ACS central science

Research Article

Cite This: ACS Cent. Sci. 2019, 5, 57−64

http://pubs.acs.org/journal/acscii

Transferable Machine-Learning Model of the Electron Density

Andrea Grisafi,[†,§] Alberto Fabrizio,[‡,§] Benjamin Meyer,[‡,§] David M. Wilkins,[†] Clemence Corminboeuf,[‡,§] and Michele Ceriotti*[,†]

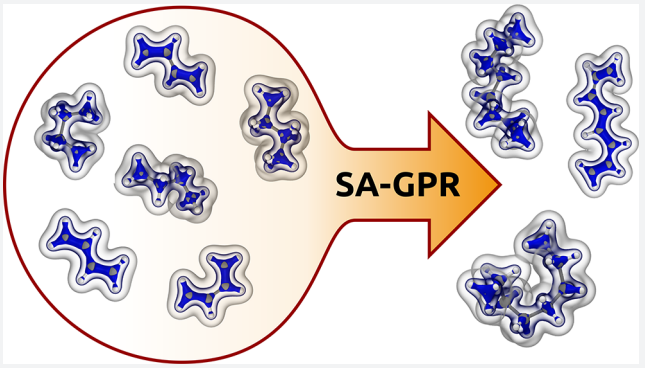
[†]Laboratory of Computational Science and Modeling, IMX, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland

[‡]Laboratory for Computational Molecular Design, Institute of Chemical Sciences and Engineering, École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland

[§]National Centre for Computational Design and Discovery of Novel Materials (MARVEL), École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland

Supporting Information

ABSTRACT: The electronic charge density plays a central role in determining the behavior of matter at the atomic scale, but its computational evaluation requires demanding electronic-structure calculations. We introduce an atom-centered, symmetry-adapted framework to machine-learn the valence charge density based on a small number of reference calculations. The model is highly transferable, meaning it can be trained on electronic-structure data of small molecules and used to predict the charge density of larger compounds with low, linear-scaling cost. Applications are shown for various hydrocarbon molecules of increasing complexity and flexibility, and demonstrate the accuracy of the model when predicting the density on octane and octatetraene after training exclusively on butane and butadiene. This transferable, data-driven model can be used to interpret experiments, accelerate electronic structure calculations, and compute electrostatic interactions in molecules and condensed-phase systems.



INTRODUCTION

The electron density $\rho(\mathbf{r})$ is a fundamental property of atoms, molecules, and condensed phases of matter. $\rho(\mathbf{r})$ can be measured directly by high-resolution electron diffraction[1,2] and transmission electron microscopy,[3] and can be analyzed to identify covalent and noncovalent patterns.[4−8] On the basis of density-functional theory (DFT), in the framework of the first Hohenberg−Kohn theorem,[9] knowledge of $\rho(\mathbf{r})$ gives access, in principle, to any ground-state property. Especially for large systems, however, the computation of $\rho(\mathbf{r})$ requires considerable effort, involving the solution of an electronic structure problem with a more or less approximate level of theory. Sidestepping these calculations and directly accessing the ground-state electron density for a given configuration of atoms would have broad implications, including real-time visualization of chemical fingerprints based on the electron density,[7] acceleration of DFT calculations by providing an estimate of the self-consistent charge density, and an exact treatment of the electrostatic interactions within an atomistic simulation. Another field of application involves the analysis and interpretation of experimental techniques that probe the electron density, such as transmission electron microscopy[3] and X-ray crystallography.[1,2] In the latter, the decomposition of the density in pseudoatomic contributions that is often performed to resolve the structure[10,11] foreshadows some of the ideas we will use here.

Following a number of successful applications of machine-learning methods to predict materials properties,[12−16] a recent landmark paper by Brockherde et al. showed that it is also possible to predict the ground-state electron density in a way that mimics the Hohenberg−Kohn mapping between the nuclear potential and the density.[17] A smoothed representation of the nuclear potential was used as a fingerprint to describe molecular configurations and to carry out individual predictions of the expansion coefficients of $\rho(\mathbf{r})$ represented in a plane-wave basis. Though in principle it is very effective, the structure of the model imposes significant constraints on its transferability to large and flexible systems. Indeed, the use of a global representation of the structure, and of an orthogonal basis to expand the density, means that the model is limited to interpolation between conformers of relatively rigid, small molecules.

In this paper, we show how to overcome these limitations by constructing a machine-learning model of the valence electron density that can be used on both large and flexible systems by predicting the density of large molecules based on training on smaller compounds. This is possible, in a nutshell, thanks to the combination of a local basis set to represent $\rho(\mathbf{r})$, which is reminiscent of local expansions of the wave function[18] and of

Received: August 10, 2018
Published: December 26, 2018

DOI: 10.1021/acscentsci.8b00551
ACS Cent. Sci. 2019, 5, 57−64

the atom density multipole analysis of X-ray diffraction,[19−23] and thanks to a recently introduced regression model which allows us to predict the local components of $\rho(\mathbf{r})$ in a symmetry-adapted fashion without the need to make simplifying assumptions on the description of molecular environments.

The method is tested on the carbon series C_2, C_4, and C_8 of both fully saturated and unsaturated hydrocarbons, having increasing complexity because of the exponentially growing number of conformers. In particular, interpolation of the electron density is first shown for ethene (C_2H_4), ethane (C_2H_6), butadiene (C_4H_6), and butane (C_4H_{10}). As a major result, the electron density of the corresponding C_8 molecules, namely, octa-tetraene (C_8H_{10}) and octane (C_8H_{18}), is instead predicted by extrapolating the information learned on the local environments of the corresponding C_4 molecules.

SYMMETRY-ADAPTED GAUSSIAN PROCESS REGRESSION FOR THE CHARGE DENSITY

Several widely adopted machine-learning schemes applied to materials rely on an additive decomposition of the target property in atom-centered contributions.[24−29] These approaches are very effective in achieving transferability across systems of different composition and size. An additive ansatz is justified by the exponential decay of the electronic density matrix (the so-called nearsightedness principle[30]) for insulators and metals at finite temperature, which underlies a plethora of linear-scaling, embedding, and fragment decomposition electronic structure methods.[18,31−38] Many methods exist to decompose the density in atom-centered contributions,[39,40] which however cannot be defined uniquely.[41] Rather than imposing that the machine-learning model should be consistent with a specific choice of density decomposition, we introduce locality only by expanding the density as a sum of atom-centered basis functions (for further details see section *Basis set optimization* in Supporting Information),

$$\rho(\mathbf{r}) = \sum_i \rho_i(\mathbf{r}) = \sum_{ik} c_k^i \phi_k^i(\mathbf{r}) = \sum_{ik} c_k^i \phi_k(\mathbf{r} - \mathbf{r}_i) \tag{1}$$

where k runs over the basis functions centered on each atom, and atoms of different species can have different kinds of functions. We then write a regression model for the combination coefficients c_k^i based exclusively on the knowledge of the positions of the nuclei, but we only use as the target property the total electron density $\rho(\mathbf{r})$. In this way, the model determines simultaneously the regression coefficients and the most convenient (and otherwise arbitrary) decomposition of $\rho(\mathbf{r})$ into atom-centered contributions.

