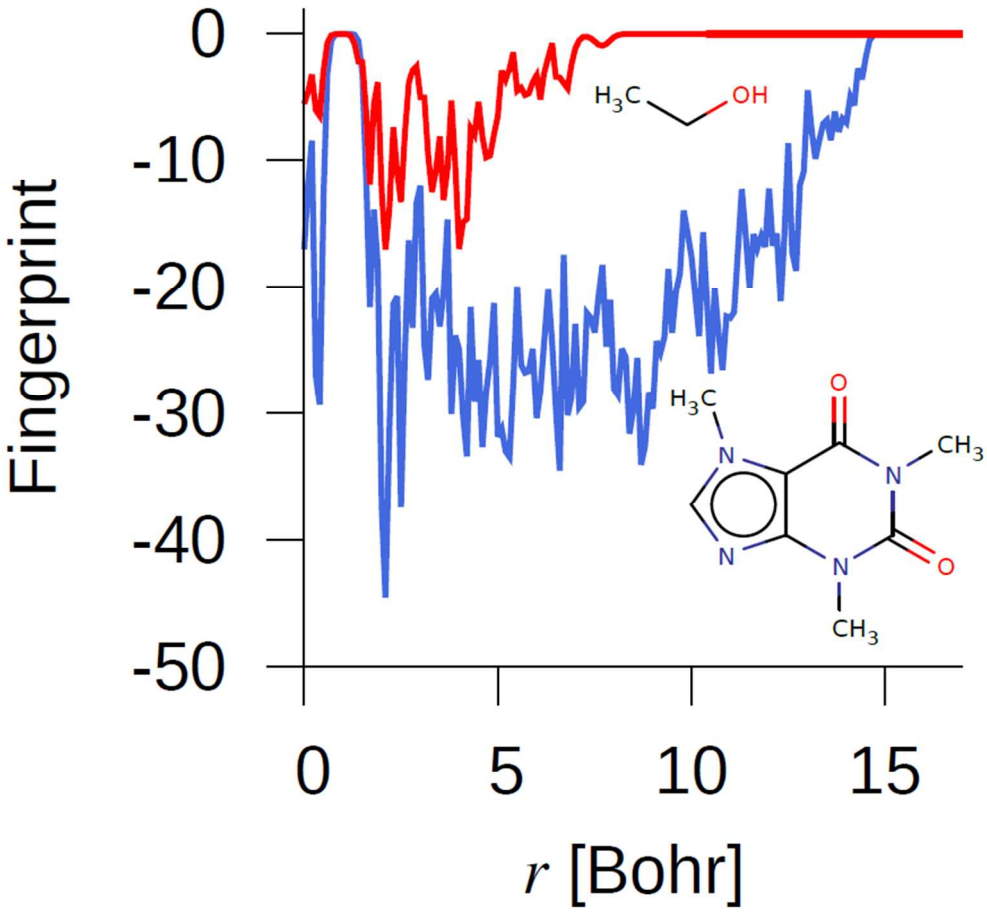
Fourier series of atomic radial distribution functions: A molecular fingerprint for machine learning models of quantum chemical properties

Journal:	*International Journal of Quantum Chemistry*
Manuscript ID:	QUA-2015-0055.R1
Wiley - Manuscript type:	Full Paper
Date Submitted by the Author:	21-Feb-2015
Complete List of Authors:	von Lilienfeld, Anatole; Argonne National Laboratory, Argonne Leadership Computing Facility Ramakrishnan, Raghunathan; University of Basel, Chemistry (Physical Chemistry) Rupp, Matthias; University of Basel, Institute of Physical Chemistry Knoll, Aaron; UT Austin, Texas Advanced Computing Center
Keywords:	machine learning, representation, descriptor, quantum chemistry, molecules
Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.	
main3.tex	



311x301mm (72 x 72 DPI)

The accuracy of machine learning models of quantum mechanical observables of molecules hinges on the quality of the molecular representation.
This article discusses necessary and desirable properties, as well as a Fourier series of atomic radial distribution function, potentially useful as unique molecular fingerprint.
For heats of atomization of over hundred thousand organic molecules this fingerprint is shown to reach density functional theory level of accuracy.

Fourier series of atomic radial distribution functions: A molecular fingerprint for machine learning models of quantum chemical properties

O. Anatole von Lilienfeld,[1,2,*] Raghunathan Ramakrishnan,[1] Matthias Rupp,[1] and Aaron Knoll[3,4]

[1] *Institute of Physical Chemistry and National Center for Computational Design and Discovery of Novel Materials, Department of Chemistry, University of Basel, Switzerland.*

[2] *Argonne Leadership Computing Facility, Argonne National Laboratory, 9700 S. Cass Avenue, Lemont, IL 60439, USA*

[3] *Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois 60439, USA*

[4] *Texas Advanced Computing Center, University of Texas Austin, Texas, USA*

We introduce a fingerprint representation of molecules based on a Fourier series of atomic radial distribution functions. This fingerprint is unique (except for chirality), continuous, and differentiable with respect to atomic coordinates and nuclear charges. It is invariant with respect to translation, rotation, and nuclear permutation, and requires no pre-conceived knowledge about chemical bonding, topology, or electronic orbitals. As such it meets many important criteria for a good molecular representation, suggesting its usefulness for machine learning models of molecular properties trained across chemical compound space. To assess the performance of this new descriptor we have trained machine learning models of molecular enthalpies of atomization for training sets with up to 10 k organic molecules, drawn at random from a published set of 134 k organic molecules with an average atomization enthalpy of over 1770 kcal/mol. We validate the descriptor on all remaining molecules of the 134 k set. For a training set of 10k molecules the fingerprint descriptor achieves a mean absolute error of 8.0 kcal/mol, respectively. This is slightly worse than the performance attained using the Coulomb matrix, another popular alternative, reaching 6.2 kcal/mol for the same training and test sets.

I. INTRODUCTION

For all but the most restricted problems and subsets of chemical compound space (CCS), screening, even when using high-throughput methods, becomes rapidly prohibitive due to the combinatorial explosion of possible arrangements of atom types and positions. The number of small stable organic molecules, for example, was estimated to exceed 10^{60}.[1] The formal dimensionality of CCS corresponds to $4N$ degrees of freedom, associated to the three Cartesian coordinates and the one nuclear charge of N atoms. This clearly illustrates the "curse of dimensionality" from which many first-principles inverse design efforts suffer[2]. Consequently, more compact representations of CCS are desirable, in particular if they can be more intuitively dealt with.

Recent machine learning (ML) efforts leverage modern data analysis methods for atomistic simulations. The basic idea is to develop algorithms that *infer* the solution of the electronic structure problem for a new material, rather than investing in the computational time to numerically solve it, with increasing accuracy as more training data are added[3]. Given sufficient data, these approaches are among the most promising avenues towards efficient exploration of CCS from first principles. A more profound and rigorous understanding of CCS would greatly help computational design and optimization of new materials with desirable properties. As such, efficient navigation of CCS is at the heart of all first principles based materials and bio design efforts.

To infer properties based on their correlations with compounds is akin to what Hammett accomplished in the 1930s through the exploitation of linear free energy relationships[4,5]. Such approaches have already delivered convincing results for highly relevant applications, such as enhanced sampling[6], screening of heterogeneous catalyst candidates based on Sabatier's principle[7], and, devising simple materials design rules leading to topological insulators, semi-conductors, and others[8].

