Predicting the electronic density response of condensed-phase systems to electric field perturbations

Alan M. Lewis ✉ iD ; Paolo Lazzaroni iD ; Mariana Rossi iD

Check for updates

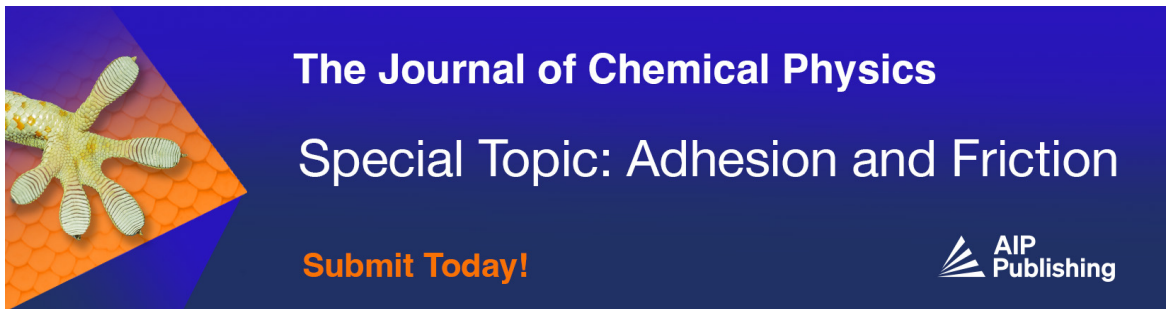View Online

Export Citation

CrossMark

Predicting the electronic density response of condensed-phase systems to electric field perturbations

View Online   Export Citation   CrossMark

Alan M. Lewis,[a] 🄳 Paolo Lazzaroni, 🄳 and Mariana Rossi[b] 🄳

AFFILIATIONS

Max Planck Institute for the Structure and Dynamics of Matter, Luruper Chaussee 149, 22761 Hamburg, Germany

[a]Author to whom correspondence should be addressed: alan.lewis@mpsd.mpg.de
[b]Electronic mail: mariana.rossi@mpsd.mpg.de

ABSTRACT

We present a local and transferable machine-learning approach capable of predicting the real-space density response of both molecules and periodic systems to homogeneous electric fields. The new method, Symmetry-Adapted Learning of Three-dimensional Electron Responses (SALTER), builds on the symmetry-adapted Gaussian process regression symmetry-adapted learning of three-dimensional electron densities framework. SALTER requires only a small, but necessary, modification to the descriptors used to represent the atomic environments. We present the performance of the method on isolated water molecules, bulk water, and a naphthalene crystal. Root mean square errors of the predicted density response lie at or below 10% with barely more than 100 training structures. Derived polarizability tensors and even Raman spectra further derived from these tensors show good agreement with those calculated directly from quantum mechanical methods. Therefore, SALTER shows excellent performance when predicting derived quantities, while retaining all of the information contained in the full electronic response. Thus, this method is capable of predicting vector fields in a chemical context and serves as a landmark for further developments.

Machine-learning (ML) tools are now widely used in computational chemistry to reduce the cost of simulations and aid the interpretation of data.[1–4] ML models are often employed to predict energies and forces[5–10] and drive molecular dynamics simulations.[11–14] This has enabled *ab initio* quality simulations on timescales and system sizes, which are completely inaccessible by conventional *ab initio* molecular dynamics (AIMD) approaches.[14–18] However, this approach comes at the cost of losing access to other properties directly related to the electronic density. These must be obtained by some independent method, which triggered the development of many recent models,[19–23] including a number of models capable of predicting the electronic polarizability.[10,24–26] We have recently developed SALTED (Symmetry Adapted Learning of Three-dimensional Electron Densities), a machine-learning method that predicts the real-space electronic density of molecular and condensed phase systems,[27,28] based on earlier work that predicted the electron densities of molecules.[29,30] SALTED may be applied in

tandem with machine-learned potentials, providing access to every ground-state electronic-structure property that would be available in a traditional AIMD calculation.

In this article, we present SALTER (Symmetry-Adapted Learning of Three-dimensional Electron Responses). SALTER builds on SALTED to predict the static real-space *response* of the electron density of a molecule or material to a homogeneous electric field. This response defines the global dielectric susceptibility of the system, which is needed to simulate many flavors of vibrational Raman and sum-frequency spectra,[31–37] and defines the high-frequency electronic screening that impacts a wide range of static and transport properties of molecules and materials.[38–42] The computational burden of calculating this electronic response is considerable—the most commonly employed method, density-functional perturbation theory (DFPT),[43] is typically 4–10 times more expensive than the underlying density functional theory (DFT) calculation.[42,44] As such, this calculation is a bottleneck when evaluating dielectric

properties along long molecular dynamics trajectories. Machine-learning approaches have recently been used to predict the non-local density-density response function, from which the density response to an applied field can be derived.[45,46] Here, we target the density response itself, developing a machine-learning method that provides more information than direct predictions of the polarizability while being less cumbersome than predictions of the density-density response function.

As in our previous work,[27,28] SALTER is based on an atom-centered expansion of the density response, with the coefficients of the expansion predicted using the symmetry-adapted Gaussian process regression (SA-GPR) approach.[47] This produces a method that is local, transferable, and capable of treating molecular and periodic systems on the same footing.[27] The key modification introduced in this work is the inclusion of a "dummy" atom in the representation of each atomic environment, which encodes information about the direction of the applied perturbation in a simple way while retaining all of the symmetry properties of the representation.