From an atom-centered description, it is natural to factorize each basis function $\phi_k(\mathbf{r} - \mathbf{r}_i)$ into a product of radial functions $R_n(r_i)$ and spherical harmonics $Y_m^l(\hat{\mathbf{r}}_i)$ (with $r_i = |\mathbf{r} - \mathbf{r}_i|$ and $\hat{\mathbf{r}}_i = (\mathbf{r} - \mathbf{r}_i)/r_i$). The subscript k refers to the combination nlm, and we will use the compact or the extended notation based on convenience. For every atom-centered environment $\boldsymbol{\mathcal{X}}_i$, which defines the structure of a neighborhood of atom i, and for each radial function R_n, the coefficients can be grouped according to their value of angular momentum l in a set of spherical multipoles $\mathbf{c}_{nl}^i$ of dimension $2l + 1$, which transform as vector spherical harmonics $\mathbf{Y}_l$ under a rigid rotation of the environment. This choice has the advantage of highlighting the tensorial nature of the density components, meaning that a significant portion of the variability of $\mathbf{c}_{nl}^i$ can be attributed to the orientation of the local environments $\boldsymbol{\mathcal{X}}_i$, rather than to an actual structural distortion of the molecule.

Dealing with the regression of tensorial properties raises nontrivial issues in terms of setting up an effective machine-learning model that takes into account the proper covariances in three dimensions. For rigid molecules, one could eliminate this geometric variability by expressing the coefficients in a fixed molecular reference frame, analogously to what has already been done in the context of electric multipoles and response functions.[42,43]

This problem has long been known in the context of the determination of electron densities from experimental X-ray diffraction data.[19−23] One of the most widely used methods is the multipole model proposed by Stewart[10,44] and by Hansen and Coppens,[11] which models the valence charge density with both a spherical and multipolar component;[45] this is essentially equivalent to the expansion (1). In practice, existing pseudoatom methods are constructed from tabulated multipolar parameters (e.g., the libraries ELMAM,[46−49] ELMAM2,[50,51] UBDB,[52,53] Invariom,[54] and SBFA[55]), that are based on the determination of molecular fragments, that also provide a local reference frame to describe the anisotropy of the density. In most cases, these fragment decompositions are used as an initial guess for the density. The nuclear coordinates and the local multipoles are both optimized to match the experimental diffraction pattern during structural refinement.[56]

Our goal here is more ambitious, as we aim to predict the charge density based exclusively on nuclear coordinates. Furthermore, we aim at a scheme that does not rely on the definition of discrete molecular fragments and captures the density modulation by structural distortions and nonbonded interactions in arbitrarily complex and flexible molecules. As shown recently, Gaussian process regression can be modified to naturally endow the machine-learning model of vectors[57] and tensors of arbitrary order[58] with the symmetries of the three-dimensional (3D) rotation group SO(3). Within this method, called symmetry adapted Gaussian process regression (SA-GPR), the machine-learning prediction of the tensorial density components is

$$c_{nlm}^i(\mathbf{x}) = \sum_{j\in M} \sum_{|m'|<l} k_{mm'}^l(\boldsymbol{\mathcal{X}}_i, \boldsymbol{\mathcal{X}}_j) x_{nlm'}^j \delta_{\alpha_i\alpha_j} \tag{2}$$

In this expression, $\mathbf{k}^l(\boldsymbol{\mathcal{X}}_i, \boldsymbol{\mathcal{X}}_j)$ is a rank-2 kernel matrix of dimension $(2l + 1) \times (2l + 1)$ that expresses, at the same time, both the structural similarity and the geometric relationship between the atom-centered environment $\boldsymbol{\mathcal{X}}_i$ of the target molecule and a set M of reference environments $\boldsymbol{\mathcal{X}}_j$. The (tensorial) regression weights $x_{nlm'}^j$ are determined from a set of N training configurations and their associated electron densities.

According to eq 2, the prediction of the density expansion coefficients $c_{nlm}^i(\mathbf{x})$ is performed independently for each radial channel n, angular momentum value l and atomic species α. However, working with a nonorthogonal basis implies that the density components belonging to different atoms of the molecule are not independent of each other. One can indeed evaluate the projections of the density on the basis functions

$$w_k^i = \langle \rho | \phi_k^i \rangle = \int \mathrm{d}\mathbf{r}\, \rho(\mathbf{r}) \phi_k(\mathbf{r} - \mathbf{r}_i) \tag{3}$$

DOI: 10.1021/acscentsci.8b00551
ACS Cent. Sci. 2019, 5, 57−64

but these differ from the expansion coefficients c^i_k. In fact, $\mathbf{w}$ and $\mathbf{c}$ are related by $\mathbf{Sc} = \mathbf{w}$, where $S^{ii'}_{kk'} = \langle \phi^i_k \mid \phi^i_{k'} \rangle$ is the overlap between basis functions. For a given density, the coefficients could therefore be determined by inverting $\mathbf{S}$, so that each individual $nl\alpha$ component could be machine-learned separately. We observed, however, that doing so led to poor regression performance and unstable predictions. Applying $\mathbf{S}^{-1}$ on $\mathbf{w}$ corresponds to a partitioning of the charge which is, most of the time, affected by numerical noise. This is connected to the fact that $\mathbf{S}$ is often ill-conditioned, and so small numerical errors in the determination of $\mathbf{w}$ translate into large instabilities in the coefficients $\mathbf{c}$, making it hard for the machine-learning algorithm to find a unique relationship between the nuclear coordinates of the molecule and the density components. To avoid this issue and improve the accuracy of the physically relevant total density, the basis set decomposition and the construction of the machine-learning model need to be combined into a single step. This essentially consists of building a regression model that, of the many nearly equivalent decompositions of ρ, is able to determine the one which best fits the target density associated with a given structure.

