Non-covalent interactions across organic and biological subsets of chemical space: Physics-based potentials parametrized from machine learning

Tristan Bereau, Robert A. DiStasio, Alexandre Tkatchenko, and O. Anatole von Lilienfeld

Citation: The Journal of Chemical Physics **148**, 241706 (2018); doi: 10.1063/1.5009502
View online: https://doi.org/10.1063/1.5009502
View Table of Contents: http://aip.scitation.org/toc/jcp/148/24
Published by the American Institute of Physics

Non-covalent interactions across organic and biological subsets of chemical space: Physics-based potentials parametrized from machine learning

Tristan Bereau,[1,a)] Robert A. DiStasio, Jr.,[2] Alexandre Tkatchenko,[3] and O. Anatole von Lilienfeld[4]

[1]*Max Planck Institute for Polymer Research, Ackermannweg 10, 55128 Mainz, Germany*
[2]*Department of Chemistry and Chemical Biology, Cornell University, Ithaca, New York 14853, USA*
[3]*Physics and Materials Science Research Unit, University of Luxembourg, L-1511 Luxembourg, Luxembourg*
[4]*Institute of Physical Chemistry and National Center for Computational Design and Discovery of Novel Materials (MARVEL), Department of Chemistry, University of Basel, Klingelbergstrasse 80, CH-4056 Basel, Switzerland*

(Received 17 October 2017; accepted 18 January 2018; published online 15 March 2018)

Classical intermolecular potentials typically require an extensive parametrization procedure for any new compound considered. To do away with prior parametrization, we propose a combination of physics-based potentials with machine learning (ML), coined IPML, which is transferable across small neutral organic and biologically relevant molecules. ML models provide on-the-fly predictions for *environment-dependent local* atomic properties: electrostatic multipole coefficients (significant error reduction compared to previously reported), the population and decay rate of valence atomic densities, and polarizabilities across conformations and chemical compositions of H, C, N, and O atoms. These parameters enable accurate calculations of intermolecular contributions—electrostatics, charge penetration, repulsion, induction/polarization, and many-body dispersion. Unlike other potentials, this model is transferable in its ability to handle new molecules and conformations without explicit prior parametrization: All local atomic properties are predicted from ML, leaving only eight global parameters—optimized once and for all across compounds. We validate IPML on various gas-phase dimers at and away from equilibrium separation, where we obtain mean absolute errors between 0.4 and 0.7 kcal/mol for several chemically and conformationally diverse datasets representative of non-covalent interactions in biologically relevant molecules. We further focus on hydrogen-bonded complexes—essential but challenging due to their directional nature—where datasets of DNA base pairs and amino acids yield an extremely encouraging 1.4 kcal/mol error. Finally, and as a first look, we consider IPML for denser systems: water clusters, supramolecular host-guest complexes, and the benzene crystal. *Published by AIP Publishing.* https://doi.org/10.1063/1.5009502

I. INTRODUCTION

Our understanding of the physical laws that govern molecular interactions have led to an ever-improving description of the high-dimensional potential energy surface of condensed molecular systems. A variety of computational methods provide various approximations thereof: while high-level methods (e.g., coupled cluster) are restricted to a small number of atoms, other electronic-structure methods (e.g., density functional theory—DFT) can reach larger system sizes of up to 10^2–10^3 atoms. Beyond this limit, classical potentials and force fields provide a much faster estimate of the interactions, enabling the calculation of thermodynamic and even kinetic properties for complex materials.

Many classical potentials and force fields are often termed *physics-based* because they encode assumptions about the governing physics of the interactions via their functional forms. Despite their widespread interest by the community, classical potentials are currently limited to a narrow set of molecules and materials, due to tedious and non-systematic parametrization strategies. Additive (i.e., non-polarizable) atomistic force fields are typically parametrized from a combination of *ab initio* calculations and experimental measurements, e.g., pure-liquid density, heat of vaporization, or NMR chemical shifts. Ensuring the accurate reproduction of various molecular properties, from conformational changes to thermodynamic properties (e.g., free energy of hydration), but also consistency across all other molecules parametrized remains challenging, time consuming, and difficult to automate.

Recently, a number of studies have brought forward the idea of more automated parametrizations. For instance, QMDFF is based on reference DFT calculations to parametrize a set of classical potentials.[1] We also point out the automatic generation of intermolecular energies[2] extracted from reference symmetry-adapted perturbation theory[3] (SAPT) calculations. Interestingly, recent efforts have aimed at parametrizing potentials and force fields from atom-in-molecule (AIM) properties. Van Vleet *et al.*[4] and Vandenbrande *et al.*[5] showed

a)Electronic mail: bereau@mpip-mainz.mpg.de

that a systematic use of AIMs can significantly reduce the number of global parameters to scale the individual energetic contributions. Overall, they propose AIMs as a means to more systematically parametrize models. Similar conclusions were reached for the additive OPLS force field,[6] for which the missing polarization effects make a systematic scheme all the more challenging. These methodologies still require a number of *a priori* reference electronic-structure calculations to optimize various parameters of any new molecule encountered.

In the context of developing classical potentials for *in silico* screening across large numbers of compounds, the necessary computational investment for the parametrization procedures of each new molecule can become daunting. A radically different strategy consists in *predicting* the potential energy surface of a system from machine learning (ML).[7–9] ML encompasses a number of statistical models that improve their accuracy with data. Recent studies have reported unprecedented accuracies in reproducing reference energies from electronic-structure calculations, effectively offering a novel framework for accurate intramolecular interactions freed from molecular-mechanics-type approximations (e.g., harmonic potential).[10–12] While they do away with free parameters that need optimization (i.e., unlike force fields), they typically suffer from limited transferability: an ML model is inherently limited to interpolating across the training samples. A model trained on water clusters can be remarkably accurate toward describing liquid-state properties (e.g., pair-correlation functions) but remains specific to interactions solely involving water.[13] Transferability of an ML model that would predict interactions across chemical compound space (i.e., the diversity of chemical compounds) stands nowadays as computationally intractable. Part of the reason is the necessity to interpolate across all physical phenomena for any geometry, as these models are driven by experience, rather than physical principles. Symmetries and conservation laws will require large amounts of data to be appropriately satisfied if they are not correctly encoded *a priori*.

In this work, we propose a balance between the aforementioned physics-based models and an ML approach, coined IPML. To best take advantage of both approaches, we choose to rely on a physics-based model, where most parameters are predicted from ML. This approach holds two main advantages: (i) Leverage our understanding of the physical interactions at hand, together with the associated symmetries and functional forms, and (ii) alleviate the reference calculations necessary to optimize the parameters of each new molecule.

The aforementioned AIM-based classical potentials, in this respect, offer an interesting strategy: they largely rely on perturbation theory to treat the long-range interactions (i.e., electrostatics, polarization, and dispersion), while overlap models of spherically symmetric atomic densities describe the short-range interactions. Both theoretical frameworks estimate interaction energies from *monomer* properties—thereby significantly reducing the ML challenge from learning interactions between any combination of molecules to the much simpler prediction of (isolated) atomic properties. Incidentally, learning atomic and molecular properties has recently been the subject of extended research, providing insight into the appropriate representations and ML models.[12,14–16] Parametrizing small-molecule force fields based on ML has already shown advantageous at a more coarse-grained resolution.[17] At the atomistic level, Bereau *et al.* had shown early developments of learning AIM properties, namely, distributed multipole coefficients to describe the electrostatic potential of a molecule.[18] The study was aiming at an accurate prediction of multipole coefficients across the chemical space of small organic molecules. These coefficients provide the necessary ingredients to compute the electrostatic interaction between molecules via a multipole expansion.[19] Here, we extend this idea by further developing physics-based models parametrized from ML to all major interaction contributions: electrostatics, polarization, repulsion, and dispersion. We base our method on a few ML models of AIM properties: distributed multipoles, atomic polarizabilities from Hirshfeld ratios, and the population and decay rate of valence atomic densities. The combination of physics-based potentials and ML reduces the number of global parameters to only 8 in the present model. We optimize our global parameters once and for all such that a new compound requires no single parameter to be optimized (because the ML needs no refitting), unlike most other aforementioned AIM- and physics-based models.[1,2,4] Vandenbrande *et al.* did present results using frozen global parameters, but their model still requires quantum-chemistry calculations on every new compound to fit certain parameters (e.g., point charges).[5] After parametrization on parts of the S22x5 small-molecule dimer dataset,[20] we validate IPML on more challenging dimer databases of small molecules, DNA base pairs, and amino-acid pairs. We later discuss examples beyond small-molecule dimers toward the condensed phase: water clusters, host-guest complexes, and the benzene crystal.

II. IPML: PHYSICS-BASED POTENTIALS PARAMETRIZED FROM MACHINE LEARNING

A. Learning of environment-dependent local atomic properties

The set of intermolecular potentials is based on ML of local (i.e., atom in molecule) properties targeted at predicting electrostatic multipole coefficients, the decay rate of atomic densities, and atomic polarizabilities, which we present in the following.

1. Electrostatic multipole coefficients

The prediction of atomic multipole coefficients up to quadrupoles was originally presented in the work of Bereau *et al.*[18] DFT calculations at the M06-2X level[21] followed by a Gaussian distributed multipole analysis (GDMA)[19] (i.e., wavefunction partitioning scheme) provided reference multipoles for several thousands of small organic molecules. ML of the multipoles was achieved using kernel-ridge regression. The geometry of the molecule was encoded in the Coulomb matrix,[14] $\mathbf{C}$, such that for two atoms i and j,

$$C_{ij} = \begin{cases} Z_i^{2.4}/2 & i = j, \\ Z_i Z_j / r_{ij} & i \neq j. \end{cases} \tag{1}$$

Though the Coulomb matrix accounts for translational and rotational symmetry, it does not provide sufficient information to unambiguously encode non-scalar, orientation-dependent quantities, such as dipolar (i.e., vector) and quadrupolar (i.e., second-rank tensor) terms. A consistent encoding of these terms had been achieved by rotating them along a local axis system, provided by the molecular moments of inertia. To improve learning, the model aimed at predicting the difference between the reference GDMA multipoles and a simple physical, parameter-free baseline that helped identify symmetries in vector and tensor components (hereafter mentioned as delta learning). The large memory required to optimize kernel-ridge regression models led us to construct one ML model per chemical element.