We start by briefly presenting the method. We follow the same general approach outlined in our previous work learning the electron density, ρ,[27,28] and highlight the key differences to the previous method. For a more detailed account, we refer the reader to Refs. 27 and 28. The density response $\rho_\beta^{(1)}$ to a field applied along a Cartesian axis β can be expanded using a set of auxiliary basis functions ϕ, such that

$$\frac{d\rho(\mathbf{r})}{de_\beta} \equiv \rho_\beta^{(1)}(\mathbf{r})$$
$$\approx \tilde{\rho}_\beta^{(1)}(\mathbf{r}) = \sum_{i,\sigma,\mathbf{U}} c_{i\sigma\beta}\phi_{i\sigma}(\mathbf{r} - \mathbf{R}_i + \mathbf{T}(\mathbf{U})), \quad (1)$$

where $\mathbf{R}_i$ is the position of atom i, and the basis function $\phi_{i\sigma}$ is centered on atom i and is the product of a radial part $R_n(r)$ and a real spherical harmonic $Y_{\lambda\mu}(\theta, \phi)$; for brevity, we use a composite index $\sigma \equiv (an\lambda\mu)$, where a labels the atomic species. $\mathbf{T}(\mathbf{U})$ is a translation vector to a point removed from the central reference unit cell by an integer multiple $\mathbf{U} = (U_x, U_y, U_z)$ of the lattice vectors, present only when the system is periodic. The optimal coefficients for this expansion can be found by minimizing the loss function

$$\epsilon(\mathbf{c}_\beta^{RI}) = \int d\mathbf{r} \left| \tilde{\rho}_\beta^{(1)}(\mathbf{r}; \mathbf{c}_\beta^{RI}) - \rho_\beta^{(1)}(\mathbf{r}) \right|^2, \quad (2)$$

which yields

$$\mathbf{c}_\beta^{RI} = \mathbf{S}^{-1}\mathbf{w}_\beta^{(1)}, \quad (3)$$

where $\mathbf{S}$ is the overlap matrix of the basis functions, and $\mathbf{w}_\beta^{(1)}$ is a vector of the projections of the self-consistent density response $\rho_\beta^{(1),QM}(\mathbf{r})$ onto the basis,

$$w_{i\sigma\beta}^{(1)} = \sum_{\mathbf{U}}^{\mathbf{U}_{cut}} \left\langle \phi_{i\sigma}(\mathbf{r} - \mathbf{R}_i + \mathbf{T}(\mathbf{U})) \middle| \rho_\beta^{(1)}(\mathbf{r}) \right\rangle_{u.c.}. \quad (4)$$

These optimal coefficients $\mathbf{c}_\beta^{RI}$ serve as the references against which the accuracy of coefficients predicted by the ML model $\mathbf{c}_\beta^{ML}$ are evaluated.

We approximate $\mathbf{c}_\beta^{ML}$ as a linear combination of the regression weights $b_{j\sigma\beta}$ associated with M reference environments and a

covariant kernel $k_{\sigma\beta}(A_i, M_j)$ that describes the similarity between the reference environments $\{M_j\}$ and the atomic environments that comprise the target structure $\{A_i\}$,

$$c_{i\sigma\beta}^{ML} \approx \sum_j^M b_{j\sigma\beta}k_{\sigma\beta}(A_i, M_j), \quad (5)$$

where $c_{i\sigma\beta}^{ML}$ and $b_{j\sigma\beta}$ are vectors of length $2\lambda + 1$ and $k_{\sigma,\beta}(A_i, M_j)$ is a square matrix of the same dimension. Defining and minimizing a loss function analogous to Eq. (2) allows us to determine the regression weights $\mathbf{b}_\beta$,

$$\epsilon(\mathbf{b}_\beta) = \sum_{A=1}^N \int_{u.c.} d\mathbf{r} \left| \tilde{\rho}_\beta^{(1),ML}(\mathbf{r}; \mathbf{b}_\beta) - \rho_\beta^{(1),RI}(\mathbf{r}) \right|^2 + \eta \mathbf{b}^T \mathbf{K}_{MM}\mathbf{b}. \quad (6)$$

The index A runs over each of the N structures in the training set, the matrix $\mathbf{K}_{MM}$ couples the reference atomic environments to one another, and we introduce the hyperparameter η to control the regularization of the minimization. In all the examples presented, we use a regularization parameter of $\eta = 10^{-8}$, finding only a weak dependence of the observed error on η. Equation (6) is solved iteratively using the conjugate gradient algorithm; for a detailed description of the implementation, see Ref. 28.

