The Journal of Chemical Physics

Predicting molecular dipole moments by combining atomic partial charges and atomic dipoles

Cite as: J. Chem. Phys. **153**, 024113 (2020); https://doi.org/10.1063/5.0009106
Submitted: 27 March 2020 . Accepted: 21 May 2020 . Published Online: 09 July 2020

Max Veit, David M. Wilkins, Yang Yang, Robert A. DiStasio, and Michele Ceriotti

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

Predicting molecular dipole moments by combining atomic partial charges and atomic dipoles

Cite as: J. Chem. Phys. 153, 024113 (2020); doi: 10.1063/5.0009106
Submitted: 27 March 2020 • Accepted: 21 May 2020 •
Published Online: 9 July 2020

Max Veit,[1] David M. Wilkins,[1,a)] Yang Yang,[2] Robert A. DiStasio Jr.,[2,b)] and Michele Ceriotti[1,b)]

AFFILIATIONS

[1] Laboratory of Computational Science and Modeling, IMX, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland
[2] Department of Chemistry and Chemical Biology, Cornell University, Ithaca, New York 14853, USA

Note: This paper is part of the JCP Special Topic on Machine Learning Meets Chemical Physics.
[a)] **Current address:** Atomistic Simulation Centre, School of Mathematics and Physics, Queen's University Belfast, Belfast BT7 1NN, Northern Ireland, United Kingdom.
[b)] **Authors to whom correspondence should be addressed:** distasio@cornell.edu and michele.ceriotti@epfl.ch

ABSTRACT

The molecular dipole moment ($\boldsymbol{\mu}$) is a central quantity in chemistry. It is essential in predicting infrared and sum-frequency generation spectra as well as induction and long-range electrostatic interactions. Furthermore, it can be extracted directly—via the ground state electron density—from high-level quantum mechanical calculations, making it an ideal target for machine learning (ML). In this work, we choose to represent this quantity with a physically inspired ML model that captures two distinct physical effects: local atomic polarization is captured within the symmetry-adapted Gaussian process regression framework which assigns a (vector) dipole moment to each atom, while the movement of charge across the entire molecule is captured by assigning a partial (scalar) charge to each atom. The resulting "MuML" models are fitted together to reproduce molecular $\boldsymbol{\mu}$ computed using high-level coupled-cluster theory and density functional theory (DFT) on the QM7b dataset, achieving more accurate results due to the physics-based combination of these complementary terms. The combined model shows excellent transferability when applied to a showcase dataset of larger and more complex molecules, approaching the accuracy of DFT at a small fraction of the computational cost. We also demonstrate that the uncertainty in the predictions can be estimated reliably using a calibrated committee model. The ultimate performance of the models—and the optimal weighting of their combination—depends, however, on the details of the system at hand, with the scalar model being clearly superior when describing large molecules whose dipole is almost entirely generated by charge separation. These observations point to the importance of simultaneously accounting for the local and non-local effects that contribute to $\boldsymbol{\mu}$; furthermore, they define a challenging task to benchmark future models, particularly those aimed at the description of condensed phases.

I. INTRODUCTION

The dipole moment $\boldsymbol{\mu}$ of a molecule quantifies the molecule's first-order response to an applied electric field. It is a key ingredient in the calculation of infrared (IR)[1] and sum-frequency generation (SFG)[2,3] spectra as well as the understanding of intermolecular interactions.[4] Despite its importance, the dipole moment presents a challenge for calculation, often depending significantly on the level of theory and the basis set used.[5–7] Furthermore, while the molecular dipole moment gives information about the distribution of charge in the molecule, it is determined by the interplay of several physical effects, such as long-range charge transfer and local polarization, which cannot be disentangled based on the knowledge of $\boldsymbol{\mu}$ alone. A number of methods for unraveling these different contributions exist and are generally based on partitioning the electron density into localized atomic charges and dipoles (accounting for charge

transfer and polarization). While these methods are attractive for understanding the underlying physics responsible for $\boldsymbol{\mu}$, they are usually poorly transferable between different molecules or classes of molecules (see Sec. II).

In this work, we design a new framework for the prediction of gas-phase molecular dipole moments, which unifies the atomic charge–atomic dipole description rooted in physics with the conformational and chemical sensitivity afforded by kernel-based machine learning (ML). We begin in Sec. II with an overview of existing methods to describe and predict molecular dipoles. In Sec. III, we formulate the different models we propose to learn and predict polarization, and we use a general symmetry-adapted framework to give environment-centered dipole predictions,[8] along with a partial-charge model in the vein of existing neural-network models,[9,10] to combine good chemical transferability with general conformational dependence. In Sec. IV, we discuss the training of three models—partial charges, environment-centered dipoles, and a combination of the two—which we collectively refer to as MuML. The models are fitted to reference calculations from high-end linear-response coupled-cluster calculations with single and double excitations (LR-CCSD) and yield $\boldsymbol{\mu}$ with an accuracy that is comparable to that of hybrid density functional theory (DFT). Next, a showcase set of larger and more complex molecules is used to test these models rigorously. Finally, we make a critical comparison of the performance of the different MuML models, which reveals the interplay of the different terms that contribute to molecular polarization.

II. THEORY

The molecular dipole moment is defined as the first moment of the total electric charge density,

$$\boldsymbol{\mu} = -\int \mathbf{r}\,\rho_e(\mathbf{r})\,\mathrm{d}^3\mathbf{r} + \sum_i \mathbf{r}_i Z_i, \tag{1}$$

where $\rho_e(\mathbf{r})$ is the electronic charge density, $\mathbf{r}_i$ is the position of the ith nucleus, and Z_i is its charge. Usually, we are concerned with the permanent dipole moment—that is, the first moment of the total charge density in the molecule's ground-state. However, this expression remains valid for non-equilibrium geometries as well as excited states.

This expression can be simplified by making the approximation that $\rho_e(\mathbf{r})$ is concentrated at individual atomic sites, that is, each atom i has an associated partial charge q_i resulting from the difference between Z_i and the partitioned electron density. The approximated *total* charge density is thus $\rho(\mathbf{r}) = \sum_i q_i \delta(\mathbf{r} - \mathbf{r}_i)$, and we can write

$$\boldsymbol{\mu} = \sum_i \mathbf{r}_i q_i, \tag{2}$$

which is uniquely defined with respect to the origin of the molecular coordinate system if the total molecular charge is zero. Charged molecules can be accommodated by setting the origin of the molecule to its centroid such that $\sum_i \mathbf{r}_i = 0$; this makes the dipole moment invariant to a collective shift of the q_i.

The problem then becomes the determination of the $\{q_i\}$ that best reproduces $\boldsymbol{\mu}$—often in addition to other physicochemical metrics, such as reproducing the molecular electrostatic potential (ESP) or characterizing chemical bonding. There are many existing methods to determine these charges, varying with the objectives of the model. Many methods are based directly on the ground-state charge density (or even the wavefunction), such as Mulliken[11] and Löwdin[12] population analyses, Hirshfeld decomposition[13] (and its iterative extension[14]), atoms-in-molecules [AIM, also known as quantum chemical topology (QCT)],[15] and iterative stockholder atoms (ISA).[16]

Another major class of atomic charge assignment methods, known collectively as ESP fitting methods, focuses directly on reproducing the molecular ESP rather than simply decomposing its charge density. One can immediately see the relevance of such methods to Eq. (2), as the far-field limit of the electrostatic potential is dominated by the dipolar term. ESP fitting methods were developed by Momany,[17] Cox and Williams,[18] Singh and Kollman,[19] and Breneman and Wiberg;[20] each of these methods finds the charges through a least-squares fit in order to reproduce the ESP at a grid of sites fairly close to the molecule but well outside the van der Waals radius. Notably, Momany[17] also fitted the total molecular dipole moment in order to satisfy Eq. (2). Many subsequent methods incorporate similar information into a fit that makes a compromise between chemical information (the charge density) and far-field electrostatics, such as the DDEC[21] and Hirshfeld-E[22] methods. However, such a compromise becomes a disadvantage when one is only interested in reproducing the molecular dipole moment.

Although the methods above are all motivated by physical and chemical principles, different methods can yield quite different results for the partial charges;[4,23] even worse, the results of certain methods may be very sensitive to the details of the underlying electronic structure calculation, such as the basis set used.[18,19,24,25] Furthermore, collapsing the total charge density to a set of points is often too severe an approximation to obtain an accurate description of the ESP.[4,23] One can therefore augment the expression in Eq. (2) to include information based on the atom-localized anisotropy or, informally, local polarization of the charge distribution by adding dipoles (or higher multipole moments) onto the atomic sites. This is the central idea behind the distributed multipole analysis (DMA) approach,[26] which gives for the total dipole

$$\boldsymbol{\mu} = \sum_{j\in\mathcal{C}} \left(\mathbf{r}_j q_j + \boldsymbol{\mu}_j\right), \tag{3}$$

where $\mathcal{C}$ is a list of centers (or points in real space) that include both atoms and interatomic positions, q_j is the partial charge associated with the jth center, and $\boldsymbol{\mu}_j$ is the associated partial dipole. We note in passing that higher multipole moments do not contribute to $\boldsymbol{\mu}$ and are therefore excluded from Eq. (3).

Several other methods use this idea of representing the molecular ESP with both charges and higher multipole moments assigned to atomic sites, such as the FOHI-D model[27] and the fullerene polarization model of Mayer,[28] the latter recently modified and incorporated into a QM/MM context (where the accurate reproduction of the far-field ESP is essential) as the FqFμ model.[29] The authors of FOHI-D, in particular, separated *intrinsic* atomic polarization, which can be calculated directly for the isolated atoms in the same iterative spirit as the classic iterative Hirshfeld method, from atomic *charge transfer*, which is described using the point-charge model.