With increasingly available simulation data stemming from routine applications of first principles methods, such as Born-Oppenheimer or Car-Parrinello molecular dynamics[9], statistical ML methods can be applied, in the hope of detecting trends and relationships that hitherto were difficult, if not impossible, to spot for the human expert. Applications of such approaches include data-mining for crystal structure discovery[10], regression for reorganization energies that enter Marcus charge transfer rates[11,12], learning of potential energy surfaces (PES)[13–17], learning of density functionals[18], and learning of the electronic Schrödinger equation of organic molecules[19]. The success of the latter, i.e., the success of predicting PESs across CCS without human bias yet in an accurate and reliable fashion, hinges on how well the input variables are represented for use by the ML algorithm. This representation, also known as "descriptor", encodes chemical identity in terms of chemical composition and atomic configuration. As such, descriptors are a crucial ingredient for the development of predictive ML models of PESs across CCS.

Conventionally, descriptors encode some prior knowledge about electronic structure effects. A frequently made assumption is that number and order of covalent bonds in compounds are known *a priori* and fixed. For example, Faulon's Signature descriptor encodes the graph of covalent bonding in a molecule[20]. Among others, this descriptor was applied to inverse QSAR[21], prediction of protein interactions[22], and to predict reorga-

nization energies of poly-aromatic hydrocarbons[12]. The underlying assumption severely limits the realm of applications when it comes to modeling processes where the bonding is not known *a priori*. Examples for which this assumption is not valid include basically all processes commonly referred to as "chemical change", i.e., bond breaking and formation, metal ligand exchanges, diffusion of defects in solids, proton hopping in aqueous solvents (Grotthuss-mechanism) or even simple tautomeric equilibria. The assumption also breaks down for interactions less localized than covalent binding, such as supramolecular van der Waals complexes, bound due to hydrogen bonds or (many-body) London dispersion forces[23]. The latter are particularly crucial for biological function, as recently illustrated for the selective self-assembly of hydrogen-bonded nano-structures[24]. For a comprehensive overview and comparative analysis of over 600 different descriptors see the 2005 study carried out by Meringer and coworkers[25]. The limitation due to such inherent assumptions might possibly explain the current state of affairs in QSAR-based drug discovery efforts[26]. It is therefore desirable to devise more general "first principles-like" descriptors that conserve the rigor of *ab initio* methods[27], such as wave function[28], density functional[29], or quantum Monte Carlo methods[30], and that consequently can also, at least in principle, account for *any* chemical scenario or compound[1,31].

In this study we introduce a molecular descriptor that uniquely (except for chirality) represents any molecule as a fingerprint, here a univariate function in terms of geometric distance. Within the Born-Oppenheimer view on the CCS of molecules, *any* molecular geometry is uniquely characterized within its $4N$-6 degrees of freedom, subtracting three rotational (two if linear) and three translational degrees of freedom. Our descriptor is unique, differentiable, and invariant with respect to rotation, translation, and indexing of atoms. In full analogy to the information entering the electronic Schrödinger equation, this descriptor requires *only* atomic coordinates and nuclear identities. Thus, any composition and geometry is accounted for in a way amenable to ML. The descriptor might even be suitable for modeling of nuclear quantum effects through *ab initio* path-integral molecular dynamics[32], relevant for instance in the case of of Watson-Crick tautomers[33], if energies and forces for all the replicas can be learned with sufficient predictive accuracy.

This paper is structured as follows. In the methods section, we first outline the conceptual framework and discuss desirable properties of descriptors. Then, starting with the external potential, we proceed with a step by step discussion of translational, rotational and atom-indexing invariances, as well as uniqueness requirements, which have guided us to the specific form of our descriptor. In the results section, the descriptor's performance is assessed and compared to the Coulomb matrix, another popular descriptor, using heats of atomization of up to 134 k organic molecules taken from Ref. 34.

II. METHOD

A. Descriptor properties

The defining purpose of a descriptor D is to represent a compound, defined through input variables, in a form that can be correlated to a property of interest $\mathcal{P}$, i.e., its form should be amenable to statistical learning. More specifically, D should rigorously *and* in a convenient fashion represent the variables that occur in the equation being modeled via ML.

Many descriptors and classification schemes for them have been proposed. For the purpose of modeling results derived from Schrödinger's equation, one could consider the following three cases

First principles: Descriptors that encode the relevant information in the quantum Hamiltonian without loss of information. As such, they should be applicable to the learning of any quantum observable, such as energies, forces, or electronic properties. Examples include the sorted Coulomb-matrix[19], Gaussian shapes[35], bispectrum, power spectrum, or angular distribution functions[17,36]. The challenge consists of removing redundancies and encoding invariances, i.e., to render them maximally compact without losing information. Note however, that some observables might require more degrees of freedom than others. In ML models of atomization energies, for example, the chirality of the molecule is not relevant. For ML models of the optical activity in circular dichroism, however, it is.

Coarsened: Descriptors that reflect important structural features typically work for a range of properties but not for all of them. Examples include the number of hydrogen-bond donors or acceptors (used in Lipinski's rule of five[37]), number of aromatic units, the diagonalized Coulomb matrix[38,39], the bag-of-bond descriptor[40], or the signature descriptors[20]. Such descriptors are not bijective, i.e., they do not allow reconstruction of the compound in general; in practice, some allow reconstruction given enough constraints. The challenge consists of finding a form for which the loss of information is minimal while maintaining the advantages of coarsening.

Integrated: Descriptors that explicitly encode integrated properties correlating well with the property of interest. Examples are adsorption energies for catalytic activity[7], logP octanol/water partition coefficients or Lipinski's rule of 5[1] for oral bio-availability, electrophilic superdelocalizability for pK_a prediction[41], HOMO eigenvalues[42,43] and other simple property descriptors commonly used in high-throughput screening[8]. The challenge consists of gauging their transferability to other compound classes, properties, or even environmental

conditions.

In this study, we restrict ourselves to **First principles** kind of descriptors which can be used for the construction of ML models of quantum mechanical observables. For **Coarsened** or **Integrated** descriptors the reader is referred to the above cited literature.

Uniqueness

We believe uniqueness, up to invariants that leave the modeled observable $\mathcal{O}$ unchanged, to be crucial. In other words, we consider a descriptor to be unique if there is no pair of molecules that produces the *same* descriptor. Here, we do not refer to the reverse case, namely that any given *single* molecule can have more than one descriptor. For example, our criterion of uniqueness is still met by any descriptor consisting simply of the set of atomic nuclei and associated Cartesian coordinates due to the mapping between molecular Hamiltonian and unperturbed wave-function Ψ of the ground-state: While no pair of molecules exists with the exact same sets of atomic nuclei and coordinates, there are many different sets of coordinates which merely differ by molecular symmetry operations (translation, rotation, or complete nuclear permutation)[44]. Removal (or reduction) of such invariant degrees of freedom is relevant for the efficiency of the machine learning model but less crucial on a conceptual level.

The reason for the uniqueness requirement can be shown by *reductio ad absurdum* in three steps—in analogy to the first Hohenberg-Kohn theorem[45]—for any quantum mechanical observable $\mathcal{O} = \langle\Psi|\hat{O}|\Psi\rangle$. Here the unperturbed ground-state Hamiltonian H is defined by its external potential, determined by $\{Z_I, \mathbf{R}_I\}$, the set of nuclear charges and coordinates, as well as number of electrons N_e. The variational principle yields the system's many-body wavefunction Ψ for any given H.

(i) Let D denote a descriptor that is *not* unique. Then two systems $H_1 \neq H_2$ exist that differ in excess of the invariants, but they are mapped to the same descriptor value d, $H_1 \to d$ and $H_2 \to d$.