Equation (6) contains no explicit constraint to ensure that $\int \tilde{\rho}_\beta^{(1),ML}(\mathbf{r}; \mathbf{b}_\beta) d\mathbf{r} = 0$. Nevertheless, we find that the charge conservation error in the examples presented is negligible: applying an electric field of 0.1 atomic unit (5.14 V/Å) leads to charge conservation errors typically below 10^{-4} per electron. This is consistent with our previous observations using SALTED to predict the electron density,[27] in which case there is also no enforcement of charge conservation.[48]

The key difference between SALTER and SALTED lies in the similarity kernel $k_{\sigma\beta}(A_i, M_j)$. In Refs. 27 and 28, kernels are constructed from the λ-SOAP representation of the atomic environments A_i and M_j.[47] This captures the symmetry covariant transformations of the spherical harmonics under rotation; the coefficients must also possess this symmetry since they expand auxiliary basis functions whose angular parts are spherical harmonics. As such, they are independent of the laboratory frame. However, in the present case, the similarity kernel $k_{\sigma\beta}(A_i, M_j)$ must retain all of the properties exploited in previous work, while also including the laboratory frame direction along which the perturbing field is applied.

Our solution is simple. An additional "dummy" atom of atomic number zero is added to every atomic environment on the β axis, differentiating this axis from the others. The kernels are then calculated using λ-SOAP representations of these modified atomic environments in the normal way. This dummy atom is included in the kernels for all values of λ, ensuring that the kernels retain all of the required symmetry properties, while including (symmetry-adapted) information about the laboratory frame direction of the applied field. This idea is illustrated in Fig. 1 and defined mathematically below.

The unmodified λ-SOAP kernel is defined as[47]

$$k_\sigma(A_i, M_j) = \int d\hat{R} \mathbf{D}^\lambda(\hat{R}) \left| \int n_i(\mathbf{r})n_j(\hat{R}\mathbf{r})d\mathbf{r} \right|^2, \quad (7)$$

where the first integral is performed over all rotations of order λ, and $\mathbf{D}^\lambda(\hat{R})$ is the Wigner matrix corresponding to rotation $\hat{R}$. The
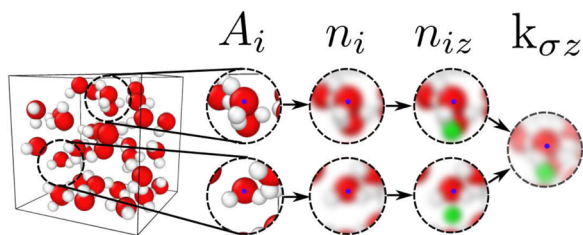
FIG. 1. An illustration of the modified λ-SOAP descriptors and kernels. Environments A_i are defined by a cutoff radius around a central atom i, labeled with a blue dot. The usual SOAP environments n_i [see Eq. (8)] are constructed from a Gaussian function centered on every atom within the cutoff radius. Our modified environments $n_{i\beta}$ [see Eq. (9)] add a Gaussian corresponding to an additional "dummy" atom, shown here in green, which encodes the direction of the applied field in the descriptor. The λ-SOAP kernels are then built from the symmetry-adapted overlap of these modified environments, as in Eq. (7).

atomic environment around atom i is described as a sum of Gaussian functions of width v centered at the position of every atom $\mathbf{R}_j$ that lies within some radius r_c of the central atom,

$$n_i(\mathbf{r}) = \sum_j g_v(\mathbf{r} - \mathbf{R}_j) \Theta(r_c - |\mathbf{R}_j - \mathbf{R}_i|). \quad (8)$$

$\Theta(x)$ is the Heaviside step function. To include information about the direction of an applied field, we include a "dummy" atom in the atomic environment, adding an additional term to Eq. (8),

$$n_{i\beta}(\mathbf{r}) = \sum_j g_v(\mathbf{r} - \mathbf{R}_j) \Theta(r_c - |\mathbf{R}_j - \mathbf{R}_i|) + g_\zeta(\mathbf{r} - x\hat{\mathbf{r}}_\beta r_c). \quad (9)$$

x is a parameter between 0 and 1 specifying the position of the dummy atom and $\hat{\mathbf{r}}_\beta$ is a unit vector in the Cartesian direction β. Note that the width of the Gaussian associated with the dummy atom is not required to be the same as that associated with the physical atoms; in general, $\zeta \neq v$. The λ-SOAP kernel $k_{\sigma\beta}$ that accounts for the direction of the applied field is then obtained by replacing $n_i(\mathbf{r})$ with $n_{i\beta}(\mathbf{r})$ in Eq. (7).

We define the % root mean square error (RMSE) across the datasets as

$$\frac{\% \text{ RMSE}}{100} = \sqrt{\frac{\frac{1}{N_t}\sum_A^{N_t} \int_{u.c.} d\mathbf{r} \left| \tilde{\rho}_{A\beta}^{(1),\text{ML}}(\mathbf{r}) - \tilde{\rho}_{A\beta}^{(1),\text{RI}}(\mathbf{r}) \right|^2}{\frac{1}{N_t-1}\sum_A^{N_t} \Delta\bar{\mathbf{c}}_{A\beta}^T \Delta\bar{\mathbf{w}}_{A\beta}^{(1)}}}. \quad (10)$$

The denominator is the standard deviation of the density response across the N_t structures in the test set, with $\Delta\bar{\mathbf{c}}_A = \mathbf{c}_A - \bar{\mathbf{c}}$ and $\Delta\bar{\mathbf{w}}_A^{(1)} = \mathbf{w}_A^{(1)} - \bar{\mathbf{w}}^{(1)}$; $\bar{\mathbf{c}}$ and $\bar{\mathbf{w}}^{(1)}$ are the mean values for the vectors of coefficients and density projections in the dataset.