However, they noted that the agreement of their model with the ESP is generally worse when dipoles are included, although this could have been due to their choice of grid points much closer to the molecule than is usually used for ESP-fitting methods. Mayer,[28] on the other hand, discussed the physical idea from the opposite perspective, that of adding atomic charges, derived from a procedure similar to the electronegativity equalization (EEQ) known in chemistry, to an atomic-dipole model in order to describe nonlocal polarization. The polarization of carbon nanostructures (nanotubes and fullerenes) is much better described by adding atomic charges to the description, as they can describe the large-scale flow of charge across the conjugated π-systems typical of these nanostructures.[28]

A key limitation of most of these methods is their inability to describe the dependence of electrostatic quantities across conformational and chemical space without performing additional *ab initio* calculations or fitting empirical parameters, which severely limits their ability to model experimental spectra and make transferable predictions for new molecules. A natural way to incorporate the required conformational and chemical sensitivity is to draw on the large body of work over the last two decades that uses ML to predict molecular properties[10,30–37] or molecular and intermolecular potential energy surfaces.[38–44]

Many existing methods are explicitly targeted to reproduce $\boldsymbol{\mu}$ or produce it as a side effect. The earliest of these is the neural network method of Darley, Handley, and Popelier[30] (see also Ref. 31), where a neural network is fitted to reproduce the multipole moments of a molecule or fragment computed via QCT (also known as atoms-in-molecules theory).[15] The two main drawbacks of this strategy, which are common to many of the other methods discussed here, are the following: (1) the need to define a local reference frame, which limits the method's transferability to other chemical compounds, and (2) the need to fit to a precomputed set of atomic charges and multipoles, the choice of which is ultimately arbitrary. The QCT charges and multipole moments, in particular, are known to be poorly convergent due to the irregular shapes of the partitioned atomic volumes.[4]

Techniques for fitting local electrostatic properties have evolved considerably since then, but most of the proposed methods retain these two key drawbacks. For example, the IPML model of Bereau *et al.*[34] predicts intermolecular interaction energies accurately by systematically treating several different physical energy contributions. The dipole moments themselves, on the other hand, are not as well predicted, given that their accurate reproduction is not the primary goal of the model. Part of the error may have come from using environment-local axis systems to predict the higher-order multipole moments, which is a less general and robust approach than the symmetry-adapted regression introduced by Grisafi *et al.*[8] Furthermore, the model retains the same drawback of being fitted to a specific partitioning scheme—in this case, the minimal-basis iterative stockholder method,[45] which was chosen for its accuracy in modeling electrostatic interactions and not for reproducing $\boldsymbol{\mu}$.

The neural network model of Gastegger, Behler, and Marquetand,[9] on the other hand, does explicitly target $\boldsymbol{\mu}$. It predicts the set of environment-dependent partial charges that best reproduces the total dipole moment, thereby bypassing the need to choose an arbitrary charge partitioning scheme, and uses the conformational sensitivity gained through the neural network to accurately predict infrared spectra. The PhysNet model of Unke and Meuwly[10] uses the same idea and additionally introduces a new representation to span a large swath of chemical space, as does the HIP-NN model of Sifain *et al.*,[46] which also incorporates enough conformational dependence to be able to predict infrared spectra. All three of these models only predict scalar atomic properties, neglecting contributions from atomic polarization, which we will see are important to achieving the best accuracy and transferability.

There are several approaches to fitting properties, such as $\boldsymbol{\mu}$, that transform as tensors—in particular, approaches that are covariant (rather than invariant) to rotations. The local-axis approach used by Bereau *et al.*[34] has already been mentioned; another approach is the covariant kernel approach introduced by Glielmo, Sollich, and De Vita[47] and developed into a general symmetry-adapted regression method for any tensor order by Grisafi *et al.*[8] This method was successfully tested on dipole moments of small molecules and clusters, as well as on accurately predicting higher-order tensors such as the polarizability.[37]

Finally, Christensen, Faber, and von Lilienfeld[48] developed a formalism (OQML) for incorporating electric field gradients into a ML fit. They used a modified version of the "FCHL" descriptor[33] augmented with a system of arbitrary, though usually realistic, partial charges in order to define an implicit local reference frame for each atomic environment, which can then be used to fit local dipole moments. The modified descriptor is called "FCHL*." While their formulation is quite different from the method developed below, we believe it is fundamentally similar to assigning an environment-dependent partial dipole to each atom, as described in Sec. III B.

III. METHODS

A. Partial-charge model

We begin by building a ML model that incorporates local environment sensitivity into the simple partial-charge model of Eq. (2) using Gaussian process regression (GPR).[49] To do this, we exploit the fact that GPR uses a linear fit in the kernel space and can therefore be used to fit the result of any linear operator applied to atomic quantities.[50] The vector of weights $\boldsymbol{w}$ is required, which minimizes the regularized loss function,

$$\mathcal{L}^2 = \|\mathbf{L}\mathbf{K}_{PM}\boldsymbol{w} - \mathbf{y}\|^2_{\Lambda^{-1}} + \|\boldsymbol{w}\|^2_{\mathbf{K}_{MM}}, \tag{4}$$

where Λ is a diagonal matrix whose entries σ^2_μ—a quantity known as the "dipole regularization," usually kept the same for all molecules—are chosen to optimize the error of the fit along with its transferability to new molecular databases, $\mathbf{L}$ is a linear operator, and

$$\boldsymbol{w} = (\mathbf{K}_{MM} + (\mathbf{L}\mathbf{K}_{PM})^T\Lambda^{-1}\mathbf{L}\mathbf{K}_{PM})^{-1}\mathbf{L}\mathbf{K}_{PM}\Lambda^{-1}\mathbf{y}. \tag{5}$$

The fit uses an "active set" of M basis functions (which, in practice, is a small fraction of the total number P of atoms in the database). Following the notation introduced in Ref. 50, we use M and P to indicate both the sets and the number of entries. The kernel matrices K_{MM} and K_{PM} contain the kernel evaluated between all sparse points (M) and themselves, as well as with all atoms in the training database (P). In principle, any sufficiently representative set of configurations

could be used to form the active set of basis functions, but, in practice, they are almost always chosen from the environments present in the molecules in the training set using an algorithm such as farthest point sampling (FPS) or a CUR decomposition.[51] The entries of the kernel matrix are $(K_{IJ})_{ij} = k(\mathcal{X}^{(i)}, \mathcal{X}^{(j)})$, where the SOAP kernel[41] is used as the similarity function $k(\cdot,\cdot)$ between two atomic environments $\mathcal{X}^{(i)}$ and $\mathcal{X}^{(j)}$.

To build up a model for $\boldsymbol{\mu}$, we predict partial charges $q(\mathcal{A}_i)$ for atom i in molecule $\mathcal{A}$,

$$q(\mathcal{A}_i) = \sum_{j\in M} w_j k(\mathcal{X}^{(i)}, \mathcal{X}^{(j)}) \tag{6}$$

(where the sum runs over all basis points, i.e., environments in the active set) such that

$$\boldsymbol{\mu}(\mathcal{A}) = \sum_{i\in\mathcal{A}} \mathbf{r}_i q(\mathcal{A}_i). \tag{7}$$

We can then define the transformed kernel matrix between dipoles and basis points as

$$(\mathbf{L}_\mu \mathbf{K}_{PM})_{\mathcal{A},j} = \sum_{i\in\mathcal{A}} \mathbf{r}_i k(\mathcal{X}^{(i)}, \mathcal{X}^{(j)}), \tag{8}$$

allowing us to use Eq. (5) to determine the weights. The expression in Eq. (8) represents a block of three rows from the $3N \times M$ matrix $\mathbf{L}_\mu \mathbf{K}_{PM}$, where N is the number of molecules in the training set, one row for each Cartesian component of the dipole moment of molecule $\mathcal{A}$. The columns index j runs over the M environments in the active set. The $\mathbf{r}_i$ are defined with respect to the coordinate system in which the dipole is given, with the origin set to the centroid of the respective molecule so that the prediction is insensitive to a shift in the total charge. The target data $\mathbf{y}$ are then defined as a concatenation of the Cartesian components of the training-set dipole moments.

The insensitivity of the model to the total molecular charge is advantageous because the model's total charges, $Q_\mathcal{A} = \sum_{i\in\mathcal{A}} q_i$, need not be constrained to reproduce exactly the total molecular charge. As noted by Unke and Meuwly,[10] applying this constraint to the training set would not guarantee that the model gives the correct charges for prediction on a new molecule. Furthermore, we found that including the exact total-charge constraints into the fit via Lagrange multipliers severely reduced the quality of the fit—in most cases simply giving all partial charges as zero—because the procedure used to select the sparse active set of M environments also discarded the basis functions necessary to satisfy this constraint while also satisfactorily reproducing $\boldsymbol{\mu}$.

However, it is usually beneficial to include some sort of restraint (even if not an exact *constraint*) on the total charge, as a model insensitive to this quantity can predict unreasonably large total charges, ultimately compromising its transferability to other datasets. We therefore include the total charge as extra information to the fit by appending to $\mathbf{y}$ the list of total charges of the molecules in the training set, and appending to the transformed kernel matrix $\mathbf{L}_\mu \mathbf{K}_{PM}$ the extra N rows representing the sums of the model's partial charges,

$$(\mathbf{L}_Q \mathbf{K}_{PM})_{\mathcal{A},j} = \sum_{i\in\mathcal{A}} k(\mathcal{X}^{(i)}, \mathcal{X}^{(j)}), \tag{9}$$

and extending the diagonal regularization matrix Λ with an extra N entries σ_Q^2—the charge regularizer—in order to be able to regularize the two target quantities separately.

B. Partial dipole model

An alternative method for predicting $\boldsymbol{\mu}$ is to build up the prediction as a sum of atom-centered dipole moment predictions using symmetry-adapted Gaussian process regression (SA-GPR),[8] a modification of standard GPR that allows tensor properties to be learned. A SA-GPR prediction of the dipole moment $\boldsymbol{\mu}$ of a test molecule $\mathcal{A}$ is given by

$$\boldsymbol{\mu}(\mathcal{A}) = \sum_{j\in M}\sum_{i\in\mathcal{A}} \mathbf{k}^V(\mathcal{X}^{(i)}, \mathcal{X}^{(j)})\boldsymbol{w}_j, \tag{10}$$

where $\mathbf{k}^V(\mathcal{X}_i, \mathcal{X}_j)$ is an element of an extended kernel matrix, being the tensor (concretely a 3×3 matrix) whose components $k^V_{\alpha\beta}(\mathcal{X}_i, \mathcal{X}_j)$ give the coupling between the Cartesian components $\mu_\alpha^{(i)}$ associated with environment $\mathcal{X}_i$ and $\mu_\beta^{(j)}$ associated with $\mathcal{X}_j$. Each environment j in the active set now requires a set of three weights (represented by the vector $\boldsymbol{w}_j$) to represent the three independent components of the vector quantity assigned to each atom.

Since the dipole moment is a vector quantity that is related by a linear transformation to the spherical harmonics with $L = 1$, the vector kernel $\mathbf{k}^V(\mathcal{X}, \mathcal{X}')$ can be obtained directly from the $\lambda = 1$-order λ-SOAP kernel of SA-GPR, $\boldsymbol{k}^{\lambda=1}(\mathcal{X}, \mathcal{X}')$, by the transformation,

$$\boldsymbol{k}^V(\mathcal{X}, \mathcal{X}') = \boldsymbol{M}^\dagger \boldsymbol{k}^{\lambda=1}(\mathcal{X}, \mathcal{X}')\boldsymbol{M}, \tag{11}$$

where $\mathbf{M}$ transforms from the Cartesian basis to the basis of $\lambda = 1$ spherical tensors (see, e.g., Ref. 4 for an explicit formula).