(ii) Because H_1 and H_2 differ by more than their property's invariances, they have different wavefunctions, $\Psi_1 \neq \Psi_2$, yielding two different observables, $\mathcal{O}_1 = \langle\Psi_1|\hat{O}|\Psi_1\rangle$ and $\mathcal{O}_2 = \langle\Psi_2|\hat{O}|\Psi_2\rangle$. Here, we deliberately ingore the obvious exception and special situation of all observables which happen to be degenerate.

(iii) A trained statistical model predicts any observable $\mathcal{O}$ solely based on descriptor input d leading to identical predictions $\mathcal{O}_1^{\text{pred}} = \mathcal{O}_2^{\text{pred}}$. In the limit of infinite training data, these predictions will be exact, implying $\mathcal{O}_1 = \mathcal{O}_2$, in contradiction to (ii).

Consequently, non-unique descriptors can yield absurd results for *any* observable. In other words, artificial degeneracies in the descriptor imply prediction errors that can not be remedied through addition of more training data. As such, non-unique descriptors defy the very idea of using ML in quantum mechanics.

Uniqueness up to invariances is necessary, but not sufficient for the design of a good descriptor. Consider the case of the invariant degrees of freedom, for which a manifold of unique descriptors could be constructed: Descriptors which depend on rotations, translations, and nuclear permutations could in principle be used, the obvious example being the $4N$-dimensional vector with four entries corresponding to nuclear charge and three Cartesian coordinates, $Zxyz$. For example, translational invariance could be imposed by including shifted copies of $Zxyz$ vectors in the training set. While representations of internal degrees of freedom, such as atom-atom distance matrices or the Z-matrix, popular in quantum chemistry communities, encode rotational and translational invariances, they still suffer from lack of nuclear permutational invariance. It is possible to obtain invariance with respect to nuclear permutation by simply representing each molecule not by one but rather by a set of $Zxyz$ vectors, each vector containing the same elements but in different order (see Ref. 46 for a successful application of this idea). However, in general such descriptors lead to substantial overhead for the statistical learning, since in order to obtain a transferable ML model, the training set would have to be constructed (and extended) to explicitly reflect all these invariances. Furthermore, the model's transferability would also be inherently limited to those ranges of the redundant degrees of freedom that have been covered in training. Also, when it comes to measuring similarity between descriptors of two molecules (the ultimate feature of any ML model) absence of translational, rotational, and nuclear permutation invariance can aggravate alignment problems, with multiple minima, and numerically difficult and challenging optimization problems, as recently reviewed by Zadeh and Ayers[47]. Another reason for aiming to remove all invariances can be given by analogy to the definition of the property. In the case of the electronic Schrödinger equation, position and orientation of the external potential in the Hamiltonian are arbitrary, and the external potential, a sum over all nuclei, is permutationally invariant. Since the descriptor is meant to represent the independent variables in the Hamiltonian, it suggests itself that it be invariant with respect to all the redundant degrees of freedom. Finally, absence of invariance with respect to nuclear permutations might also become cumbersome for modeling the energy when it comes to simulation regimes in which Heisenberg's uncertainty principle applies to atoms, such as collisions at high temperature, and when atoms of the same type and weight can become indistinguishable. We conclude that a lack of invariances can present severe challenges in practice, it appears therefore desirable to map all invariant structures to the same descriptor, i.e., for the descriptor

to obey all invariances. The challenge consists of removing as many of these redundant degrees of freedom as possible, *without* losing uniqueness, i.e. without turning the **First principles** descriptor into a **Coarsened** descriptor. Below we will encounter an example for a coarsened descriptor (FD) that meets all invariances but has lost uniqueness [see Eq. (6) and Fig. (2)].

We reiterate that the uniqueness vs. invariance issue is strongly dependent on property. For example, atomization energies of stereo-isomers, calculated within non-relativistic Born-Oppenheimer-approximated time-independent electronic structure theory, do not violate parity and are therefore also invariant with respect to choice of enantiomers[48]. If the goal was to also account for parity-violating effects in the potential energy, the descriptor would have to be enantio-selective.

Desirable properties

The descriptor's size-extensive and symmetry behavior is also relevant. In analogy to the external potential in Schrödinger's equation, atoms or groups that are symmetric should contribute in equal ways to the descriptor; and changes in system size should lead to corresponding changes in the descriptor (e.g., its size or range of component values). Another important feature of a descriptor is its completeness, or global nature, meaning that it encodes the whole of a compound, as opposed to only a local part of it. Local descriptors in terms of expansions over atomic contributions were successfully used with neural networks[49] and kernel ridge regression[17,36] to learn potential energy hypersurfaces and forces across configuration space, and, for enhanced sampling using molecular dynamics. While local descriptions form the basis for linear-scaling electronic structure software, and might be appropriate for some properties (e.g., properties of atoms in molecules, such as nuclear magnetic resonance chemical shifts or atomic forces) and systems (e.g., insulators and semi-conductors where the electronic "nearsightedness principle"[50] can be exploited), this can not be assumed in general, and might be limiting for energies of long-range electron/phonon coupling, electron transfer, or metals.

Other desirable features of descriptors include a closed and analytic form for analysis and rapid evaluation, differentiability (with respect to nuclear charges and coordinates) to account for response properties and use of advanced learning techniques, uniform length for finite sets of compounds to conveniently compare molecules that differ strongly in size (number of atoms), and, a functional form that can cope well with all the various ranges relevant to physical chemistry, i.e. nuclear charges ranging from 0 to ∼100, and interatomic distances ranging from tenths to dozens of Å (or even a thousand Å in the case of more exotic molecules[51]).

An overview of crucial and desirable properties is given in Table I for various relevant descriptors, including Signature (σ) [20], nuclear charges and Cartesian coordinates ($Zxyz$), Coulomb matrix (CM), diagonalized CM matrix (Eig(CM))[19], and the Fourier series of atomic radial distribution functions (FR), introduced here. CM, recently introduced[19], satisfies many of the aforementioned requirements, but not all. In particular, it lacks invariance with respect to nuclear permutation (fixed in practice by sorting atom indices with respect to the norm of it's rows or columns), and its dimensionality scales quadratically with number of atoms. We note that the dimensionality listed in Table I, however, is a rather formal construct: The $(N^2-N)/2$ entries in the Coulomb matrix are not independent variables. This statement is clearly also true for the m grid-points representing "dimensions" of the FR descriptor.

TABLE I. Properties of various descriptors including Signature (σ) [20], nuclear charges and Cartesian coordinates ($Zxyz$), Coulomb matrix (CM), diagonalized CM matrix (Eig(CM))[19], and the radial distribution Fourier series descriptor (FR) introduced here. N denotes number of atoms. The upper part contains requirements used for the design of FR. ✓ and ¬ indicate whether a requirement is fulfilled or not.

Property	σ	$Zxyz$	CM	Eig(CM)	FR
Unique	✓	✓	✓	¬	✓
First principles	¬	✓	✓	✓	✓
Transl. invariant	✓	¬	✓	✓	✓
Rotat. invariant	✓	¬	✓	✓	✓
Index. invariant	¬[a]	¬	¬	✓	✓
Differentiable	¬.	N.A.	✓	✓	✓
Symmetry	✓	¬	✓	✓	✓
Size extensive	✓	✓	✓	✓	✓
Complete/global	¬ [a]	✓	✓	¬	✓
Dimensionality	N.A.	$4N$	$(N^2-N)/2$	N	m [b]
Analytical	✓	✓	✓	✓	✓
Uniform length	¬	¬	¬	¬	✓ [c]
Variable ranges	✓	✓	✓	✓	✓

[a] unless taken to full height $h = N$.
[b] $m \geq (N^2-N)/2$ being the number of grid elements required for discretizing the largest interatomic distance
[c] If damped by a Gaussian

Another aspect is the smoothness of the property as a function of the descriptor. Smoothness is a prerequisite for machine learning (to enable meaningful selection from the infinitely many models that are compatible with the training data), related to regularization. However, the function's smoothness might vary along different directions in descriptor space (as an example, consider ligand binding, where steric constraints of the host might cause abrupt changes in affinity upon certain geometrical changes of the ligand, as opposed to more gradual changes not conflicting with the host's geometry, e.g., "magic methyls", or, more generally, "activity cliffs"[52]). Reducing the models smoothness requires more training data than necessary in smooth data regions, whereas in-

creasing the models smoothness reduces prediction accuracy in data regions with more rapid changes. A potential solution might be models with local smoothness [53].