We used the FHI-aims software package[49] to obtain the training data $\mathbf{S}$ and $\mathbf{w}_\beta^{(1)}$ and reference density responses $\rho_\beta^{(1),\text{QM}}$ using density functional theory[50,51] employing the local-density approximation (LDA) functional.[50] In all cases, "light" basis sets were used, and the auxiliary basis functions introduced in Eq. (1) are constructed by taking all possible products of these basis functions and then eliminating linear dependencies from the resulting basis set.[52] The molecular dynamics simulations were performed using i-PI.[53] We note that the "light" settings used are not the most accurate for

the prediction of these properties but are sufficient to test the quality of the ML model. Increasing the model accuracy only requires calculating new training data, which we will show below is not very costly.

First, we tackle the simple example of an isolated water molecule. We learn the density response to an electric field for a dataset of 1000 distorted configurations taken from Ref. 47 using both the usual kernels k_σ and the modified kernels $k_{\sigma\beta}$. Each molecule in the dataset is aligned to a reference molecule, as shown in Fig. 2. We use the SOAP parameters $r_c = 4.0$ Å, $v = 0.3$ Å, as is commonly used to describe water,[47] and chose $\zeta = v$ and $x = 0.9$ to define the dummy atom. The choice of these parameters will be discussed in more detail below and makes little difference to the results in this simple example. The dataset of was split into a training set of 250 and a test set of 750 structures; $M = 300$ reference environments were used.

The accuracy of SALTER density responses for this simple dataset are summarized in the table in Fig. 2. The first row shows that while aligning the molecules is sufficient to allow reasonably accurate learning of the density response to a field applied along the y axis, the learning entirely fails for fields applied along the x or z axes. This can be explained by symmetry: since the molecules lie in the xy plane, the SOAP descriptors are invariant to reflections in that plane; there is no differentiation between a field applied in the $+z$ and $-z$ directions, resulting in errors approaching 100%. By contrast, once the dummy atom is introduced, this ambiguity is removed. As a result, using the modified kernels, we see errors of around 1% regardless of the field direction.[54]

This example also displays a feature of SALTER. Since the molecules are aligned, the Cartesian axes are nonequivalent. As a result, a separate machine-learning model must be trained for each non-equivalent direction. This additional cost is offset by the conceptual simplicity of the approach, the computational simplicity of calculating the descriptors, and that this drawback only applies in the specific case of nonequivalent axes. It is not necessary to align

Kernel	x	y	z
k_σ	115.5	7.93	99.9
$k_{\sigma\beta}$	1.04	0.63	0.96
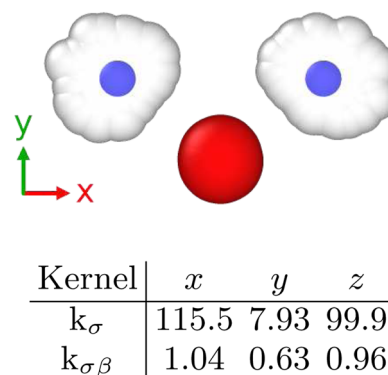
FIG. 2. Top: A visualization of the 1000 water molecule dataset in which all structures are overlayed. Each molecule lies in the xy plane, with the O atom at the origin; the positions of the H atoms are varied around the equilibrium geometry, which is highlighted in blue and whose center of mass lies on the y axis. Bottom: The % RMSE across the test set of 750 structures using λ-SOAP kernels built using atomic environments that exclude (first row) and include (second row) a dummy atom.

systems for this approach to succeed; the alignment here has a purely pedagogical purpose.

Having established that our approach works in principle, we turn to a more realistic example: a cubic water cell of side 9.67 Å containing 32 water molecules. Our dataset consists of 500 configurations, divided into a training set of 400 structures and a test set of 100 structures; this is the same dataset as used in Ref. 28. Because there are now physical atoms distributed throughout each spherical atomic environment, it is not clear *a priori* what the optimal values for x and ζ in Eq. (9) would be. We performed a search of the parameter space using a subset of 50 structures from the full training set, finding the lowest errors when $x = 0.3$ and $\zeta = 1.0$ Å. The results show only a weak dependence on x, since the role of the dummy atom is simply to break the symmetry of the atomic environment—the precise position of the atom in the environment, as long as it lies along the relevant Cartesian direction, is unimportant. Full details of this optimization are provided in the supplementary material.

The learning curves for the density response to a field applied along each Cartesian axis are shown in Fig. 3. Learning improves monotonically with increasing training set size, with the error reaching a plateau of around 12% for $N \geq 100$. We have used $M = 3000$ reference environments for these calculations (see convergence with respect to M in the supplementary material). As the Cartesian directions are all equivalent, the machine-learning model trained on the density response to a field applied along the x axis is also used to predict the response to a field along the y and z axes; the errors observed are very similar regardless of the field direction, as would be expected by symmetry.