In this work, we both simplify the protocol and significantly improve the model's accuracy. Reference multipoles are now extracted from DFT calculations at the PBE0 level. Rather than using GDMA multipoles, we now rely on the minimal basis iterative stockholder (MBIS) partitioning scheme. While Misquitta *et al.* recently recommended the use of the iterated stockholder atom (ISA) multipoles,[22] we use MBIS multipoles for their consistency with the abovementioned atomic-density parameters and the small magnitude of the higher multipoles, easing the learning procedure. We have also found MBIS multipoles to yield reasonable electrostatic energies at long ranges (data not shown). MBIS multipoles were computed using Horton.[23] Instead of relying on the molecular moments of inertia as a local axis system, we project each non-scalar multipole coefficient into a basis set $\{\mathbf{e}_{ij}, \mathbf{e}_{ik}, \mathbf{e}_{il}\}$ formed by three non-collinear vectors $\mathbf{e}$ from the atom of interest i to its three closest neighbors: j, k, and l [e.g., $\mathbf{e}_{ij} = (\mathbf{r}_j - \mathbf{r}_i)/\|\mathbf{r}_j - \mathbf{r}_i\|$, where $\mathbf{r}_i$ denotes the Cartesian coordinates of atom i]. The vectors $\mathbf{e}_{ik}$ and $\mathbf{e}_{il}$ are further adjusted to form a right-handed orthonormal basis set.

Further, the representation used for the ML model of electrostatic multipoles is now the atomic Spectrum of London and Axilrod-Teller-Muto (aSLATM) potentials.[24,25] aSLATM represents an atomic sample and its environment through a distribution of (i) chemical elements, (ii) pairwise distances scaled according to London dispersion, and (iii) triplet configurations scaled by the three-body Axilrod-Teller-Muto potential. We point out that aSLATM is atom-index invariant and as such does not suffer from discontinuities other representations may have. We used the qml implementation.[26] Point charges are systematically corrected so as to yield an exactly neutral molecule.

2. Atomic-density overlap

Exchange-repulsion and other short-ranged interactions are proportional to the overlap of the electron densities,[4,27]

$$S_{ij} = \int \mathrm{d}^3\mathbf{r}\; n_i(\mathbf{r}) n_j(\mathbf{r}). \tag{2}$$

Van Vleet *et al.*[4] presented a series of short-ranged intermolecular potentials based on a Slater-type model of overlapping valence atomic densities. They approximated the atomic density using the iterated stockholder atom (ISA) approach.[22,28] The atomic density of atom i, $n_i(\mathbf{r})$, is approximated by a single exponential function centered around the nucleus,

$$n_i(r) \propto \exp(-\sigma_i r), \tag{3}$$

where σ_i characterizes the rate of decay of the valence atomic density. The short-ranged interactions proposed by Van Vleet *et al.* rely on combinations of the decay rates of atomic densities, i.e., $\sigma_{ij} = \sqrt{\sigma_i \sigma_j}$, for the atom pair i and j. While the decay rates were obtained from reference DFT calculations, atom-type-dependent prefactors were fitted to short-range interaction energies. Vandenbrande *et al.* more recently applied a similar methodology to explicitly include the reference populations as normalization, $N_i = \int \mathrm{d}\mathbf{r}\; n_i(\mathbf{r})$, i.e., the volume integrals of the valence atomic densities.[5] Their method allowed reducing the number of unknown prefactors per dimer: a single value for repulsion and short-range polarization and no free parameter for penetration effects (*vide infra*).

We constructed an ML model of N and σ using the same representations and kernel as for Hirshfeld ratios (see above). Reference coefficients N and σ were computed using Horton[23,29] for 1102 molecules using PBE0, amounting to 16 945 atom-in-molecule properties. Instead of the ISA approach, we followed Verstraelen *et al.* and relied on the MBIS partitioning method.[29]

3. Atomic polarizabilities

The Hirshfeld scheme provides a partitioning of the molecular charge density into atomic contributions (i.e., an atom-in-molecule description).[30–33] It consists of estimating the change of atomic volume of atom p due to the neighboring atoms, as compared to the corresponding atom in free space

$$\frac{V_p^{\mathrm{eff}}}{V_p^{\mathrm{free}}} = \frac{\int \mathrm{d}\mathbf{r} r^3 w_p(\mathbf{r}) n(\mathbf{r})}{\int \mathrm{d}\mathbf{r} r^3 n_p^{\mathrm{free}}(\mathbf{r})}, \tag{4}$$

where $n_p^{\mathrm{free}}(\mathbf{r})$ is the electron density of the free atom, $n(\mathbf{r})$ is the electron density of the molecule, and $w_p(\mathbf{r})$ weighs the contribution of the free atom p against all free atoms at $\mathbf{r}$

$$w_p(\mathbf{r}) = \frac{n_p^{\mathrm{free}}(\mathbf{r})}{\sum_q n_q^{\mathrm{free}}(\mathbf{r})}, \tag{5}$$

where the sum runs over all atoms in the molecule.[31] The static polarizability is then estimated from the free-atom polarizability scaled by the Hirshfeld ratio, h,[34]

$$\alpha_p = \alpha_p^{\mathrm{free}} \left(\frac{V_p^{\mathrm{eff}}}{V_p^{\mathrm{free}}} \right)^{4/3} = \alpha_p^{\mathrm{free}} h^{4/3}. \tag{6}$$

Reference Hirshfeld ratios were provided from DFT calculations of 1000 molecules using the PBE0[35] functional and extracted using postg.[36,37] The geometry of the molecule was encoded in the Coulomb matrix [Eq. (1)]. An ML model of the Hirshfeld ratios was built using kernel-ridge regression and provided predictions for atomic polarizabilities of atoms in molecules for the chemical elements H, C, O, and N. For all ML models presented here, datasets are split between training and test subsets at an 80:20 ratio, in order to avoid overfitting.

B. Intermolecular interactions from physics-based models

In the following, we present the different terms in our interaction energy and how they rely on the abovementioned ML properties.

1. Distributed multipole electrostatics

The description of atom-distributed multipole electrostatics implemented here follows the formalism of Stone.[19] A Taylor series expansion of the electrostatic potential of atom i gives rise to a series of multipole coefficients

$$\phi_i(\mathbf{r}) = \frac{1}{4\pi\epsilon_0}\left[q_i\left(\frac{1}{r}\right) - \mu_{i,\xi}\nabla_\xi\left(\frac{1}{r}\right) + \frac{1}{3}\Theta_{i,\xi\zeta}\nabla_\xi\nabla_\zeta\left(\frac{1}{r}\right) - \cdots\right], \tag{7}$$

where ξ and ζ indices run over coordinates and the Einstein summation applies throughout. We lump the multipole coefficients in a vector $M_i = (q_i, \mu_{i,1}, \mu_{i,2}, \mu_{i,3}, \ldots)^t$ and derivatives of $1/r$ into the interaction matrix $\mathbf{T}^{ij} = (T^{ij}, T^{ij}_1, T^{ij}_2, T^{ij}_3, T^{ij}_{11}, \ldots)^t$ for the interaction between atoms i and j, where the number of indices indicates the order of the derivative [e.g., $T^{ij}_\xi = \nabla_\xi(1/r_{ij})$]. In this way, the multipole electrostatic interaction energy is given by

$$E_{\text{elec}} = \sum_{ij} M_i T^{ij} M_j. \tag{8}$$

More details on the formalism and implementation of multipole electrostatics can be found elsewhere.[19,38,39] Multipole coefficients are provided by the ML model for electrostatics originally presented in the work of Bereau *et al.*[18] and improved herein (see Methods Sec. II A 1 above).

2. Charge penetration

The abovementioned multipole expansion explicitly assumes no wavefunction overlap between molecules. At short range, the assumption is violated, leading to discrepancies in the electrostatic energy, denoted *penetration* effects. The link between penetration and charge-density overlap[19] has been leveraged before by separating an atomic point charge into an effective core and a damped valence electron distribution.[40–43] An extension has later been proposed by Vandenbrande *et al.* to efficiently estimate the correction *without* any free parameter.[5] This is achieved by including the atomic-density population N_i of atom i—the normalization term in Eq. (3). Penetration is modeled by correcting the monopole-monopole interactions in a pairwise fashion,

$$\begin{aligned} E_{\text{pen}} &= \sum_{ij}\frac{q^{\text{c}}_i N_j}{r}g(\sigma_j, r) + \frac{N_i q^{\text{c}}_j}{r}g(\sigma_i, r) \\ &\quad - \frac{N_i N_j}{r}\left(f(\sigma_i, \sigma_j, r) + f(\sigma_j, \sigma_i, r)\right), \\ g(\sigma, r) &= \left(1 + \frac{r}{2\sigma}\right)\exp\left(-\frac{r}{\sigma}\right), \\ f(\sigma_i, \sigma_j, r) &= \frac{\sigma_i^4}{(\sigma_i^2 - \sigma_j^2)^2}\left(1 + \frac{r}{2\sigma_i} - \frac{2\sigma_j^2}{\sigma_i^2 - \sigma_j^2}\right)\exp\left(-\frac{r}{\sigma_i}\right). \end{aligned} \tag{9}$$

The present expression for $f(\sigma_i, \sigma_j, r)$ is problematic when $\sigma_i \approx \sigma_j$ given the denominator, but Vandenbrande *et al.* derived corrections for such cases.[5] The parameter q^c corresponds to a core charge that is not subject to penetration effects, i.e., $q = q^c - N$, where q is determined from the multipole expansion.

We note the presence of three terms when considering electrostatics together with penetration [Eq. (9)]: the core-core interaction [part of E_{elec}, Eq. (8)], the damping term between the core and smeared density, and the last is the overlap between two smeared density distributions. In most existing approaches, the damping functions aim at modeling the outer Slater-type orbitals of atoms—e.g., note the presence of exponential functions in Eq. (9). Unfortunately, penetration effects due to the higher moments are not presently corrected. Conceptually, a separation between core and smeared contributions of higher multipoles is unclear. Rackers *et al.* proposed an interesting framework that assumes a simplified functional form for the damping term and factors out of the entire interaction matrix T^{ij}_ξ.[44] We have not attempted to express Eq. (9) for the interaction matrix T^{ij}_ξ of all multipoles.