In the following we discuss the sequence of steps that has led us to the FR descriptor which meets all the required and desired features listed in Table I, i.e. (i) first principles and nuclear permutation invariance, (ii) translational invariance, (iii) rotational invariance and mirror symmetries (Euclidean symmetries), (iv) uniqueness, and (v) differentiability.

B. First principles *Ansatz*: The external potential

The first Hohenberg-Kohn theorem shows that the electron density $n(\mathbf{r})$ of a given system, as determined by its external potential $v(\mathbf{r})$ through application of the variational principle, is as unique as its electronic wavefunction $\Psi(\mathbf{r})$ obtained through solution of Schrödinger's equation[45]. After application of the variational principle (yielding the electron density that minimizes the energy), the total potential energy is commonly given as an integral containing electron density and external potential,

$$E[n(\mathbf{r})] = F_{ee}[n(\mathbf{r})] - \int d\mathbf{r}\; n(\mathbf{r})v(\mathbf{r}) + \frac{1}{2}\sum_{IJ}\frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|}, \tag{1}$$

with F_{ee} the universal functional encoding all contributions to energy coming from electron-electron interactions, the second term representing the Coulomb attraction between nuclear charges and electrons, and the last term corresponding to the nuclear Coulomb repulsion between all atoms. The total potential energy of *any* molecule is therefore determined, independent of translations, rotations, or nuclear permutations, determined by its unique electron density.

The electron density can be viewed as a "quantum" molecular descriptor, used to predict molecular energies through the map $n(\mathbf{r}) \mapsto E$. Already three decades ago, Carbó et al. proposed to use the overlap integral of electron densities of different molecules to quantify molecular similarity.[54] In fact, the electron density is already used as a descriptor in practice when density functionals are trained empirically to reproduce the energies of a training set. If the electron density were not unique, density functional theory as we know it would not exist. The Hohenberg-Kohn theorem in this sense underscores the importance of the descriptor's uniqueness when it comes to the training of potential energy surface models.

The external potential, conversely, is in a unique relationship with atomic Cartesian coordinates $\{\mathbf{R}_I\}$ and nuclear charges $\{Z_I\}$, $v(\mathbf{r}) = \sum_I Z_I/|\mathbf{r} - \mathbf{R}_I|$. Due to its translational and rotational dependence, however, the external potential itself does not qualify as a promising descriptor. In a first step we replace the system's representation in form of its external potential by a model of nuclear charge densities, namely a sum of Gaussians located at atomic coordinates with atom type-specific heights Z_I,

$$P(\mathbf{r}) = \sum_I Z_I e^{-a|\mathbf{r}-\mathbf{R}_I|^2}, \tag{2}$$

where the sum runs over all N atoms in the molecule, and a is a global parameter for all atoms and all molecules, for now simply fixing the nuclear width for *all* atoms independent of type. Note that a could be defined in an atom-type specific way, and that P does no longer integrate to N_p, the total number of protons present in the system, except for infinitely small width of the Gaussians. Similar to Gaussian type orbitals as a basis for molecular orbitals, we thereby deliberately forego any physical non-Gaussian like features in favor of computational convenience. Note, however, that $P(\mathbf{r})$ is still in a one-to-one relationship with the external potential, and that it is still atom index invariant.

C. 3-D Fourier transform

An appealing characteristic of using plane wave basis sets in electronic structure calculations is their invariance with respect to atomic position (translational invariance). In contrast to atomic basis sets, Pulay forces and basis set superposition errors, i.e., additional force terms due to basis set incompleteness, can be avoided, which makes the implementation of geometry optimization or molecular dynamics methods more straight-forward. Analogously, we can remove translational degrees of freedom of a charge distribution by changing to the Fourier frequency domain[55]. The Fourier transform of the Gaussian charge distribution,

$$\mathcal{F}(P) = \frac{1}{(2a)^{3/2}} e^{-\frac{\boldsymbol{\omega}^2}{4a}} \sum_I Z_I e^{i\boldsymbol{\omega}^T \mathbf{R}_I}, \tag{3}$$

can be multiplied with its conjugate to yield a real function in the three dimensions of the Fourier domain $\boldsymbol{\omega}$,

$$F(\boldsymbol{\omega}) = \frac{1}{(2a)^3} e^{-\frac{\boldsymbol{\omega}^2}{2a}} \sum_{J,I} Z_I Z_J \cos[\boldsymbol{\omega}^T(\mathbf{R}_I - \mathbf{R}_J)], \tag{4}$$

after simplification using Euler's formula. Eq. (4) is a translation-invariant representation of the nuclear charge distribution in Eq. (2). $F(\boldsymbol{\omega})$ can be viewed as a sum over all elements of a symmetric atom-atom pairwise matrix $\mathbf{M}$ with elements

$$M_{IJ} = Z_I Z_J \cos[\boldsymbol{\omega}^T(\mathbf{R}_I - \mathbf{R}_J)]. \tag{5}$$

This matrix is reminiscent of the Coulomb matrix[19]. At $\omega = 0$ and for $a = (1/4)^{1/3}$, its diagonal elements become identical to those of a preliminary version of the Coulomb matrix, $0.5\ Z_I^2$, the potential energy of the hydrogenic atom. While Eq. (4) has appealing features, it still lacks rotational invariance. Furthermore, preliminary tests with machine learning models of atomization energies based on this descriptor, after alignment

of all molecule pairs in the Fourier domain, resulted in rather disappointing predictive accuracy for out-of-sample molecules.

D. 1-D version

To remove rotational dependence, we project the Fourier transform (Eq. 4) onto one dimension by replacing the argument of the cosine function by the scalar product of a frequency and the interatomic distance:

$$FD(r) = \frac{1}{(2a)^3} e^{-\frac{\omega^2}{2a}} \sum_{J,I} Z_I Z_J \cos[\omega \times r_{IJ}], \quad (6)$$

where $r_{IJ} = |\mathbf{R}_I - \mathbf{R}_J|$ is interatomic distance. Eq. 4 is a double sum over atoms that maintains invariance with respect to nuclear permutation, translation, and rotation. Fig. 1 illustrates the meaning of this "hack" for various interatomic distances of H_2, LiH, and HF. Changes in interatomic distances induce changes in oscillatory frequency, while changes in elemental composition affect overall amplitude. Differences in atomic numbers within any of these three diatomics are expressed through the width of the oscillatory band, i.e., the larger the difference, the narrower the band. The Gaussian prefactor dampens the descriptor towards zero for large frequencies.