The problem can be cast into a single least-squares optimization of a loss function that measures the discrepancy between the reference and the model densities,

$$l(\mathbf{x}) = \sum_{\mathcal{A}\in N} \int \mathrm{d}\mathbf{r} \left| \rho_{\mathcal{A}}(\mathbf{r}) - \sum_{i\in\mathcal{A}} \sum_k c^i_k(\mathbf{x}) \phi_k(\mathbf{r} - \mathbf{r}_i) \right|^2 + \eta |\mathbf{x}|^2 \quad (4)$$

Here the index $\mathcal{A}$ runs over the training set N, while i runs over the environments of a given training structure. The second term in the loss is a regularization, which avoids overfitting. In this context, η represents an adjustable parameter that is related to the intrinsic noise of the training data set. The coefficients $\mathbf{c}$ depend parametrically on the regression weights $\mathbf{x}$ via eq 2; by differentiating the loss with respect to x^j_{nlm} one obtains a set of linear equations that makes it possible to evaluate the weights in practice. In compact notation, the solution of this problem reads

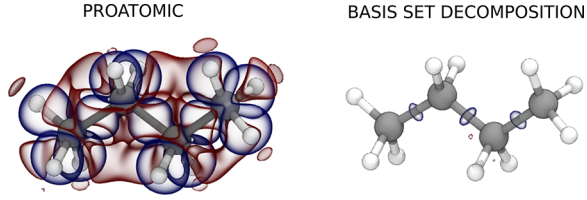
$$\mathbf{x} = (\mathbf{K}^T\mathbf{S}\mathbf{K} + \eta\mathbf{1})^{-1}\mathbf{K}^T\mathbf{w} \quad (5)$$

where $\mathbf{x}$ and $\mathbf{w}$ are vectors containing the regression weights and the density projections on the basis functions, while $\mathbf{K}$ and $\mathbf{S}$ are sparse matrix representations containing the symmetry-adapted tensorial kernels and the spatial overlaps between the basis functions. The details of this derivation and the resulting expressions are given in the Supporting Information. It should, however, be stressed that the final regression problem is highly nontrivial. The kernels that involve environments within the same training configuration are coupled by the overlap matrix, so that all the regression weights $\mathbf{x}$ for different elements, radial and angular momentum values must be determined simultaneously. An efficient implementation of a ML model based on eq 5 requires the optimization of a basis set for the expansion, the evaluation of $\rho(\mathbf{r})$ on dense atom-centered grids, the sparsification of the descriptors that are used to evaluate the kernels, and the determination of a diverse, minimal set of reference environments $\boldsymbol{X}_j$. All of these technical aspects are discussed extensively in the Supporting Information.

RESULTS AND DISCUSSION

Charge Decomposition Analysis. It is instructive to inspect the decomposition of the charge density in terms of the optimized basis, obtained from density projections on the basis functions $\mathbf{w}$ and the overlap matrix $\mathbf{S}$ as $\mathbf{c} = \mathbf{S}^{-1}\mathbf{w}$, which corresponds to the best accuracy that can be obtained with a given basis. With a basis set of four contracted radial functions, and angular momentum components up to $l = 3$, the typical error in the density decomposition can be brought down to about 1%. In Table 1 we compare, for the case of a butane molecule, the residual in the expansion with the typical error that can be expected by taking a superimposition of free-atom densities, between 16 and 20%.

Table 1. Mean Absolute Errors in the Representation of the Electron Density Using a Superimposition of Free Atoms (Proatomic Density) and the Optimized Basis Set Used in This Work (Basis Set Decomposition), Averaged over the Whole Training Set for the C_2 and C_4 Molecules[a]



	$\langle \varepsilon_\rho \rangle$ (%)			
	C_2H_4	C_2H_6	C_4H_6	C_4H_{10}
proatomic	18.06	19.23	16.79	18.13
basis set	1.04	1.14	0.98	1.19

[a]The graphic shows isosurfaces for the error in the electron density for proatomic (left) and basis set (right) representation, for a typical configuration of butane (red and blue isosurfaces correspond to an error of ±0.005 electrons Bohr^{-3}, respectively).

It is also possible to compute separately the contributions to the charge carried by each angular momentum channel l, e.g., $\rho_l(\mathbf{r}) = \sum_{inm} c^i_{nlm}\phi_{nlm}(\mathbf{r} - \mathbf{r}_i)$. As exemplified in Figure 1, while the isotropic $l = 0$ functions determine the general shape of the density, the $l = 1$ functions primarily describe the gradient of electronegativity in the region close to C−H bonds. Furthermore, the $l = 2$ functions describe the charge modulation associated with the C−C bonds along the main chain as well as the π-cloud along the conjugated backbone, while the $l = 3$ functions act as a further modulation that captures the nontrivial anisotropy. The figure also shows the collective contribution to the charge variability carried by each angular momentum channel l and atomic type α, i.e., $\sigma(l, \alpha) = \sqrt{\sum_n \langle |\mathbf{c}^i_{\mathrm{ln}} - \langle \mathbf{c}^i_{\mathrm{ln}} \rangle|^2 \rangle_{\alpha_i=\alpha}}$, with the average $\langle\cdot\rangle$ involving all the atoms of the same type included in the data set.

After having subtracted the mean atomic density of pure $l = 0$ character, the $l = 1$ components largely dominate the charge density variability associated with hydrogen atoms. As previously demonstrated,[46] functions with $l = 2$ symmetry also carry a substantial contribution, particularly for the carbon atoms of alkenes, while $l = 3$ functions appear to be dominant for carbon atoms of alkanes and almost irrelevant for hydrogen atoms in all the four molecules. In comparison to an atom-centered expansion of the wave function ψ, the choice of using a larger basis set is justified by the greater complexity in

DOI: 10.1021/acscentsci.8b00551
ACS Cent. Sci. 2019, 5, 57−64

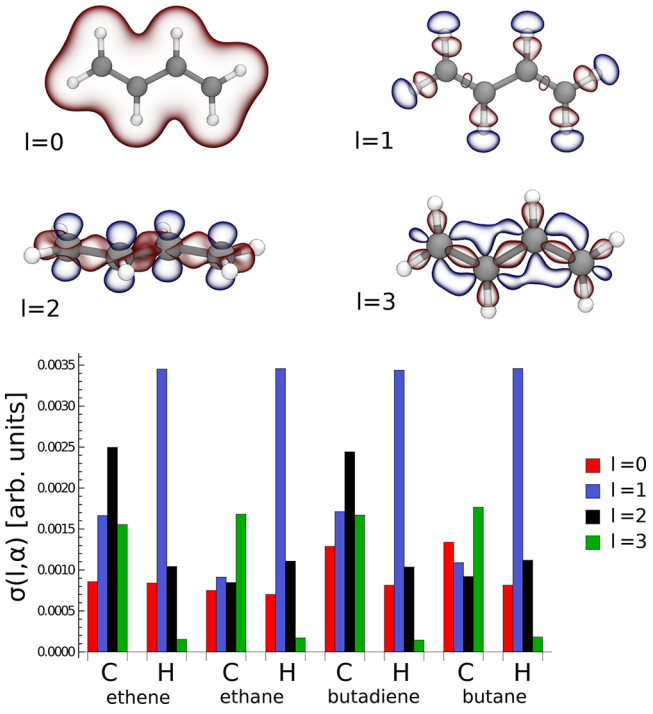


Figure 1. (Top) representation of the angular momentum decomposition of the electron density. Red and blue isosurfaces refer to ±0.01 electrons Bohr^{-3} respectively. (Bottom) angular momentum spectrum of the valence electron density of C_2 and C_4 data sets. The isotropic contributions $l = 0$ express the collective variations with respect to the data set's mean value, while the mean is statistically zero for $l > 0$.