As has been shown by recent work,[8,37,52] SA-GPR performs very well for the response properties of different orders in a wide variety of systems. Furthermore, one can see from the atom-centered formulation of Eq. (10) that the atom-centered dipoles, analogous to the atomic partial charges of Eq. (6), can easily be extracted,

$$\boldsymbol{\mu}(\mathcal{A}_i) = \sum_{j\in M} \mathbf{k}^V(\mathcal{X}^{(i)}, \mathcal{X}^{(j)})\boldsymbol{w}_j. \tag{12}$$

Although an SA-GPR prediction of $\boldsymbol{\mu}$ does not require charge constraints, it is computationally more expensive than a partial-charge model, requiring the inversion of a square matrix with three times the number of rows ($3M$ rows, where M is the number of basis functions in the active set).

C. Combined model

We now consider the partial-charge model and partial-dipole model as two separate models for the same system, encoding two different physical effects. It should then be possible to get a better prediction simply by fitting the *sum* of the two models to the training data. We call the matrix of Eq. (8) the "transformed scalar kernel,"

$$\mathbf{K}^S_{NM} := \mathbf{L}_\mu \mathbf{K}_{PM}, \tag{13}$$

and the analogous "transformed vector kernel" $\mathbf{K}^V_{NM'}$ whose rows are the atom-wise summations of the kernel from Eq. (10),

$$(\mathbf{K}^V_{NM'})_{\mathcal{A},j} = \sum_{i\in\mathcal{A}} \mathbf{k}^V(\mathcal{X}^{(i)}, \mathcal{X}^{(j)}). \tag{14}$$

Because the models are of different dimensions and correspond to different physical effects, we assign each a different weight vector: $\boldsymbol{w}^S$ for the scalar weights and $\boldsymbol{w}^V$ for the vector weights. Also note

that this means we do not need to use the same set of basis functions for the scalar and vector models; they can be chosen independently. Then, in order to find the best combined sum model, we optimize

$$\mathcal{L}^2 = \|\delta_S^2 \mathbf{K}^S_{NM} \boldsymbol{w}^S + \delta_V^2 \mathbf{K}^V_{NM'} \boldsymbol{w}^V - \mathbf{y}\|^2_{\Lambda^{-1}} + \delta_S^2 \|\boldsymbol{w}^S\|^2_{\mathbf{K}^S_{MM}} + \delta_V^2 \|\boldsymbol{w}^V\|^2_{\mathbf{K}^V_{M'M'}} \quad (15)$$

with respect to both sets of weights $\boldsymbol{w}^S$ and $\boldsymbol{w}^V$ simultaneously. (The $\mathbf{K}^S_{MM}$ and $\mathbf{K}^V_{M'M'}$ are the matrices of non-transformed kernels of all the basis functions with each other.) The result can be expressed using the inversion of a square matrix with $M + M'$ rows, where M' is the number of vector weights (three times the number of vector basis functions). Since the number of basis functions is usually kept the same for both scalar and vector models, the matrix to be inverted has $4M$ rows, making the combined model the most expensive of the three models discussed here. In practice, however, the cost is typically manageable.

Furthermore, the charge restraint can be incorporated as discussed in Sec. III A, where the transformed scalar kernel is appended with the matrix from Eq. (9), and the transformed vector kernel is appended with the same number of rows of zeros (since the vector model does not contribute to the total molecular charge). Also note that we have introduced weights δ_S and δ_V to modify the overall relative amount that the scalar and vector components contribute to the combined model. The δ-weights effectively allow for different regularizations of the scalar and vector components of the model, which is equivalent to assuming different variances for the dipole components modeled by the scalar and vector models.[54]

IV. RESULTS AND DISCUSSION

We optimized and trained the scalar, vector, and combined models on the QM7b dataset,[55] which contains 7211 small organic molecules with up to seven heavy/non-hydrogen atoms (specifically C, N, O, S, and Cl) with varying degrees of H saturation. The dipoles were computed using the methods described by Yang *et al.*,[56] namely, DFT with the hybrid B3LYP functional[57,58] and linear-response coupled-cluster theory with single and double excitations (LR-CCSD,[59] hereafter just "CCSD"). In both cases, the doubly augmented double-ζ d-aug-cc-pVDZ basis set[60] (hereafter referred to as "daDZ") was employed during all calculations. We then demonstrate the transferability of this model on the QM9[61] dataset, comparing with the state-of-the-art results from Ref. 48, and on a "MuML showcase" dataset of larger molecules. Finally, we push the models to their limits by studying different polymers composed of or derived from the glycine amino acid.

A. Model optimization

We first optimize the models for space and computational requirements by subsampling the SOAP feature matrices (which are multiplied and raised to an integer entry-wise power to obtain the SOAP kernel) using the FPS selection algorithm described by Imbalzano *et al.*[51] Descriptors are first subsampled in the feature space dimension, allowing for fewer SOAP components (N_F) to be used in calculating the kernel, then in the environment space dimension, allowing for fewer representative environments (M) to be used when performing the fit. The convergence of the final fitting error with respect to these parameters, as well as other kernel convergence parameters such as the number of radial channels ($n_{\max}$) and the maximum angular momentum ($l_{\max}$) of the expansion, is shown in the supplementary material; final values are given in Table I.

We chose the model's overall distance-based cutoff as 5 Å to encompass all atom pairs in the QM7b dataset. The actual radial dependence of the kernels, however, is optimized using the radial-scaling function from Willatt, Musil, and Ceriotti.[53] Together with the SOAP atom width and the regularization parameters, this leaves us with several continuous hyperparameters whose optimal values need to be determined. In a Bayesian approach, these would be considered priors; they would ideally be integrated over using a previously known prior distribution. Here, however, we do not have much prior knowledge about the distribution of these parameters—in contrast to the study of potential energy surfaces, where good values can be guessed quite accurately based on prior experience and physical knowledge.[62] Instead, we use optimization to find the best values of these parameters for our problem, along with cross-validation (CV) to guard against the problem of overfitting (which is otherwise introduced by hyperparameter optimization techniques).

First, the hyperparameters for the scalar and vector models are each independently optimized on a randomized fourfold CV split of 5400 randomly selected molecules of the QM7b test set.[55] The results of this optimization are given in Table I. The combined model is then obtained as follows: since there are only three free parameters between the overall scalar weight δ_S, the overall vector weight δ_V, the dipole regularization σ_μ, and the total charge regularization σ_Q, we set the dipole regularization to 1 and scale the rest of the parameters accordingly: If σ^S_μ is the optimal dipole regularizer, σ^S_Q the optimal charge regularizer for the scalar model, and σ^V_μ is the optimal dipole regularizer for the vector model, then we take $\delta_S \mapsto 1/\sigma^S_\mu$, $\sigma_Q \mapsto \sigma^S_Q/\sigma^S_\mu$, and $\delta_V \mapsto 1/\sigma^V_\mu$. Further details of the optimization procedure are discussed in the supplementary material.

Finally, once the model's hyperparameters are converged and optimized, model training and testing are quite fast. For example, computing scalar and vector training and testing kernels for the set of $N = 20\,000$ molecules of QM9 used in Sec. IV E (with a test set of $T = 1000$ molecules; $M = M' = 2000$) required just over 1 h and 95 GiB of memory on a modern 24-core machine, with almost all of the time and memory used to compute the training kernels; the test kernels required less than 2 min and 3 GiB. Once the kernels were computed, fitting the combined (most expensive) model required only 2 min and 20 GiB of memory, and computing test-set predictions was almost negligible in comparison, taking 2 s and 1 GiB of memory. This means that the regularizers can be optimized quite cheaply once optimal kernels have been computed.

Table I. Convergence parameters for the scalar and vector kernels: n_{max} is the number of radial basis functions and l_{max} is the angular momentum band limit for the SOAP kernel, N_F is the number of selected sparse features, and M is the number of selected sparse environments for each model. Note that the scalar and tensor *power spectrum* components of the vector SOAP kernel use the same parameters.

Model	$n_{\max}$	$l_{\max}$	N_F	M
Scalar	8	6	200	2000
Vector	4	2	200	2000

TABLE II. Optimal hyperparameters for the pure scalar and pure vector models, obtained using a Nelder–Mead optimization. θ_a: Gaussian width for SOAP atom smearing; r_0 and m: radial scaling parameters (see Ref. 53); σ_μ: dipole regularization (unitless since $\delta = 1$ for the pure fits) and σ_Q: total charge regularization. Parameters for the combined model are derived as indicated in the text. All numbers truncated to three significant figures.

Model	θ_a (Å)	r_0 (Å)	m	σ_μ (10^{-3})	σ_Q (10^{-3} Å^{-1})
Scalar (CCSD)	0.375	2.32	4.41	4.38	35.5
Scalar (B3LYP)	0.375	2.32	4.41	4.41	78.8
Vector (CCSD)	0.256	2.75	3.34	1.47	
Vector (B3LYP)	0.256	2.75	3.34	1.15	

B. Error measures

Throughout this work, we use two different error measures. The "per-atom" RMSE (root-mean-squared error)

$$\mathrm{RMSE} = \sqrt{\frac{1}{N_{\mathrm{test}}}\sum_{j\in\mathrm{test}}\left\|\frac{\boldsymbol{\mu}^{(j)}_{\mathrm{predicted}} - \boldsymbol{\mu}^{(j)}_{\mathrm{actual}}}{N_j}\right\|_2^2} \tag{16}$$

reports on both the magnitude and the orientation of the predicted dipoles. The residuals are normalized by the number of atoms N_j in the respective molecule before taking the RMSE. This scaling posits a generally linear trend of the dipole moment norm as a function of the number of atoms. Such a trend would be expected from an additive model where each atom contributes a certain, locally dependent amount. This is the case with the vector model, but not with the scalar model, where the contribution additionally depends on its distance from the molecular origin, making the scaling depend on the molecular geometry. Therefore, to provide an alternative assessment of the error of the total dipole and to facilitate comparison with other studies, we additionally plot the MAE (mean absolute error) of the norm of the *total* dipole moment,

$$\mathrm{MAE} = \frac{1}{N_{\mathrm{test}}}\sum_{j\in\mathrm{test}}\left|\|\boldsymbol{\mu}^{(j)}_{\mathrm{predicted}}\|_2 - \|\boldsymbol{\mu}^{(j)}_{\mathrm{actual}}\|_2\right|. \tag{17}$$

For the QM7b dataset these two measures provide similar information, but for transferability testing on other datasets these measures provide complementary information.