E. Uniqueness

The 1-D Fourier fingerprint (Eq. (6)) is invariant with respect to translations, rotations, and nuclear permutations. However, it is no longer unique due to the information lost in modifying the argument of the cosine. This can easily be seen for the task of distinguishing homometric molecules, i.e., molecules with identical sets of interatomic distances.[57] Note that while in Ref. 39 it is mentioned that all enantiomers are homometric, there exist also homometric compounds that are not enantiomers. While potentially of interest for the ML modeling of parity violation[48], for the electronic Schrödinger equation within the Born-Oppenheimer approximation any mirror symmetries (leading to enantiomers) represent only redundant degrees of freedom, which need not be distinguished by the descriptor. However, all pairs of homometric molecules that are not enantiomers should be distinguished by the descriptor. An example of such a compound pair, proposed in Ref. 56, is on display in Fig. 2. Note that any two homometric compounds would have exactly the same potential energy if modeled by an exclusively pair-wise interatomic potential, no matter how well parametrized to effectively account for many-body effects. As such homometric compound pairs exemplify the importance of many-body effects in interatomic potentials, effects recently shown to be sizeable not only for covalent bonding, but also for intermolecular van der Waals forces[23,58].

Homometric compounds do not differ by the number of "bonds" (interatomic distances), but rather by their distribution: In the rectangular compound all four atoms have a short, medium, and long bond, (s, m, l). In the triangular compound, two atoms (lower corners of the triangle) have the same distribution of bonds (s, m, l), but the lower middle and upper atoms have different distributions (s, m, s) and (l, m, l). In Ref. 59 it has been shown that any simplex can be represented without loss of information, i.e., uniquely, using such distributions of distances between vertices.

A continuous version of such a distribution of interatomic distances can be obtained by replacing the scalar $\omega \times r_{IJ}$ argument in the cosine in the Fourier series by an atomic radial distribution function RDF_I for each atom I,

$$FDR(r) = \frac{1}{(2a)^3} e^{-\frac{r^2}{2a}} \sum_{I} Z_I^2 \cos[RDF_I(r)]. \quad (7)$$

In other words, the 1-D frequency domain ω has been turned into a 1-D real space interatomic distance domain. Any functional form of atomic radial distribution functions, numerical or analytical such as "softened" Coulomb potentials $\sum_J Z_J/(|r - r_{IJ}| + 1)$, or Slater (or Laplace) functions $\sum_J e^{-\alpha|r-r_{IJ}|}$ can be used. Other smoothening functions, such as Gaussian radial distributions, were already proposed as descriptors in the past (See Refs.[17,36,60–62]). To the best of our knowledge, however, they were not used as arguments in Fourier series expansions.

We have used ML models to test several variants of radial distribution functions. The Gaussian radial distribution function, $\sum_J Z_J^n e^{-b(r-r_{IJ})^2}/Z_I^m$, included in the Fourier series in the following form,

$$FR(r) = \sum_I Z_I^l \left(\cos \left[Z_I^m \sum_J Z_J^n e^{-b(r-r_{IJ})^2} \right] - 1 \right) \quad (8)$$

has resulted in the best performance. Here, we have omitted the Gaussian prefactor from Eq. (7) to keep the descriptor complete, notwithstanding that it could still be used to obtain finite length, or to localize the descriptor. Parameters $l, m, n, b \in \mathbb{R}$ are global hyperparameters which we optimize via cross-validation when training the machine learning model. Further flexibility could still be introduced by making these parameters atom type Z_I-dependent. In this study, however, we have not investigated these degrees of freedom.

FR is a fingerprint as a function of interatomic distance. It decays to zero for all interatomic distances larger than the molecule. The linear independence of the atomic terms in the Fourier summation, measurable by the Wronskian, guarantees that no atoms' RDF's linearly add (or cancel) each other—unless they *all* have exactly the same radial distribution. As such, the Fourier series introduces the linear independence of the radial distribution around each atom I. Only if all atoms in two

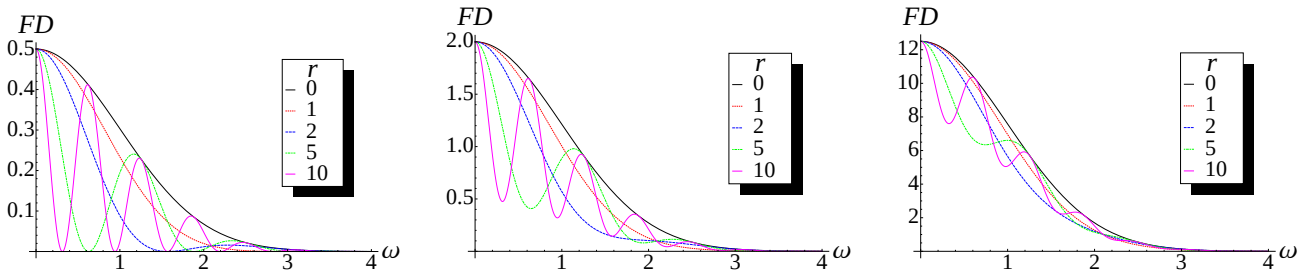


FIG. 1. Illustration of Fourier descriptor FD (units of charge squared, Eq. (6)) for three diatomics, H_2 (left), LiH (center), and HF (right) for five interatomic distances r. The hyperparameter a is set to 1. Note that for $r = 0$, FD corresponds to $(Z_1^2 + 2Z_1Z_2 + Z_2^2)/8$.

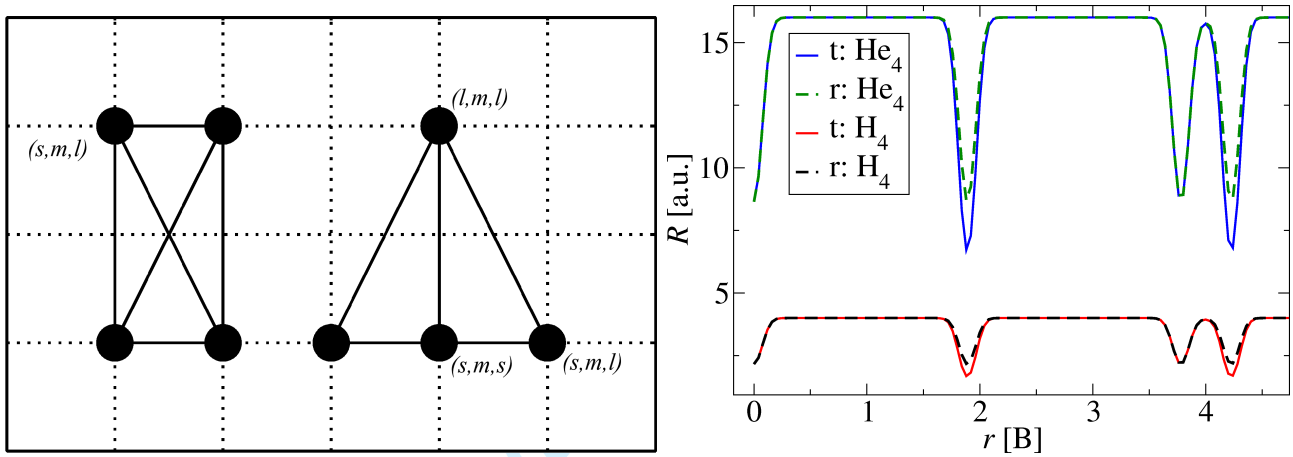


FIG. 2. LEFT: Sketch of two homometric molecules (same atom types, same sum of interatomic distances) from Ref. 56. The atomic distribution of distances (s short, m medium, l long) are indicated. The sorted Coulomb matrix can distinguish these two molecules[38,39,57]. s and l are set to 1 and 2 Å, respectively. RIGHT: Illustration of Fourier series with Gaussian radial distribution function based descriptor R (according to Eq. (7)) for the homometric, rectangular (r) and triangular (t), geometries displayed in Fig. 2.

molecules have the same RDF will the two molecules yield the same FR and therefore represent the same point on the potential energy surface. Note that in Ref. 36 an angular Fourier series-based descriptor that sums over individual angles (as opposed to distributions of angles) has been investigated. This descriptor, however, has been introduced in the context of modeling the potential energy surface of a single compound, not for training across CCS.