describing an electron density field rather than the $N_e/2$ occupied orbitals being the solution of an effective single particle Hamiltonian. The need for high angular momentum components can be also justified by the fact that—even neglecting the overlap between adjacent atoms—the squaring of ψ that yields $\rho(\mathbf{r})$ would introduce nonzero components with up to twice the maximum l used to expand the wave function.

Density Learning with SA-GPR. Having optimized the basis set and analyzed the variability of the electron density when expanded in this optimized basis, we now proceed to test the SA-GPR regression scheme. The difficulty of the learning exercise largely depends on the structural flexibility of the molecules. Small, rigid systems such as ethene and ethane require little training, and could be equivalently learned through a machine-learning framework based on a pairwise comparison of aligned molecules. Butadiene data, containing both cis and trans conformers, as well as distorted configurations approaching the isomerization transition state, poses a more significant challenge, due to an extended conjugated system that makes the electronic structure very sensitive to small molecular deformations. The case of butane is also particularly challenging because of the broad spectrum of intramolecular noncovalent interactions spanned by the many different conformers contained in the data set. Being fully flexible, this kind of system is expected to benefit most from a ML scheme that can adapt its kernel similarity measure to different orientations of molecular subunits. Figure 2 shows the performance of the method in terms of prediction accuracy of the electron density as a function of the number of training molecules. The number M of reference environments has been fixed to the 1500 most diverse, FPS-selected, environments contained in each data set. The convergence with respect to M is discussed in the Supporting Information. The symmetry adapted similarity measure which enters in the regression formula of eq 5 is given by the tensorial λ-SOAP kernels of ref 58. This generalizes the scalar ($\lambda = 0$) smooth overlap of atomic positions framework[59] that has been used successfully for constructing interatomic potentials[25,60] and predicting molecular properties.[61,62] In constructing these kernel functions, we chose a radial cutoff of 4.5 Å for the definition of atomic environments (further details are in the Supporting Information). Learning curves are then obtained by varying the number of training molecules up to 800 randomly selected configurations out of the total of 1000. The remaining 200 molecules for each of these random selections are used to estimate the error in the density prediction.

We express the error in terms of the mean absolute difference between the predicted and quantum mechanical densities, i.e., $\epsilon_\rho(\%) = 100 \times \langle \int d\mathbf{r} |\rho_{QM}(\mathbf{r}) - \rho_{ML}(\mathbf{r})| \rangle / N_e$. The prediction errors of ethene and ethane saturate to the limit imposed by the basis set representation, which is around 1% for all molecules, with as few as 10 training points. As expected, given the greater flexibility, learning the charge density of butadiene and butane is more challenging, requiring the inclusion of more than 100 training structures in order to approach the basis set limit. This level of accuracy (an error which is almost 20 times smaller than that obtained with a

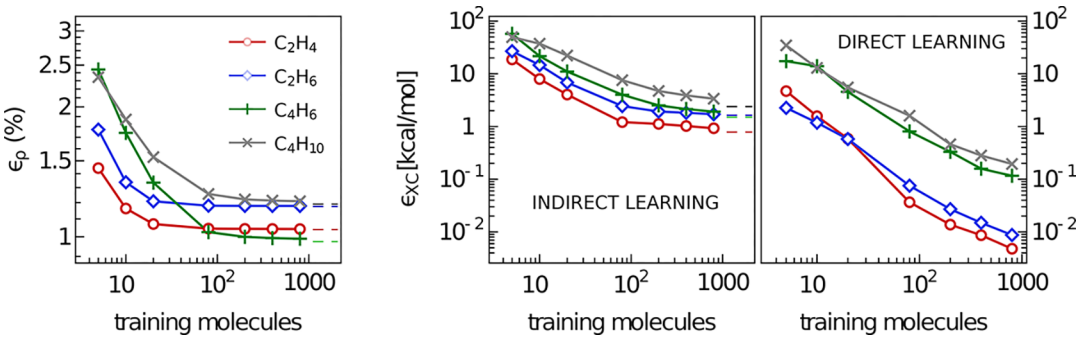


Figure 2. Learning curves for C_2 and C_4 molecules. (Left) % mean absolute error of the predicted SA-GPR densities as a function of the number of training molecules. The error normalization is provided by the total number of valence electrons. (Right) root-mean-square errors of the exchange-correlation energies indirectly predicted from the SA-GPR densities and directly predicted via a scalar SOAP kernel, as a function of the number of training molecules. Dashed lines refer to the error carried by the basis set representation.

DOI: 10.1021/acscentsci.8b00551
ACS Cent. Sci. 2019, 5, 57−64

superposition of rigid atomic densities, as discussed above) was demonstrated to be sufficient[46] for most applications that rely on the accuracy of the density representation, such as the modeling of X-ray and transmission electron microscopy,[1−3] or the evaluation of density-based fingerprints of chemical interactions.[4−8]

Using the predicted $\rho(\mathbf{r})$ as the basis for a density-functional calculation is more challenging. As a benchmark for this application, we use the SA-GPR predictions for $\rho(\mathbf{r})$ to evaluate the PBE exchange-correlation functional $E_{XC}[\rho]$ used for the reference quantum-mechanical calculations. Depending on the gradient of the density, this quantity is very sensitive to small density variations, especially localized around the atomic nuclei. Figure 2 shows the root-mean-square error for the exchange-correlation energies ε_{XC}. Using the full set of 800 training molecules, we reach a, RMSE of 0.9 and 1.7 kcal/mol for ethene and ethane, 1.9 kcal/mol for butadiene, and 3.5 kcal/mol for butane, basically matching the basis set limit. It is clear that the ML scheme has the potential to reach higher accuracy with a small number of reference configurations, but a significant reduction of the basis set error is necessary to reach chemical accuracy (roughly 1 kcal/mol RMSE) in the prediction of E_{XC}. At the same time, it is not obvious that computing E_{XC} indirectly, by first predicting the electron charge density, is the most effective strategy to obtain an ML model of DFT energetics. As shown in the figure, applying a direct, scalar regression based on conventional SOAP kernels to learn the relationship between the molecular structure and E_{XC} leads to vastly superior performance while requiring a much simpler machine-learning model.