C. Uncertainty quantification

We can estimate the uncertainty in the model predictions using a calibrated committee model, as described by Musil *et al.*[63] We train n_{comm} models $\tilde{\boldsymbol{\mu}}^{(k)}(\mathcal{A})$ using the same active set but choosing a different random subset of the full training set in each model. The predictions of these models are then rescaled around their mean,

$$\begin{aligned}\bar{\boldsymbol{\mu}}(\mathcal{A}) &= \frac{1}{n_{\mathrm{comm}}}\sum_k \tilde{\boldsymbol{\mu}}^{(k)}(\mathcal{A}),\\ \boldsymbol{\mu}^{(k)}(\mathcal{A}) &= \bar{\boldsymbol{\mu}}(\mathcal{A}) + \alpha\left(\tilde{\boldsymbol{\mu}}^{(k)}(\mathcal{A}) - \bar{\boldsymbol{\mu}}(\mathcal{A})\right),\end{aligned} \tag{18}$$

by a calibration factor α that is determined using the "internal validation procedure" described in Ref. 63. The best estimate of the committee model is given by its mean, $\bar{\boldsymbol{\mu}}(\mathcal{A})$, and uncertainty is then computed as the standard deviation of the rescaled predictions. Individual members of the calibrated committee can be used to separately compute the derived quantities (e.g., the norm of the dipole moment), which greatly simplifies the propagation of uncertainty (see, e.g., Ref. 52). While the use of a committee model for a sparse Gaussian process model entails virtually no computational overhead when making a new prediction, the training process is somewhat more cumbersome. For this reason, we only use a committee model when making predictions for the showcase dataset in Sec. IV G. More systematic tests performed on benchmark datasets use a single regression model, without error estimation, which usually also achieves a higher accuracy than the ensemble average (see the supplementary material) because it is trained on all training points together.

D. Training on QM7b

Figure 1 shows the learning curves of the MuML models, with the kernel parameters fixed to the values optimized on 5400 points. Errors are computed on a test set of 1811 randomly selected molecules from the QM7b dataset.[55] Note that both the pure scalar and pure vector models achieve similar performance in the limit of a large amount of data, while the combined model clearly outperforms both (by a factor of about 20%) in the same regime.

This figure reports the results for models trained on CCSD/daDZ dipoles. The results for B3LYP/daDZ-trained models are very similar (see the supplementary material). For reference, the discrepancy between B3LYP/daDZ and CCSD/daDZ molecular dipole moments in the QM7b database amounts to an RMSE = 0.011 D per atom or MAE = 0.087 D. It should be stressed that contrary to the case of the polarizability,[37,56] the performance of DFT is usually quite satisfactory when predicting molecular dipole moments. When trained on 5400 QM7b structures, the combined model delivers better accuracy (RMSE = 0.0086 D per atom and MAE = 0.054 D) at a dramatically reduced computational cost.

E. Testing on QM9

In order to test the extrapolation capabilities of the MuML models, we selected 1000 random samples from the QM9 dataset[61] and computed the dipole moments following the same protocol used for the QM7b dataset.[37,56] Due to the high computational cost of CCSD, we used B3LYP/daDZ as the reference in this case and the corresponding models trained on QM7b at the B3LYP/daDZ level. The learning curves of these models are shown in Fig. 2. The combined model outperforms the scalar and vector models in terms of the per-atom RMSE measure but performs worse than the vector model using the norm MAE. The errors are much larger than those seen when testing on QM7b, and the asymptotic behavior of the learning curves indicates saturation and even overfitting. In order to determine whether the saturation in model performance is due to limitations in the models or just insufficient training data, we also computed learning curves for models trained on QM9 dipoles using a set of 20 000 additional molecules drawn from the QM9 set and dipoles computed at the B3LYP/daDZ level. The scalar and vector regularizers were re-optimized using 15 000 training points. The QM9-trained models, in contrast to the QM7b-trained models, do not saturate early. The QM9-trained combined model reaches an MAE of 0.084 D; this is more accurate than the QM9-trained scalar

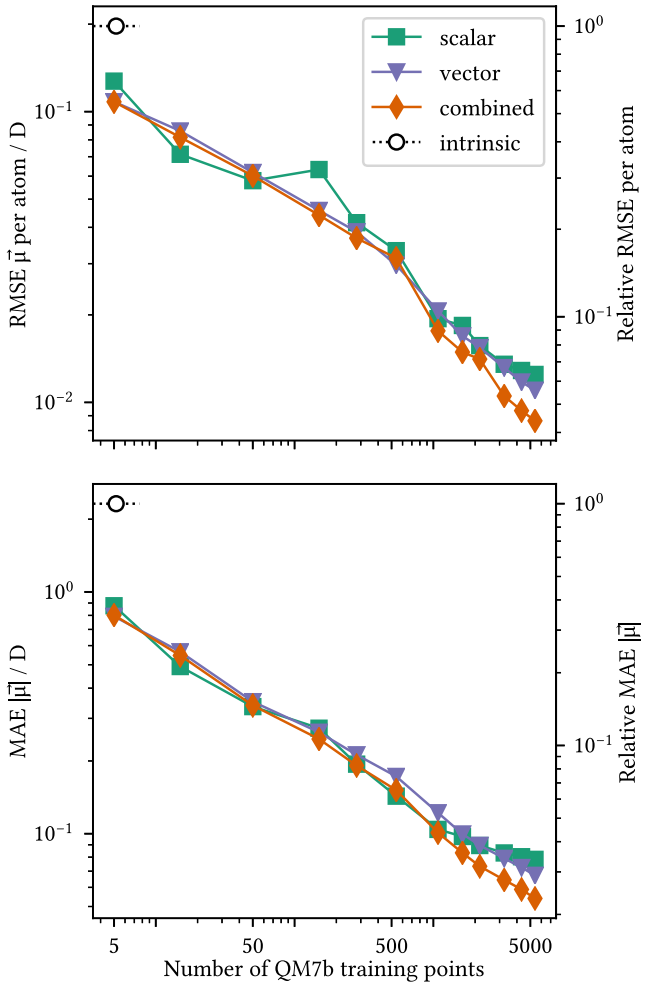

FIG. 1. Learning curves of the scalar (green squares), vector (purple triangles), and combined (orange diamonds) models for CCSD/daDZ dipoles computed on 1811 randomly selected molecules of the QM7b dataset. The models were trained on subsets of the remaining 5400 molecules. The top plot has per-atom RMSEs, and the bottom plot has per-molecule dipole moment norm MAEs. The open circle denotes the intrinsic variation of the dataset, i.e., the error of a zero model.

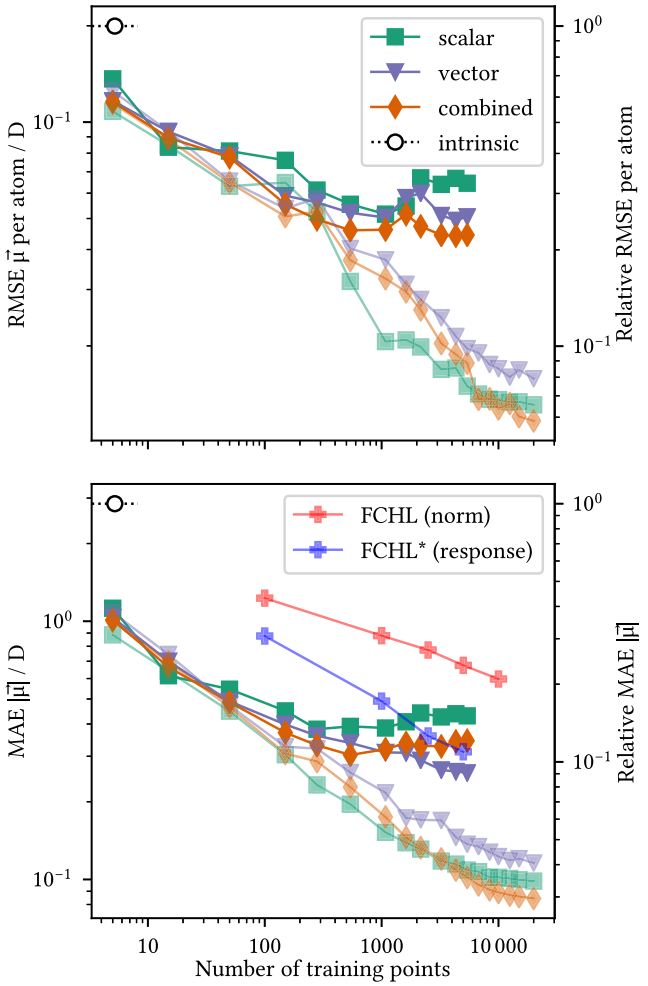

FIG. 2. Learning curves on a random sample of 1000 molecules from the QM9 dataset.[61] Reference dipoles were computed with B3LYP/daDZ; MuML models were re-trained on QM7b dipoles (solid symbols) and QM9 dipoles (semi-transparent symbols) computed at the same level of theory. Top: per-atom RMSEs. The QM7b combined fit narrowly outperforms the pure charge and pure dipole models, with significant saturation apparent in all QM7b models. No such saturation is apparent in the QM9 models. Bottom: MAE of the error of the dipole moment norms for each molecule. The FCHL (norm-only) and FCHL* (vector response) curves are reproduced from Christensen, Faber, and von Lilienfeld;[48] both models were trained on QM9 dipoles. Using this error measure, the QM7b pure vector fit has a clear advantage, even outperforming the FCHL* response learning. The QM9 fits again perform significantly better than the QM7b fits; the QM9 combined fit retains the best performance, reaching an MAE of 0.084 D at 20 000 training points.

model (MAE 0.099 D), which is, in turn, more accurate than the QM9-trained vector model (MAE 0.12 D). In fact, this is comparable to the performance of the SchNet neural network model,[35] which reaches an accuracy of 0.033 D using 110 000 training molecules. It is likely that the QM9-trained combined model would reach the same accuracy if the slight saturation in the MAE curve were corrected, e.g., by increasing the SOAP convergence parameters (which were set for 5400 QM7b molecules) and re-optimizing the hyperparameters. The comparison of the QM7b-trained models to those trained on QM9 clearly shows that the QM7b scalar model especially suffers in the extrapolative regime. Together with the degrading performance of both RMSE and MAE as the number of training points approaches the full training set size, this indicates that the scalar model has a strong tendency to overfit. As for the combined model, it seems that its poor performance is a result of its inclusion of too much of the overfitted scalar component.