The uniqueness of FR can be recognized from a *Gedankenexperiment*: Imagine two FRs corresponding to two molecules. In order for them to be the same, for each atomic term in the Fourier sum of one molecule there has to be an identical atomic term in the Fourier sum of the other molecule. This is only possible if the corresponding atoms in the respective molecules happen to have the same RDF_I (see below Eq. (8) for examples of atomic RDFs). Now, only if for each atomic RDF_I in one molecule there is an identical atomic RDF_I in the other molecule will the two FR be the same, in which case the two molecules are identical (see also Boutin and Kemper[59]).

III. RESULTS

A. Organic molecules

To illustrate the Fourier series of radial distribution function descriptor for realistic systems, Fig. (3) features FR for three iso-electronic organic molecules, drawn at random from the GDB data base[63]. The nature of a molecular fingerprint, reminiscent of a spectrum, becomes evident for these more complex molecules. Compound (B) has a different stoichiometry while (A) and (C) are constitutional isomers, differing merely in their covalent bonding pattern. Clearly, the fingerprints in Fig. (3) reflect the differences in molecular structure, in particular for larger distances. For smaller r, FR can more easily be understood. For very small r they look very similar, the first peak at $r < 0.5$ Bohr is due to the stoichiometry (nuclear charges) only, with (B) being slightly off from the FR of (A) and (C) which are superimposed. The subsequent three peaks (at 2 to 2.5 Bohr) reflect contributions from the first neighbor shell in the

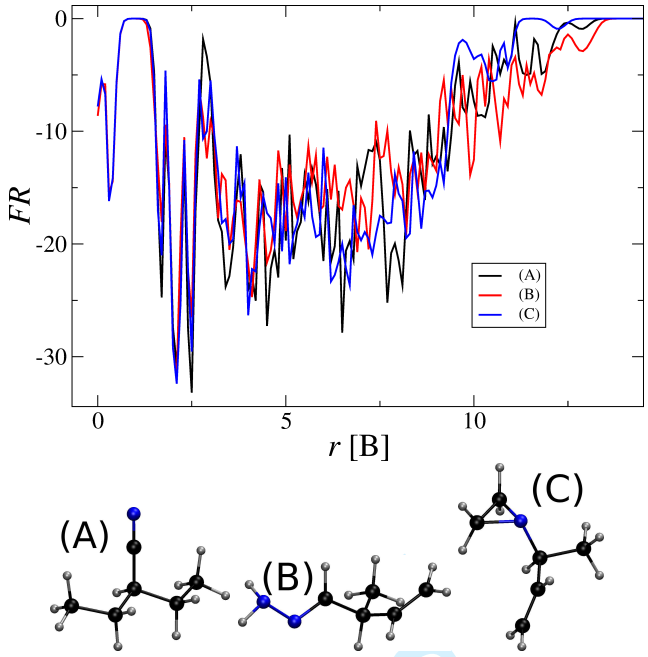


FIG. 3. FR fingerprints [Eq. (8)] with optimized parameters l, n, m, b (from 3k ML model) for three iso-electronic organic molecules (A), (B) and (C) with seven atoms (not counting hydrogens), drawn at random from the GDB-7 data base[63]. White, blue, and black atoms denote hydrogen, carbon, and nitrogen, respectively. (A) and (C) are constitutional isomers while (B) differs in stoichiometry.

atomic radial distributions, being also very similar, albeit not identical, for all three molecules with single, double, cyclic, or even triple bonds (for (A)) between CH, NH, CC, CN, and NN atom pairs. Note that the gap between stoichiometry and structural peaks in Fig. (3) can be expected to be conserved throughout CCS since there are no covalent bonds that are shorter than bonds formed with hydrogen. The fingerprints shown correspond to optimized hyperparameters settings in Eq. (8), alternative parameter combinations would lead to different appearance.

B. Machine Learning models

Any inductive approach requires us to measure distances in terms of input variables. In order to compare chemical compounds, we use the Euclidean norm between the FR descriptors of the two compounds as a proper metric. More specifically, we consider the integral over the squared differences of two FR-descriptors corresponding to molecules i and j,

$$D_{ij}(r_{IJ}^{max}) = \sqrt{\int_0^{r_{IJ}^{max}} dr\ |FR_i(r) - FR_j(r)|^2}.$$

In our implementation, we discretized the FR descriptor and employed an optimal value of 0.1 Bohr for the grid spacing, dr. Note that for $\lim_{r_{IJ}^{max}\to\infty} D_{ij}$ converges for any molecular pair. Here, we used 20 Bohr as the integration upper bound, r_{IJ}^{max}, for all molecules.

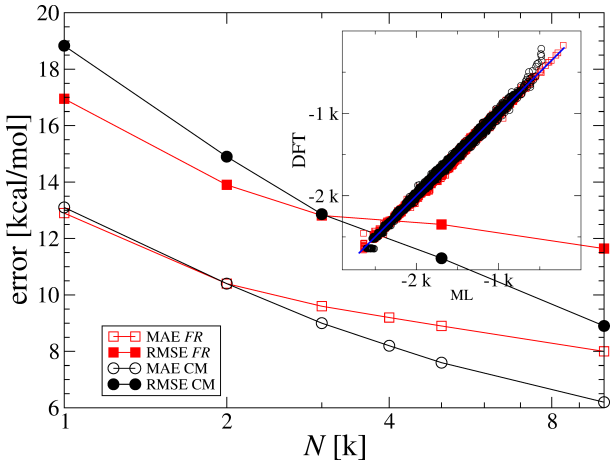


FIG. 4. Mean absolute error (MAE) and root mean square error (RMSE) for out-of-sample predictions of atomization enthalpies at $T = 298.15$ K, as a function of training set size, N, for FR and sorted CM descriptor. Training and test sets consist of enthalpies of atomization at $T = 298.15$ K of the 134 k molecules in the GDB-9 data base[63], calculated at the B3LYP level of theory[34]. The inset shows the corresponding scatter plot for the 10k machine for predicted (ML) versus actual (DFT) enthalpies of atomization in kcal/mol using sorted CM (black) and FR (red) descriptors.

In order to have an idea of the FR's performance of a descriptor, we have built ML models using the enthalpies of atomization for 134 k organic molecules taken from the GDB-17 database, recently published in Ref. 34. The GDB data base represents an exhaustive list of all organic molecules that can be constructed from up to 17 heavy atoms, containing as atom types C, N, O, F, S, Cl, Br, or I and saturating valencies with hydrogen atoms[64,65]. All GDB molecules are expected to be stable and synthetically accessible according to organic chemistry rules[63]. We have drawn at random training sets of sizes 1 k, 2 k, 3 k, 4 k, 5 k, and 10 k.

We then solved the kernel ridge regression problem for the given training sets following the recipes set out in Ref. 66. The solution yields the coefficients $\{\alpha_i\}$ in the ML model of the atomization enthalpy H with FR as an input for any out of sample molecule j,

$$H(FR_j) = \sum_i^N \alpha_i k(D_{ij}). \quad (9)$$

Similarly, we have also trained ML models of the sorted Coulomb matrix CM for the same training and testing sets. In the case of FR, a Gaussian kernel function k with Euclidean norm proved to lead to the best predictive performance. By contrast, in the case of CM we used the Laplacian kernel function with a Manhattan norm, following the findings in Ref. 66.