The learning curve for this dataset saturates at a relatively large % RMSE: around three times larger than the error in the predicted densities we reported in our previous work.[28] The reason for this could be a deficiency of the descriptor or a lack of quality in the training data. Our descriptors are short ranged: any correlations in the density response, which are longer ranged than the 4 Å cutoff radius of the atomic environment, will be lost, limiting the accuracy of predictions. Alternatively, more accurate predictions may be possible if the training data were obtained using a larger basis set, since this allows a more accurate expansion of the density response and
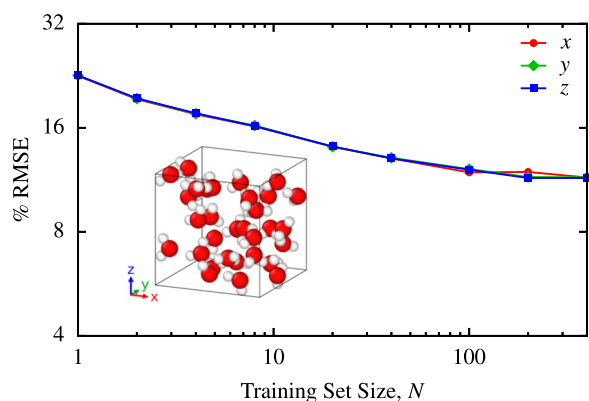
reduces the error in the training data, which in turn reduces the noise in the machine-learning model.

We then tested the performance of SALTER on a derived quantity from the electronic response, the static dielectric susceptibility tensor. This quantity is related to the polarizability of individual molecules and is defined by[42]

$$\alpha_{\gamma\delta} = -\int_{\mathrm{u.c.}} r_\gamma \rho_\delta^{(1)}(\boldsymbol{r})\mathrm{d}\boldsymbol{r}. \tag{11}$$

When considering a non-metallic periodic system, instead of evaluating Eq. (11), we exploit the commutation relation between the unperturbed Hamiltonian and the position operator to avoid an ill-defined result,[42,43] having obtained the unperturbed Kohn–Sham orbitals from a ground-state calculation or an ML prediction.[55] We denote the tensor defined by the density response found by a self-consistent DFPT calculation as $\boldsymbol{\alpha}^{\mathrm{QM}}$, while $\boldsymbol{\alpha}^{\mathrm{ML}}$ is derived from the predicted density response; the error in $\boldsymbol{\alpha}^{\mathrm{ML}}$ relative to $\boldsymbol{\alpha}^{\mathrm{QM}}$ provides a measure of the accuracy of the predicted density responses.

Table I gives the errors in different elements of $\boldsymbol{\alpha}^{\mathrm{ML}}$ derived from the density response predicted by the machine-learning model with $N = 400$, $M = 3000$. In all cases, errors are around 5%; this compares favorably to previous direct machine-learning predictions of $\boldsymbol{\alpha}^{\mathrm{ML}}$, which typically find errors >10% for models trained on a few hundred structures, reducing to 5% only when trained on a few thousand structures.[24–26] Indeed, a direct prediction of the dielectric susceptibilities of this dataset yielded errors between 20% and 30% using the same training set and learning parameters. Therefore, for a given accuracy, our indirect method of predicting $\boldsymbol{\alpha}^{\mathrm{ML}}$ reduces the computational effort required to generate training data by an order of magnitude relative to a direct prediction of $\boldsymbol{\alpha}^{\mathrm{ML}}$.

The smaller errors in $\boldsymbol{\alpha}^{\mathrm{ML}}$ than in $\boldsymbol{\rho}^{(1),\mathrm{ML}}$ can be understood from Eq. (11). If errors in the density response accumulate in regions of space, which do not contribute significantly to the integral, one would expect to see lower errors in $\boldsymbol{\alpha}^{\mathrm{ML}}$ than the density response from which it is derived. We discussed a similar dependence of the error in properties derived from electron density on the spatial distribution of errors in the electron density itself in Ref. 28.

Finally, we use SALTER to predict a quantity that is "twice removed" from the target prediction of the model: the vibrational Raman spectrum of the P2$_1$/a naphthalene crystal at 80 K. We used five 10 ps NVE *ab initio* molecular dynamics trajectories of a $2 \times 2 \times 1$ crystal supercell (eight molecules) using the Perdew–Burke–Ernzerhof (PBE) functional[56] with many-body dispersion corrections.[57,58] To train the machine-learning model, we selected 100 structures from each of these trajectories, dividing these structures into a training set of 400 configurations and a test set of the remaining 100. The learning curves for the density response to a field applied along each Cartesian direction are shown in Fig. 4. These curves are converged with respect to M (see

FIG. 3. Learning curves for the density response to a field along each Cartesian direction for the bulk water dataset, $M = 3000$.

TABLE I. % RMSE in each component of the dielectric susceptibility derived from the density responses predicted by a machine-learning model trained using $N = 400$, $M = 3000$ for the bulk water test set.

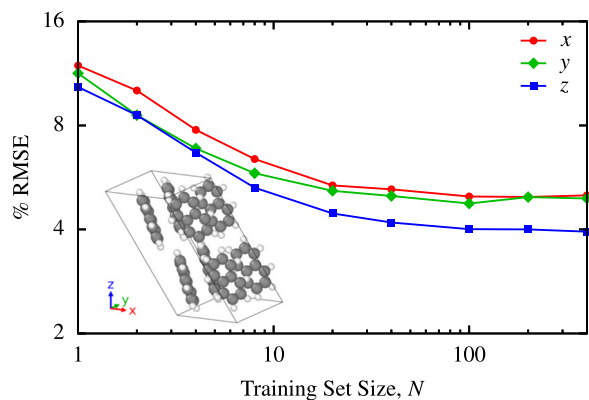| Component	xx	yy	zz	xy	xz	yz
% RMSE	5.03	4.88	4.39	5.11	4.91	4.31

FIG. 4. Learning curves for the density response to a field along each Cartesian direction for the naphthalene test set, $M = 2000$.

the supplementary material) and we see no significant difference in performance when predicting the response to a field applied along each Cartesian axis. The dataset is relatively homogeneous: a single training structure is sufficient to produce an ML model accurate to within 12%, and no improvement is seen when the training set size is increased beyond 100.

