\} \\
\bar{\lambda}_{(0.04)}^{(0)} &= \{\rho, M_{(0.02,000)}^{(0)}, M_{(0.04,000)}^{(0)}\} \\
\bar{\lambda}_{(0.04)}^{(1)} &= \{\rho, M_{(0.02,000)}^{(0)}, M_{(0.02,100)}^{(1)}, M_{(0.04,000)}^{(0)}, M_{(0.04,100)}^{(1)}\}
\end{aligned}
\tag{24}
$$

The total number of descriptors for each of the descriptor sets are listed in Table II.

C. Regression models for exchange-correlation energy

The functional form linking the MCSH descriptors and the exchange correlation (XC) energy density is not known. While reductionist approaches may be feasible, the empirical approach is more pragmatic since the physical meaning of the descriptors is not obvious. In this work we employ a machine-learning strategy based on a

function with two terms: a local-density term based on a re-parameterization of the VWN functional form of LDA (r-VWN), and a descriptor-based term using a NN with ReLU activation functions. The r-VWN term is static for all regression models, and the NN is trained using the residuals of the r-VWN model (Eq. 2); this is similar to the Δ machine learning strategy proposed previously[81]. This section first discusses the results of the r-VWN model and a NN based solely on the local density, and subsequently addresses the performance of NNs based on the convolutional descriptors. It is worth to note that the models built in this study are not self-consistent. Instead, they are trying to directly predict B3LYP SCF converged XC energy density from converged electron density.

1. *r-VWN and NN LDA model*

The domain and range of the electron density and corresponding XC energy density span over 12 orders of magnitude, as seen in Fig. 3. This creates a substantial numerical challenge for machine-learning models since most implementations rely on double precision floats with a machine epsilon of $\sim 10^{-16}$. Practically, the situation is somewhat better, since the vast majority of the distribution (99.3%) falls between $10^{-6.5} - 10^{0}$ (Fig. 3). However, even relatively small errors in the high-density region can have a substantial impact on the system-level energy, and training a machine-learning model that is accurate across this span is challenging. Nonetheless, physical models are known to approximate the energy density across this large domain/range. The VWN parameterization of the LDA model achieves this by using an analytical function that reproduces the behavior of the homogeneous electron gas (HEG)[82]:

$$\begin{aligned} e_{XC,VWN} &= e_{X,VWN} + e_{C,VWN} \\ &= \rho\frac{C_1}{r_s} + \rho G(r_s, \gamma, \alpha_1, \beta_1, \beta_2, \beta_3, \beta_4) \end{aligned} \tag{25}$$

where $C_1, \gamma, \alpha_1, \beta_1, \beta_2, \beta_3, \beta_4$ are the parameters, r_s is the Wigner–Seitz radius defined as:

$$r_s = (3/4\pi\rho)^{1/3} \tag{26}$$

G is defined as:

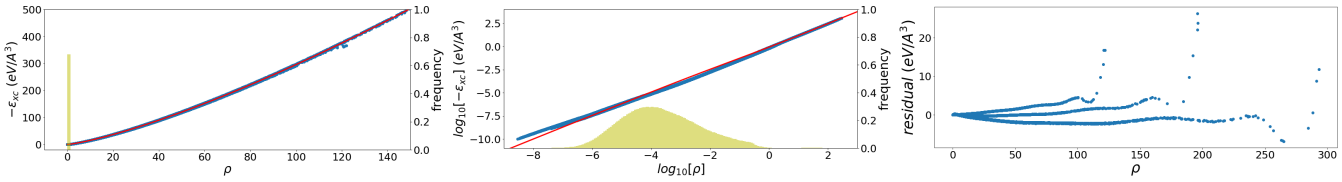
$$\begin{aligned} G(r_s, \gamma, \alpha_1, \beta_1, \beta_2, \beta_3, \beta_4) = -2\gamma(1+\alpha_1 r_s) \\ \times \ln\left\{1 + \frac{1}{2\gamma r_s^{1/2}(\beta_1 + r_s^{1/2}(\beta_2 + r_s^{1/2}(\beta_3 + r_s^{1/2}\beta_4)))}\right\} \end{aligned} \tag{27}$$

The behavior of the HEG is qualitatively similar to the B3LYP training data, but there are quantitative discrepancies. These residuals were minimized using nonlinear optimization with the VWN parameters as initial guesses (see Sec. II) providing an improved LDA approximation (r-VWN) to B3LYP for molecular systems of interest. The refitted model is illustrated in Fig. 3b, and the magnitude of the residuals is reduced by an order of magnitude as compared to the original energy density (Fig. 3a,c).

The r-VWN model is applied to the test and training sets, and the results are compared to the VWN and common PBE GGA functionals in Fig. 4. The results show that the energy prediction error stays approximately constant ($MAE_{VWN} = 67.44$ eV, $MAE_{r-VWN} = 69.18$ eV). Interestingly, both VWN and r-VWN have lower energy errors than PBE, and the "test" set has higher errors on average for both VWN and PBE despite the fact that they are not trained on the training set (i.e. the test set has inherently larger system-level errors). The trend between VWN and r-VWN is similar for the formation energy errors, where both systematically underestimate the B3LYP formation energy, and the r-VWN model has less systematic error ($MAE_{VWN} = 1.69$ eV, $MAE_{r-VWN} = 1.23$ eV). The magnitude of formation energy errors are also ∼2 orders of magnitude smaller than the system-level errors, due to cancellation of error. The formation energy errors for PBE are somewhat lower than even the r-VWN model, opposite of the trend for system-level energies, indicating that PBE relies more on cancellation of error between systems than the LDA models. The findings regarding the relative accuracy of VWN and PBE are counter to prior results[99], likely owing to the fact that the calculations in this work are not self-consistent and rely on B3LYP energies as the ground truth.

The residual learning framework is also applied to the LDA model space ($NN[\bar{\lambda}^{(0)}_{(0.00)}]$) to provide a control for the convolutional descriptor models. Although the $r-VWN$ model doesn't show significant improvement in accuracy, it provides a good approximation to B3LYP XC energy density, and learning the residual is easier than learning the energy density directly due to a reduction in the range of the dependent variable. The results, also shown in Fig. 4, show a significant improvement in the system-level energy and a less dramatic but still significant improvement in the formation energy. Interestingly the formation energies from the NN model are more accurate than the GGA results, despite the fact that the GGA model space contains more information. The cancellation of error indicates that the NN model is not over-fit, and the improved performance indicates that there is room for improvement of LDA models by increasing flexibility, consistent with prior studies[65]. However, as shown in Figure 3c, the local energy residuals for the $r-VWN$ model shows three tails, which are attributed to the core regions of C, O, and N. In both cases the main error arises from the fact that the local energy is not a single-valued function of the local density, especially in the core regions. This suggests that the limiting factor to further improve accuracy is not the flexibility of the XC model, but the information contained in the model space. This motivates the inclusion of additional descriptors (Sec. III C 2).

To get more insight into the origins of the errors for



(a) ϵ_{xc} vs. ρ (b) ϵ_{xc} vs. ρ log scale (c) r-VWN residual vs. ρ

Figure 3: Plots of 3 million randomly sampled data points from training set to show the distribution, plus 10000 uniformly subsampled data points to highlight the high-density (core) regions. (a) Plot of ϵ_{xc} vs. ρ in linear scale. The yellow bar plot shows the distribution of points. (b) Plot of ϵ_{xc} vs. ρ in log scale. The yellow bar plot shows the distribution of points. (c) Plot of r-VWN model residual vs. ρ in linear scale. It is clear that the domain and range of both the electron and energy density span many orders of magnitude, and that the energy density is multi-valued with respect to electron density.