3. Repulsion

Following Vandenbrande *et al.*,[5] we parametrize the repulsive energy based on the overlap of valence atomic densities:

$$\begin{aligned} E_{\text{rep}} &= U_i^{\text{rep}} U_j^{\text{rep}} \sum_{ij}\frac{N_i N_j}{8\pi r}\left(h(\sigma_i, \sigma_j, r) + h(\sigma_j, \sigma_i, r)\right), \\ h(\sigma_i, \sigma_j, r) &= \left(\frac{4\sigma_i^2\sigma_j^2}{(\sigma_j^2 - \sigma_i^2)^3} + \frac{\sigma_i}{(\sigma_j^2 - \sigma_i^2)^2}\right)\exp\left(-\frac{r}{\sigma_i}\right), \end{aligned} \tag{10}$$

where U_i^{rep} is an overall prefactor that depends only on the *chemical element* of i. The multiplicative mixing rule we apply leads to U_i^{rep} having units of (energy)$^{1/2}$. Here again, corrections for $h(\sigma_i, \sigma_j, r)$ when $\sigma_i \approx \sigma_j$ can be found elsewhere.[5]

4. Induction/polarization

Polarization effects are introduced via a standard Thole-model description.[45] Induced dipoles, μ^{ind}, are self-consistently converged against the electric field generated by both multipoles and the induced dipoles themselves,

$$\mu^{\text{ind}}_{i,\xi} = \alpha_i\left(\sum_j T^{ij}_\xi M_j + \sum_{j'} T^{ij'}_{\xi\zeta}\mu^{ind}_{j',\zeta}\right), \tag{11}$$

where we follow the notation of Ren and Ponder:[38] the first sum (indexed by j) only runs over atoms *outside* of the molecule containing i—a purely intermolecular contribution—while the second sum (indexed by j') contains all atoms except for i. We self-iteratively converge the induced dipoles using an overrelaxation coefficient $\omega = 0.75$ as well as a smeared charge distribution, n', following Thole's prescription[45] and the AMOEBA force field,[38]

$$n' = \frac{3a}{4\pi}\exp\left(-au^3\right), \tag{12}$$

where $u = r_{ij}/(\alpha_i\alpha_j)^{1/6}$ and a controls the strength of damping of the charge distribution. The smeared charge distribution n' leads to a modified interaction matrix, as described by Ren and Ponder.[38] The electrostatic contribution of the induced dipoles is then evaluated to yield the polarization energy. In this scheme, polarization thus relies on both the predicted atomic polarizabilities and predicted multipole coefficients.

5. Many-body dispersion

Many-body dispersion[46] (MBD) relies on the formalism of Tkatchenko and co-workers.[47] It consists of a computationally efficient cast of the random-phase approximation into a system of quantum harmonic oscillators.[48] In Appendix A, we briefly summarize the MBD implementation and suggest the interested reader to Ref. 32 for additional details.

6. Overall model

To summarize, our intermolecular IPML model is made of five main contributions: (i) electrostatics, (ii) charge penetration, (iii) repulsion, (iv) induction/polarization, and (v) many-body dispersion. Our use of ML to predict AIM properties yields only eight global parameters to be optimized: (i) none; (ii) none; (iii) $U_{\mathrm{H}}^{\mathrm{rep}}$, $U_{\mathrm{C}}^{\mathrm{rep}}$, $U_{\mathrm{N}}^{\mathrm{rep}}$, $U_{\mathrm{O}}^{\mathrm{rep}}$; (iv) a; and (v) β, γ, d. We will optimize these parameters simultaneously across different compounds to explore their transferability.

We provide a Python-based implementation of this work at https://gitlab.mpcdf.mpg.de/trisb/ipml for download. The ML models relied on kernel ridge regression, implemented here using NUMPY routines.[49] Different atomic properties were trained on different datasets. These datasets are also provided in the repository. While a single training set for all properties would offer more consistency, different properties require very different training sizes to reach an accuracy that is satisfactory. Molecular configurations were generated from SMILES strings using OPEN BABEL.[50] These approximate configurations were purposefully *not* further optimized to obtain a more heterogeneous training set of configurations, thereby improving the interpolation of the ML.

III. TRAINING AND PARAMETRIZATION OF IPML

We show the accuracy of the prediction of the multipole coefficients, the Hirshfeld ratios, and the atomic-density decay rates, followed by the assessment of experimental molecular polarizabilities. We then parametrize the different terms of the intermolecular potentials against reference total energies on parts of the S22x5 dataset and validate it against various other intermolecular datasets.

A. Training of multipole coefficients

We performed ML of the multipole coefficients trained on up to 20 000 atoms in molecules—limited to neutral compounds. While our methodology allows us to learn all compounds together, we chose to train an individual ML model for each chemical element. Figure 1 shows the correlation between reference and predicted components for $\sim 10^3$ atoms

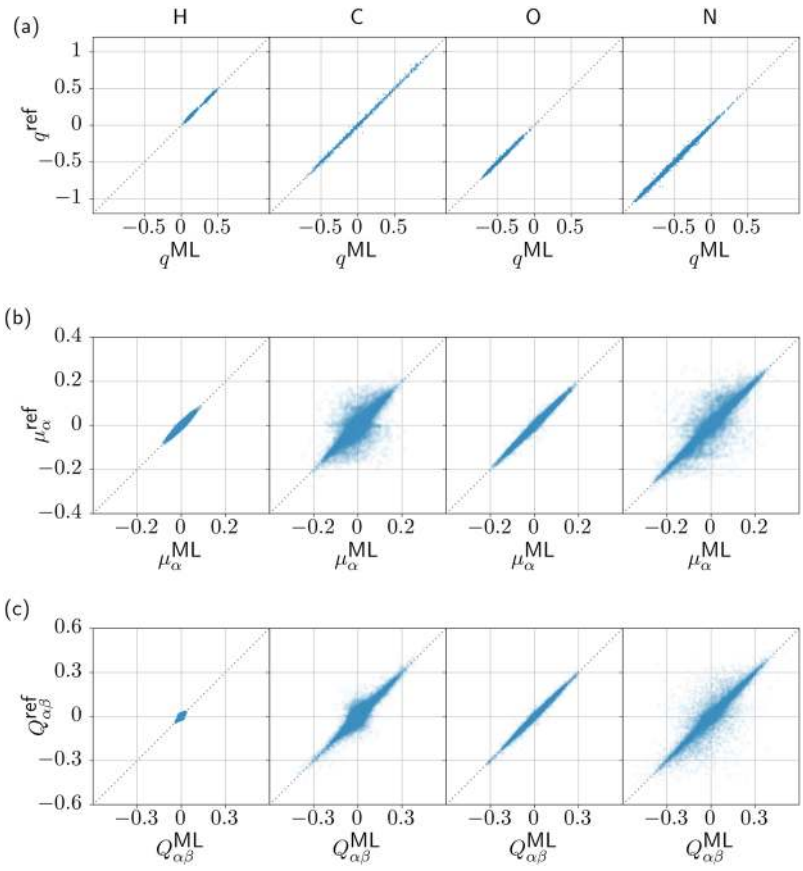

FIG. 1. ML of the multipole coefficients of neutral molecules. Scatter correlation plots (out-of-sample predictions) for all components of (a) monopoles, (b) dipoles, and (c) quadrupoles of each chemical element, as predicted by the ML model with 80% training fraction. All quantities are expressed in units $e\text{Å}^l$, where l is the rank of the multipole.

in the test set. Compared to our previous report,[18] the accuracy of the learning procedure is strongly improved for all ranks, i.e., mean-absolute errors (MAEs) of 0.01 e, 0.01 eÅ, and 0.02 eÅ^2 instead of 0.04 e, 0.06 eÅ, and 0.13 eÅ^2 for monopoles, dipoles, and quadrupoles, respectively. The basis-set projection used here yields significantly more accurate predictions compared to the previously reported local-axis system augmented by a delta-learning procedure.[18] We also point out the strong improvement due to aSLATM (see below). Finally, we draw the reader's attention to the much smaller MBIS multipoles, as compared to GDMA, thereby helping reaching lower MAEs.

Figure 2 displays learning curves for the different multipole moments of each chemical element. It compares the two representations considered in this work: (a) Coulomb matrix and (b) aSLATM. The latter performs significantly better for point charges. Though we reach excellent accuracy for the monopoles, some of the higher multipoles remain more difficult, namely, C and N. On the other hand, H and O both display excellent accuracy. The main difference between these two types of elements lies in their valency: H and O are often found as terminal atoms, while N and C display much more complex local environments. This likely affects the performance of the basis-set projection used in this work. The similar learning efficiency between the Coulomb matrix and aSLATM for dipoles and quadrupoles further suggests the need for larger training sets (e.g., Faber *et al.*, went up to 120 000 samples[51]) or better local projections. We note the existence of ML methodologies that explicitly deal with tensorial objects, though only applied to dipoles so far.[52,53] In Appendix B, we extend Glielmo *et al.*'s covariant-kernel description to quadrupoles using atom-centered Gaussian functions. Tests on small training sets indicated results on par with Fig. 2. We suspect that while covariant kernels offer a more robust description of the rotational properties of tensorial objects, the Coulomb matrix and aSLATM offer more effective representations, offsetting overall the results. Furthermore, the construction of covariant kernels is computationally involved: it requires several outer products of rotation matrices to construct a 9 × 9 matrix [Eqs. (B6) and (B7)] for a quadrupole alone. This significant computational overhead led us to use aSLATM with the basis-set projection for the rest of this work. Covariant kernels for multipoles up to quadrupoles are nonetheless implemented in our Python-based software.

B. Training of valence atomic densities

The accuracy of prediction of the populations and decay rates of valence atomic densities, N and σ, respectively, for

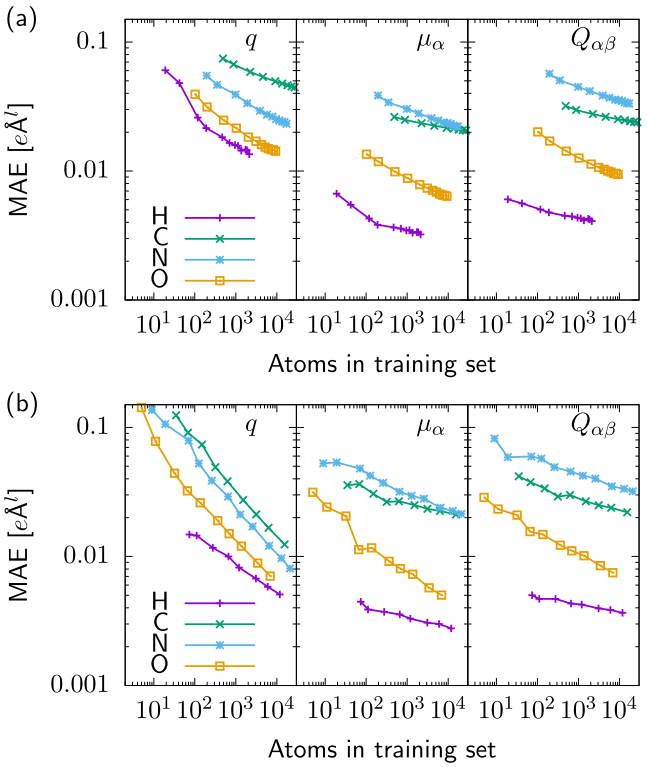

FIG. 2. ML of the multipole coefficients of neutral molecules. Comparison of representations: (a) Coulomb matrix and (b) aSLATM. Saturation curves of the mean-absolute error (MAE) for monopoles, dipoles, and quadrupoles of each chemical element.