We therefore investigate the dependence of the model error on the scalar–vector mixing to see if the combined model can be improved by including less of the scalar component. The scalar–vector mixing is parameterized here by varying the scalar model's variance δ_S^2 from zero to its pure-scalar equivalent value: $\delta_S^2(t) = t(\sigma_\mu^S)^{-2}$ while simultaneously varying the vector model's

variance from its pure-vector equivalent value to zero: $\delta_V^2(t) = (1-t)(\sigma_\mu^V)^{-2}$. The dipole regularization is kept at one, as the regularization is encoded in the model variances; the total-charge regularization σ_Q is likewise kept constant, at its optimal scalar-model value, as it only applies to the scalar model. This parameterization reproduces the pure scalar and pure vector predictions at each endpoint while smoothly transferring the total model's variance from the vector to the scalar model. The value of $t = 0.5$ corresponds to the combined model (modulo a factor of 2 in the regularizer, which is negligible in practice). Note that varying the scalar and vector weights δ_S and δ_V is more than a simple post-processing adjustment; it requires recomputing the model weights $\boldsymbol{w}$ via Eq. (15) as well. We plot such a scalar–vector scan in Fig. 3 for the models trained on either QM7b or QM9 and tested on either QM7b (QM7b-trained model only) or QM9 (both models). Both models were trained on 5400 molecules using dipoles computed at the B3LYP/daDZ level.

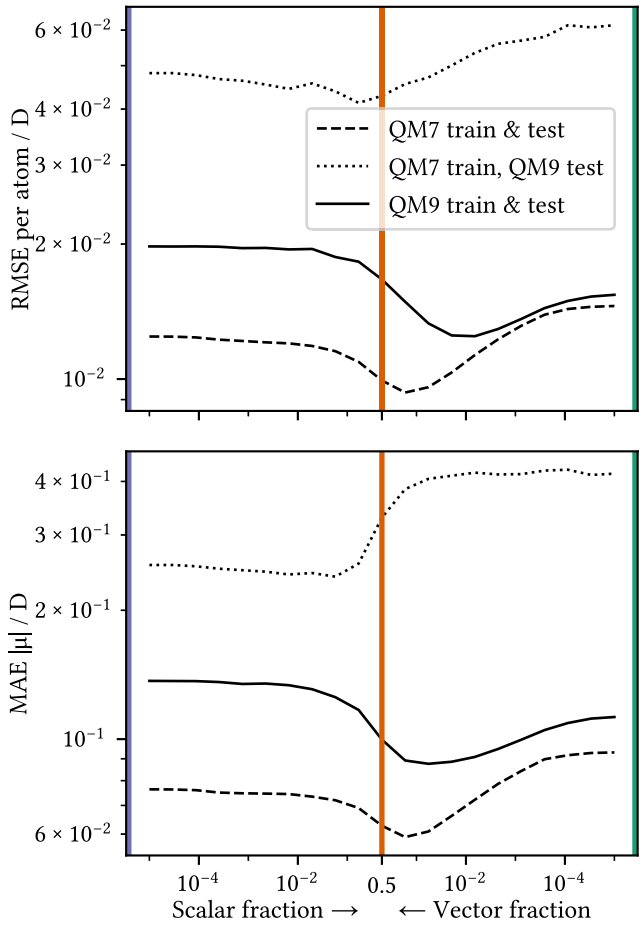

FIG. 3. Errors when interpolating between the pure scalar and pure vector B3LYP/daDZ MuML models. The pure vector model is located at the left (as the scalar fraction $t \to 0$), the pure scalar model is located at the right (as the vector fraction $1 - t \to 0$), and the combined model is located in the middle at 0.5. The models are trained on either 5400 points of QM7b or 5400 points of QM9 and tested on either the remaining 1811 molecules of QM7b or the 1000 randomly selected molecules of QM9.

We see that the optimum for models tested in the interpolative regime—that is, QM7b tested on QM7b and QM9 tested on QM9—does not in fact lie at $t = 0.5$, but closer to the pure-scalar model (vector fraction of 0.1 or 0.01, depending on the model and whether one wants to optimize MAE or RMSE). The naïve QM7b combined model at $t = 0.5$ is still better than either the pure scalar or pure vector models (this is also the case with the QM9 model once we add more training points). On the other hand, for the QM7b model in its extrapolative regime (i.e., tested on QM9), the situation is the opposite: the optimal model has a *scalar* fraction of around 0.1, and the naïve combined model at $t = 0.5$ is even worse in MAE than the pure vector model, as shown in Fig. 2.

These observations confirm our suspicions that the scalar model is prone to overfitting, as it achieves very good performance in the interpolative regime, but relatively poor performance in the extrapolative regime. Models with a higher fraction of the vector contribution, on the other hand, may not achieve the same accuracy in the interpolative regime, but they are better at extrapolating (i.e., they are more transferable). Following these observations, it may be possible to derive a strategy for adjusting the combined weights to achieve the best accuracy on a variety of testing sets. Although we do not explore such a strategy in this work, we do comment further on the interplay between these two contributions in Sec. IV G.

F. Comparison with OQML

It is interesting to compare the performance of our models to that of the operator quantum machine learning (OQML) scheme in Ref. 48. In OQML, a formal dependence on an applied electric field is included in the definition of the (scalar) kernel by assigning fictitious charges to each atom. This makes it possible to define derivatives of the kernels relative to an applied field, which are naturally covariant and serve as a basis to fit molecular dipoles. It should be stressed that, even though the scheme relies on formal atomic charges, it amounts effectively to learning local dipoles and is therefore similar to our vector model. While OQML couples the energy and regression dipole models through a scalar constant, our approach allows every property to be trained independently.

As shown in Fig. 2, the QM7b vector model (the most transferable of the QM7b models) outperforms the FCHL* OQML model by approximately 20%. This is particularly remarkable because the OQML model of Ref. 48 was trained on 5000 structures from QM9; the QM7b models, on the other hand, are trained on smaller structures, and are therefore functioning in the much more challenging extrapolative regime. This is in contrast to the QM9 scalar, vector, and combined models, which are functioning in the interpolative regime in this test. Here, we observe that the slopes of the QM9–QM9 learning curves are approximately the same as that of the FCHL* (response) curve, but they have a large offset. In other words, the MuML models achieve an MAE of about 1/3 that of FCHL* with the same amount of data.

G. MuML showcase dataset

Similar to Ref. 37, we now turn from standard, systematically generated benchmark datasets to a showcase dataset in which

chemically relevant molecules have been specifically chosen to test the sensitivity of the ML model to subtler variations in chemical structure and bonding. To this end, we assembled the so-called MuML showcase dataset, which is depicted in Fig. 4 and comprised of the first 29 molecules of the AlphaML showcase dataset[56,64] (and includes the nucleobases, amino acids, sugars, and common drug molecules). The C_8H_n isomers from the AlphaML showcase dataset were discarded (because they all have very small dipole moments) and substituted with 31 $C_4H_nNH_2COOH$ amino acid derivatives, with dipole moments spanning a broad range from 0.5 D to 6 D. Molecular geometries and dipole moments for these new molecules were obtained using the same protocol described in Refs. 37 and 56. For reference, the dipole moment norms computed with CCSD/daDZ on the MuML showcase are shown in the top panel of Fig. 5(b).

The learning curves of the three dipole models on the MuML showcase dataset are shown in Fig. 5(a). All three models achieve an accuracy comparable, in absolute terms, to that on QM9. The (unadjusted) combined model narrowly outperforms both the scalar and vector models. Even in this extrapolative regime, the accuracy of MuML is competitive with that of B3LYP: for the largest training set size, MuML achieves errors (RMSE = 0.029 D per atom, MAE = 0.24 D) that are only 30% larger in RMSE (56% larger in MAE) than those of B3LYP relative to CCSD (RMSE = 0.019 D per atom, MAE = 0.19 D). The dramatic increase in accuracy observed when training on the larger QM9 molecules (see Fig. 2) suggests that it is possible to train a MuML model that will outperform DFT on this showcase dataset. Unfortunately, the cost of performing LR-CCSD calculations on thousands of QM9-sized molecules is still prohibitive at the current time.

Due to the relatively small number of molecules in the MuML showcase set, we can examine the performance of the MuML models for each of the molecules individually. Furthermore, we can also benchmark the uncertainty quantification scheme discussed in Sec. IV C. Each of the eight models in the committee model was trained on a sample of 2700 molecules (50% of the full training set) drawn without replacement from the QM7b training set. The calibrated error predictions were then validated against the QM7b test set; additional details are found in the supplementary material. The overall errors of the ensemble averages are comparable to (if slightly higher than) those of the model trained on all 5400 points.

Figure 5(b) shows the breakdown of the errors of the ensemble average, along with the uncertainties predicted from the ensemble. Note that the errors are shown reversed from the usual convention—they are shown as reference minus predicted—and the error bars are shown centered about zero. Both the predicted uncertainties and the ensemble-average residuals show no apparent systematic patterns across this set of molecules, although there are some outliers. All three MuML models perform particularly poorly on molecule 14 (cysteine), and the uncertainty estimate is also relatively high for this molecule. The evidence suggests that the high errors and uncertainties are a consequence of the highly polarizable nature of sulfur, given that the models also give large overpredictions and high uncertainties in the case of methionine, the only other S-containing molecule in the MuML showcase dataset. Other relatively large errors (and large uncertainties) are seen on all models for molecule 1 (guanine), molecule 21 (caffeine), molecule 24 (metformin), and molecule 25 (acyclovir); the vector and combined models additionally give large errors and uncertainties for molecule 4 (cytosine) and molecule 23 (aspirin).

Overall, the prediction errors are consistent with the error bars, with 88% of the scalar predictions, 55% of the vector predictions, and 72% of the combined predictions falling within one error bar

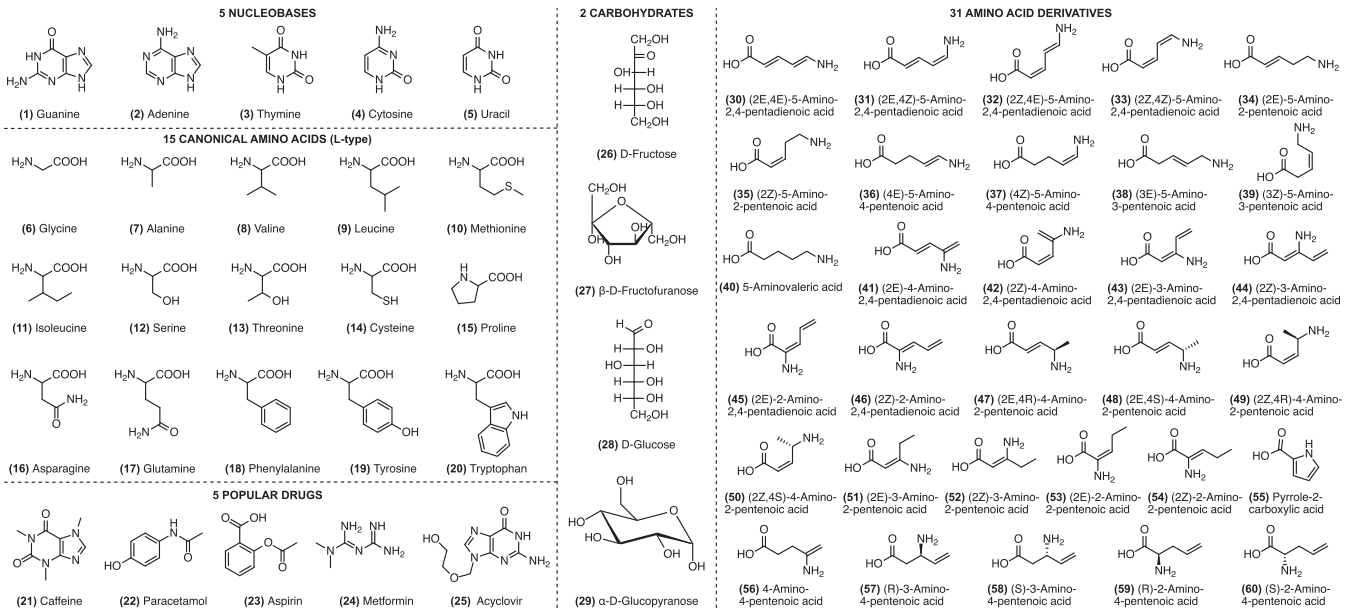

FIG. 4. List of molecules included in the MuML showcase dataset. The numerical key is used to identify the various compounds in other figures.