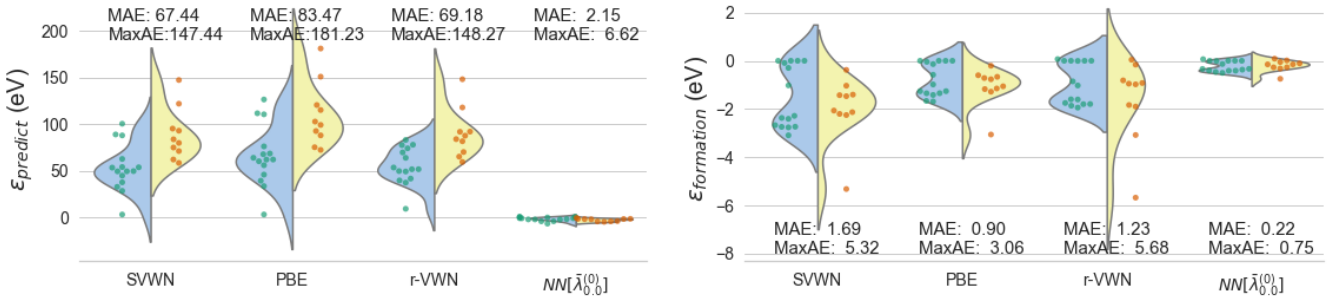
the $r-VWN$ and $NN[\bar{\lambda}^{(0)}_{(0.00)}]$ functionals we examine the contribution to the system-level error as a function of density. From Figure 3 it is clear that the domain and range of both the electron and energy density span many orders of magnitude, and that there is a wide range in the number of points that occur at different electron densities, with the vast majority falling between $10^{-6.5}-10^{0}$. The system-level energy is ultimately computed by integrating (approximated by summation) the local energy density, hence the system-level error will depend on a trade-off between the size of the error at a given density and the number of points with that density. This is illustrated in Fig. 5, where the contribution to the system-level error is plotted as a function of the electron density. The results indicate that for the r-VWN model nearly all of the system-level error occurs in the density region of $10^{-3}-10^{1}$ eV/Å^3, corresponding to the valence/bonding regions of the molecular system. This is intuitive, since this is where chemical bonding occurs, and where there are an appreciable number of data points (29.5%), the energy density is relatively large ($10^{-4}-10^{1}eV/$Å^3), and is multi-valued (see Fig. 3c). In comparison, the error from the NN[LDA] model is much smaller, and concentrated in the region of $10^{-1}-10^{1}eV/$Å^3. This is attributed to the near-core regions where the energy density is relatively large and multi-valued, making it impossible for the neural network to capture the behavior.

This trade-off between the error and the number of sample points highlights the importance of the inclusion of randomly sampled data in the sub-sampling routine, and the selection of a proper choice of objective function during training. The randomly sampled data effectively weights the error at each density by the number of points similar to the weight that will be used to compute the system-level energy. Enough randomly sampled data must be included to ensure that regions with low electron/energy density contribute to the objective function, but if too much randomly-sampled data is included it will overwhelm the contribution from the tails of the distribution. The ratio of random to uniform sub-samples is chosen heuristically to minimize the system level error in this work. A multi-step training process is also employed, with multiple types of objective function. The details of the training procedure can be found in the Supplementary Information. Ultimately, the results of the $r-VWN$ and $NN[\bar{\lambda}^{(0)}_{(0.00)}]$ models indicate that information beyond the local electron density must be included in the model space to significantly improve accuracy.

2. *NN models with MCSH descriptors*

To capture more non-local information about the electronic environments MCSH descriptors are used, and NN models are applied to connect the descriptors to the energy density. For each set a NN model with 2 hidden layers of 100 nodes each and ReLU activation functions is used as the non-linear model for the XC energy density, denoted as NN$[\bar{\lambda}^{(n)}_{R}]$ where R denotes the cutoff radius and n denotes the order of the MCSHs used (see Eq. 18). Each NN model is trained using a consistent training procedure, as described in the Supplementary Information. The hypothesis that including more semi-local information in the model space will systematically improve the model accuracy is tested by comparing system-level sum of local absolute error ($\varepsilon_{absolute}$), energy prediction error ($\varepsilon_{predict}$) and formation energy prediction error ($\varepsilon_{formation}$) as defined in Section II with a consistent NN architecture. Systematic improvement is defined as improvement for each individual system without exception, while general improvement is defined as a decrease in the mean absolute error.

The results for general improvement are shown in Figure 6, indicating that the general accuracy always improves as more descriptors are added. The detailed result for each model and system are given in the Supporting Information. In this section we focus on the models' performances for the 15 training and 7 test systems; the 3 extrapolation systems are described in Sec. III C 3. Based on Figure 6d, general improvement in the model accuracy is observed as angular features from zero-th order to first order are included and the range is increased from 0 to 0.2 Å. The inclusion of the first-order angular



(a) Energy prediction error distribution

(b) Formation energy prediction error distribution

Figure 4: Results for local density based models. Error distributions for system-level energy prediction error (a) and formation energy prediction error (b). Blue points/curves correspond to the training molecules, orange points/yellow curves correspond to the test and extrapolation molecules. MAE denotes the mean absolute error of all systems, and $MaxAE$ denotes the maximum of absolute error. Results show that the flexiblity of the neural network for an LDA-like model reduces the error, but that formation energy errors still exceed chemical accuracy.

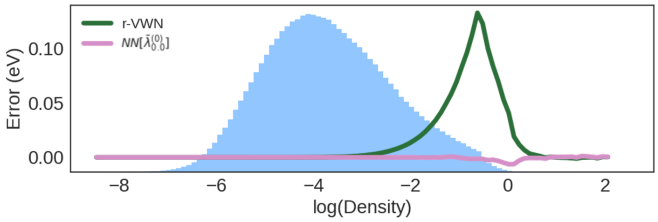


Figure 5: Plot of the sum of error for the models across different density scales for the same randomly sampled data points. The plot indicates that for the r-VWN model nearly all of the system-level error occurs in the density region of $10^{-3} - 10^{1}$ eV/Å^3, corresponding to the valence/bonding regions of the molecular system.

features has a drastic impact, where the accuracy of the first-order model with a range of 0.02 Å is comparable to the zeroth-order model with a range of 0.08 Å. The first-order angular feature is needed to express the reduced gradient, and the grid spacing is 0.02 Å, so the NN$[\bar{\lambda}^{(1)}_{0.02}]$ model is analogous to the GGA model space. A further and substantial improvement is observed as the range is increased, with a minimum MAE of 0.061 eV achieved at a range of 0.2 Å. The inclusion of descriptors of second-order angular features further improves the model, particularly at short ranges, but the improvement is less drastic. The high accuracy of these models with explicit ranges of $\leq$ 0.2 Å is interesting, since prior studies of the XC-hole of similar systems have suggested that the scale of the X-hole is much wider than 0.2 Å[100,101]. The fact that the XC energy can be reproduced with the relatively short range of 0.2 Å suggests that the longer range behavior of XC-hole is predictable from the short-range information that was used to train the models. This is similar to the finding that semi-local mGGA functionals are able to describe intermediate-range van der Waals forces[102,103].

The results for the systematic improvability test for the sum of absolute error are shown in Figure 7, where the number represents the maximum increase in error for any given system when the order of the angular feature or spatial range is increased; a value of 0 represents a systematic improvement since the error of every system is decreased without exception. The results show that systematic improvement is often, but not always, observed when additional descriptors are added. In particular, systematic improvability is not observed for zeroth-order angular features as range is increased from 0 to 0.08 Å. This is hypothesized to occur because the added descriptors contain relatively little additional information, causing statistical noise to play a larger role in training. The randomly-selected training data represents only 0.55% of the total data, and neural networks are initialized with random weights, causing the resulting models to favor some electronic environments over others due to randomness. This could possibly be overcome by using a static training set and a systematic strategy for initializing the neural networks, or by adding descriptors with more information. The latter strategy is shown to work here, as systematic improvability is achieved when higher-order angular descriptors and/or longer-range information are included. One exception to this trend is observed when moving from descriptors of first-order angular features to that of second-order angular features at a range of 0.2 Å. This is attributed to the fact that the dimensionality of the descriptor space increases substantially from 21-41, but the flexibility of the NN model is not increased. The points will be more separated in a higher-dimensional space, as supported by the fact that the number of uniformly-sampled data from the $\bar{\lambda}^{(n)}_{(0.2)}$ sets are an order of magnitude higher than for the $\bar{\lambda}^{(n)}_{(\leq 0.08)}$ descriptors (see Supplementary Information). This forces the model to interpolate over larger distances, and will generally require a more complex NN model. These re-

sults provide evidence supporting the hypothesis that systematic XC model improvement can be achieved by systematically increasing the range and rotational order of convolutional descriptors, though optimization of the regression model architecture and training procedure is an important consideration.