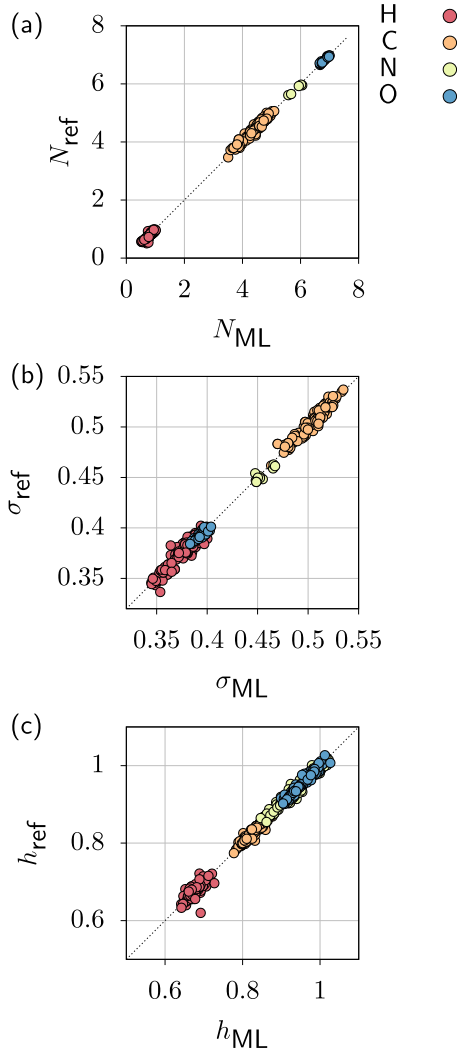

FIG. 3. Correlation plots of out-of-sample predictions. (a) ML of the populations (i.e., volume integral) of the valence atomic densities, N (units in e). (b) ML of the decay rate of the valence atomic densities, σ (units in a.u.$^{-1}$). (c) ML of the Hirshfeld ratios, h.

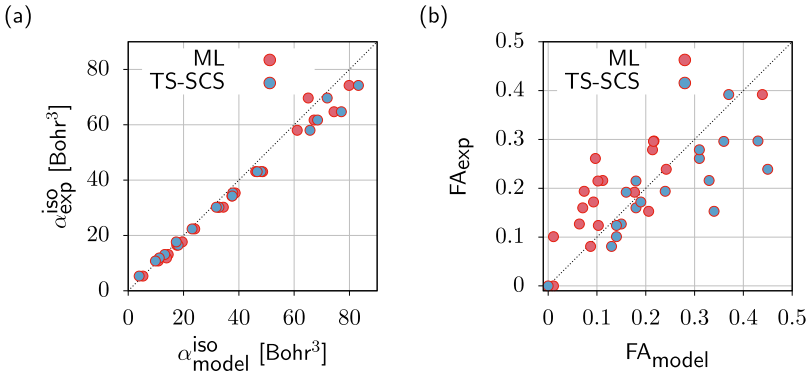

FIG. 4. Correlation plot between (a) isotropic and (b) fractional anisotropies of molecular polarizabilities predicted from the current ML model (blue) and Tkatchenko-Scheffler polarizabilities after SCS procedure[31,54] (red) against experimental values for the set of 18 compounds proposed in Ref. 45.

a size of the Coulomb matrix $n = 6$ is shown in Figs. 3(a) and 3(b). The model was trained against 13 500 atoms in 800 molecules and tested against a separate set of 3400 atoms in 200 molecules. The model shows high accuracy with MAEs of only 0.04 e and 0.004 a.u.$^{-1}$, respectively. Both models yield correlation coefficients above 99.5%.

C. Training of Hirshfeld ratios

Figure 3(c) shows a correlation plot of the predicted and reference Hirshfeld ratios using the $n = 12$ (i.e., size of the Coulomb matrix) model trained against 12 300 atoms in 1000 small organic molecules. We test the prediction accuracy on a different set of 17 100 atoms. We find high correlation (coefficient of determination $R^2 = 99.5\%$) and a small MAE of 0.006.

D. Molecular polarizabilities

Predictions of the Hirshfeld ratios were further assessed by calculating (anisotropic) molecular polarizabilities. Reference experimental values of 18 small molecules were taken from the work of Thole,[45] for both the isotropic molecular polarizability as well as the fractional anisotropy, as defined elsewhere.[32] Figure 4 shows both the isotropic [panel (a)] and fractional anisotropy [panel (b)], comparing the present ML prediction with the calculations using the Tkatchenko-Scheffler method after solving the self-consistent screening (SCS) equation.[31,54] We find excellent agreement between the ML prediction and experiment for the isotropic component: an MAE of 3.2 bohr^3 and a mean-absolute relative error (MARE) of 8.6%, both virtually identical to the Tkatchenko-Scheffler calculations after SCS.[54] The fractional anisotropy tends to be underestimated, though overall the agreement with experiment is reasonable, as compared to previous calculations that explicitly relied on DFT calculations for each compound.

E. Parametrization of the intermolecular energies

To optimize the abovementioned free parameters, we aimed at reproducing the intermolecular energies of a representative set of molecular dimers. The collection of global parameters optimized during this work are reported in Table I. The parameters, shown in Table I, were optimized simultaneously using basin hopping[55,56] to reproduce the total intermolecular energy from reference calculations. We also provide a rough estimate of the sensitivity of these parameters through the standard deviation of all models up to 20% above the identified global minimum. We introduce chemical-element-specific prefactors for the repulsion interaction. The repulsive interaction is thus scaled by the *product* of element specific prefactors for each atom pair. The apparent lack of dependence of the dispersion parameter d led us to fix it to the value $d = 3.92$.[32]

A better understanding of the variability of our global parameters led us to consider two sets of reference datasets for fitting, coined below, model 1 and model 2. While model 1 only considers small-molecule dimers, model 2 also incorporates host-guest complexes. For both models, we rely on the S22x5 small-molecule dataset[20,57] at the equilibrium distance (i.e., 1.0× distance factor). In addition, model 1 also considers configurations at the shorter distance factor 0.9× to help improve the description of the curvature of the potential energy landscape. Model 2, on the other hand, adds to S22x5 at 1.0× a series of host-guest complexes: the S12L database.[58] All the results presented below will be derived from model 1, unless otherwise indicated. The comparison with model 2 aims at showing (i) the robustness of the fit from the relatively low variability of global parameters (except possibly for U_H) and (ii) an outlook toward modeling condensed-phase systems.

TABLE I. Optimized global parameters determined from two different training sets. Model 1: fitting to the S22x5 at distances 0.9× and 1.0×. Model 2: fitting to the S22x5 at distance 1.0× and S12L. Parameters U_X correspond to the repulsion of chemical element X, expressed in (kcal/mol)$^{1/2}$. "Value" corresponds to the optimal parameter, while "sensitivity" reflects the standard deviation of parameters around (up to 20% above) the identified global minimum. Sensitivity is not provided for d (see the main text).

		Model 1		Model 2	
Interaction	Parameter	Value	Sensitivity	Value	Sensitivity
Polarization	a	0.0187	0.09	0.0193	0.03
Dispersion	γ	0.9760	0.04	0.9772	0.04
	β	2.5628	0.08	2.2789	0.04
	d	3.92		3.92	
Repulsion	U^{rep}_H	27.3853	1	23.5936	1
	U^{rep}_C	24.6054	0.5	24.0509	0.5
	U^{rep}_N	22.4496	0.6	21.4312	0.3
	U^{rep}_O	16.1705	0.8	16.0782	0.2