Size-Extensive Extrapolation. While incremental improvements of the underlying density representation framework are desirable to use the predicted density as the basis of DFT calculations, we can already demonstrate the potential of our SA-GPR scheme in terms of transferability of the model. From the prediction formula of eq 2, it is clear that no assumption is made about the identity of the molecule for which the electron density is predicted. Practically speaking, the regression weights x^j_{nlm} are associated with representative environments that could be taken from any kind of compound, not necessarily the same as that for which the density is being predicted. As long as the training set is capable of describing different chemical environments, and contains local configurations similar to the ones of our prediction target, accurate densities can be obtained simply by computing the kernels $\mathbf{k}^l(\boldsymbol{X}_i, \boldsymbol{X}_j)$ between the environments $\boldsymbol{X}_i$ of an arbitrarily large molecule and the reference environments $\boldsymbol{X}_j$. The cost of this prediction is proportional to the number of environments, making this method of evaluating the electron charge density strictly linear scaling in the size of the target molecule.

As a proof of concept of this extrapolation procedure, we use environments and training information from the butadiene and butane configurations already discussed to construct the electron density of octatetraene (C_8H_{10}) and octane (C_8H_{18}), respectively. It is important to stress that the transferability is because on a local scale the larger molecules are similar to those used for training, and so the prediction is effectively an *interpolation* in the space of local environments. This is emphasized by the observation that the optimal extrapolation accuracy is obtained using a machine learning cutoff of r_{cut} = 3 Å, versus a value of r_{cut} = 4.5 Å that was optimal for same-molecule predictions. On a scale larger than 3 Å, the environments present in C8 molecules differ substantially from those in the corresponding C4 compound, which negatively affects the transferability of the model. Ideally, as the training data set is extended to include larger and larger molecules, this locality constraint can be relaxed until no substantial difference can be appreciated between the prediction accuracy of the interpolated and extrapolated density.

For both octane and octatetraene, the extrapolation is carried out on a challenging data set made of the 100 most diverse structures extracted by farthest point sampling from the 300 K replica of a long replica exchange molecular dynamics (REMD) run. When learning on the full data set of butadiene and butane, we obtain a low density mean absolute error of 1.8% for octatetraene and of 1.4% for octane. As shown in Figure 3 for two representative configurations, the size-extensive SA-GPR prediction accurately reproduces the structure of the electron density for both octane and octatetraene. Because of the high sensitivity of the electronic π-cloud to the molecular identity and configuration, major difficulties arise in predicting the electron density of octatetraene, particularly in the middle regions, for which no analogous examples are contained in the butadiene training data set.

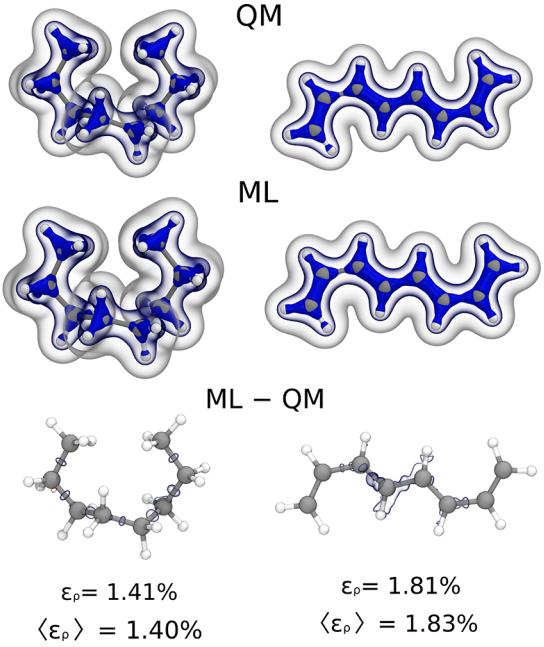


Figure 3. Extrapolation results for the valence electron density of one octane (left) and one octatetraene (right) conformer. (Top) DFT/PBE density isosurface at 0.25, 0.1, 0.01 electrons Bohr^{-3}, (middle) machine-learning prediction isosurface at 0.25, 0.1, 0.01 electrons Bohr^{-3}, (bottom) machine-learning error, red and blue isosurfaces refer to ±0.005 electrons Bohr^{-3} respectively. Relative mean absolute errors averaged over 100 conformers are also reported for both cases.

The SOAP representation can be easily extended to more complex molecules and condensed phases,[63] and has been shown to be remarkably effective in making predictions on larger molecules based on training on very simple compounds.[64] Achieving similar results for the charge density involves some technical challenges, connected with the presence of correlations between coefficients due to the nonorthogonal basis expansion, that makes the cost of training

DOI: 10.1021/acscentsci.8b00551
ACS Cent. Sci. 2019, 5, 57−64

(but not of predicting) the density scale unfavorably with the system size. In the presence of large electric fields, or long-range charge transfer, it will be necessary to extend the scheme to be compatible with a description of the underlying physical process. One can look for inspiration to existing self-consistent equilibration schemes for atomic charges,[65] or to the use of local electric fields as part of the input representation.[43]

CONCLUSIONS

Machine-learning the electronic charge density of molecular systems as a function of nuclear coordinates poses great technical and conceptual challenges. Transferability across molecules of different size and stoichiometry calls for a scheme based on a local decomposition, which should be performed without relying on arbitrary charge partitioning or discarding the fundamental physical symmetries of the problem. The framework we present here overcomes these hurdles by decomposing the density in optimized atom-centered basis functions, exploiting a symmetry-adapted regression scheme to incorporate geometric covariances, and by designing a loss function that relies only on the total charge density as a physically meaningful constraint. The atom-centered decomposition means the ML model can predict the density of large molecules or condensed phases with a cost that scales linearly with the number of atoms. For instance, learning the chemical environment of all the functional units of the 20 natural amino acids in all their protonation states and forms (N-terminal, nonterminal, C-terminal), one possible perspective for our method will be the prediction of the charge density of proteins.

We have demonstrated the viability and accuracy of this scheme by learning the ground-state valence electron density of saturated and unsaturated hydrocarbons with two and four carbon atoms, achieving in all cases an error of the order of 1% on the reconstructed density. Given that this estimate is based exclusively on the nuclear position, it could be used for structural determination, e.g., in the analysis of X-ray[46] and transmission electron microscopy experiments. What is more, models trained on C4 compounds can be used to predict the electronic charge of their larger, C8 counterparts, providing a first example of the transferability that is afforded by a symmetry-adapted local decomposition scheme.