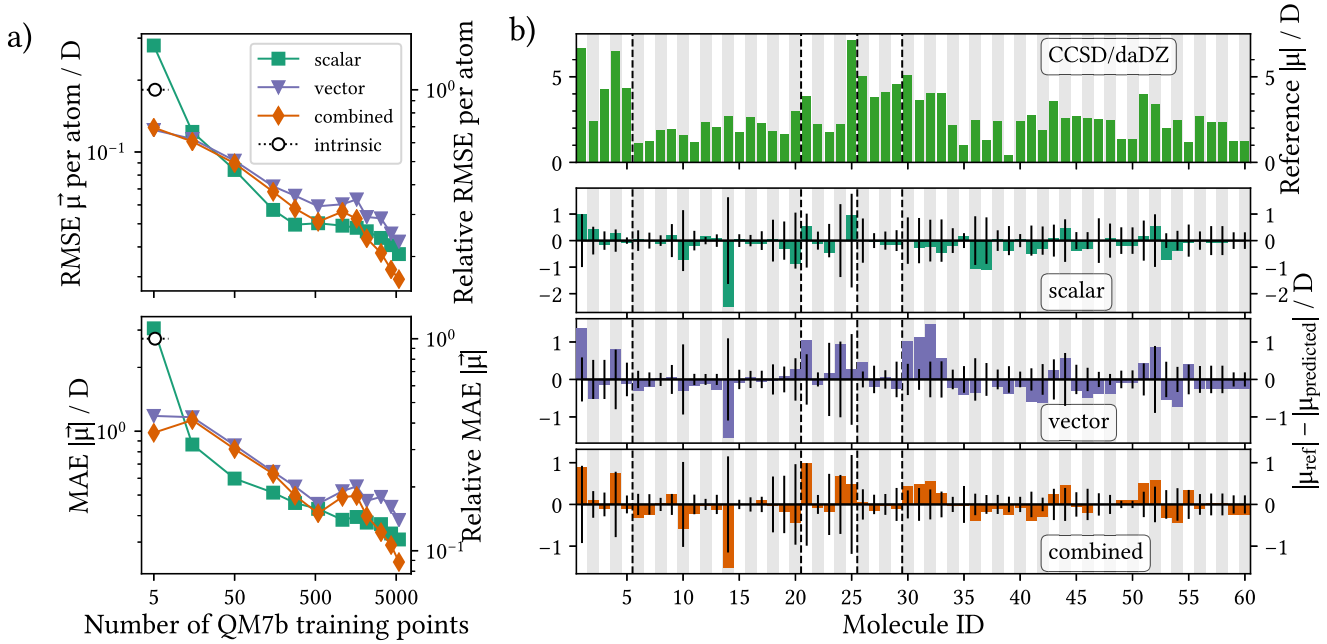

FIG. 5. Performance of the MuML models on the MuML showcase dataset: (a) learning curves by per-atom RMSE (top) and MAE (bottom), (b) per-molecule breakdown of the MuML models trained on the full QM7b training set (5400 molecules): norms of the reference dipole moments computed with CCSD/daDZ on the MuML showcase dataset (top) and errors in the norms of the dipole moment predictions across the same set (bottom three). Prediction errors are shown along with error bars from an ensemble of models trained on subsets of the full training set.[63] The molecule ID is in reference to Fig. 4.

of the reference (compare this with the 68% expected if the prediction errors were to follow a Gaussian distribution with a standard deviation equal to the error bar). Thus, the uncertainty quantification scheme applied herein provides a reliable estimate of the model accuracy, improving our interpretation of the model results in the extrapolative regime where the errors can be several times larger than those in the original testing set.

The only cases in which the predictions are farther than two error bars from the reference are that of molecules 30, 31, and 32: these show large errors but small uncertainties in the vector model. Together with the similar structure of these molecules—they are all polyenoic amino acids, effectively an amine group and a carboxylic acid joined by a fully conjugated polyacetylene chain/linker—these deviations suggest a *systematic* error in the vector model predictions. The delocalized nature of the conjugated chains in these molecules suggests that the error could derive from a non-local effect that the vector model, with its finite cutoff and strictly local environmental dependence of the atomic dipoles, fails to capture.

In order to provide a more systematic, and far more stringent, test of our models' extrapolative capabilities, as well as to investigate the effect of non-local effects on each of the models, we designed four new "challenge" test sets, each of which consists of a series of approximately linear (pseudo-1D) molecules with polar groups and (in three of the four sets) large separations of charge, thereby giving rise to large dipole moments. More specifically, we considered polymers of the glycine amino acid, in both the α-helix and β-strand configurations, as well as a series of polyenoic amino acids, with an amine group and a carboxylic acid group separated by a polyacetylene spacer. Finally, a set of *n*-amino carboxylic acids (the saturated analogs of the polyenoic amino acids) was included to investigate the effect of saturation in the spacer on the molecular dipole moments and the model predictions. Because of the large size of these molecules (up to 122 atoms, of which 69 were heavy/non-H atoms for the longest α helical configuration), we used B3LYP/daDZ references and models. Figure 6 contrasts the growth of the dipole with the chain length with the predictions of the scalar, vector, and combined models. In the case of polyglycine, the three models capture at least qualitatively the trend, with the vector model usually under-predicting the slope, and the combined model performing substantially better than either the scalar or the vector model. In the case of the polyenoic amino acids, however, the vector model breaks down completely, predicting a constant dipole as a function of chain length. The scalar model most closely approaches the correct slope, and the combined model shows the correct trend, but with a smaller slope to the pure scalar model. The saturated *n*-amino carboxylic acids showed a completely different trend, with the total dipole leveling off to a constant small value and the model predictions essentially following this trend. This contrast points to the conjugated, non-local nature of the polyenoic acids as an essential ingredient to their large dipole moments: indeed, their saturated counterparts have stronger charge locality and cannot transfer/delocalize charge across the whole molecule, like the unsaturated chains can.

To gain deeper insight into the performance of the different MuML models as well as the physical effects that determine the

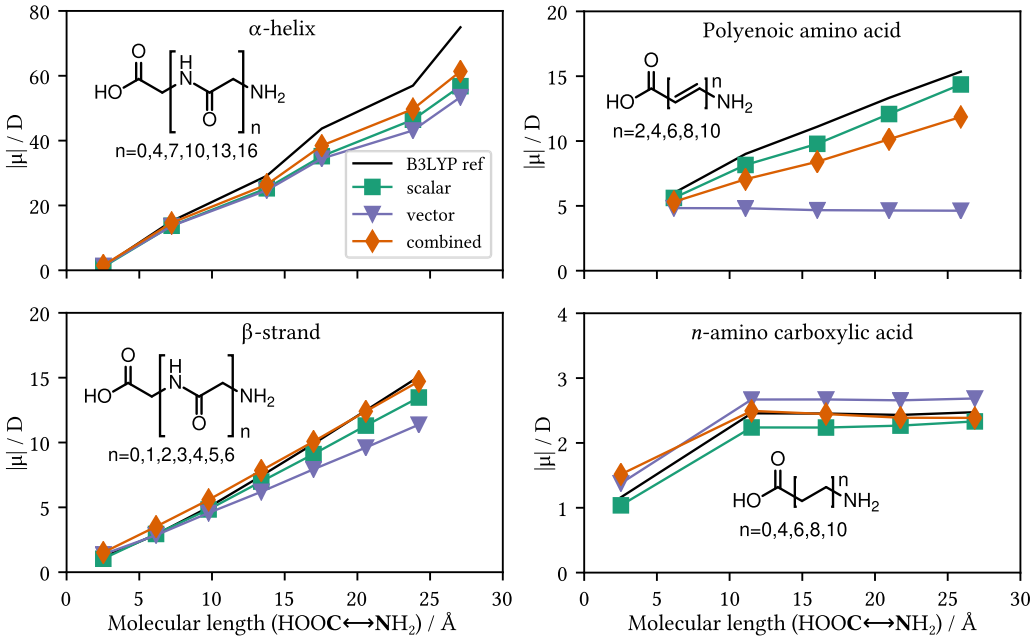

FIG. 6. Dipole moment predictions for the four "challenge" cases: polyglycine in the α-helix conformation, polyglycine in the β-strand conformation, polyenoic amino acids (*trans*-polyacetylene-bridged glycine), and *n*-amino carboxylic acids (polyethylene-bridged glycine). All three models perform fairly well on the polypeptides, where the charge polarization is a mostly local phenomenon. On the unsaturated bridged glycine, however, the vector model completely fails, with only the scalar model maintaining accuracy (and the combined model suffering from the inclusion of the unphysical asymptotic behavior of the vector model). The saturated bridged glycine has completely different behavior, with the dipole saturating to a small constant value; all three models predict this trend accurately. All predictions are from the MuML models trained on 5400 QM7b molecules (not the ensemble models).

breakdown of the vector model, we computed atomic contributions to the dipole moment—both the vector predictions and the partial charges (for the models that use them) for each atom—and represented them together with the molecular structure in Fig. 7. Here, we discuss only the β-strands and the polyenoic amino acids, as the observations for the α-helical structures are very similar to those for the β-strands. The per-atom breakdown for the α-helices and the *n*-amino carboxylic acids are found in the supplementary material. In the case of the polyglycine β-strand, each monomeric unit is polar. Since the total dipole is almost entirely made up of these local monomeric contributions, the vector model based on local atomic dipoles captures the correct scaling behavior with system size. The scalar model also captures the correct behavior, as each molecular unit is (approximately) neutral and contributes a roughly constant term (even though individual atomic dipoles grow larger for atoms that are farther away from the molecular center). The *n*-amino carboxylic acid also exhibits strongly localized physics, with the molecular dipole moment being mostly generated by local polarization of the end groups, and all three MuML models are able to give accurate predictions. Clearer differences between the MuML models arise in the case of the polyenoic amino acids. The non-polar spacer is (correctly) predicted to contribute very little to the total dipole, while the amine and carboxylic acid functional groups each bear a (roughly) constant dipole, which results in a prediction that is independent of the length of the spacer. The scalar model, on the other hand, predicts net positive and negative scalar charges on the amine and carboxylic acid groups, and as a consequence predicts a total dipole that scales linearly with the length of the spacer. Even if it underpredicts the total dipole, the combined model most closely reflects conventional chemical wisdom: it predicts negligible charges along the polyacetylene spacer with only the polar end groups contributing to the total dipole. Since the end groups carry a net positive and negative charge, the total dipole increases with their separation.