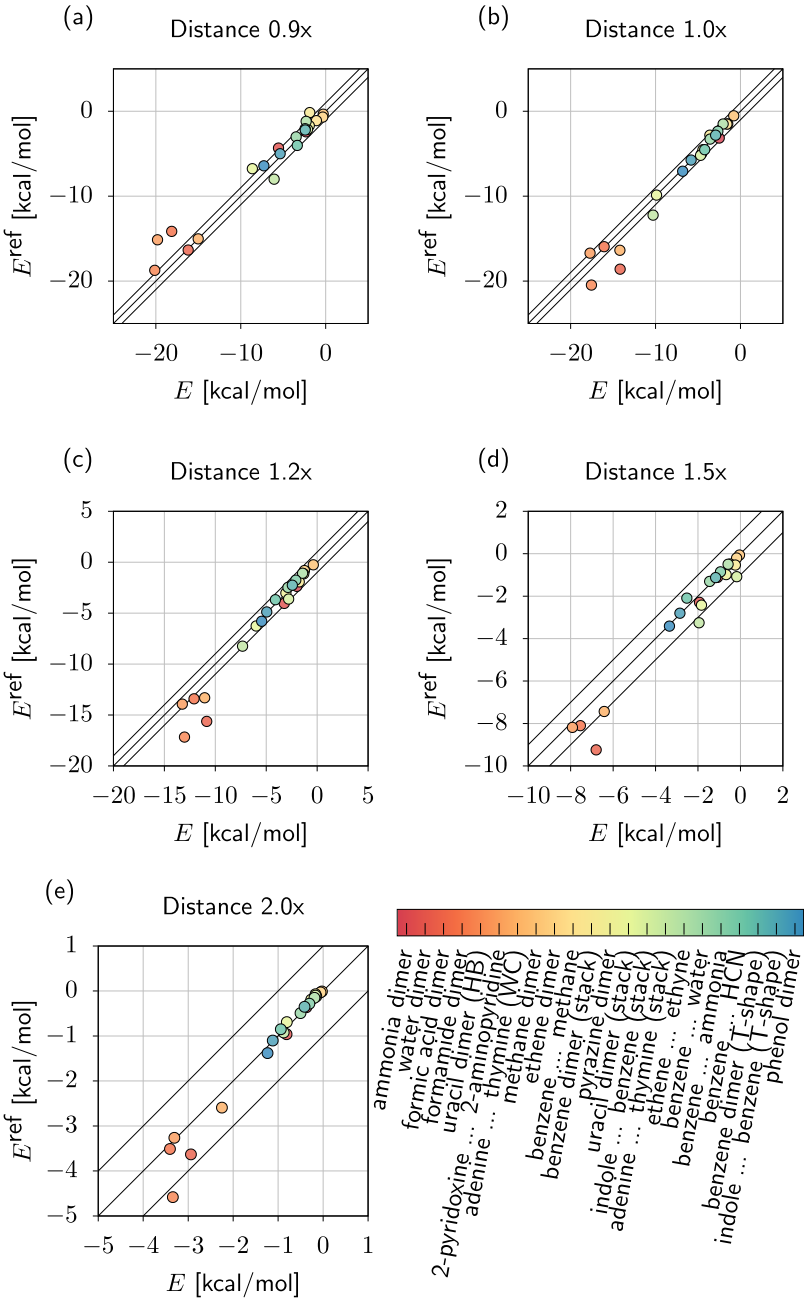

FIG. 5. Correlation of intermolecular energies for S22x5. The different panels describe the interactions at specific distance factors (i.e., from 0.9× to 2.0×). Color coding corresponds to the compound ID—hydrogen bonding compounds correspond to low values, while van der Waals compounds correspond to the larger values. The different diagonals bracket the ±1 kcal/mol area of accuracy.

While the overall MAE averaged over all distance factors is 0.7 kcal/mol, the error clearly drops with distances 1.0, 0.8, 0.8, 0.5, and 0.2 for distance factors 0.9×, 1.0×, 1.2×, 1.5×, and 2.0×, respectively (Fig. 5). This illustrates that the model yields robust asymptotics, with significant improvement compared to a cruder model that only included multipole electrostatics and many-body dispersion.[32] Outliers from the ±1 kcal/mol accuracy region are composed of strongly hydrogen-bonding complexes (e.g., 2-pyridoxine with 2-aminopyridine), which depend significantly on the quality of the electrostatic description. The correlation achieved here depends critically on the accuracy of the multipole moments. Indeed, the few global parameters included in our model provide little room for error compensations. For instance, we found that a poorer ML model of the multipole moments yielded significant artifacts on the partial charges of hydrogen cyanide, leading to an artificially strong polarization of the hydrogen.

We also point out the small value of the polarization parameter, a (Table I), leading effectively to small polarization energies. Rather than an imbalance in the model, we suspect that significant short-range polarization energy is absorbed in the repulsion terms. Indeed, several AIM- and physics-based force fields use the same overlap model to describe repulsion and short-range polarization.[4,5] Since we optimize all terms directly against the total energy rather than decompose each term, such cancellations may well occur. We also expect that including systems in which strong non-additive polarization effects would play a role in outweighing effective pairwise polarization. In addition, we note that the pairwise scheme is optimized per chemical element, while the Thole model is not.

IV. PERFORMANCE OF THE IPML MODEL

A. Non-equilibrium geometries (S66a8)

A recent extension of the S66 dataset of molecular dimers provides angular-displaced non-equilibrium geometries, i.e., S66a8 (66 × 8 = 528 dimers).[59] The correlation between our model and reference calculations using coupled cluster singles doubles perturbative triples at the complete basis-set limit (CCSD(T)/CBS) are presented in Fig. 6(a). Excellent agreement is found for most samples, with an MAE of only 0.4 kcal/mol across a larger, representative set of molecular dimers, as compared to the S22 used for training. Model 2 performs virtually on par with an MAE of only 0.5 kcal/mol.

We compare our results with the MEDFF model whose overlap model is used in the present work but relies on point-charge electrostatics and a pairwise dispersion model.[5] They report root-mean squared errors of 0.36 kcal/mol for the dispersion-dominated complexes of the S66 dataset at equilibrium distances. Given that hydrogen-bonded complexes are typically more challenging,[1,5] our model likely compares favorably, keeping in mind that the dataset and error measurement are different. They also report a reduced 0.26 kcal/mol error over the entire S66 dataset when each parameter is optimized specifically for each complex. Given our focus on model transferability, we did not attempt a similar measurement. For the same dataset and error measurement, the QMDFF model reports a larger 1.1 kcal/mol error.[5]

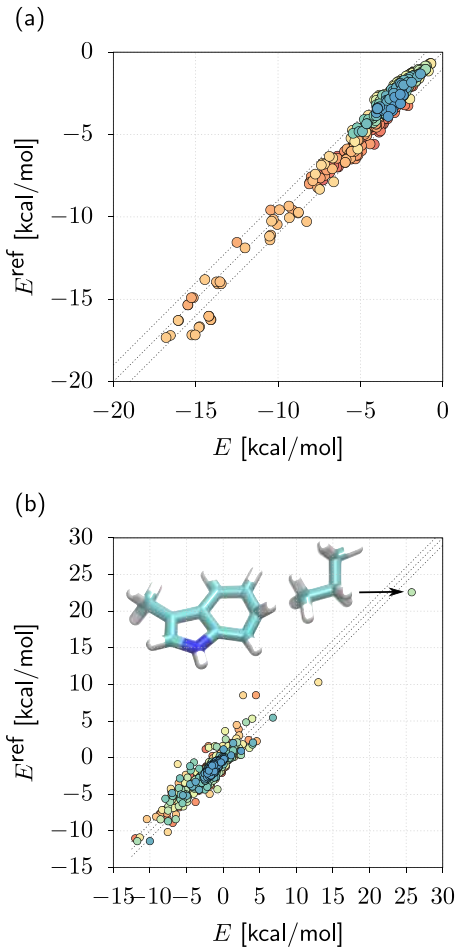

FIG. 6. Correlation plots for the total intermolecular energy between reference and present calculations for (a) the S66a8 dataset of dimers translated and rotated away from their equilibrium geometry and (b) the SSI dataset of amino acids (only dimers involving neutral compounds made of HCON atoms). Inset: strongly-repulsive tryptophan-glutamine dimer.

B. Amino-acid side chains (SSI dataset)

The SSI dataset contains pairs of amino-acid side chains extracted from the protein databank.[60] We removed dimers containing charged compounds and sulfur-containing side chains (i.e., cysteine and methionine), for a total of 2216 dimers. We computed intermolecular energies using the present method and compare them with reference CCSD(T) at the complete basis set limit. In Fig. 6(b), we compare the total energy with reference energies. We find again excellent agreement throughout the much larger range. We note the presence of a high-energy dimer at +23 kcal/mol, corresponding to a tryptophan-glutamine dimer [inset of Fig. 6(b)]. The strong deformation of the tryptophan ring illustrates the robustness of our model in accurately reproducing intermolecular interactions for a variety of conformers. Model 1 yields overall an MAE of 0.37 kcal/mol. Interestingly, this accuracy is on par with additive force fields, such as GAFF and CGenFF (0.35 and 0.23 kcal/mol, respectively), and better than certain semi-empirical methods, e.g., AM1 (1.45 kcal/mol).[60] Model 2 yields virtually the same MAE, 0.38 kcal/mol, but underpredicts the high-energy dimer highlighted in Fig. 6(b), 3.6 instead of 22.6 kcal/mol. It highlights how widening the training set of the model to both small molecules and host-guest complexes decreases the accuracy on the former.

C. DNA-base and amino-acid pairs (JSCH-2005)

The JSCH-2005 dataset offers a benchmark of representative DNA base and amino-acid pairs.[20] Again, we focus on neutral molecules only, for a total of 127 dimers. The correlation of total interaction energies is shown in Fig. 7(a). We find a somewhat larger MAE of 1.4 kcal/mol. This result remains extremely encouraging, given the emphasis of strong hydrogen-bonded complexes present in this dataset. While others have pointed out the challenges associated with accurately modeling these interactions,[1,5] we have not found reference benchmarks on specific datasets such as this one for similar physics-based models. Given the prevalence of hydrogen bonds in organic and biomolecular systems, we hope that this work will motivate a more systematic validation on these interactions.

Representative examples are shown on Fig. 7. While the Watson-Crick complex of the guanine (G) and cytosine (C) dimer [panel (b)] leads to one of the strongest binders, weak hydrogen bonds can still lead to the dominant contribution, as seen in (f) for the methylated GC complex. We find two outliers, shown in (d) and (e), where π-stacking interactions dominate the interaction energy. The discrepancies likely arise from an inadequate prediction of some quadrupole moments, especially involving nitrogen (see Fig. 1). Note the structural similarity between (d)–(f): the weak hydrogen bonds in the latter case dominate the interaction and resolve any apparent discrepancy with the reference energy. For this dataset, model 2 performs significantly worse, with an MAE of 2.3

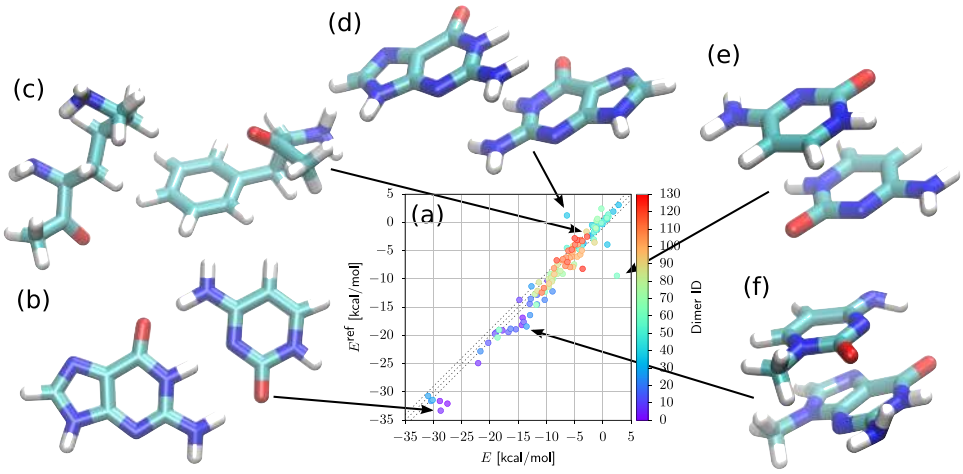

FIG. 7. (a) Correlation plots for the total intermolecular energy between reference and present calculations for the JSCH-2005 dataset[20] of DNA-base and amino-acid pairs (dimers involving charged compounds are not shown); (b) GC in a Watson-Crick geometry; (c) lysine and phenylalanine; (d) GG complex; (e) CC complex; (f) methylated GC complex.

kcal/mol, indicating that forcing transferability across both small-molecule dimers and host-guest complexes strains the accuracy of the model for challenging small molecules exhibiting significant π-stacking and hydrogen-bonding behavior. This significant change in performance contrasts the very similar parameters between the two models, highlighting a sensitive parameter dependence.