Further improvements of the accuracy are likely to be possible, by better optimization of the basis set, by simultaneously fine-tuning the representation of environments by λ-SOAP kernels and the representation of the density in terms of projections on a local basis set, and also by using inexpensive semiempirical methods to provide a baseline for the electron density prediction. In fact, this work can be seen as a first, successful attempt to apply machine learning in a transferable way to molecular properties that cannot be simply decomposed as the sum of atom-centered values, but exhibit a richer, more complex geometric structure. The Hamiltonian, the density matrix, vector fields, and density response functions are other examples that will require careful consideration of both the representation of the input structure, and of the property one wants to predict, and that can benefit from the framework we have introduced in the present work.

METHODS

As a demonstration of our framework, we consider hydrocarbons, using a data set of 1000 independent structures of ethene, ethane, butadiene, and butane. Atomic configurations are generated by running REMD simulations at the density functional tight binding level,[66] using a combination of the DFTB+[67] and i-PI[68] simulation software.[69] In order to construct a realistic and challenging test of the ML scheme, we chose the replica at T = 300 K and selected a diverse set of 1000 configurations, by a farthest point sampling (FPS) algorithm based on the SOAP metric.[61,70] For each selected configuration we computed the valence electron pseudo density at the DFT/PBE level with SBKJC effective core potentials. Further details of the data set construction are given in the Supporting Information.

The problem of representing a charge density in terms of a nonorthogonal localized basis set shares many similarities with that of expanding the wave function. For this reason, we resort to many of the tricks used in quantum chemistry codes, including the use of Gaussian type orbitals (GTOs) to compute the basis set overlap analytically, and the contraction of 12 regularly spaced radial GTOs down to four optimized functions. We find that angular momentum channels up to l = 3 functions are needed to obtain a decomposition error around 1% for the density. The coefficients of the contraction are optimized to minimize the mean charge decomposition error and the condition number of the overlap matrix for the four molecules,[71] as discussed in the Supporting Information. A systematic analysis of the interplay between the details of the basis set and the performance of the ML model goes beyond the scope of this work. It is likely however that substantial improvements of this approach could be achieved by further optimization of the basis.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acscentsci.8b00551.

Supporting Information of the manuscript contains further details about the derivation, implementation and benchmarks of the method, including: a summary of the symmetry-adapted Gaussian process regression (SA-GPR) method; a detailed derivation of the regression formula for the learning of the charge density; the definition of the primitive basis set together with the recipe to compute the overlap matrix analytically; the details about the numerical evaluation of density projections over the basis functions; the definition of the grid used to estimate the error associated with the predicted densities; the procedure adopted to optimize the primitive basis set; a detailed description of how the molecular configurations are selected; details about the quantum-mechanical calculations used to produce the training densities; the machine-learning parameters used to compute λ-SOAP kernels; the details about the selection of the reference environments used for the density learning; a comparison between the prediction accuracy obtained with our data set of ethane (C_2H_6) and the data set of ethane generated in ref 17; a summary of regression performances (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: michele.ceriotti@epfl.ch.

DOI: 10.1021/acscentsci.8b00551
ACS Cent. Sci. 2019, 5, 57−64

ORCID

Alberto Fabrizio: 0000-0002-4440-3149
Clemence Corminboeuf: 0000-0001-7993-2879
Michele Ceriotti: 0000-0003-2571-2832

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors are grateful to G. Csányi, J. VandeVondele, and M. Willat for insightful discussion. M.C. and D.M.W. were supported by the European Research Council under the European Union's Horizon 2020 research and innovation programme (Grant Agreement No. 677013-HBMAP), and benefited from generous allocation of computer time by CSCS, under Project ID s843. A.G. acknowledges funding by the MPG-EPFL Center for Molecular Nanoscience and Technology. C.C., B.M., A.F. and M.C. acknowledge the National Centre of Competence in Research (NCCR) Materials Revolution: Computational Design and Discovery of Novel Materials (MARVEL) of the Swiss National Science Foundation (SNSF) for financial support and the EPFL for computing time.