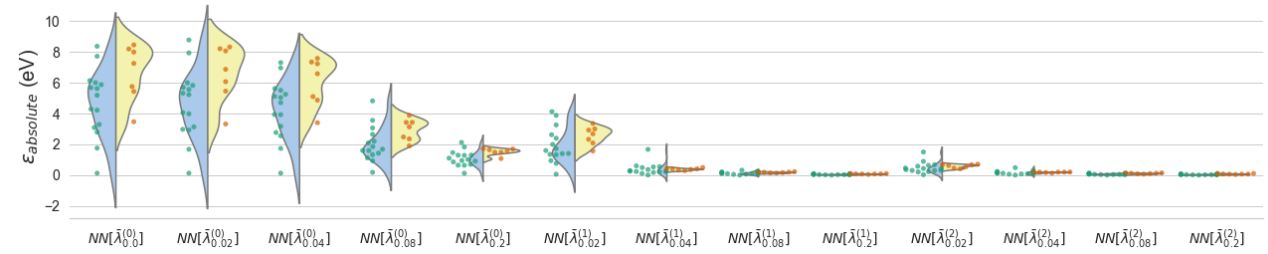
Systematic and general improvement of the absolute error is promising, but the physical quantities of energy prediction error and formation energy prediction error are of practical interest. The absolute error provides an upper bound for these quantities, which can take advantage of cancellation of error within a single system (energy prediction error) and across systems (formation energy prediction error). Indeed, the general accuracy of these quantities is greatly improved as compared to the absolute error as shown in Figure 6. The MAE of the prediction error reaches "chemical accuracy" (0.043 eV) at a spatial range of 0.08 Å for the first-order rotational descriptors, and 0.04 Å with second-order rotational descriptors. A longer range is needed to reduce the maximum error to below chemical accuracy, but this can be achieved with both first- and second-order rotational descriptors at a range of 0.2 Å. While this general decrease in error is promising, it should be noted that the systematic improvement is not observed at the prediction energy level. This arises due to the fact that cancellation of error plays a varying role in different chemical systems depending on the frequency with which different electronic environments occur. This variation in cancellation of error will also occur with other types of XC functionals, and explains the tremendous difficulty of achieving systematic improvement in the field of XC functional design. The counter-intuitive nature of cancellation of error is even more apparent when comparing prediction energy errors and formation energy errors. Formation energies generally benefit from cancellation of error across different systems, particularly in the core regions since the atomic composition of a molecule is utilized to compute the formation energy. This is apparent in the general improvement of between the prediction energy and the formation energy. However, when examining systematic improvement it is clear that some systems exhibit drastically larger formation-energy errors than prediction errors. This arises due to the anti-cancellation of error between a reference system and the system of interest, and highlights an additional consideration for the design of functionals with systematic improvements in formation energies. Nonetheless, several models ($\bar{\lambda}^{(1)}_{(0.2)}$, $\bar{\lambda}^{(2)}_{(0.08)}$, $\bar{\lambda}^{(2)}_{(0.2)}$) are capable of reducing even the maximum formation-energy error of the convolutional descriptor models to within chemical accuracy.

3. *Outliers and Extrapolation*

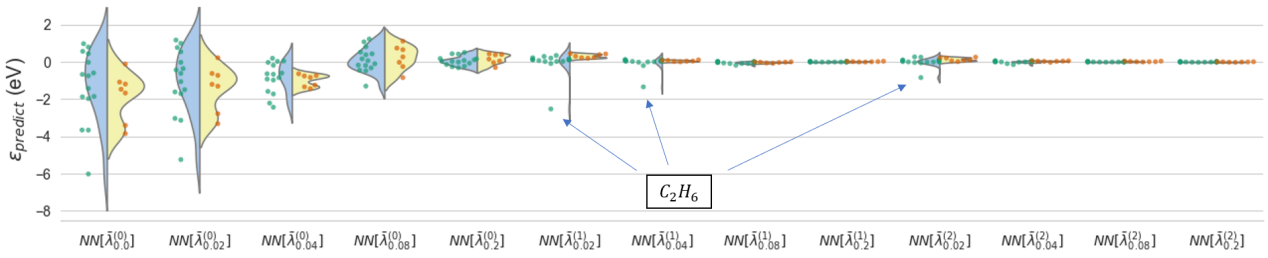
The results discussed in Sec. III C 2 are generally positive, although there is a noticeable outlier in the training set for some models, and the "universality" of the model is not clear since the test set contains similar chemistry to the training set. In this section we examine the outlier (C_2H_6) to gain insight into where the approach fails, and attempt to extrapolate the model to three systems with different chemistry: CH_3NO_2, *glycine* and $NCCN$. These compounds contain nitro groups, amine groups, and multiple cyano groups that are not present in the train or test data and hence probe the machine-learning model's ability to generalize to new chemistries.

First, we address the C_2H_6 system that appears as an outlier in the training set. This is generally surprising, since machine-learning models tend to perform well on data they are trained to. Notably, the C_2H_6 system is not a clear outlier for $\varepsilon_{absolute}$ (Figure 6a), which is directly related to the objective functions used in training (see SI), confirming that this is not a failure of the NN training procedure. However, C_2H_6 becomes an outlier in $\varepsilon_{predict}$ (Figure 6b), and even more substantially in $\varepsilon_{formation}$ (Figure 6c). This indicates that the issue arises due to a lack of cancellation of error in the prediction energy, and/or anti-cancellation of error in the formation energy. This is attributed to a combination of two factors: electronic environments that are not sufficiently distinguished by descriptors and under-representation of these electronic environments. These factors are illustrated using the $NN[\bar{\lambda}^{(1)}_{(0.04)}]$ model. Since domain decomposition method is used in model training (see Sec. II), it is possible to determine that most of the prediction error (1.51 eV out of 1.67 eV for $\varepsilon_{absolute}$ and -1.32 eV out of -1.34 eV for $\varepsilon_{predict}$) can be attributed to the $C-C$ bonding region of the system. The electronic environments in this region were projected onto a low-dimensional space using principal component analysis (PCA) and compared to the environments in the entire training set. Figure 8 shows the model error as a function of two principal components, and illustrates that several points have substantially smaller errors for the training set as compared to C-C bonding region of C_2H_6. This indicates that the objective function is multi-valued at these locations in descriptor space, forcing the model to make a tradeoff in accuracy between the two possible outcomes. This tradeoff will depend on the relative frequency of the two types of environments that are present in the training set. In this study C_2H_6 is the only molecule in the training set with a C-C single bond, causing these environments to be under-represented and not favored by the model. This could be remedied to some extent by including more examples of C-C bonds in the training set, though this would simply balance the error between systems rather than reducing it. Alternatively, the inclusion of more descriptors enables the model to distinguish between these environments and reduce the error for both; this is evident from the fact that both the prediction and formation energy errors for the C_2H_6 molecule reduce substantially as higher-order and longer-range descriptors are included (Fig. 6).

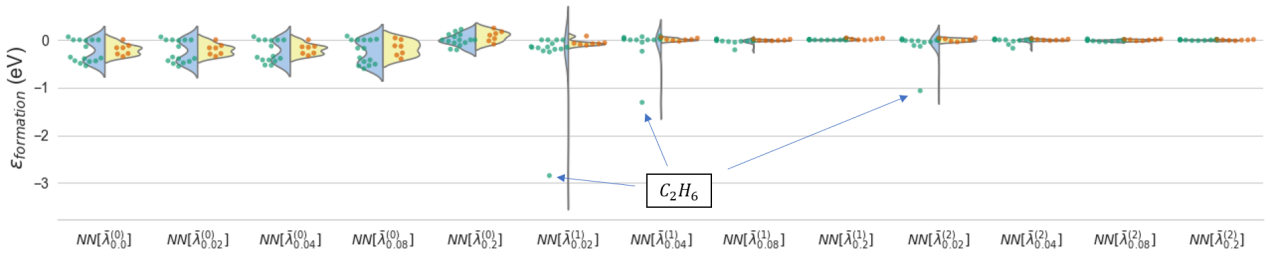
The results for the extrapolation set (CH_3NO_2,



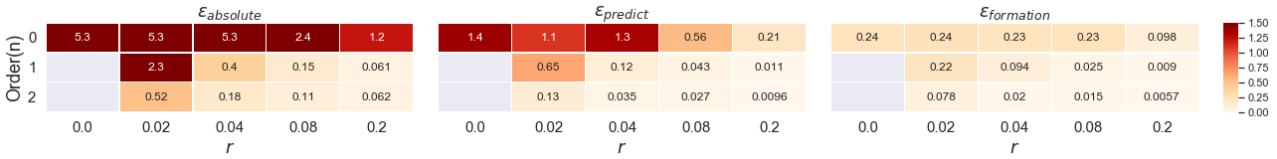
(a) Sum of absolute energy prediction error distribution