D. Water clusters

Beyond dimers, we test the ability of our potentials to reproduce energies of larger clusters. Figure 8(a) shows the correlation of the total energy between the present work and CCSD(T) calculations at the complete basis set limit of water clusters involving from 2 to 10 molecules.[61] The model's energies correlate highly with the reference but progressively overstabilize. This shift results from compounding errors that grow with cluster size, amounting to an MAE of 8.1 kcal/mol. Note that we can correct the slope by including a single water cluster in the above-mentioned parametrization (data not shown). Model 2 performs virtually on par with model 1.

IPML recovers the overall trend of energies for complexes of various sizes, but there is still room for improvements. This is notable given that the many-body polarization term was optimized to zero in both models (see Table I). It indicates that a pairwise description captures the main effects even for the larger complexes considered here. Improving the results would require forcing the parametrization to rely more significantly on many-body polarization. Improving the modeling of other terms, such as repulsion, may also help reduce incidental cancellations of errors.

E. Supramolecular complexes (S12L)

Moving toward more complex systems, we test the ability to reproduce intermolecular energies of host-guest complexes. Figure 8(b) shows the correlation of the total intermolecular energy against diffusion Monte Carlo.[58] Although we find high correlation, the MAE is substantial: 9.7 kcal/mol. A comparison with model 2, which significantly improves the agreement, demonstrates the benefit of including larger complexes in the fit of the global parameters. Still, one outlier remains: the glycine anhydride-macrocycle, with an overstabilization of 8 kcal/mol, despite being fitted into the global parameters. This compound (displayed in Fig. 8 of Ref. 32) displays sites at which multiple hydrogen bonds coincide. It further suggests the role of inaccurate multipoles, as well as an inadequate electrostatic penetration model (i.e., missing higher-order multipoles beyond monopole correction), and possibly many-body repulsion interactions.

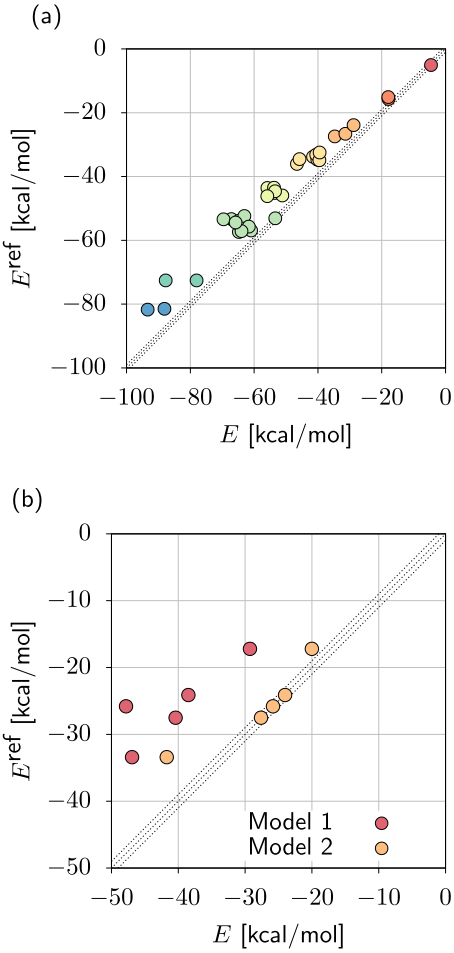

FIG. 8. Correlation plots for the total intermolecular energy between the reference and present calculations for (a) the water-clusters dataset and (b) the host-guest complexes in the S12L database. The colors in (a) indicate the number of molecules involved in the cluster: from two (red) to 10 (blue) molecules.

F. Benzene crystal

As another example leading to condensed-phase properties, we evaluate the model's ability to reproduce the cohesive energy of the benzene crystal. We scale the lattice unit cell around the equilibrium value, as detailed in previous work.[32] The various contributions of the energy are shown in Fig. 9(c). For reference, we compare the cohesive energy with the experimental results[62] and dispersion-corrected atom-centered potentials (DCACP).[63]

As reported before,[32,64] we find the benzene crystal to display significant dispersion interactions. Though the overall curvature against density changes agrees reasonably well with DCACP, we find that the method overstabilizes the molecular crystal. Model 1 yields a cohesive energy of −17.2 kcal/mol at equilibrium, as compared to the experimental value of −12.2 kcal/mol.[62] For reference, we show the potential energy landscapes of the benzene dimer in the stacked (a) and T-shaped (b) conformations. Excellent agreement is found in the latter case, while the former shows an overstabilization.

Interestingly, while model 2 seems to understabilize these two dimer configurations, it better reproduces the cohesive energy of the crystal, with a value at equilibrium density of −14.3 kcal/mol, only 2 kcal/mol away from the experimental value. We conclude that the inclusion of host-guest complexes in the optimization of the global parameters helps describe systems toward or in the condensed phase. Still, the compounding errors present in the model limit a systematic extension to molecular crystals. We again point at the necessity for extremely accurate multipole moments, where any discrepancy can have significant effects in the condensed phase.

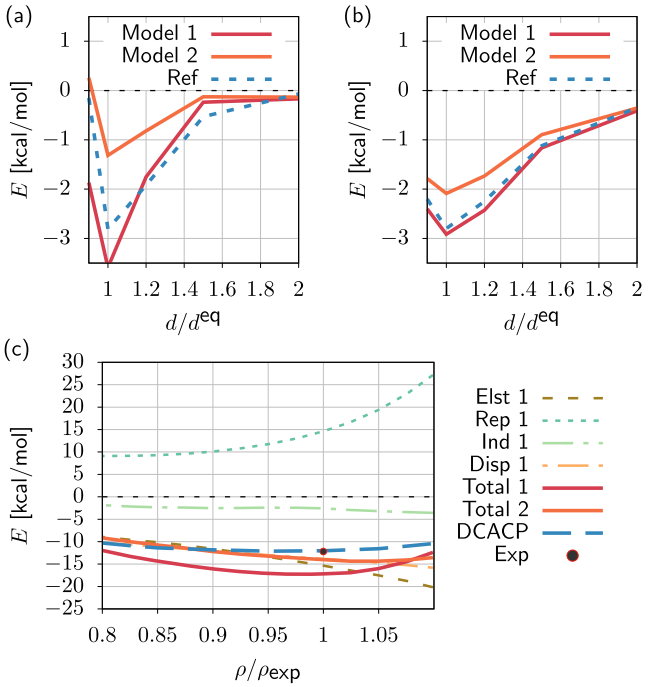

FIG. 9. Comparison of the intermolecular energy as a function of the dimer distance for the benzene dimer in the (a) parallel-displaced and (b) T-shaped conformations. (c) Cohesive binding energy of the benzene crystal as a function of the scaling factor of the unit cell.

Further improving the prediction of the multipole moments will strongly contribute to an improved accuracy of the present energy model.

V. CONCLUSIONS AND FUTURE OUTLOOK

We have presented a set of classical potentials to describe the intermolecular interactions of small molecules, coined IPML. Notably, we present a methodology that readily provides parameters for a large range of small molecules by relying on atom-in-molecule properties predicted from machine learning (ML). Predictions for distributed multipoles, Hirshfeld ratios, valence atomic density decay rate, and population provide the necessary parameters for electrostatics, polarization, repulsion, and many-body dispersion. Remarkably, our methodology provides a first attempt at transferable intermolecular potentials with few global parameters optimized across a subset of chemical space containing H, C, N, and O atoms only. In contrast to other studies, we do not reoptimize the global parameters for every new compound. We rationalize this by the use of more sophisticated physical models, e.g., many-body rather than pairwise dispersion, multipole rather than point-charge electrostatics, and non-additive rather than pairwise additive polarization.

As compared to purely data-driven methodologies, IPML starts from physics-based interactions and only relies on ML to predict parameters thereof. Perturbation theory and the short-range overlap method offer an appealing framework to describe interactions based on monomer properties—effectively simplifying greatly the training of ML models of parameters. Conceptually, inserting physical constraints in an ML model would ideally take the form of specific prior probability distributions. As an example, reproducing kernel Hilbert space can fit a potential energy surface by imposing the asymptotics at long range.[65,66]

Extensions of the present work to a force field would amount to computing derivatives. Analytical derivatives of the potentials with respect to atomic coordinates are either straightforward (e.g., pairwise repulsion and charge penetration) or already available (e.g., many-body dispersion[67] or electrostatics and induction[68]). Our ML models being conformationally dependent, computation of the forces would also entail a derivative with respect to the atom-in-molecule properties. While not implemented here, this information can readily be extracted from derivatives of the kernel used in the ML.[69] How to optimize such a conformationally dependent force field to best balance the extra accuracy with the additional computational overhead remains an open problem.

Even though we did not aim at a performance optimization, the present implementation can help us gain insight into the computational cost of each term. Compared to standard classical force fields, the inclusion of explicit polarization and many-body dispersion leads to larger evaluation times: 1–100 s for systems composed of 10–100 atoms on a single core, respectively. Notably, roughly 90% of this time is spent predicting the multipoles, due to the large training set and complexity of the aSLATM representation. While such an evaluation time is significant, several strategies may be devised in the context of a molecular dynamics simulation. For instance,

multipoles may remain frozen and only get updated when large conformational changes are detected.

We presented electrostatic calculations using distributed multipole—up to quadrupole—models. In comparison with other atomic properties, an accurate prediction of multipole electrostatics proves all the more challenging and critical for the accurate estimation of various molecular systems. Improvements will require more accurate models, and possibly the incorporation of more advanced physical interactions, such as anisotropic[70] or many-body repulsion interactions. Our framework paves the way toward significantly more transferable models that blend in the physical laws and symmetries relevant for the phenomena at hand with a data-driven approach to infer the variation of environmentally dependent local atomic parameters across chemical space. We expect such models that are transferable across chemical composition to be of use in systems of interest in chemistry, biology, and materials science.