Resulting mean absolute errors (MAE) and root mean square errors (RMSE) as a function of training set size N, as measured on the remaining molecules in the 134 k set, is shown for both models in Fig. 4. These error estimates have been obtained for out-of-sample predictions (not part of training set), using noise-level and length-scale hyper-parameters optimized through cross-validation runs on training sets. In the case of FR, also parameters b, l, m, n in FR [Eq. (8)] have been optimized using cross-validation for training set sizes $N = 1\,\text{k}$, $2\,\text{k}$ and $3\,\text{k}$. We have found that the training set size has relatively little influence on these parameters, and we have therefore kept them fixed for all training set sizes larger than $3\,\text{k}$. For the set of $N = 3\,\text{k}$, optimal parameters b, l, m, n amount to 7.0052, 0.0852, 1.2395, and -0.1626, respectively. Note for the construction of FR that these parameters refer to interatomic distances in Bohr.

The systematic decay of the mean absolute error with increasing training set size (see Fig. 4) is encouraging. When compared to the current state of the art, the sorted Coulomb matrix, the FR descriptor starts off at a slightly smaller MAE, and significantly smaller RMSE, for a training set size of $1\,\text{k}$. Up to $2\,\text{k}$, CM and FR model errors decay with similar off-set and speed until they reach an accuracy with MAE $\sim$11 kcal/mol, an accuracy similar to the early generalized gradient approximated density functionals in Kohn-Sham DFT[67]. For larger training set sizes the MAE and RMSE of the FR based model continue to decrease, however at a decidedly slower learning rate. The CM-model's errors, in contrast, continue to decay significantly faster. A possible explanation for the FR's change in learning rate could be that as the model's error passes 11 kcal/mol remaining energy differences are dominated by differences in geometry which, due to its high frequency oscillatory nature, the FR descriptor possibly captures only in a less efficient manner than the Coulomb matrix. These first results, however, do not yet enable us to conclusively assess the FR's performance. Merely due to some inherent selection bias of the employed data sets the ML-model's performance using one descriptor might look favorable over the other. Here, for example, we used only relaxed geometries. When attempting to model reaction barriers, however, the performance could possibly be inverted and lead to a different outcome. In any ways, the presented results do amount to numerical evidence suggesting that for the modeling of atomization enthalpies further improvements are necessary before the FR descriptor can be considered competitive with the sorted CM matrix. Further improvements could possibly be achieved by making the FR hyperparameters atom type Z_I-dependent.

C. Computational details

Hyperparameters were estimated through 5-fold cross validation (CV) on training set of size N. Accordingly, N training molecules were distributed at random into 5 bins, each containing $N/5$ molecules. Each bin was used once as the holdout set, with the remaining 4 bins as training set, and hyperparameters were optimized by minimizing the MAE for the holdout-bin. Globally optimal hyperparameters were obtained by taking the median of the 5 folds. The final kernel with globally optimized hyperparameters was subsequently used to predict atomization enthalpies for the $134\,\text{k} - N$ out-of-sample molecules which never had a part in training.

IV. CONCLUSIONS

A set of fundamental physical arguments has been introduced as to what are crucial and desirable properties of descriptors that can be expected to yield reliable performance in intelligent data analysis (IDA) methods when applied to the modeling of quantum chemical properties of molecules. Starting from the external potential in the electronic Hamiltonian, and using Fourier transforms and radial distribution functions, we have introduced an intramolecular distance based fingerprint-like descriptor, FR, corresponding to a Fourier series of atomic radial distribution functions. The FR is unique for any molecular compound (i.e. chemical composition and geometry), and invariant with respect to translation, rotation, and atom indexing. Furthermore, FR is differentiable, not only with respect to nuclear displacement for geometry optimization or molecular dynamics, but also with respect to "alchemical" changes, i.e. change in nuclear charges[31,68–71], potentially useful for computational materials design[72–74]. As such, this descriptor exhibits all the crucial and desired properties listed in Table I. The FR descriptor can be reduced to $N \times (N-1)/2$ dimensionality if it is evaluated only at those r-values that correspond to interatomic distances in a compound. A Gaussian pre-factor can be used to damp the descriptor to reduce the dimensionality further and to introduce locality. Results from preliminary ML models, yielding promising predictive power for out-of-sample compounds, suggest that the FR descriptor, or variants thereof, is likely to be well suited for the generic and systematic construction of ML models that are valid for all regions of the potential energy surface of novel compounds, as long as trained across sufficiently representative subspace of CCS. The current FR-performance, however, is not (yet) on par with the sorted Coulomb matrix. Note that the FR exclusively represents the external potential of a molecule, not the molecule's charge. Differences in the latter can easily be added to FR distances through the use of more sophisticated metrics, such as normalized Euclidean, or Mahalanobis, distances. A more comprehensive assessment, also including non-equilibrium geometries on the same potential energy surfaces, will be subject of future work.

V. ACKNOWLEDGMENTS

The authors thank J. R. Hammond, M. Hereld, and A. Vazquez-Mayagoitia for discussions and suggestions. This research used resources of the Argonne Leadership Computing Facility at Argonne National Laboratory, which is supported by the Office of Science of the U.S. DOE under Contract No. DE-AC02-06CH11357. OAvL acknowledges support from LDRD funding (Multiscale Materials Modeling using Accurate Ab Initio Approaches (M3A3)). OAvL acknowledges funding from the Swiss National Science foundation No. PP00P2_138932. Some of the calculations were performed at sciCORE (http://scicore.unibas.ch/) scientific computing core facility at University of Basel.