(b) Energy prediction error distribution



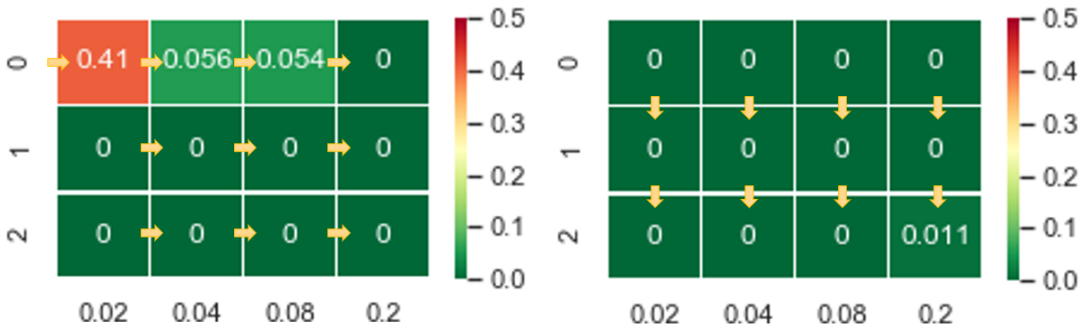
(c) Formation energy prediction error distribution



(d) Mean absolute error (MAE) of the error metrics with all models

Figure 6: Results for MCSH descriptor based models. Error distributions for sum of absolute error (a), system-level energy prediction error (b) and formation energy prediction error (c). Blue points/curves correspond to the 15 training molecular systems, orange points/yellow curves correspond to the 7 test molecular systems. *d*) Statistical analysis of the errors of the 15 training and 7 test molecular systems. The 3 extrapolation systems are not considered here. The plots show that the general accuracy always improves as more descriptors are added, and that the MAE of the prediction error reaches "chemical accuracy" (0.043 eV) at a spatial range of 0.08 Å for the first and second order rotational descriptors.

glycine and *NCCN*) are shown in Table III, and it is clear that the errors are generally larger by as much as an order of magnitude. This situation is common in machine learning, and generally arises when the training data is not representative of the test data. This occurs because the NN model can only interpolate between training examples, and will become unreliable if used for extrapolation[104]. In this case the extrapolation system contains several chemical environments that are not observed in the test systems, so it is not surprising that unique electronic environments arise. This is quantitatively illustrated in Fig. 9 in the case of the $NN[\bar{\lambda}^{(1)}_{(0.04)}]$ model for CH_3NO_2, where it is clear that there are a large number of electronic environments that fall outside the domain of the training data. This is found to



(a) Range

(b) Angular feature

Figure 7: Systematic improvability test. The numbers denote the maximum deviation from systematic improvement for sum of absolute error as compared to previous models. A value of 0 indicates that the model improves systematically since no system gets worse. a) each model is compared to all other models with same order of angular features and shorter range as indicated by the arrows (e.g. $NN[\bar{\lambda}^{(0)}_{(0.04)}]$ model is compared with $NN[\bar{\lambda}^{(0)}_{(0.02)}]$ and $NN[\bar{\lambda}^{(0)}_{(0.00)}]$) b) each model is compared to all other models with lower order of angular features and same range as indicated by the arrows (e.g. $NN[\bar{\lambda}^{(2)}_{(0.04)}]$ model is compared with $NN[\bar{\lambda}^{(1)}_{(0.04)}]$ and $NN[\bar{\lambda}^{(0)}_{(0.04)}]$). The results show that systematic improvement is often, but not always, observed when additional descriptors are added.

be consistent across the other extrapolation systems and in higher dimensions, as provided in the Supplementary Information. One straightforward solution to this issue is to add additional training systems that capture the chemistries of interest. This highlights the general limitation of machine-learning models that they are only as good as the data they are trained on. However, the problem can also be mitigated by increasing the dimensionality of the descriptor space. This is evident from the model performance, where the prediction error for these outlier systems is reduced to 0.08 eV, 0.19 eV and 0.12 eV, respectively, for the $NN[\bar{\lambda}^{(2)}_{(0.2)}]$ model. This occurs because as more information about the rotational and radial variations of the electron density is added these outlier systems become more similar to environments that exist in the training data. In these higher-dimensional spaces the new systems appear more like interpolations between existing environments as opposed to extrapolations beyond the domain of all training data. This phenomenon suggests that sufficiently large convolutional descriptor spaces, combined with diverse training data sets and comprehensive testing, may enable the construction of universal machine-learning XC functionals.

IV. CONCLUSIONS

This work introduces convolutional descriptors as a promising new paradigm for the construction of model spaces for XC functionals. Convolutional descriptors provide a systematically expandable and theoretically complete feature space for constructing XC functionals in a finite difference representation. They are orbital-free and can be computed with $N\log(N)$ computational complexity. Furthermore, convolutional descriptors can be combined with non-linear regression models to construct machine-learning functionals. Using neural networks is particularly promising, since the universal approximation theorem ensures that NNs can represent an arbitrarily complex function. The resulting models are conceptually similar to convNets, suggesting that deep learning approaches are a promising route for functional development. A sub-class of convolutional descriptors, Maxwell-Cartesian spherical harmonics (MCSH) descriptors, were employed to construct and test a range of machine-learned orbital-free approximations to the hybrid B3LYP functional based on data from a total of 25 small-molecule systems containing C, H, O, and N. These descriptors provide a numerically stable and rotationally invariant route to capturing rotational and radial variations in the electron density. The machine-learning models are constructed from model spaces based on the descriptors with increasing range from 0.02 Å - 0.2 Å and progressively finer angular features from zero-order to second-order. The results show that the average accuracy of the models improves as either the range or rotation symmetry is increased. A systematic improvement in the absolute error is typically observed for both training and test sets, but the improvement in system-level energy and formation energy are not systematic due to cancellation of error.

In addition to these promising initial results, this work also identifies several challenges that must be addressed in the construction of XC functionals based on convolutional descriptors and/or machine learning. One challenge that is general to any approach that utilizes localized XC energy density is the ability to generate train-

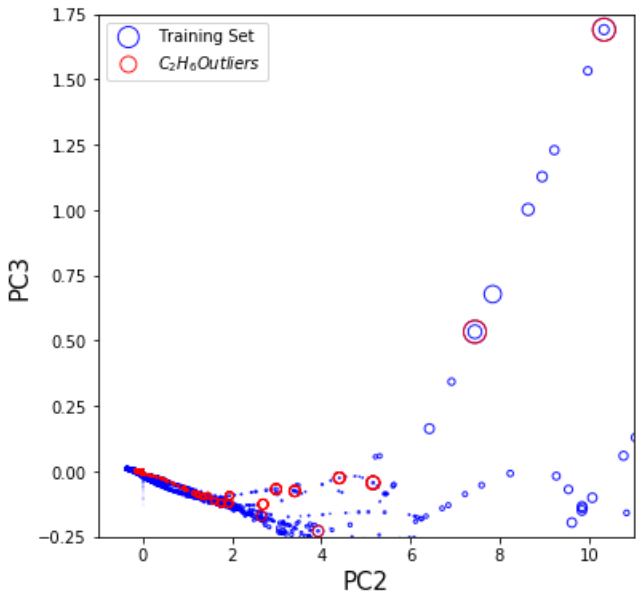


Figure 8: Comparison between the electronic environment of $C-C$ bonding region and that of the whole training set as characterized by the $\bar{\lambda}^{(1)}_{(0.04)}$ descriptor set. The training set is represented by uniformly sampled points plus 3,000,000 randomly sampled points. Principle component analysis (PCA) model is trained with $C-C$ bonding region data points and applied to both datasets. The plot of 2^{nd} and 3^{rd} principle components are shown here, where the red circles correspond to $C-C$ bonding environments and blue circles correspond to training data, the sizes of the circles correspond to the absolute prediction error of the $NN[\bar{\lambda}^{(1)}_{(0.04)}]$ model. The results show that the function is multi-valued in the $C-C$ bonding region, and the model is forced to make a tradeoff between general accuracy and accuracy in the $C-C$ region.