ACKNOWLEDGMENTS

We thank Denis Andrienko, Omar Valsson, and Alessandro de Vita for critical discussions and Lori A. Burns and C. David Sherrill for access to the SSI database.

T.B. acknowledges funding from an Emmy Noether Fellowship of the German Research Foundation (DFG). R.D. acknowledges partial support from Cornell University through startup funding and the Cornell Center for Materials Research with funding from the NSF MRSEC Program (No. DMR-1719875). A.T. acknowledges funding from the European Research Council (ERC Consolidator Grant BeStMo). O.A.v.L. acknowledges funding from the Swiss National Science Foundation (Nos. PP00P2_138932 and 407540_167186 NFP 75 Big Data). This research was partly supported by the NCCR MARVEL, funded by the Swiss National Science Foundation. Part of this research was performed during the long program Understanding Many-Particle Systems with Machine Learning at the Institute for Pure and Applied Mathematics (IPAM).

APPENDIX A: MANY-BODY DISPERSION

The following summarizes the many-body dispersion (MBD) method[31,46,47] as implemented elsewhere.[32] We start with the atomic polarizability α_p of atom p. The frequency dependence of α_p allows for an estimation of the pairwise dispersion coefficient via the Casimir-Polder integral,

$$C_{6pq} = \frac{3}{\pi}\int_0^\infty \mathrm{d}\omega \alpha_p(i\omega)\alpha_q(i\omega), \tag{A1}$$

where $i\omega$ are imaginary frequencies and p and q are a pair of atoms. Given reference free-atom values for C_{6pp}, we can estimate the characteristic frequency of atom $p\omega_p = 4C_{6pp}/3\alpha_p^2$.[71]

The atomic polarizabilities and characteristic frequencies yield the necessary ingredients for the system of coupled quantum harmonic oscillators with N atoms,

$$C_{pq}^{\mathrm{QHO}} = \omega_p^2\delta_{pq} + (1-\delta_{pq})\omega_p\omega_q\sqrt{\alpha_p\alpha_q}\mathcal{T}_{pq}, \tag{A2}$$

where $\mathcal{T}_{pq} = \nabla_{\mathbf{r}_p}\otimes\nabla_{\mathbf{r}_q}W(r_{pq})$ is a dipole interaction tensor with modified Coulomb potential

$$W(r_{pq}) = \frac{1-\exp\left[-\left(\frac{r_{pq}}{R_{pq}^{\mathrm{vdW}}}\right)^\beta\right]}{r_{pq}}. \tag{A3}$$

In this equation, β is a range-separation parameter and $R_{pq}^{\mathrm{vdW}} = \gamma(R_p^{\mathrm{vdW}} + R_q^{\mathrm{vdW}})$ is the sum of effective van der Waals radii scaled by a chemistry-independent fitting parameter. The effective van der Waals radius is obtained by scaling its reference free-atom counterpart: $R_p^{\mathrm{vdW}} = (\alpha_p/\alpha_p^{\mathrm{free}})^{1/3}R_p^{\mathrm{vdW,\,free}}$. An expression for $\mathcal{T}_{pq}$ is provided in the work of Bereau and von Lilienfeld.[32] In particular, we apply a range separation to the dipole interaction tensor by scaling it by a Fermi function[72]

$$f(r_{pq}) = \frac{1}{1+\exp\left[-d(r_{pq}/R_{pq}^{\mathrm{vdW}}-1)\right]}. \tag{A4}$$

Diagonalizing the $3N\times 3N$ matrix C_{pq}^{QHO} yields its eigenvalues $\{\lambda_i\}$, which in turn provide the MBD energy,

$$E_{\mathrm{MBD}} = \frac{1}{2}\sum_{i=1}^{3N}\sqrt{\lambda_i} - \frac{3}{2}\sum_{p=1}^{N}\omega_p. \tag{A5}$$

The methodology depends on three chemistry-independent parameters: β, γ, and d.

APPENDIX B: COVARIANT KERNELS

Glielmo *et al.*[52] recently proposed a covariant kernel $\mathbf{K}^\mu$ for vector quantities—suitable here to predict dipoles—such that two samples ρ and ρ' subject to rotations $\mathcal{S}$ and $\mathcal{S}'$, respectively, will obey

$$\mathbf{K}^\mu(\mathcal{S}\rho,\mathcal{S}'\rho') = \mathbf{S}\mathbf{K}^\mu(\rho,\rho')\mathbf{S}'^{\mathrm{T}}. \tag{B1}$$

The atom i from sample ρ is encoded by a set of atom-centered Gaussian functions

$$\rho(\mathbf{r},\{\mathbf{r}_i\}) = \frac{1}{(2\pi\sigma^2)^{3/2}}\sum_i \exp\left(-\frac{||\mathbf{r}-\mathbf{r}_i||^2}{2\sigma^2}\right), \tag{B2}$$

and the covariant kernel is analytically integrated over all 3D rotations to yield[52]

$$\mathbf{K}^\mu(\rho,\rho') = \frac{1}{L}\sum_{ij}\phi(r_i,r_j)\mathbf{r}_i\otimes\mathbf{r}_j'^{\mathrm{T}},$$

$$\phi(r_i,r_j) = \frac{\exp\left(-\alpha_{ij}\right)}{\gamma_{ij}^2}\left(\gamma_{ij}\cosh\gamma_{ij}-\sinh\gamma_{ij}\right),$$

$$L = (2\sqrt{\pi\sigma^2})^3,\quad \alpha_{ij} = \frac{r_i^2+r_j^2}{4\sigma^2},\quad \gamma_{ij} = \frac{r_ir_j}{2\sigma^2}, \tag{B3}$$

where $\otimes$ denotes the outer product.

In the present work, we extend the construction of covariant kernels to predict quadrupole moments. Following a similar procedure adapted to second-rank tensors, we enforce the relation

$$\mathbf{K}^{\mathrm{Q}}(\mathcal{S}\rho,\mathcal{S}'\rho') = \mathbf{S}'\mathbf{S}^{\mathrm{T}}\mathbf{K}^{\mathrm{Q}}(\rho,\rho')\mathbf{S}\mathbf{S}'^{\mathrm{T}} \tag{B4}$$

onto a base pairwise kernel of diagonal form $\mathbf{K}^{\mathrm{b}}(\rho,\rho') = \mathbb{1}k^{\mathrm{b}}(\rho,\rho')$, where $k^{\mathrm{b}}(\rho,\rho')$ is independent of the reference

frame. The covariant kernel is constructed by integrating the base kernel over all 3D rotations

$$\mathbf{K}^{\mathrm{Q}}(\rho,\rho') = \frac{1}{L}\sum_{ij}\int \mathrm{d}\mathcal{S}\mathbf{S}\otimes\mathbf{S}^{\mathrm{T}}k^{\mathrm{b}}(\rho,\mathcal{S}^{\mathrm{T}}\rho'), \tag{B5}$$

which leads to the expression

$$\begin{aligned}\mathbf{K}^{\mathrm{Q}}(\rho,\rho') &= \frac{1}{L}\sum_{ij}\left(\mathbf{R}_j^{\mathrm{T}}\otimes\mathbf{R}_i\right)\Phi(r_i,r_j)\left(\mathbf{R}_i{}^{\mathrm{T}}\otimes\mathbf{R}_j\right),\\ \Phi(r_i,r_j) &= \int \mathrm{d}\tilde{\mathcal{R}}\tilde{\mathbf{R}}^{\mathrm{T}}\otimes\tilde{\mathbf{R}}k^{\mathrm{b}}(\tilde{\mathbf{r}}_i,\tilde{\mathbf{R}}\tilde{\mathbf{r}}'_j),\end{aligned} \tag{B6}$$

where $\mathbf{R}_i$ and $\mathbf{R}_j$ are the rotation matrices that align $\mathbf{r}_i$ and $\mathbf{r}_j$ onto the z axis to form $\tilde{\mathbf{r}}_i$ and $\tilde{\mathbf{r}}'_j$, respectively.[52] We analytically integrate all 3D rotations

$$\begin{aligned}\Phi(r_i,r_j) &= \mathrm{e}^{\frac{-\alpha_{ij}^2}{4\sigma^2}}\int \mathrm{d}\alpha\int \mathrm{d}\beta\int \mathrm{d}\gamma\frac{\sin\beta}{8\pi^2}\\ &\quad\times\mathbf{R}^{\mathrm{T}}(\alpha,\beta,\gamma)\otimes\mathbf{R}(\alpha,\beta,\gamma)\mathrm{e}^{\frac{r_ir_j\cos\beta}{2\sigma^2}}\\ &= \begin{pmatrix}\varphi_1 & 0 & 0 & 0 & \varphi_2 & 0 & 0 & 0 & 0\\ 0 & \varphi_1 & 0 & -\varphi_2 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & \varphi_3 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & -\varphi_2 & 0 & \varphi_1 & 0 & 0 & 0 & 0 & 0\\ \varphi_2 & 0 & 0 & 0 & \varphi_1 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & \varphi_3 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & \varphi_3 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \varphi_3 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \varphi_4\end{pmatrix},\end{aligned} \tag{B7}$$

where

$$\begin{aligned}\varphi_1 &= \frac{\mathrm{e}^{\frac{-\alpha_{ij}^2}{4\sigma^2}}}{4\gamma_{ij}^2}\left(\gamma_{ij}^2\sinh\gamma_{ij}-\gamma_{ij}\cosh\gamma_{ij}+\sinh\gamma_{ij}\right),\\ \varphi_2 &= \frac{\mathrm{e}^{\frac{-\alpha_{ij}^2}{4\sigma^2}}}{4\gamma_{ij}}\left(\gamma_{ij}\cosh\gamma_{ij}-\sinh\gamma_{ij}\right),\\ \varphi_3 &= \frac{\mathrm{e}^{\frac{-\alpha_{ij}^2}{4\sigma^2}}}{2\gamma_{ij}^2}\left(\gamma_{ij}\cosh\gamma_{ij}-\sinh\gamma_{ij}\right),\\ \varphi_4 &= \frac{\mathrm{e}^{\frac{-\alpha_{ij}^2}{4\sigma^2}}}{\gamma_{ij}^2}\left(\frac{\gamma_{ij}^2}{2}\sinh\gamma_{ij}-\gamma_{ij}\cosh\gamma_{ij}+\sinh\gamma_{ij}\right).\end{aligned} \tag{B8}$$