We then used the predicted density responses to calculate $\boldsymbol{\alpha}^{\mathrm{ML}}$ for each structure in the test set. Curiously, the errors given in Table II are significantly larger than those found in bulk water (Table I), despite the % RMSE of the density response itself being significantly lower for naphthalene. In contrast to the water dataset, these indirect predictions also underperform direct predictions of the dielectric susceptibility, which show errors of around 7% using the same training data and machine-learning parameters. Furthermore, the errors in the components of $\boldsymbol{\alpha}^{\mathrm{ML}}$ derived from a response to a field applied in the z direction are the largest, despite the errors in this response being the lowest (see crystal cell orientation in Fig. 4). This counter-intuitive result is again due to the fact that the errors in $\boldsymbol{\alpha}^{\mathrm{ML}}$ depend not just on the magnitude of the error but also its spatial distribution. Specifically, for a given magnitude of error, the greater the distance between the regions of positive and negative error, the greater the error in the dielectric susceptibility derived from the density response—as we observe along the z direction of naphthalene. This is illustrated in more detail in the supplementary material. This also suggests a possible solution: adding auxiliary basis functions with larger values of λ will likely improve the description of the electron density response far from the nuclei, resulting in a reduction in the error in the dielectric susceptibility.

We compared the Raman spectrum obtained using $\boldsymbol{\alpha}^{\mathrm{QM}}$ with one obtained indirectly using SALTER predictions of $\rho^{(1),\mathrm{ML}}$ to

obtain $\boldsymbol{\alpha}^{\mathrm{ML}}$. The anharmonic Raman spectrum of a disordered naphthalene sample (the powder spectrum) was calculated using ensemble-averaged time-correlation functions, as in Ref. 32. The spectra are shown in Fig. 5. Good agreement is seen at the polymorph-sensitive low frequency modes, with small discrepancies in the peak intensities appearing at higher frequencies. Using SALTER, we obtain Raman spectra of comparable quality to those obtained using *ab initio* methods and to those obtained from a direct learning of $\boldsymbol{\alpha}^{\mathrm{QM}}$,[24] while dramatically reducing the simulation cost and retaining access to all the information included in the real-space distribution of $\rho^{(1)}(\boldsymbol{r})$.

In conclusion, we presented a simple modification to the widely used λ-SOAP descriptors, which allows us to use machine learning to predict a vector field; specifically, the response of the electron density to a static electric field. By adding a "dummy" atom to each atomic environment that indicates the direction of the applied field, we retain all of the beneficial properties of the λ-SOAP kernels, which allowed us to predict the scalar electron density field in previous work,[27,28] while including (symmetry-adapted) information about the laboratory frame direction of the applied field. Therefore, we expect the method developed here to be straightforward to generalize to the prediction of other vector fields. To the best of our knowledge, this is the first reported machine-learning model capable of predicting vector fields in a chemical context. We note that, after the initial submission of this paper, a very similar model was developed independently and used to predict the electron density of systems under a finite field.[59]

We applied SALTER to two condensed phase systems: liquid water and the $P2_1/a$ naphthalene crystal. A single machine-learning model successfully predicted the density response of water to a field applied along each of the Cartesian axes; we derived the dielectric susceptibility from these predictions, finding values within 5% of the reference DFPT calculations. For naphthalene, we constructed machine-learning models of similar accuracy regardless of the direction of the applied field. We found a counter-intuitive relationship between the errors in the density response, which were smallest when the field was applied along the z axis, and the errors in the corresponding dielectric susceptibilities, which were the largest for
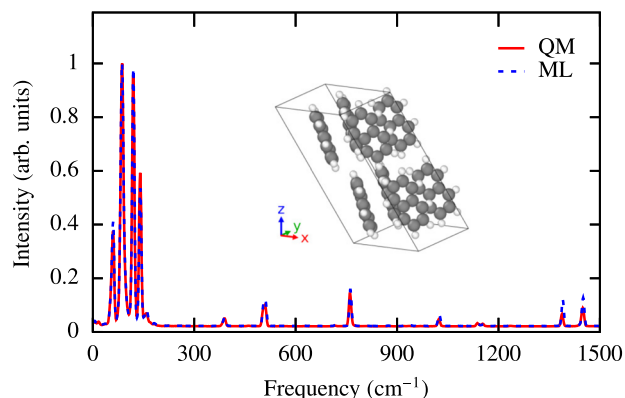

FIG. 5. The powder Raman spectrum of the $P2_1/a$ naphthalene crystal at 80 K, calculated from time-correlation functions of dielectric susceptibilities *ab initio* (QM) and via an indirect SALTER prediction.