Model	Error	CH_3NO_2	*glycine*	$NCCN$	MAE
$NN[\bar{\lambda}^{(1)}_{0.04}]$	$\varepsilon_{abs.}$	2.01	0.95	0.94	0.4
	$\varepsilon_{pred.}$	1.38	0.28	-0.21	0.12
	$\varepsilon_{form.}$	1.27	0.19	-0.22	0.09
$NN[\bar{\lambda}^{(1)}_{0.08}]$	$\varepsilon_{abs.}$	1.24	0.76	0.49	0.15
	$\varepsilon_{pred.}$	0.18	0.11	0.24	0.04
	$\varepsilon_{form.}$	0.24	0.16	0.22	0.02
$NN[\bar{\lambda}^{(1)}_{0.2}]$	$\varepsilon_{abs.}$	0.24	0.55	0.13	0.06
	$\varepsilon_{pred.}$	0.08	0.37	0.03	0.01
	$\varepsilon_{form.}$	0.08	0.38	0.02	0.01
$NN[\bar{\lambda}^{(2)}_{0.04}]$	$\varepsilon_{abs.}$	3.66	0.56	0.49	0.18
	$\varepsilon_{pred.}$	-3.14	0.12	0.12	0.03
	$\varepsilon_{form.}$	-3.19	0.06	0.12	0.02
$NN[\bar{\lambda}^{(2)}_{0.08}]$	$\varepsilon_{abs.}$	1.42	0.61	0.14	0.09
	$\varepsilon_{pred.}$	1.08	0.3	0.02	0.01
	$\varepsilon_{form.}$	1.05	0.26	-0.02	0.01
$NN[\bar{\lambda}^{(2)}_{0.2}]$	$\varepsilon_{abs.}$	0.35	0.7	0.12	0.06
	$\varepsilon_{pred.}$	0.08	0.19	-0.01	0.01
	$\varepsilon_{form.}$	0.09	0.19	-0.01	0.01

Table III: Errors of the outlier systems in the test set. The unit is eV. MAE is the mean absolute error of the corresponding error metric for the 15 training molecular systems plus 7 test systems

ing data. Machine-learning approaches are most powerful when they are based on data from high-level methods for which no analytical form exists; however, these approaches are typically based on non-local integrals, so projecting the XC energy density to a finite difference grid is challenging. Approaches for this have been reported[88,105,106], but implementations are not openly available. Another related challenge is the fact that these high-level methods are typically all-electron, resulting in rapidly varying electron/energy density near the core regions. Accurately representing this with a finite difference grid requires very fine grid spacings (0.02 Å in this work). In this case although the theoretical scaling of convolutions is $N\log(N)$, the size of N is so large that the approach is much slower than the underlying B3LYP calculation. Similar concerns will be faced for other models seeking to reproduce the results of wavefunction-based theories, and routes to extract the XC contribution of valence electrons will be critical to training models that are consistent with pseudopotentials commonly used in practical DFT calculations. Moreover, integrating the approach with a full SCF cycle will require either direct learning of the XC potential, or an accurate approach to obtaining derivatives of the energy density. The latter has proven challenging for other machine-learning XC models[107], but recent advances suggest that directly learning forces[108] or applying automatic differentiation[109] are promising strategies. Finally, the optimization and numerical performance of machine-learning models must be considered. In this work NN models were used, leading to challenges in deconvoluting the error due to an insufficient model space and error due to sub-optimal hyperparameters or training procedures. The machine-learning models must also achieve a high accuracy over a large numerical range due to the large number of points with relatively low energy/electron density, and the quantities of interest (e.g. formation energy) rely on cancellation of error and may require specialized objective functions. This may present challenges in the application of out-of-the-box machine-learning models to the problem of XC functionals.

This work indicates that the combination of convolutional descriptors and machine learning models is a theoretically appealing framework for XC functional design. Despite practical challenges, the framework provides a route to empirically investigate fundamental questions

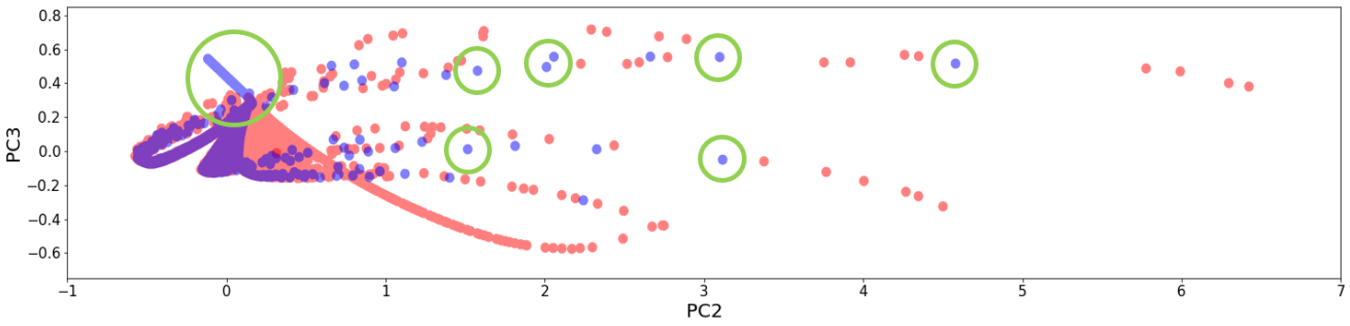


Figure 9: Comparison between the electronic environment of CH_3NO_2 system (blue) and that of the whole training set as characterized by the $\bar{\lambda}^{(1)}_{(0.04)}$ descriptor set (red). The CH_3NO_2 system is represented by the uniformly sampled points of the system, and the training set is represented by uniformly sampled points plus 3,000,000 randomly sampled points. Principle component analysis (PCA) model is trained with the training data points and applied to both datasets. The plot of 2^{nd} and 3^{rd} principle components are shown here. Green circles highlight regions where the CH_3NO_2 system is outside the domain of the training data. The plot shows that the electronic environments of the "extrapolation" systems are outside the domain of the training data.

about the nature of the XC energy. For example, this work provides empirical evidence that the exact exchange contribution of the B3LYP functional can be represented to within chemical accuracy (of system-level energies) for an orbital-free functional with a spatial range of <0.2 Å for small C, H, O, N molecules. Further examination of these numerical approximations may provide inspiration for new physical or empirical XC models with improved accuracy and practicality. In addition, there are many possible routes to improvement of the accuracy of these machine learning models. The choice of convolutional descriptors could be improved by inclusion of higher-order spherical harmonics, longer radial distances, decreasing grid-point space, integration with pseudopotentials, or the use of deep-learning convNets to automatically extract the optimal convolutional descriptors from the data. Training data can be extracted from high-level wavefunction theories, and convolutional models can be easily implemented in solid-state codes, providing a data-driven alternative to wavefunction embedding[110–113]. Moreover, the MCSH descriptor sets introduced in this study are not specific to electron density, but could be applied generally to any 3D functions that are inherently rotationally invariant. This includes many problems in physics since rotational and translational invariance are common. These exciting possibilities suggest that further research into convolutional-based machine-learning functionals is a worthwhile addition to the already numerous strategies for density functional design.

V. ACKNOWLEDGEMENTS

We acknowledge Daniel G. A. Smith and David Sherrill for assistance in extracting grid-resolved B3LYP XC densities, and Polo Chau and Fred Hohman for assistance in visualizing electron densities and descriptors. We also acknowledge Patrick F. Riley for his suggestions regarding spherical harmonics. This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences Computational Chemical Sciences program under Award Number DE-SC0019410. The authors are also grateful for a GPU generously provided by the NVIDIA GPU Grant program that was used to train the neural networks.

* ajm@gatech.edu

[1] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136:B864–B871, Nov 1964. doi: 10.1103/PhysRev.136.B864.

[2] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140:A1133–A1138, Nov 1965. doi: 10.1103/PhysRev.140.A1133.

[3] John P. Perdew. Jacob's ladder of density functional approximations for the exchange-correlation energy. In *AIP Conference Proceedings*. AIP, 2001. doi: 10.1063/1.1390175.

[4] David C. Langreth and M. J. Mehl. Easily implementable nonlocal exchange-correlation energy functional. *Phys. Rev. Lett.*, 47:446–450, Aug 1981. doi: 10.1103/PhysRevLett.47.446. URL https://link.aps.org/doi/10.1103/PhysRevLett.47.446.

[5] A. D. Becke. Correlation energy of an inhomogeneous electron gas: A coordinate space model. *The Journal of Chemical Physics*, 88(2):1053–1062, January 1988. ISSN 0021-9606.

[6] John P. Perdew and Wang Yue. Accurate and simple density functional for the electronic exchange energy: Generalized gradient approximation. *Physical Review B*, 33 (12):8800–8802, June 1986. ISSN 0163-1829.