* anatole.vonlilienfeld@unibas.ch

[1] P. Kirkpatrick and C. Ellis, Nature **432**, 823 (2004).
[2] A. Franceschetti and A. Zunger, Nature **402**, 60 (1999).
[3] S. Curtarolo, D. Morgan, K. Persson, J. Rodgers, and G. Ceder, Phys. Rev. Lett. **91**, 135503 (2003).
[4] L. P. Hammett, Chem. Rev. **17**, 125 (1935).
[5] L. P. Hammett, J. Am. Chem. Soc. **59**, 96 (1937).
[6] S. Roy, S. Goedecker, and V. Hellmann, Phys. Rev. B **77**, 056707 (2008).
[7] J. K. Nørskov, T. Bligaard, J. Rossmeisl, and C. H. Christensen, Nature Chemistry **1**, 37 (2009).
[8] S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, Nature Mater **12**, 191 (2013).
[9] R. Iftimie, P. Minary, and M. E. Tuckerman, Proc. Natl. Acad. Sci. USA **102**, 6654 (2005).
[10] G. Hautier, C. C. Fischer, A. Jain, T. Mueller, and G. Ceder, Chem. Mater. **22**, 3762 (2010).
[11] G. R. Hutchison, M. A. Ratner, and T. J. Marks, J. Am. Chem. Soc. **127**, 2339 (2005).
[12] M. Misra, D. Andrienko, B. Baumeier, J.-L. Faulon, and O. A. von Lilienfeld, J. Chem. Theory Comput. **7**, 2549 (2011).
[13] B. G. Sumpter and D. W. Noid, Chemical Physics Letters **192**, 455 (1992).
[14] S. Lorenz, A. Gross, and M. Scheffler, Chem. Phys. Lett. **395**, 210 (2004).
[15] S. Manzhos and T. Carrington, Jr., J. Chem. Phys. **125**, 084109 (2006).
[16] J. Behler and M. Parrinello, Phys. Rev. Lett. **98**, 146401 (2007).
[17] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Phys. Rev. Lett. **104**, 136403 (2010).
[18] J. C. Snyder, M. Rupp, K. Hansen, K.-R. Müller, and K. Burke, Phys. Rev. Lett. **108**, 253002 (2012).
[19] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, Phys. Rev. Lett. **108**, 058301 (2012).
[20] J.-L. Faulon, D. P. Visco, Jr., and R. S. Pophale, J. Chem. Inf. Comp. Sci. **43**, 707 (2003).
[21] J. Visco, R. S. Pophale, M. D. Rintoul, and J. L. Faulon, J. Mol. Graph. Model. **20**, 429 (2002).
[22] S. Martin, D. Roe, and J.-L. Faulon, BIOINFORMATICS **21**, 218 (2005).
[23] R. A. DiStasio, O. A. von Lilienfeld, and A. Tkatchenko, Proc. Natl. Acad. Sci. USA **109**, 14791 (2012).
[24] G. Brunklaus, A. Koch, D. Sebastiani, and H. Spiess, Phys. Chem. Chem. Phys. **9**, 4545 (2007).
[25] J. Braun, A. Kerber, M. Meringer, and C. Rücker, MATCH **54**, 163 (2005).
[26] G. Schneider, Nature Reviews **9**, 273 (2010).
[27] K. Burke, (2011), "Any *ab initio* method must either be void of empirical parameters, or at least have parameters that do not depend on the system being studied." Oral communication, IPAM, UCLA.
[28] T. Helgaker, P. Jørgensen, and J. Olsen, *Molecular Electronic-Structure Theory* (John Wiley & Sons, LTD, 2000).
[29] Y. Zhao and D. G. Truhlar, Acc. Chem. Res. **41**, 157 (2008).
[30] J. Ma, A. Michaelides, and D. Alfé, J. Chem. Phys. **134**, 134701 (2011).
[31] O. A. von Lilienfeld and M. E. Tuckerman, J. Chem. Phys. **125**, 154104 (2006).
[32] M. E. Tuckerman, *Statistical mechanics: Theory and molecular simulation* (Oxford University Press, 2010).
[33] A. Pérez, M. E. Tuckerman, H. P. Hjalmarson, and O. A. von Lilienfeld, J. Am. Chem. Soc. **132**, 11510 (2010).
[34] R. Ramakrishnan, P. Dral, M. Rupp, and O. A. von Lilienfeld, Scientific Data **1**, 140022 (2014).
[35] J. A. GRANT, M. A. GALLARDO, and B. T. PICKUP, Journal of Computational Chemistry **17**, 1653 (1996).
[36] A. P. Bartók, R. Kondor, and G. Csányi, Phys. Rev. B **87**, 184115 (2013).
[37] C. Lipinski, F. Lombardo, B. Dominy, and P. Feeney, **23**, 3 (1997).
[38] J. E. Moussa, Phys. Rev. Lett. **109**, 059801 (2012).
[39] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, Phys. Rev. Lett. **109**, 059802 (2012).
[40] K. Hansen, F. Biegler, O. A. von Lilienfeld, K.-R. Müller, and A. Tkatchenko, (2015), submitted to Nature Comm.
[41] B. Tehan, E. Lloyd, M. Wong, W. Pitt, J. Montana, D. Manallack, and E. Gancia, **21**, 457 (2002).
[42] L. Hu, X. Wang, L. Wong, and G. Chen, J. Chem. Phys. **119**, 11501 (2003).
[43] X. Zheng, L. Hu, X. Wang, and G. Chen, Chem. Phys. Lett. **390**, 186 (2004).
[44] P. R. Bunker and P. Jensen, in *Molecular Symmetry and Spectroscopy* (NRC Research Press, 2006).
[45] P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864 (1964).
[46] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, New Journal of Physics **15**, 095003 (2013).
[47] F. H. Zadeh and P. W. Ayers, Journal of Mathematical Chemistry **51**, 927 (2013).
[48] M. Quack, Angew. Chem. Int. Ed. **41**, 4619 (2002).
[49] J. Behler, Phys. Chem. Chem. Phys. **13**, 17930 (2011).
[50] E. Prodan and W. Kohn, Proceedings of the National Academy of Sciences of the United States of America **102**, 11635 (2005), http://www.pnas.org/content/102/33/11635.full.pdf+html.
[51] W. Li, T. Pohl, J. M. Rost, S. T. Rittenhouse, H. R. Sadeghpour, J. Nipper, B. Butscher, J. B. Balewski, V. Bendkowsky,

R. Lw, and T. Pfau, Science **334**, 1110 (2011), http://www.sciencemag.org/content/334/6059/1110.full.pdf.

[52] G. M. Maggiora, J. Chem. Inf. Model. **46**, 1535 (2006).

[53] C. Plagemann, K. Kersting, and W. Burgard, in *Proceedings of the 2008 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD2008), Antwerp, Belgium, September 15–19*, Lecture Notes in Artificial Intelligence, Vol. 5212, edited by W. Daelemans, B. Goethals, and K. Morik (Springer, 2008) pp. 204–219.

[54] R. Carbó, L. Leyda, and M. Arnau, Int. J. Quantum Chem. **17**, 1185.

[55] Q. Wang, O. Ronneberger, and H. Burkhardt, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE **31**, 1715 (2009).

[56] S. Doraiswamy, J. Bender, G. V. Candler, Y. Paukku, K. Yang, Z. Varga, and D. G. Truhlar, (2012), to be published.

[57] A. L. Patterson, Nature **143**, 939 (1939).

[58] O. A. von Lilienfeld and A. Tkatchenko, J. Chem. Phys. **132**, 234109 (2010).

[59] M. Boutin and G. Kemper, (2008), http://arxiv.org/abs/0710.1870.

[60] M. C. Hemmer, V. Steinhauer, and J. Gasteiger, Vibrational Spectroscopy **19**, 151 (1999).

[61] R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors* (Wiley-VCH, Weinheim, 2009).

[62] M. P. Gonzlez, C. Tern, M. Teijeira, and A. M. Helguera, European Journal of Medicinal Chemistry **41**, 56 (2006).

[63] L. C. Blum and J.-L. Reymond, J. Am. Chem. Soc. **131**, 8732 (2009).

[64] T. Fink, H. Bruggesser, and J.-L. Reymond, Angew. Chem. Int. Ed. **44**, 1504 (2005).

[65] T. Fink and J.-L. Reymond, J. Chem. Inf. Model. **47**, 342 (2007).

[66] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. von Lilienfeld, A. Tkatchenko, and K.-R. Müller, Journal of Chemical Theory and Computation **9**, 3404 (2013), http://pubs.acs.org/doi/pdf/10.1021/ct400195d.

[67] W. Koch and M. C. Holthausen, *A Chemist's Guide to Density Functional Theory* (Wiley-VCH, 2002).

[68] F. Weigend, C. Schrodt, and R. Ahlrichs, J. Chem. Phys. **121**, 10380 (2004).

[69] V. Marcon, O. A. von Lilienfeld, and D. Andrienko, J. Chem. Phys. **127**, 064305 (2007).

[70] O. A. von Lilienfeld and M. E. Tuckerman, J. Chem. Theory Comput. **3**, 1083 (2007).

[71] O. A. von Lilienfeld, J. Chem. Phys. **131**, 164102 (2009).

[72] O. A. von Lilienfeld, R. Lins, and U. Rothlisberger, Phys. Rev. Lett. **95**, 153002 (2005).

[73] D. Sheppard, G. Henkelman, and O. A. von Lilienfeld, J. Chem. Phys. **133**, 084104 (2010).

[74] O. A. von Lilienfeld, International Journal of Quantum Chemistry **113**, 1676 (2013).