TABLE II. % RMSE in each component of the dielectric susceptibility derived from the density responses predicted by machine-learning models trained using $N = 400$, $M = 3000$ for the naphthalene test set.

Component	xx	yy	zz	xy	xz	yz
% RMSE	9.13	14.40	19.74	9.38	12.81	20.11

the yz and zz components of the tensor. This apparent contradiction can be explained by analyzing the spatial distribution of the error in the predicted density response. Nevertheless, the Raman powder spectrum obtained from these indirectly predicted dielectric susceptibilities was very similar to that obtained using DFPT, and at a fraction of the computational cost. In general, this method produces accurate derived quantities while retaining the full information contained in the electron density response to an applied field.

The supplementary material contains details of the optimization of the dummy atom parameters, the convergence of the machine-learning models with respect to the number of reference environments, a description of additional machine-learning parameters, details of the auxiliary basis employed, and an analysis of the origin of the error in the polarizability or dielectric susceptibility derived from a predicted density response.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Alan M. Lewis: Investigation (lead); Methodology (lead); Software (lead); Writing – original draft (lead); Writing – review & editing (equal). **Paolo Lazzaroni**: Data curation (lead); Investigation (equal); Software (equal); Writing – review & editing (equal). **Mariana Rossi**: Conceptualization (lead); Methodology (equal); Supervision (lead); Writing – original draft (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The machine-learning datasets and MD trajectories can be found at dx.doi.org/10.5281/zenodo.7940570. The naphthalene training data can also be found in the NOMAD repository dx.doi.org/10.17172/NOMAD/2023.04.13-1. The code to build similarity kernels using the modified λ-SOAP descriptors can be found at github.com/alanmlewis/TENSOAP, with the release for this paper at dx.doi.org/10.5281/zenodo.7941362.

REFERENCES

[1] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, "Machine learning for molecular and materials science," Nature **559**, 547–555 (2018).

[2] M. Ceriotti, "Unsupervised machine learning in atomistic simulations, between predictions and understanding," J. Chem. Phys. **150**, 150901 (2019); arXiv:1902.05158.

[3] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, "Machine learning and the physical sciences," Rev. Mod. Phys. **91**, 045002 (2019); arXiv:1903.10563.

[4] V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti, and G. Csányi, "Gaussian process regression for materials and molecules," Chem. Rev. **121**, 10073–10141 (2021).

[5] A. P. Thompson, L. P. Swiler, C. R. Trott, S. M. Foiles, and G. J. Tucker, "Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials," J. Comput. Phys. **285**, 316–330 (2015).

[6] A. Kamath, R. A. Vargas-Hernández, R. V. Krems, T. Carrington, and S. Manzhos, "Neural networks vs Gaussian process regression for representing potential energy surfaces: A comparative study of fit quality and vibrational spectrum accuracy," J. Chem. Phys. **148**, 241702 (2018).

[7] R. Drautz, "Atomic cluster expansion for accurate and transferable interatomic potentials," Phys. Rev. B **99**, 014104 (2019).

[8] V. L. Deringer, M. A. Caro, and G. Csányi, "Machine learning interatomic potentials as emerging tools for materials science," Adv. Mater. **31**, 1902765 (2019).

[9] J. Behler, "Four generations of high-dimensional neural network potentials," Chem. Rev. **121**, 10037–10072 (2021).

[10] M. Cools-Ceuppens, J. Dambre, and T. Verstraelen, "Modeling electronic response properties with an explicit-electron machine learning potential," J. Chem. Theory Comput. **18**, 1672–1691 (2022).

[11] F. Noé, A. Tkatchenko, K.-R. Müller, and C. Clementi, "Machine learning for molecular simulation," Annu. Rev. Phys. Chem. **71**, 361–390 (2020).

[12] P. Friederich, F. Häse, J. Proppe, and A. Aspuru-Guzik, "Machine-learned potentials for next-generation matter simulations," Nat. Mater. **20**, 750–761 (2021).

[13] O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko, and K.-R. Müller, "Machine learning force fields," Chem. Rev. **121**, 10142–10186 (2021); arXiv:2010.07067.

[14] A. Musaelian, S. Batzner, A. Johansson, L. Sun, C. J. Owen, M. Kornbluth, and B. Kozinsky, "Learning local equivariant representations for large-scale atomistic dynamics," Nat. Commun. **14**, 579 (2023).

[15] G. C. Sosso, D. Donadio, S. Caravati, J. Behler, and M. Bernasconi, "Thermal transport in phase-change materials from atomistic simulations," Phys. Rev. B **86**, 104301 (2012).

[16] V. Botu, R. Batra, J. Chapman, and R. Ramprasad, "Machine learning force fields: Construction, validation, and outlook," J. Phys. Chem. C **121**, 511–522 (2017); arXiv:1610.02098.

[17] F. Maresca, D. Dragoni, G. Csányi, N. Marzari, and W. A. Curtin, "Screw dislocation structure and mobility in body centered cubic Fe predicted by a Gaussian approximation potential," npj Comput. Mater. **4**, 69 (2018).

[18] V. L. Deringer, N. Bernstein, A. P. Bartók, M. J. Cliffe, R. N. Kerber, L. E. Marbella, C. P. Grey, S. R. Elliott, and G. Csányi, "Realistic atomistic structure of amorphous silicon from machine-learning-driven molecular dynamics," J. Phys. Chem. Lett. **9**, 2879–2885 (2018).