[7] Perdew and Yue. Erratum: Accurate and simple density functional for the electronic exchange energy: Generalized gradient approximation. *Physical review. B, Condensed matter*, 40(5), August 1989. ISSN 0163-1829.

[8] Michael G. Medvedev, Ivan S. Bushmarinov, Jianwei Sun, John P. Perdew, and Konstantin A. Lyssenko. Density functional theory is straying from the path toward the exact functional. *Science*, 355(6320):49–52, jan 2017. doi: 10.1126/science.aah5975.

[9] A.D. Becke. Density-functional exchange-energy approximation with correct asymptotic behavior. *Physical Review A*, 38(6):3098–3100, 1988. ISSN 10502947.

[10] A.D. Becke. Density-functional thermochemistry. iii. the role of exact exchange. *The Journal of Chemical Physics*, 98(7):5648–5652, 1993. ISSN 00219606.

[11] C. Lee, W. Yang, and R.G. Parr. Development of the colle-salvetti correlation-energy formula into a functional of the electron density. *Physical Review B*, 37(2):785–789, 1988. ISSN 01631829.

[12] Jochen Heyd, Gustavo E. Scuseria, and Matthias Ernzerhof. Hybrid functionals based on a screened coulomb potential. *The Journal of Chemical Physics*, 118(18):8207–8215, May 2003. ISSN 0021-9606.

[13] Aliaksandr V. Krukau, Oleg A. Vydrov, Artur F. Izmaylov, and Gustavo E. Scuseria. Influence of the exchange screening parameter on the performance of screened hybrid functionals. *The Journal of Chemical Physics*, 125(22), December 2006. ISSN 0021-9606.

[14] Y Zhao and DG Truhlar. 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. *Journal Of Physical Chemistry A*, 110(49):13126–13130, December 2006. ISSN 1089-5639.

[15] Axel D. Becke. A real-space model of nondynamical correlation. *The Journal of Chemical Physics*, 119(6):2972–2977, August 2003. ISSN 0021-9606.

[16] R. Armiento and A. E. Mattsson. Functional designed to include surface effects in self-consistent density functional theory. *Physical Review B*, 72(8), aug 2005. doi: 10.1103/physrevb.72.085108.

[17] Piotr de Silva and Clémence Corminboeuf. Communication: A new class of non-empirical explicit density functionals on the third rung of jacob's ladder. *The Journal of Chemical Physics*, 143(11):111105, sep 2015. doi: 10.1063/1.4931628.

[18] O. Gunnarsson, M. Jonson, and B. I. Lundqvist. Exchange and correlation in inhomogeneous electron systems. *Solid State Communications*, 24:765–768, December 1977. doi:10.1016/0038-1098(77)91185-1.

[19] J. A. Alonso and L. A. Girifalco. Nonlocal approximation to the exchange potential and kinetic energy of an inhomogeneous electron gas. *Physical Review B*, 17(10):3735–3743, may 1978. doi:10.1103/physrevb.17.3735. URL https://doi.org/10.1103%2Fphysrevb.17.3735.

[20] O. Gunnarsson, M. Jonson, and B. I. Lundqvist. Descriptions of exchange and correlation effects in inhomogeneous electron systems. *Physical Review B*, 20(8):3136–3164, oct 1979. doi:10.1103/physrevb.20.3136. URL https://doi.org/10.1103%2Fphysrevb.20.3136.

[21] Yan Zhao, Benjamin J. Lynch, and Donald G. Truhlar. Multi-coefficient extrapolated density functional theory for thermochemistry and thermochemical kinetics. *Physical Chemistry Chemical Physics*, 7(1):43–52, December 2005. ISSN 1463-9076.

[22] Janos G. Angyan, Iann C. Gerber, Andreas Savin, and Julien Toulouse. van der waals forces in density functional theory: Perturbational long-range electron-interaction corrections. *Physical Review. A*, 72(1), 2005. ISSN 1050-2947.

[23] Stefan Grimme. Semiempirical hybrid density functional with perturbative second-order correlation. *The Journal of Chemical Physics*, 124(3), January 2006. ISSN 0021-9606.

[24] Qin Wu and Weitao Yang. Empirical correction to density functional theory for van der waals interactions. *The Journal of Chemical Physics*, 116(2):515–524, January 2002. ISSN 0021-9606.

[25] Stefan Grimme. Accurate description of van der waals complexes by density functional theory including empirical corrections. *Journal of Computational Chemistry*, 25 (12):1463–1473, September 2004. ISSN 0192-8651.

[26] Stefan Grimme. Semiempirical gga type density functional constructed with a long range dispersion correction. *Journal of Computational Chemistry*, 27(15):1787–1799, November 2006. ISSN 0192-8651.

[27] Y. Andersson, D.C. Andersson, and B.I. Lundqvist. Van der waals interactions in density-functional theory. *Physical Review Letters*, 76(1):102–105, 1996. ISSN 00319007.

[28] J.F. Dobson and B.P. Dinte. Constraint satisfaction in local and gradient susceptibility approximations: Application to a van der waals density functional. *Physical Review Letters*, 76(11):1780–1783, 1996. ISSN 0031-9007.

[29] M Dion, H Rydberg, E Schroder, DC Langreth, and Bi Lundqvist. Van der waals density functional for general geometries. *Physical Review Letters*, 92(24), June 2004. ISSN 0031-9007.

[30] T. Thonhauser, V.R. Cooper, S. Li, A. Puzder, P. Hyldgaard, and D.C. Langreth. Van der waals density functional: Self-consistent potential and the nature of the van der waals bond. *Physical Review B - Condensed Matter and Materials Physics*, 76(12), September 2007. ISSN 10980121.

[31] Andrew P. Witkin. Scale-space filtering. In *Readings in Computer Vision*, pages 329–332. Elsevier, 1987. doi: 10.1016/b978-0-08-051581-6.50036-2.

[32] Tony Lindeberg. *Scale-Space Theory in Computer Vision*. Springer, 1993. ISBN 9780792394181.

[33] Yann LeCun, Bernhard E. Boser, John S. Denker, Donnie Henderson, R. E. Howard, Wayne E. Hubbard, and Lawrence D. Jackel. Handwritten digit recognition with a back-propagation network. In D. S. Touretzky, editor, *Advances in Neural Information Processing Systems 2*, pages 396–404. Morgan-Kaufmann, 1990.

[34] Yann Lecun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324, 1998.

[35] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. ISSN 0001-0782. doi:10.1145/3065386.

[36] P.Y. Simard, D. Steinkraus, and J.C. Platt. Best practices for convolutional neural networks applied to visual

document analysis. volume 2003-, pages 958–963, USA, 2003. IEEE. ISBN 0769519601.

[37] Régis Vaillant, Christophe Monrocq, and Yann Le Cun. An original approach for the localization of objects in images, 1994.

[38] Steven J. Nowlan and John C. Platt. A convolutional neural network hand tracker. In *Advances in Neural Information Processing Systems 7*, pages 901–908. Morgan Kaufmann, 1995.

[39] S. Lawrence, C.L. Giles, Ah Chung Tsoi, and A.D. Back. Face recognition: a convolutional neural-network approach. *Neural Networks, IEEE Transactions on*, 8(1): 98–113, January 1997. ISSN 1045-9227.

[40] J.P. Perdew, K. Burke, and M. Ernzerhof. Generalized gradient approximation made simple. *Physical Review Letters*, 77(18):3865–3868, 1996. ISSN 0031-9007.

[41] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple [phys. rev. lett. 77, 3865 (1996)]. *Physical Review Letters*, 78(7): 1396–1396, February 1997. ISSN 0031-9007.

[42] Lucian A. Constantin, John P. Perdew, and J. M. Pitarke. Collapse of the electron gas to two dimensions in density functional theory. *Physical Review Letters*, 101(1), jul 2008. doi:10.1103/physrevlett.101.016406. URL https://doi.org/10.1103%2Fphysrevlett.101.016406.

[43] Lucian A. Constantin. Dimensional crossover of the exchange-correlation energy at the semilocal level. *Physical Review B*, 78(15), oct 2008. doi:10.1103/physrevb.78.155106. URL https://doi.org/10.1103%2Fphysrevb.78.155106.

[44] Jess Wellendorff, Keld T. Lundgaard, Andreas Møgelhøj, Vivien Petzold, David D. Landis, Jens K. Nørskov, Thomas Bligaard, and Karsten W. Jacobsen. Density functionals for surface science: Exchange-correlation model development with bayesian error estimation. *Phys. Rev. B*, 85:235149, Jun 2012. doi: 10.1103/PhysRevB.85.235149.