■ REFERENCES

(1) Koritsanszky, T. S.; Coppens, P. Chemical Applications of X-Ray Charge-Density Analysis. *Chem. Rev.* **2001**, *101*, 1583−1628.
(2) Gatti, C., Macchi, P., Eds. *Modern Charge-Density Analysis*, 1st ed.; Springer Netherlands, 2012.
(3) Meyer, J. C.; Kurasch, S.; Park, H. J.; Skakalova, V.; Kunzel, D.; Gross, A.; Chuvilin, A.; Algara-Siller, G.; Roth, S.; Iwasaki, T.; Starke, U.; Smet, J. H.; Kaiser, U. Experimental Analysis of Charge Redistribution Due to Chemical Bonding by High-Resolution Transmission Electron Microscopy. *Nat. Mater.* **2011**, *10*, 209−215.
(4) Coppens, P. Charge Densities Come of Age. *Angew. Chem., Int. Ed.* **2005**, *44*, 6810−6811.
(5) Becke, A. D.; Edgecombe, K. E. A Simple Measure of Electron Localization in Atomic and Molecular Systems. *J. Chem. Phys.* **1990**, *92*, 5397−5403.
(6) Johnson, E. R.; Keinan, S.; Mori-Sanchez, P.; Contreras-Garcia, J.; Cohen, A. J.; Yang, W. Revealing Noncovalent Interactions. *J. Am. Chem. Soc.* **2010**, *132*, 6498−6506.
(7) de Silva, P.; Corminboeuf, C. Simultaneous Visualization of Covalent and Noncovalent Interactions Using Regions of Density Overlap. *J. Chem. Theory Comput.* **2014**, *10*, 3745−3756.
(8) Pastorczak, E.; Corminboeuf, C. Perspective: Found in Translation: Quantum Chemical Tools for Grasping Non-Covalent Interactions. *J. Chem. Phys.* **2017**, *146*, 120901.
(9) Parr, R.; Weitao, Y. *Density-Functional Theory of Atoms and Molecules*; International Series of Monographs on Chemistry; Oxford University Press, 1994.
(10) Stewart, R. F. Electron Population Analysis with Rigid Pseudoatoms. *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.* **1976**, *32*, 565−574.
(11) Hansen, N. K.; Coppens, P. Testing Aspherical Atom Refinements on Small-Molecule Data Sets. *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.* **1978**, *34*, 909−921.
(12) Aspuru-Guzik, A.; Lindh, R.; Reiher, M. The Matter Simulation (R)evolution. *ACS Cent. Sci.* **2018**, *4*, 144−152.
(13) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559*, 547−555.
(14) Ferguson, A. L. ACS Central Science Virtual Issue on Machine Learning. *ACS Cent. Sci.* **2018**, *4*, 938−941.
(15) Rupp, M.; von Lilienfeld, O. A.; Burke, K. Guest Editorial: Special Topic on Data-Enabled Theoretical Chemistry. *J. Chem. Phys.* **2018**, *148*, 241401.
(16) Welborn, M.; Cheng, L.; Miller, T. F., III *J. Chem. Theory Comput.* **2018**, *14* (9), 4772−4779.
(17) Brockherde, F.; Vogt, L.; Li, L.; Tuckerman, M. E.; Burke, K.; Müller, K.-R. Bypassing the Kohn-Sham Equations with Machine Learning. *Nat. Commun.* **2017**, *8*, 872.
(18) Marzari, N.; Vanderbilt, D. Maximally Localized Generalized Wannier Functions for Composite Energy Bands. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1997**, *56*, 12847−12865.
(19) Dominiak, P. M.; Volkov, A.; Dominiak, A. P.; Jarzembska, K. N.; Coppens, P. Combining Crystallographic Information and an Aspherical-Atom Data Bank in the Evaluation of the Electrostatic Interaction Energy in an Enzymesubstrate Complex: Influenza Neuraminidase Inhibition. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2009**, *65*, 485−499.
(20) Pichon-Pesme, V.; Jelsch, C.; Guillot, B.; Lecomte, C. A Comparison Between Experimental and Theoretical Aspherical-Atom Scattering Factors for Charge-Density Refinement of Large Molecules. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **2004**, *60*, 204−208.
(21) Guillot, B.; Jelsch, C.; Podjarny, A.; Lecomte, C. Charge-Density Analysis of a Protein Structure at Subatomic Resolution: The Human Aldose Reductase Case. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2008**, *64*, 567−588.
(22) Ghermani, N.-E.; Bouhmaida, N.; Lecomte, C. Modelling Electrostatic Potential from Experimentally Determined Charge Densities. I. Spherical-Atom Approximation. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **1993**, *49*, 781−789.
(23) Bouhmaida, N.; Ghermani, N.-E.; Lecomte, C.; Thalal, A. Modelling Electrostatic Potential from Experimentally Determined Charge Densities. II. Total Potential. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **1997**, *53*, 556−563.
(24) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.
(25) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, Without the Electrons. *Phys. Rev. Lett.* **2010**, *104*, 136403.
(26) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, No. 058301.
(27) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: An Extensible Neural Network Potential with DFT Accuracy at Force Field Computational Cost. *Chem. Sci.* **2017**, *8*, 3192−3203.
(28) Zhang, L.; Han, J.; Wang, H.; Car, R.; E, W. Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics. *Phys. Rev. Lett.* **2018**, *120*, 143001.
(29) Gubaev, K.; Podryabinkin, E. V.; Shapeev, A. V. Machine Learning of Molecular Properties: Locality and Active Learning. *J. Chem. Phys.* **2018**, *148*, 241727.
(30) Prodan, E.; Kohn, W. Nearsightedness of Electronic Matter. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 11635−11638.
(31) Yang, W. Direct Calculation of Electron Density in Density-Functional Theory. *Phys. Rev. Lett.* **1991**, *66*, 1438−1441.
(32) Galli, G.; Parrinello, M. Large Scale Electronic Structure Calculations. *Phys. Rev. Lett.* **1992**, *69*, 3547−3550.
(33) Goedecker, S. Linear Scaling Electronic Structure Methods. *Rev. Mod. Phys.* **1999**, *71*, 1085−1123.
(34) Ceriotti, M.; Kühne, T. D.; Parrinello, M. An Efficient and Accurate Decomposition of the Fermi Operator. *J. Chem. Phys.* **2008**, *129*, 024707.
(35) Fedorov, D. G.; Kitaura, K. Extending the Power of Quantum Chemistry to Large Systems with the Fragment Molecular Orbital Method. *J. Phys. Chem. A* **2007**, *111*, 6904−6914.
(36) Merz, K. M. Using Quantum Mechanical Approaches to Study Biological Systems. *Acc. Chem. Res.* **2014**, *47*, 2804−2811.
(37) Walker, P. D.; Mezey, P. G. Molecular Electron Density Lego Approach to Molecule Building. *J. Am. Chem. Soc.* **1993**, *115*, 12423−12430.