These observations reflect the shortcomings of a local ML model, similar to what was observed in Ref. 37 for the molecular polarizability of conjugated hydrocarbons (e.g., alkenes and acenes). SOAP features are computed with a cutoff of 5 Å and cannot therefore describe structure–property correlations beyond this limit. The scalar model circumvents this limitation by assuming that atomic charges are local and that the non-locality of the dipole moment is entirely captured by the spatial separation of the atomic charges. As shown in the supplementary material, the radial scaling functions of the two models, which reflect how quickly the influence of far-away atoms decays, are consistent with the greater non-locality of the vector model. The radial scaling of the scalar models decays rapidly, well before the neighbor list cutoff, while that of the vector model indicates that correlations beyond 5 Å would be needed to describe molecular dipoles as a sum of local contributions.

Finally, while the local partial charges and dipoles provided by this analysis bear some similarities to the electron density decomposition schemes discussed in Sec. II, they should not be confused. The partitioning scheme shown here does not use

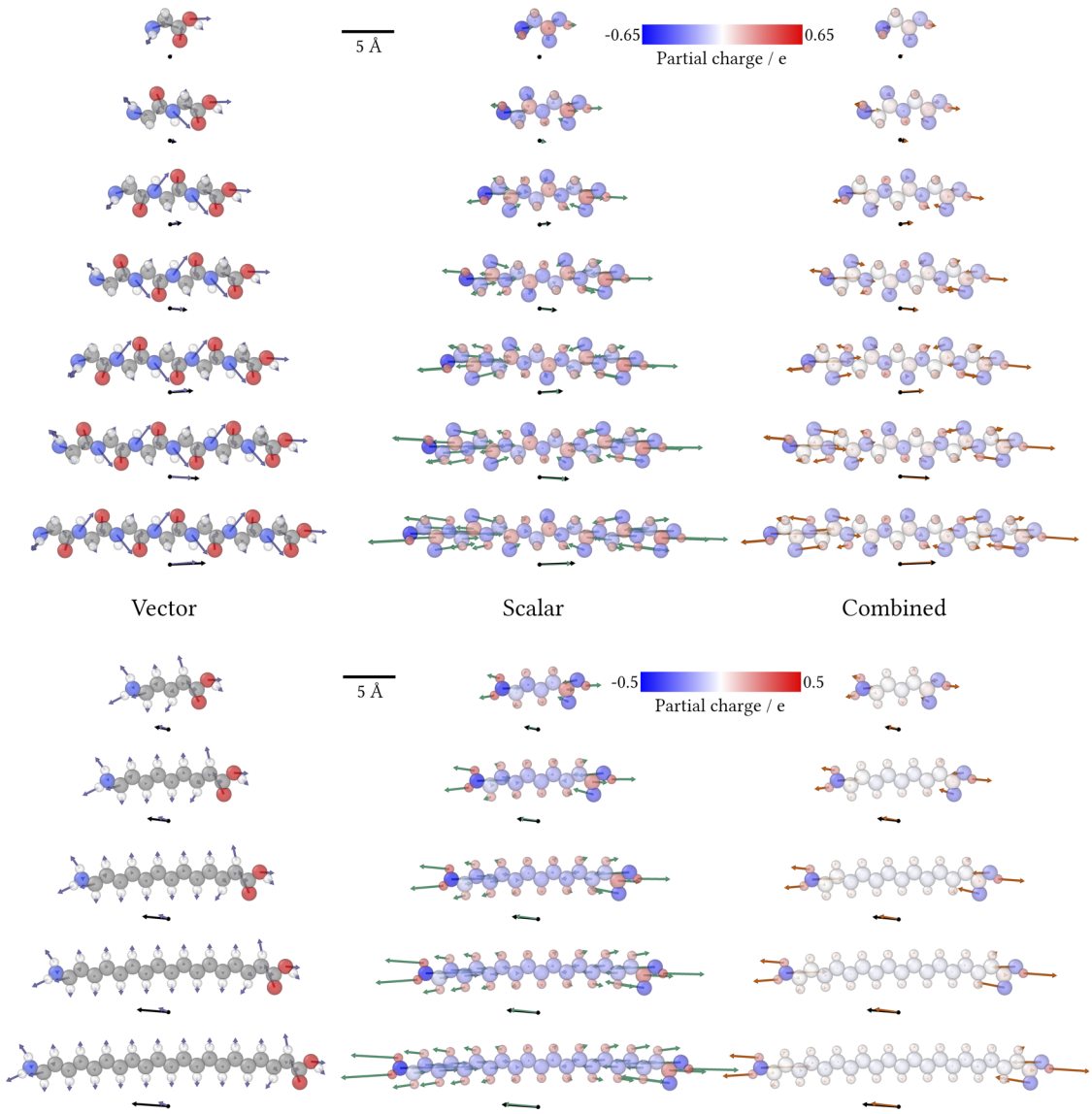

FIG. 7. A representation of the per-atom contributions to the total dipole for two of the challenge systems: polyglycine in the β-strand conformation (top) and polyenoic amino acids in the all-trans conformation (bottom). Vector per-atom dipoles are defined in Eq. (12) and plotted exaggerated by a factor of 5 for visibility; atoms are colored according to the atom type. Scalar per-atom dipoles are defined as the partial charges multiplied with the displacement vectors (referenced to the molecule's center of geometry), as in Eq. (7); atomic charges are also represented as atom colors according to the displayed color scale. Per-atom dipoles for the combined model are the (appropriately weighted) sums of the respective scalar and vector per-atom dipole predictions. The total dipoles, in each case equal to the sum of the per-atom predictions, are shown below each molecule along with (in black) the reference dipole moment computed from B3LYP/daDZ. The scale bar shows the maximum range of sensitivity (5 Å) of the partial charges and atomic dipoles to their environments. Visualizations are created with OVITO.[65]

the electron density; rather, it provides an interpretable description of how the ML model arrives at its prediction of the total dipole, allowing us to verify whether or not it includes the appropriate physics.

V. CONCLUSIONS

In this work, we have introduced a set of models for predicting molecular dipole moments that we collectively refer to as "MuML." These models rely on a local, atom-centered description of the molecular structure that fulfills the symmetries of the target property. We compare a vector model that predicts atom-centered dipolar contributions with a scalar model that predicts atomic charges entering into a physics-based expression for the contribution to the total dipole moment. Training on reference CCSD calculations performed on a set of small organic molecules, both models can achieve a similar accuracy of around 0.07 D in MAE, which is comparable

to the accuracy of DFT, with a slight improvement made possible by combining the two models. The differences between the models are more noticeable—up to 40% RMSE—when considering the transferability to larger molecules, namely, the QM9 dataset. Here, the vector model seems to be more robust, while the scalar model appears to overfit, with a model trained on 5000 small molecules giving worse performance than one trained only on 500. Even with these limitations, the vector model outperforms a state-of-the-art model based on the FCHL* framework,[48] even though the FCHL* model is trained on another subset of QM9 molecules, and is therefore operating in the interpolative regime. When we use training structures from QM9, the performance of MuML dramatically improves and we observe a threefold reduction in the error. The accuracy of the combined model can be improved by adjusting the relative weight of the scalar and vector models, with better interpolative performance observed in the limit of large scalar weights and better extrapolative performance when using large vector weights. State-of-the-art performance for MuML is also observed for a showcase dataset of even larger and more complex molecules, where the scalar model shows improved performance relative to the vector model, and the combined model approaches the accuracy of DFT. For these molecules, we also show that a calibrated committee model can accurately estimate the uncertainty in the model predictions, thereby further improving the reliability of predictions in this challenging extrapolative regime.

In this work, we finally pushed MuML to its breaking point by performing predictions on a set of polymers of increasing length, which extend far beyond the cutoff radius of the atom-centered features used to describe the molecules. In this regime, the vector model can predict reasonably well the molecular dipole moment of polyglycine for which each monomeric unit contributes a dipolar term. It fails dramatically, however, for the polyenoic amino acid series, where the increase in the molecular dipole moment arises because of charge separation by fully conjugated (but nonpolar) spacer units. The scalar model, on the other hand, recovers this effect correctly because the geometric separation between atoms is built into the form of the kernel, introducing an element of non-locality.

The combination of these two models makes it possible to improve the performance of MuML, even though the optimal combination of weights depends rather strongly on the nature of the test molecules. This suggests that, even when taken together, local vector and scalar models of the dipole only partially capture the physics of polarization, affecting the overall model's transferability. An explicit treatment of long-range effects using a charge equilibration scheme,[66] or incorporating long-range correlations by long-distance equivariant features,[67] might further improve the accuracy of MuML, which is already competitive with that of hybrid DFT calculations while being dramatically less computationally expensive.

Another direction for further research involves the modeling of condensed phases. The presence of periodic boundary conditions makes the position operator ill-defined. As a consequence, an expression like Eq. (2) cannot be used to define polarization in the condensed phase, which makes the scalar model and, by extension, the combined model inapplicable. One way around this limitation is to instead model the position of Wannier function centers so that each point in the unit cell is an integer multiple of the electron charge, thereby preserving the lattice condition for polarization in a periodic medium (see, e.g., Spaldin[68] or Resta[69]). Current implementations of the idea, however, predict the position of centers attached to an atom[70] so that the framework is effectively equivalent to learning atom-centered dipoles. Indeed, a vector model can be readily applied to bulk systems and has already been used successfully to predict the infrared spectrum of liquid water.[71] It is not obvious, however, that this methodology will work well in systems where there is significant delocalization of charge. Incorporating ideas from the modern theory of polarization,[69] learning the Born effective charge tensors, or taking a more decidedly data-driven approach by using long-range features without explicitly incorporating a physical description of electrostatics all provide possible strategies to apply to condensed phases a model that can capture, such as MuML, the different phenomena that give rise to permanent or transient polarization.