[45] Yan Zhao and Donald Truhlar. 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*, 120(1):215–241, May 2008. ISSN 1432-881X.

[46] Roberto Peverati and Donald G. Truhlar. Improving the accuracy of hybrid meta-gga density functionals by range separation. *The Journal of Physical Chemistry Letters*, 2 (21):2810–2817, November 2011. ISSN 1948-7185.

[47] Roberto Peverati and Donald G Truhlar. 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*, 8(7), July 2012. ISSN 1549-9618.

[48] Haoyu S. Yu, Xiao He, Shaohong L. Li, and Donald G. Truhlar. Mn15: A kohnsham global-hybrid exchangecorrelation density functional with broad accuracy for multireference and single-reference systems and noncovalent interactions. *Chemical Science*, 7(8):5032–5051, July 2016. ISSN 2041-6520.

[49] John C Snyder, Matthias Rupp, Katja Hansen, Klaus-Robert Müller, and Kieron Burke. Finding density functionals with machine learning. *Physical review letters*, 108 (25), June 2012. ISSN 1079-7114.

[50] Tran Doan Huan, Rohit Batra, James Chapman, Sridevi Krishnan, Lihua Chen, and Rampi Ramprasad. A universal strategy for the creation of machine learning-based atomistic force fields. *npj Computational Materials*, 3(1), 2017. ISSN 2057-3960.

[51] Kun Yao, John E Herr, DavidW Toth, Ryker Mckintyre, and John Parkhill. The tensormol-0.1 model chemistry: a neural network augmented with long-range physics. *Chemical science.*, 9(8):2261–2269, 2018. ISSN 2041-6520.

[52] Patrick Rowe, Gábor Csányi, Dario Alfè, and Angelos Michaelides. Development of a machine learning potential for graphene. *Physical review.*, 97(5), 2018. ISSN 2469-9950.

[53] Felix Brockherde, Leslie Vogt, Li Li, Mark E Tuckerman, Kieron Burke, and Klaus-Robert Müller. Bypassing the kohn-sham equations with machine learning. *Nature communications.*, 8(1), 2017. ISSN 2041-1723.

[54] Ganesh Hegde and R Chris Bowen. Machine-learned approximations to density functional theory hamiltonians. *Scientific reports.*, 7, 2017. ISSN 2045-2322.

[55] Burak Himmetoglu. Tree based machine learning framework for predicting ground state energies of molecules. *Journal of chemical physics.*, 145(13), 2016. ISSN 0021-9606.

[56] Matthias Rupp. Machine learning for quantum mechanics in a nutshell. *International journal of quantum chemistry.*, 115(16):1058–1073, 2015. ISSN 0020-7608.

[57] Katja Hansen, Franziska Biegler, Raghunathan Ramakrishnan, Wiktor Pronobis, O Anatole von Lilienfeld, Klaus-Robert Müller, and Alexandre Tkatchenko. Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. *The journal of physical chemistry letters.*, 6(12):2326–2331, 2015. ISSN 1948-7185.

[58] Kun Yao and John Parkhill. Kinetic energy of hydrocarbons as a function of electron density and convolutional neural networks. *Journal of chemical theory and computation : JCTC.*, 12(3):1139–1147, 2016. ISSN 1549-9618.

[59] Kyle Mills, Michael Spanner, and Isaac Tamblyn. Deep learning and the schrödinger equation. *Physical review.*, 96(4), 2017. ISSN 2469-9926.

[60] Florbela Pereira, Kaixia Xiao, Diogo A R S Latino, Chengcheng Wu, Qingyou Zhang, and Joao Aires-de Sousa. Machine learning methods to predict density functional theory b3lyp energies of homo and lumo orbitals. *Journal of chemical information and modeling.*, 57(1):11–21, 2017. ISSN 1549-9596.

[61] Albert P Bartók, Michael J Gillan, Frederick R Manby, and Gábor Csányi. Machine-learning approach for one- and two-body corrections to density functional theory: Applications to molecular and condensed water. *Physical review.*, 88(5), 2013. ISSN 1098-0121.

[62] Junji Seino, Ryo Kageyama, Mikito Fujinami, Yasuhiro Ikabata, and Hiromi Nakai. Semi-local machine-learned kinetic energy density functional with third-order gradients of electron density. *Journal of chemical physics.*, 148 (24), 2018. ISSN 0021-9606.

[63] Ting Gao, Hongzhi Li, Wenze Li, Lin Li, Chao Fang, Hui Li, LiHong Hu, Yinghua Lu, and Zhong-Min Su. A machine learning correction for dft non-covalent interactions based on the s22, s66 and x40 benchmark databases. *Journal of cheminformatics.*, 8(1), 2016. ISSN 1758-2946.

[64] Qin Liu, JingChun Wang, PengLi Du, LiHong Hu, Xiao Zheng, and GuanHua Chen. Improving the performance

of long-range-corrected exchange-correlation functional with an embedded neural network. *The journal of physical chemistry.*, 121(38):7273–7281, 2017. ISSN 1089-5639.

[65] David J. Tozer, Victoria E. Ingamells, and Nicholas C. Handy. Exchange-correlation potentials. *The Journal of Chemical Physics*, 105(20):9200–9213, nov 1996. doi: 10.1063/1.472753.

[66] Jess Wellendorff, Keld T Lundgaard, Karsten W Jacobsen, and Thomas Bligaard. mbeef: An accurate semi-local bayesian error estimation density functional. *Journal of chemical physics.*, 140(14), 2014. ISSN 0021-9606.

[67] Keld T. Lundgaard, Jess Wellendorff, Johannes Voss, Karsten W. Jacobsen, and Thomas Bligaard. mbeef-vdw: Robust fitting of error estimation density functionals. *Phys. Rev. B*, 93:235162, Jun 2016. doi: 10.1103/PhysRevB.93.235162.

[68] Manuel Aldegunde, James R Kermode, and Nicholas Zabaras. Development of an exchange–correlation functional with uncertainty quantification capabilities for density functional theory. *Journal of computational physics*, 311:173–195, 2016. ISSN 0021-9991.

[69] Vladimir Naumovich Vapnik. *The nature of statistical learning theory*. Springer, New York, 1995. ISBN 0387945598.

[70] Kevin P. Murphy. *Machine learning a probabilistic perspective*. Adaptive computation and machine learning. MIT Press, Cambridge, Mass., 2012. ISBN 9780262305242.

[71] Carl Edward Rasmussen. *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, Cambridge, Mass., 2006. ISBN 9786612097966.

[72] Warren McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biology*, 52(1):99–115, January 1990. ISSN 0092-8240.

[73] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989. ISSN 0893-6080.

[74] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

[75] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems (MCSS)*, 2(4):303–314, 1989. ISSN 0932-4194. doi:10.1007/BF02551274.

[76] Daniel E. Worrall, Stephan J. Garbin, Daniyar Turmukhambetov, and Gabriel J. Brostow. Harmonic networks: Deep translation and rotation equivariance. 2016.

[77] Jon Applequist. Maxwell–cartesian spherical harmonics in multipole potentials and atomic orbitals. *Theoretical Chemistry Accounts*, 107(2):103–115, 2002. ISSN 1432-881X.

[78] Robert M. Parrish, Lori A. Burns, Daniel G. A. Smith, Andrew C. Simmonett, A. Eugene DePrince, Edward G. Hohenstein, Uğur Bozkaya, Alexander Yu. Sokolov, Roberto Di Remigio, Ryan M. Richard, Jérôme F. Gonthier, Andrew M. James, Harley R. McAlexander, Ashutosh Kumar, Masaaki Saitow, Xiao Wang, Benjamin P. Pritchard, Prakash Verma, Henry F. Schaefer, Konrad Patkowski, Rollin A. King, Edward F. Valeev, Francesco A. Evangelista, Justin M. Turney, T. Daniel Crawford, and C. David Sherrill. Psi4 1.1: An open-source electronic structure program emphasizing automation, advanced libraries, and interoperability. *Journal of Chemical Theory and Computation*, 13(7):3185–3197, 2017. doi:10.1021/acs.jctc.7b00174. PMID: 28489372.

[79] Russell D. Johnson. NIST computational chemistry comparison and benchmark database, August 2011.

[80] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. URL http://www.scipy.org/. [Online; accessed Aug 03 2018].