[1] S. Grimme, J. Chem. Theory Comput. **10**, 4497 (2014).
[2] M. P. Metz, K. Piszczatowski, and K. Szalewicz, J. Chem. Theory Comput. **12**, 5895 (2016).
[3] B. Jeziorski, R. Moszynski, and K. Szalewicz, Chem. Rev. **94**, 1887 (1994).
[4] M. J. Van Vleet, A. J. Misquitta, A. J. Stone, and J. Schmidt, J. Chem. Theory Comput. **12**, 3851 (2016).
[5] S. Vandenbrande, M. Waroquier, V. V. Speybroeck, and T. Verstraelen, J. Chem. Theory Comput. **13**, 161 (2017).
[6] D. J. Cole, J. Z. Vilseck, J. Tirado-Rives, M. C. Payne, and W. L. Jorgensen, J. Chem. Theory Comput. **12**, 2312 (2016).
[7] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Phys. Rev. Lett. **104**, 136403 (2010).
[8] Z. Li, J. R. Kermode, and A. De Vita, Phys. Rev. Lett. **114**, 096405 (2015).
[9] J. Behler, J. Chem. Phys. **145**, 170901 (2016).
[10] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller, Sci. Adv. **3**, e1603015 (2017).
[11] V. Botu, R. Batra, J. Chapman, and R. Ramprasad, J. Phys. Chem. C **121**, 511 (2016).
[12] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, Nat. Commun. **8**, 13890 (2017).
[13] S. K. Natarajan, T. Morawietz, and J. Behler, Phys. Chem. Chem. Phys. **17**, 8356 (2015).
[14] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. Von Lilienfeld, Phys. Rev. Lett. **108**, 058301 (2012).
[15] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. Von Lilienfeld, A. Tkatchenko, and K.-R. Müller, J. Chem. Theory Comput. **9**, 3404 (2013).
[16] R. Ramakrishnan and O. A. von Lilienfeld, "Machine learning, quantum chemistry, and chemical space," in *Reviews in Computational Chemistry* (John Wiley & Sons, Inc., 2017), pp. 225–256.
[17] T. Bereau and K. Kremer, J. Chem. Theory Comput. **11**, 2783 (2015).
[18] T. Bereau, D. Andrienko, and O. A. von Lilienfeld, J. Chem. Theory Comput. **11**, 3225 (2015).
[19] A. Stone, *The Theory of Intermolecular Forces* (Oxford University Press, 2013).
[20] P. Jurečka, J. Šponer, J. Černỳ, and P. Hobza, Phys. Chem. Chem. Phys. **8**, 1985 (2006).
[21] Y. Zhao and D. G. Truhlar, Theor. Chem. Acc. **120**, 215 (2008).
[22] A. J. Misquitta, A. J. Stone, and F. Fazeli, J. Chem. Theory Comput. **10**, 5405 (2014).
[23] T. Verstraelen, P. Tecmer, F. Heidar-Zadeh, K. Boguslawski, M. Chan, Y. Zhao, T. D. Kim, S. Vandenbrande, D. Yang, C. E. González-Espinoza, S. Fias, P. A. Limacher, D. Berrocal, A. Malek, and P. W. Ayers, HORTON, version 2.0.1, http://theochem.github.com/horton/, accessed 01 Augest 2016.
[24] B. Huang and O. A. von Lilienfeld, J. Chem. Phys. **145**, 161102 (2016).
[25] B. Huang and O. A. von Lilienfeld, preprint arXiv:1707.04146 (2017).
[26] A. S. Christensen, F. A. Faber, B. Huang, L. A. Bratholm, A. Tkatchenko, K. R. Müller, and O. A. von Lilienfeld, QML: A Python Toolkit for Quantum Machine Learning, https://github.com/qmlcode/qml, accessed 01 July 2017.
[27] Y. S. Kim, S. K. Kim, and W. D. Lee, Chem. Phys. Lett. **80**, 574 (1981).
[28] T. C. Lillestolen and R. J. Wheatley, Chem. Commun. **0**(45), 5909 (2008).
[29] T. Verstraelen, S. Vandenbrande, F. Heidar-Zadeh, L. Vanduyfhuys, V. Van Speybroeck, M. Waroquier, and P. W. Ayers, J. Chem. Theory Comput. **12**, 3894 (2016).
[30] F. L. Hirshfeld, Theor. Chim. Acta **44**, 129 (1977).
[31] A. Tkatchenko and M. Scheffler, Phys. Rev. Lett. **102**, 073005 (2009).
[32] T. Bereau and O. A. von Lilienfeld, J. Chem. Phys. **141**, 034101 (2014).
[33] T. Bučko, S. Lebègue, J. G. Ángyán, and J. Hafner, J. Chem. Phys. **141**, 034114 (2014).
[34] V. V. Gobre, "Efficient modelling of linear electronic polarization in materials using atomic response functions," Ph.D. thesis, Technische Universität Berlin, 2016.
[35] C. Adamo and V. Barone, J. Chem. Phys. **110**, 6158 (1999).
[36] F. O. Kannemann and A. D. Becke, J. Chem. Theory Comput. **6**, 1081 (2010).
[37] A. Otero-de-la Roza and E. R. Johnson, J. Chem. Phys. **138**, 054103 (2013).
[38] P. Ren and J. W. Ponder, J. Phys. Chem. B **107**, 5933 (2003).
[39] T. Bereau, C. Kramer, and M. Meuwly, J. Chem. Theory Comput. **9**, 5450 (2013).
[40] B. Wang and D. G. Truhlar, J. Chem. Theory Comput. **6**, 3330 (2010).
[41] J.-P. Piquemal, N. Gresh, and C. Giessner-Prettre, J. Phys. Chem. A **107**, 10353 (2003).
[42] Q. Wang, J. A. Rackers, C. He, R. Qi, C. Narth, L. Lagardere, N. Gresh, J. W. Ponder, J.-P. Piquemal, and P. Ren, J. Chem. Theory Comput. **11**, 2609 (2015).
[43] C. Narth, L. Lagardère, E. Polack, N. Gresh, Q. Wang, D. R. Bell, J. A. Rackers, J. W. Ponder, P. Y. Ren, and J.-P. Piquemal, J. Comput. Chem. **37**, 494 (2016).
[44] J. A. Rackers, Q. Wang, C. Liu, J.-P. Piquemal, P. Ren, and J. W. Ponder, Phys. Chem. Chem. Phys. **19**, 276 (2017).
[45] B. T. Thole, Chem. Phys. **59**, 341 (1981).
[46] J. Hermann, R. A. DiStasio, Jr., and A. Tkatchenko, Chem. Rev. **117**, 4714 (2017).
[47] A. Tkatchenko, R. A. DiStasio, Jr., R. Car, and M. Scheffler, Phys. Rev. Lett. **108**, 236402 (2012).
[48] A. Donchev, J. Chem. Phys. **125**, 074713 (2006).
[49] S. van der Walt, S. C. Colbert, and G. Varoquaux, Comput. Sci. Eng. **13**, 22 (2011).

[50]N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, J. Cheminf. **3**, 33 (2011).
[51]F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley, and O. A. von Lilienfeld, J. Chem. Theory Comput. **13**, 5255 (2017).
[52]A. Glielmo, P. Sollich, and A. De Vita, Phys. Rev. B **95**, 214302 (2017).
[53]A. Grisafi, D. M. Wilkins, G. Csányi, and M. Ceriotti, Phys. Rev. Lett. **120**(3), 036002 (2018).
[54]R. A. DiStasio, Jr., V. V. Gobre, and A. Tkatchenko, J. Phys.: Condens. Matter **26**, 213202 (2014).
[55]D. J. Wales and J. P. Doye, J. Phys. Chem. A **101**, 5111 (1997).
[56]D. J. Wales, *Energy Landscapes: Applications to Clusters, Biomolecules and Glasses* (Cambridge University Press, 2003).
[57]L. Gráfová, M. Pitoňák, J. Řezáč, and P. Hobza, J. Chem. Theory Comput. **6**, 2365 (2010).
[58]A. Ambrosetti, D. Alfè, R. A. DiStasio, Jr., and A. Tkatchenko, J. Phys. Chem. Lett. **5**, 849 (2014).
[59]J. Rezác, K. E. Riley, and P. Hobza, J. Chem. Theory Comput. **7**, 3466 (2011).
[60]L. A. Burns, J. C. Faver, Z. Zheng, M. S. Marshall, D. G. A. Smith, K. Vanommeslaeghe, A. D. MacKerell, Jr., K. M. Merz, Jr., and D. Sherrill, J. Chem. Phys. **147**, 161727 (2017).
[61]B. Temelso, K. A. Archer, and G. C. Shields, J. Phys. Chem. A **115**, 12034 (2011).
[62]W. B. Schweizer and J. D. Dunitz, J. Chem. Theory Comput. **2**, 288 (2006).
[63]E. Tapavicza, I.-C. Lin, O. A. von Lilienfeld, I. Tavernelli, M. D. Coutinho-Neto, and U. Rothlisberger, J. Chem. Theory Comput. **3**, 1673 (2007).
[64]O. A. von Lilienfeld and A. Tkatchenko, J. Chem. Phys. **132**, 234109 (2010).
[65]T.-S. Ho and H. Rabitz, J. Chem. Phys. **104**, 2584 (1996).
[66]O. T. Unke and M. Meuwly, J. Chem. Inf. Model. **57**, 1923 (2017).
[67]M. A. Blood-Forsythe, T. Markovich, R. A. DiStasio, Jr., R. Car, and A. Aspuru-Guzik, Chem. Sci. **7**, 1712 (2016).
[68]J. W. Ponder, C. Wu, P. Ren, V. S. Pande, J. D. Chodera, M. J. Schnieders, I. Haque, D. L. Mobley, D. S. Lambrecht, R. A. DiStasio, Jr. *et al.*, J. Phys. Chem. B **114**, 2549 (2010).
[69]C. E. Rasmussen and C. K. Williams, *Gaussian Processes for Machine Learning* (MIT Press, Cambridge, 2006), Vol. 1.
[70]M. J. Van Vleet, A. J. Misquitta, and J. R. Schmidt, J. Chem. Theory Comput. **14**, 739 (2018).
[71]X. Chu and A. Dalgarno, J. Chem. Phys. **121**, 4083 (2004).
[72]A. Ambrosetti, A. M. Reilly, R. A. DiStasio, Jr., and A. Tkatchenko, J. Chem. Phys. **140**, 18A508 (2014).