SUPPLEMENTARY MATERIAL

See the supplementary material for further details about the derivation, implementation, and benchmarks of the method, including convergence of the scalar and vector models on QM7b; kernel optimization procedure; radial scaling function for scalar and vector models; uncertainty quantification calibration procedure; QM7b learning curves for B3LYP dipoles; comparison of B3LYP, CCSD, and SCAN0 dipole moment data for the MuML showcase; and per-atom breakdown of the α-helix and *n*-amino carboxylic acid predictions.

ACKNOWLEDGMENTS

D.M.W. was supported by the European Research Council under the European Union's Horizon 2020 Research and Innovation Programme (Grant No. 677013-HBMAP) and benefited from generous allocation of computer time by CSCS under Project No. s843. M.V. and M.C. were supported by the Samsung Advanced Institute of Technology and the NCCR MARVEL of the Swiss National Science Foundation (SNSF). Y.Y. and R.A.D. acknowledge support from Cornell University through start-up funding. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231.

DATA AVAILABILITY

The geometries and data used to fit the models in this paper are publicly available[72] in Materials Cloud under the DOI: https://doi.org/10.24435/materialscloud:2k-3h; the models themselves are available[73] at Zenodo under the DOI: https://doi.org/10.5281/zenodo.3820297. The code used to fit and evaluate the models is publicly available at https://github.com/max-veit/velociraptor.

REFERENCES

[1] D. A. McQuarrie, *Statistical Mechanics*, 1st ed. (University Science Books, Sausalito, CA, 2000).
[2] A. Morita and J. T. Hynes, Chem. Phys. **258**, 371 (2000).

[3]S. Roke and G. Gonella, Annu. Rev. Phys. Chem. **63**, 353 (2012).
[4]A. J. Stone, *The Theory of Intermolecular Forces*, 2nd ed. (Oxford University Press, Oxford, 2013).
[5]F. De Droft, F. Tielens, and P. Geerlings, J. Mol. Struct.: THEOCHEM **506**, 1 (2000).
[6]A. L. Hickey and C. N. Rowley, J. Phys. Chem. A **118**, 3678 (2014).
[7]D. Hait and M. Head-Gordon, J. Chem. Theory Comput. **14**, 1969 (2018).
[8]A. Grisafi, D. M. Wilkins, G. Csányi, and M. Ceriotti, Phys. Rev. Lett. **120**, 036002 (2018).
[9]M. Gastegger, J. Behler, and P. Marquetand, Chem. Sci. **8**, 6924 (2017).
[10]O. T. Unke and M. Meuwly, J. Chem. Theory Comput. **15**, 3678 (2019).
[11]R. S. Mulliken, J. Chem. Phys. **23**, 1833 (1955).
[12]P. O. Löwdin, J. Chem. Phys. **18**, 365 (1950).
[13]F. L. Hirshfeld, Theor. Chim. Acta **44**, 129 (1977).
[14]P. Bultinck, C. Van Alsenoy, P. W. Ayers, and R. Carbó-Dorca, J. Chem. Phys. **126**, 144111 (2007).
[15]R. F. W. Bader, *Atoms in Molecules: A Quantum Theory* (Clarendon Press, Oxford, 1990).
[16]T. C. Lillestolen and R. J. Wheatley, Chem. Commun. **2008**, 5909.
[17]F. A. Momany, J. Phys. Chem. **82**, 592 (1978).
[18]S. R. Cox and D. E. Williams, J. Comput. Chem. **2**, 304 (1981).
[19]U. C. Singh and P. A. Kollman, J. Comput. Chem. **5**, 129 (1984).
[20]C. M. Breneman and K. B. Wiberg, J. Comput. Chem. **11**, 361 (1990).
[21]T. A. Manz and D. S. Sholl, J. Chem. Theory Comput. **8**, 2844 (2012).
[22]T. Verstraelen, P. W. Ayers, V. Van Speybroeck, and M. Waroquier, J. Chem. Theory Comput. **9**, 2221 (2013).
[23]K. B. Wiberg and P. R. Rablen, J. Comput. Chem. **14**, 1504 (1993).
[24]A. J. Stone, J. Chem. Theory Comput. **1**, 1128 (2005).
[25]E. K. Conway, I. E. Gordon, O. L. Polyansky, and J. Tennyson, J. Chem. Phys. **152**, 024105 (2020).
[26]A. J. Stone, Chem. Phys. Lett. **83**, 233 (1981).
[27]D. Geldof, A. Krishtal, F. Blockhuys, and C. Van Alsenoy, J. Chem. Phys. **140**, 144104 (2014).
[28]A. Mayer, Phys. Rev. B **75**, 045407 (2007).
[29]T. Giovannini, A. Puglisi, M. Ambrosetti, and C. Cappelli, J. Chem. Theory Comput. **15**, 2233 (2019).
[30]M. G. Darley, C. M. Handley, and P. L. A. Popelier, J. Chem. Theory Comput. **4**, 1435 (2008).
[31]C. M. Handley and P. L. A. Popelier, J. Chem. Theory Comput. **5**, 1474 (2009).
[32]G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, New J. Phys. **15**, 095003 (2013).
[33]F. A. Faber, A. S. Christensen, B. Huang, and O. A. von Lilienfeld, J. Chem. Phys. **148**, 241717 (2018).
[34]T. Bereau, R. A. DiStasio, Jr., A. Tkatchenko, and O. A. Von Lilienfeld, J. Chem. Phys. **148**, 241706 (2018).
[35]K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, J. Chem. Phys. **148**, 241722 (2018).
[36]F. M. Paruzzo, A. Hofstetter, F. Musil, S. De, M. Ceriotti, and L. Emsley, Nat. Commun. **9**, 4501 (2018).
[37]D. M. Wilkins, A. Grisafi, Y. Yang, K. U. Lao, R. A. DiStasio, Jr., and M. Ceriotti, Proc. Natl. Acad. Sci. U. S. A. **116**, 3401 (2019).
[38]J. Behler and M. Parrinello, Phys. Rev. Lett. **98**, 146401 (2007).
[39]A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Phys. Rev. Lett. **104**, 136403 (2010).
[40]B. J. Braams and J. M. Bowman, Int. Rev. Phys. Chem. **28**, 577 (2009).
[41]A. P. Bartók, M. J. Gillan, F. R. Manby, and G. Csányi, Phys. Rev. B **88**, 054104 (2013).
[42]A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi, and M. Ceriotti, Sci. Adv. **3**, e1701816 (2017).
[43]T. Morawietz, A. Singraber, C. Dellago, and J. Behler, Proc. Natl. Acad. Sci. U. S. A. **113**, 8368 (2016).
[44]M. Veit, S. K. Jain, S. Bonakala, I. Rudra, D. Hohl, and G. Csányi, J. Chem. Theory Comput. **15**, 2574 (2019).
[45]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).
[46]A. E. Sifain, N. Lubbers, B. T. Nebgen, J. S. Smith, A. Y. Lokhov, O. Isayev, A. E. Roitberg, K. Barros, and S. Tretiak, J. Phys. Chem. Lett. **9**, 4495 (2018).
[47]A. Glielmo, P. Sollich, and A. De Vita, Phys. Rev. B **95**, 214302 (2017).
[48]A. S. Christensen, F. A. Faber, and O. A. von Lilienfeld, J. Chem. Phys. **150**, 064105 (2019).
[49]C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning* (MIT Press, Cambridge, MA, 2006).
[50]M. Ceriotti, M. J. Willatt, and G. Csányi, in *Handbook of Materials Modeling*, edited by W. Andreoni and S. Yip (Springer International Publishing, Cham, 2018), pp. 1–27.
[51]G. Imbalzano, A. Anelli, D. Giofré, S. Klees, J. Behler, and M. Ceriotti, J. Chem. Phys. **148**, 241730 (2018).
[52]N. Raimbault, A. Grisafi, M. Ceriotti, and M. Rossi, New J. Phys. **21**, 105001 (2019).
[53]M. J. Willatt, F. Musil, and M. Ceriotti, Phys. Chem. Chem. Phys. **20**, 29661 (2018).
[54]A. P. Bartók and G. Csányi, Int. J. Quantum Chem. **115**, 1051 (2015).
[55]M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, Phys. Rev. Lett. **108**, 058301 (2012).
[56]Y. Yang, K. U. Lao, D. M. Wilkins, A. Grisafi, M. Ceriotti, and R. A. DiStasio Jr., Sci. Data **6**, 152 (2019).
[57]A. D. Becke, J. Chem. Phys. **98**, 5648 (1993).
[58]K. Kim and K. D. Jordan, J. Phys. Chem. **98**, 10089 (1994).
[59]H. Koch and P. Jørgensen, J. Chem. Phys. **93**, 3333 (1990).
[60]D. E. Woon and T. H. Dunning, J. Chem. Phys. **100**, 2975 (1994).
[61]R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, Sci. Data **1**, 140022 (2014).
[62]A. P. Bartók, J. Kermode, N. Bernstein, and G. Csányi, Phys. Rev. X **8**, 041048 (2018).
[63]F. Musil, M. J. Willatt, M. A. Langovoy, and M. Ceriotti, J. Chem. Theory Comput. **15**, 906 (2019).
[64]Y. Yang, K. U. Lao, D. M. Wilkins, A. Grisafi, M. Ceriotti, and R. A. DiStasio, Jr., "Quantum mechanical static dipole polarizabilities in the QM7b and AlphaML showcase databases" (2019), Materials Cloud, https://doi.org/10.24435/materialscloud:2019.0002/v3.
[65]A. Stukowski, Modell. Simul. Mater. Sci. Eng. **18**, 015012 (2009).
[66]S. A. Ghasemi, A. Hofstetter, S. Saha, and S. Goedecker, Phys. Rev. B **92**, 045131 (2015).
[67]A. Grisafi and M. Ceriotti, J. Chem. Phys. **151**, 204105 (2019).
[68]N. A. Spaldin, J. Solid State Chem. **195**, 2 (2012).
[69]R. Resta, Eur. Phys. J. B **91**, 100 (2018).
[70]L. Zhang, M. Chen, X. Wu, H. Wang, W. E, and R. Car, arXiv:1906.11434 (2019).
[71]V. Kapil, D. M. Wilkins, J. Lan, and M. Ceriotti, J. Chem. Phys. **152**, 124104 (2020).
[72]M. Veit, D. M. Wilkins, Y. Yang, R. A. DiStasio, Jr., and M. Ceriotti, "Models and source data for MuML dipole fitting" (2020), Zenodo, https://doi.org/10.5281/zenodo.3820297.
[73]M. Veit, D. M. Wilkins, Y. Yang, R. A. DiStasio, Jr., and M. Ceriotti, "Quantum mechanical dipole moments in the QM7b, 21k molecules of QM9, and MuML showcase datasets" (2020), Materials Cloud, https://doi.org/10.24435/materialscloud:2k-3h.