[81] Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole von Lilienfeld. Big data meets quantum chemistry approximations: The δ-machine learning approach. *Journal of chemical theory and computation : JCTC.*, 11(5):2087–2096, 2015. ISSN 1549-9618.

[82] S. H Vosko, L. Wilk, and M. Nusair. Accurate spin dependent electron liquid correlation energies for local spin density calculations: a critical analysis. August 1980.

[83] Fuchang Gao and Lixing Han. Implementing the nelder-mead simplex algorithm withadaptive parameters. *Computational Optimization and Applications*, 51(1):259–277, January 2012. ISSN 0926-6003.

[84] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. December 2014.

[85] François Chollet et al. Keras. https://keras.io, 2015.

[86] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 807–814, USA, 2010. Omnipress. ISBN 978-1-60558-907-7.

[87] Brian Kolb, Levi C. Lentz, and Alexie M. Kolpak. Discovering charge density functionals and structure-property relationships with PROPhet: A general framework for coupling machine learning and first-principles methods. *Scientific Reports*, 7(1), apr 2017. doi:10.1038/s41598-017-01251-z.

[88] Kieron Burke, Federico G. Cruz, and Kin-Chung Lam. Unambiguous exchange-correlation energy density. *The Journal of Chemical Physics*, 109(19):8161–8167, nov 1998. doi:10.1063/1.477479.

[89] J. J. Mortensen, L. B. Hansen, and K. W. Jacobsen. Real-space grid implementation of the projector augmented wave method. *Phys. Rev. B*, 71(3):035109, JAN 2005. ISSN 1098-0121. doi:10.1103/PhysRevB.71.035109.

[90] Swarnava Ghosh and Phanish Suryanarayana. Sparc: Accurate and efficient finite-difference formulation and parallel implementation of density functional theory: Extended systems. *Computer Physics Communications*, 216: 109–125, 2017. ISSN 0010-4655.

[91] Swarnava Ghosh and Phanish Suryanarayana. Sparc: Accurate and efficient finite-difference formulation and parallel implementation of density functional theory: Isolated clusters. *Computer Physics Communications*, 212 (C):189–204, 2017. ISSN 0010-4655.

[92] John P. Perdew and Lucian A. Constantin. Laplacian-level density functionals for the kinetic energy density and exchange-correlation energy. *Physical Review B*, 75(15), apr 2007. doi:10.1103/physrevb.75.155109.

[93] Zidan Yan, John P. Perdew, Timo Korhonen, and Paul Ziesche. Numerical test of the sixth-order gradient expansion for the kinetic energy:application to the monovacancy in jellium. *Phys. Rev. A*, 55:4601–4604, Jun 1997. doi:10.1103/PhysRevA.55.4601. URL https://link.aps.org/doi/10.1103/PhysRevA.55.4601.

[94] Lucian A. Constantin, Aleksandrs Terentjevs, Fabio Della Sala, Pietro Cortona, and Eduardo Fabiano. Semiclassical atom theory applied to solid-state physics. *Phys. Rev. B*, 93:045126, Jan 2016. doi: 10.1103/PhysRevB.93.045126. URL https://link.aps.org/doi/10.1103/PhysRevB.93.045126.

[95] Christopher J. Cramer. *Essentials of computational chemistry : theories and models*. Wiley, Chichester, West Sussex, England ; Hoboken, NJ, 2nd ed.. edition, 2004. ISBN 0470091827.

[96] Fabio Della Sala, Eduardo Fabiano, and Lucian A. Constantin. Kinetic-energy-density dependent semilocal exchange-correlation functionals. *International Journal of Quantum Chemistry*, 116(22):1641–1694, 2016. doi:10.1002/qua.25224. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/qua.25224.

[97] Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation- and translation-equivariant neural networks for 3d point clouds. 2018.

[98] J.C. Maxwell. *A Treatise on Electricity and Magnetism*. Number v. 1 in A Treatise on Electricity and Magnetism. Clarendon Press, 1873.

[99] John P. Perdew, Jianmin Tao, Viktor N. Staroverov, and Gustavo E. Scuseria. Meta-generalized gradient approximation: Explanation of a realistic nonempirical density functional. *The Journal of Chemical Physics*, 120 (15):6898–6911, apr 2004. doi:10.1063/1.1665298. URL https://doi.org/10.1063%2F1.1665298.

[100] R. O. Jones and O. Gunnarsson. The density functional formalism, its applications and prospects. *Rev. Mod. Phys.*, 61:689–746, Jul 1989. doi: 10.1103/RevModPhys.61.689. URL https://link.aps.org/doi/10.1103/RevModPhys.61.689.

[101] Randolph Q. Hood, M. Y. Chou, A. J. Williamson, G. Rajagopal, and R. J. Needs. Exchange and correlation in silicon. *Physical Review B*, 57(15):8972–8982, apr 1998. doi:10.1103/physrevb.57.8972. URL https://doi.org/10.1103%2Fphysrevb.57.8972.

[102] Jianwei Sun, Bing Xiao, Yuan Fang, Robin Haunschild, Pan Hao, Adrienn Ruzsinszky, Gábor I. Csonka, Gustavo E. Scuseria, and John P. Perdew. Density functionals that recognize covalent, metallic, and weak bonds. *Physical Review Letters*, 111(10), sep 2013. doi: 10.1103/physrevlett.111.106401. URL https://doi.org/10.1103%2Fphysrevlett.111.106401.

[103] Mohan Chen, Hsin-Yu Ko, Richard C. Remsing, Marcos F. Calegari Andrade, Biswajit Santra, Zhaoru Sun, Annabella Selloni, Roberto Car, Michael L. Klein, John P. Perdew, and Xifan Wu. Ab initio theory and modeling of water. *Proceedings of the National Academy of Sciences*, 114(41):10846–10851, sep 2017. doi: 10.1073/pnas.1712499114. URL https://doi.org/10.1073%2Fpnas.1712499114.

[104] P.J. Haley and D. Soloway. Extrapolation limitations of multilayer feedforward neural networks. volume 4, pages 25–30. IEEE Publishing, 1992. ISBN 0780305590.

[105] Zhao, Morrison, and Parr. From electron densities to kohn-sham kinetic energies, orbital energies, exchange-correlation potentials, and exchange-correlation energies. *Physical review. A, Atomic, molecular, and optical physics*, 50(3):2138–2142, 1994. ISSN 1050-2947.

[106] Federico G. Cruz, Kin-Chung Lam, and Kieron Burke. Exchange-correlation energy density from virial theorem. *The Journal of Physical Chemistry A*, 102(25):4911–4917, 1998. doi:10.1021/jp980950v.

[107] John C Snyder, Matthias Rupp, Katja Hansen, Klaus-Robert Müller, and Kieron Burke. Finding density functionals with machine learning. *Physical review letters*, 108 (25), June 2012. ISSN 1079-7114.

[108] Stefan Chmiela, Alexandre Tkatchenko, Huziel E. Sauceda, Igor Poltavsky, Kristof T. Schütt, and Klaus-Robert Müller. Machine learning of accurate energy-conserving molecular force fields. *Science Advances*, 3(5): e1603015, may 2017. doi:10.1126/sciadv.1603015. URL https://doi.org/10.1126%2Fsciadv.1603015.

[109] Teresa Tamayo-Mendoza, Christoph Kreisbeck, Roland Lindh, and Alán Aspuru-Guzik. Automatic differentiation in quantum chemistry with applications to fully variational hartree–fock. *ACS Central Science*, 4(5):559–566, may 2018. doi:10.1021/acscentsci.7b00586. URL https://doi.org/10.1021%2Facscentsci.7b00586.

[110] Frederick R Manby, Martina Stella, Jason D Goodpaster, and Thomas F Miller. A simple, exact density-functional-theory embedding scheme. *Journal of chemical theory and computation*, 8(8), 2012. ISSN 1549-9626.

[111] Florian Libisch, Chen Huang, and Emily A. Carter. Embedded correlated wavefunction schemes: theory and applications.(report). *Accounts of Chemical Research*, 47(9): 2768–2775, 2014. ISSN 0001-4842.

[112] Chen Huang, Michele Pavone, and Emily A. Carter. Quantum mechanical embedding theory based on a unique embedding potential. *The Journal of Chemical Physics*, 134(15), 2011. ISSN 0021-9606.

[113] N. Govind, Y.A. Wang, A.J.R. Da Silva, and E.A. Carter. Accurate ab initio energetics of extended systems via explicit correlation embedded in a density functional environment. *Chemical Physics Letters*, 295(1):129–134, 1998. ISSN 0009-2614.