DOI: 10.1021/acscentsci.8b00551
ACS Cent. Sci. 2019, 5, 57−64

(38) Meyer, B.; Guillot, B.; Ruiz-Lopez, M. F.; Genoni, A. Libraries of Extremely Localized Molecular Orbitals. 1. Model Molecules Approximation and Molecular Orbitals Transferability. *J. Chem. Theory Comput.* **2016**, *12*, 1052−1067.
(39) Hirshfeld, F. L. Bonded-Atom Fragments for Describing Molecular Charge Densities. *Theor. Chim. Acta* **1977**, *44*, 129−138.
(40) Bader, R. F. W. *Atoms in Molecules: A Quantum Theory*; Oxford University Press, 1990.
(41) Gonthier, J. F.; Steinmann, S. N.; Wodrich, M. D.; Corminboeuf, C. Quantification of fuzzy Chemical Concepts: A Computational Perspective. *Chem. Soc. Rev.* **2012**, *41*, 4671.
(42) Bereau, T.; Andrienko, D.; von Lilienfeld, O. A. Transferable Atomic Multipole Machine Learning Models for Small Organic Molecules. *J. Chem. Theory Comput.* **2015**, *11*, 3225−3233.
(43) Liang, C.; Tocci, G.; Wilkins, D. M.; Grisafi, A.; Roke, S.; Ceriotti, M. Solvent Fluctuations and Nuclear Quantum Effects Modulate the Molecular Hyperpolarizability of Water. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2017**, *96*, No. 041407.
(44) Stewart, R. F. Electron Population Analysis with Generalized X-Ray Scattering Factors: Higher Multipoles. *J. Chem. Phys.* **1973**, *58*, 1668.
(45) Coppens, P.; Guru Row, T. N.; Leung, P.; Stevens, E. D.; Becker, P. t.; Yang, Y. W. Net Atomic Charges and Molecular Dipole Moments from Spherical-Atom X-Ray Refinements, and the Relation Between Atomic Charge and Shape. *Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr.* **1979**, *35*, 63−72.
(46) Pichon-Pesme, V.; Lecomte, C.; Lachekar, H. On Building a Data Bank of Transferable Experimental Electron Density Parameters Applicable to Polypeptides. *J. Phys. Chem.* **1995**, *99*, 6242−6250.
(47) Jelsch, C.; Pichon-Pesme, V.; Lecomte, C.; Aubry, A. Transferability of Multipole Charge-Density Parameters: Application to Very High Resolution Oligopeptide and Protein Structures. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **1998**, *54*, 1306−1318.
(48) Zarychta, B.; Pichon-Pesme, V.; Guillot, B.; Lecomte, C.; Jelsch, C. On the Application of an Experimental Multipolar Pseudo-Atom Library for Accurate Refinement of Small-Molecule and Protein Crystal Structures. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **2007**, *63*, 108−125.
(49) Lecomte, C.; Jelsch, C.; Guillot, B.; Fournier, B.; Lagoutte, A. Ultrahigh-Resolution Crystallography and Related Electron Density and Electrostatic Properties in Proteins. *J. Synchrotron Radiat.* **2008**, *15*, 202−203.
(50) Domagala, S.; Munshi, P.; Ahmed, M.; Guillot, B.; Jelsch, C. Structural Analysis and Multipole Modelling of Quercetin Monohydrate a Quantitative and Comparative Study. *Acta Crystallogr., Sect. B: Struct. Sci.* **2011**, *67*, 63−78.
(51) Domagala, S.; Fournier, B.; Liebschner, D.; Guillot, B.; Jelsch, C. An Improved Experimental Databank of Transferable Multipolar Atom Models ELMAM2. Construction Details and Applications. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **2012**, *68*, 337−351.
(52) Koritsanszky, T.; Volkov, A.; Coppens, P. Aspherical-Atom Scattering Factors from Molecular Wave Functions. 1. Transferability and Conformation Dependence of Atomic Electron Densities of Peptides Within the Multipole Formalism. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **2002**, *58*, 464−472.
(53) Dominiak, P. M.; Volkov, A.; Li, X.; Messerschmidt, M.; Coppens, P. A Theoretical Databank of Transferable Aspherical Atoms and Its Application to Electrostatic Interaction Energy Calculations of Macromolecules. *J. Chem. Theory Comput.* **2007**, *3*, 232−247.
(54) Dittrich, B.; Koritsanszky, T.; Luger, P. A Simple Approach to Nonspherical Electron Densities by Using Invarioms. *Angew. Chem., Int. Ed.* **2004**, *43*, 2718−2721.
(55) Hathwar, V. R.; Thakur, T. S.; Row, T. N. G.; Desiraju, G. R. Transferability of Multipole Charge Density Parameters for Supramolecular Synthons: A New Tool for Quantitative Crystal Engineering. *Cryst. Growth Des.* **2011**, *11*, 616−623.
(56) Bak, J. M.; Domagala, S.; Hubschle, C.; Jelsch, C.; Dittrich, B.; Dominiak, P. M. Verification of Structural and Electrostatic Properties Obtained by the Use of Different Pseudoatom Databases. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **2011**, *67*, 141−153.
(57) Glielmo, A.; Sollich, P.; De Vita, A. Accurate Interatomic Force Fields via Machine Learning with Covariant Kernels. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2017**, *95*, 214302.
(58) Grisafi, A.; Wilkins, D. M.; Csányi, G.; Ceriotti, M. Symmetry-Adapted Machine Learning for Tensorial Properties of Atomistic Systems. *Phys. Rev. Lett.* **2018**, *120*, No. 036002.
(59) Bartók, A. P.; Kondor, R.; Csányi, G. On Representing Chemical Environments. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2013**, *87*, 184115.
(60) Deringer, V. L.; Csányi, G. Machine Learning Based Interatomic Potential for Amorphous Carbon. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2017**, *95*, No. e094203.
(61) De, S.; Bartók, A. A. P.; Csányi, G.; Ceriotti, M. Comparing Molecules and Solids Across Structural and Alchemical Space. *Phys. Chem. Chem. Phys.* **2016**, *18*, 13754−13769.
(62) Bartók, A. A. P.; De, S.; Poelking, C.; Bernstein, N.; Kermode, J. R. J.; Csányi, G.; Ceriotti, M. Machine Learning Unifies the Modeling of Materials and Molecules. *Sci. Adv.* **2017**, *3*, No. e1701816.
(63) Musil, F.; De, S.; Yang, J.; Campbell, J. E. J.; Day, G. G. M.; Ceriotti, M. Machine Learning for the Structure-Energy-Property Landscapes of Molecular Crystals. *Chemical Science* **2018**, *9*, 1289−1300.
(64) Wilkins, D. M.; Grisafi, A.; Yang, Y.; Lao, K. U.; DiStasio, R. A., Jr.; Ceriotti, M. Accurate Molecular Polarizabilities with Coupled-Cluster Theory and Machine Learning. **2018**, ArXiv:1809.05337.
(65) Ghasemi, S. A.; Hofstetter, A.; Saha, S.; Goedecker, S. Interatomic Potentials for Ionic Systems with Density Functional Accuracy Based on Charge Densities Obtained by a Neural Network. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2015**, *92*, No. 045131.
(66) Elstner, M.; Seifert, G. Density Functional Tight Binding. *Philos. Trans. R. Soc., A* **2014**, *372*, 20120483.
(67) Aradi, B.; Hourahine, B.; Frauenheim, T. DFTB+, a Sparse Matrix-Based Implementation of the DFTB Method. *J. Phys. Chem. A* **2007**, *111*, 5678−5684.
(68) Kapil, V. et al. i-PI 2.0: A universal force engine for advanced molecular simulations. *Comput. Phys. Commun.* **2018**, DOI: 10.1016/j.cpc.2018.09.020.
(69) Petraglia, R.; Nicolaï, A.; Wodrich, M. M. D.; Ceriotti, M.; Corminboeuf, C. Beyond Static Structures: Putting Forth REMD As a Tool to Solve Problems in Computational Organic Chemistry. *J. Comput. Chem.* **2016**, *37*, 83−92.
(70) Imbalzano, G.; Anelli, A.; Giofré, D.; Klees, S.; Behler, J.; Ceriotti, M. Automatic Selection of Atomic Fingerprints and Reference Configurations for Machine-Learning Potentials. *J. Chem. Phys.* **2018**, *148*, 241730.
(71) VandeVondele, J.; Hutter, J. Gaussian Basis Sets for Accurate Calculations on Molecular Systems in Gas and Condensed Phases. *J. Chem. Phys.* **2007**, *127*, 114105.

DOI: 10.1021/acscentsci.8b00551
ACS Cent. Sci. 2019, 5, 57−64