[19] M. G. Darley, C. M. Handley, and P. L. A. Popelier, "Beyond point charges: Dynamic polarization from neural net predicted multipole moments," J. Chem. Theory Comput. **4**, 1435–1448 (2008).

[20] O. T. Unke and M. Meuwly, "PhysNet: A neural network for predicting energies, forces, dipole moments, and partial charges," J. Chem. Theory Comput. **15**, 3678–3693 (2019).

[21] M. Veit, D. M. Wilkins, Y. Yang, R. A. DiStasio, and M. Ceriotti, "Predicting molecular dipole moments by combining atomic partial charges and atomic dipoles," J. Chem. Phys. **153**, 024113 (2020).

[22] G. A. Pinheiro, J. Mucelini, M. D. Soares, R. C. Prati, J. L. F. Da Silva, and M. G. Quiles, "Machine learning prediction of nine molecular properties based on the SMILES representation of the QM9 quantum-chemistry dataset," J. Phys. Chem. A **124**, 9854–9866 (2020).

[23] J. Sun, L. Cheng, and T. F. Miller, "Molecular dipole moment learning via rotationally equivariant derivative kernels in molecular-orbital-based machine learning," J. Chem. Phys. **157**, 104109 (2022).

[24] N. Raimbault, A. Grisafi, M. Ceriotti, and M. Rossi, "Using Gaussian process regression to simulate the vibrational Raman spectra of molecular crystals," New J. Phys. **21**, 105001 (2019); arXiv:1906.07485.

[25] D. M. Wilkins, A. Grisafi, Y. Yang, K. U. Lao, R. A. DiStasio, and M. Ceriotti, "Accurate molecular polarizabilities with coupled cluster theory and machine learning," Proc. Natl. Acad. Sci. U. S. A. **116**, 3401–3406 (2019); arXiv:1809.05337.

[26]H. Shang and H. Wang, "Anharmonic Raman spectra simulation of crystals from deep neural networks," AIP Adv. **11**, 035105 (2021).

[27]A. M. Lewis, A. Grisafi, M. Ceriotti, and M. Rossi, "Learning electron densities in the condensed phase," J. Chem. Theory Comput. **17**, 7203–7214 (2021); arXiv:2106.05364.

[28]A. Grisafi, A. M. Lewis, M. Rossi, and M. Ceriotti, "Electronic-structure properties from atom-centered predictions of the electron density," J. Chem. Theory Comput. (2022); arXiv:2206.14087.

[29]A. Grisafi, A. Fabrizio, B. Meyer, D. M. Wilkins, C. Corminboeuf, and M. Ceriotti, "Transferable machine-learning model of the electron density," ACS Cent. Sci. **5**, 57–64 (2019); arXiv:1809.05349.

[30]A. Fabrizio, A. Grisafi, B. Meyer, M. Ceriotti, and C. Corminboeuf, "Electron density learning of non-covalent systems," Chem. Sci. **10**, 9424–9432 (2019).

[31]P. J. Hendra and P. M. Stratton, "Laser-Raman spectroscopy," Chem. Rev. **69**, 325–344 (1969).

[32]D. A. McQuarrie, *Statistical Mechanics* (Harper Collins, New York, 1976).

[33]A. Morita and J. T. Hynes, "A theoretical analysis of the sum frequency generation spectrum of the water surface," Chem. Phys. **258**, 371–390 (2000).

[34]A. Morita and J. T. Hynes, "A theoretical analysis of the sum frequency generation spectrum of the water surface. II. Time-dependent approach," J. Phys. Chem. B **106**, 673–685 (2002).

[35]B. M. Auer and J. L. Skinner, "IR and Raman spectra of liquid water: Theory and interpretation," J. Chem. Phys. **128**, 224511 (2008).

[36]J. Lee, K. T. Crampton, N. Tallarida, and V. A. Apkarian, "Visualizing vibrational normal modes of a single molecule with atomically confined light," Nature **568**, 78–82 (2019).

[37]N. Raimbault, V. Athavale, and M. Rossi, "Anharmonic effects in the low-frequency vibrational modes of aspirin and paracetamol crystals," Phys. Rev. Mater. **3**, 053605 (2019).

[38]T. Sohier, M. Calandra, and F. Mauri, "Density-functional calculation of static screening in two-dimensional materials: The long-wavelength dielectric function of graphene," Phys. Rev. B **91**, 165428 (2015).

[39]P. Cudazzo, I. V. Tokatly, and A. Rubio, "Dielectric screening in two-dimensional insulators: Implications for excitonic and impurity states in graphane," Phys. Rev. B **84**, 085406 (2011).

[40]J. H. Skone, M. Govoni, and G. Galli, "Self-consistent hybrid functional for condensed systems," Phys. Rev. B **89**, 195112 (2014).

[41]N. P. Brawand, M. Vörös, M. Govoni, and G. Galli, "Generalization of dielectric-dependent hybrid functionals to finite systems," Phys. Rev. X **6**, 041002 (2016).

[42]H. Shang, N. Raimbault, P. Rinke, M. Scheffler, M. Rossi, and C. Carbogno, "All-electron, real-space perturbation theory for homogeneous electric fields: Theory, implementation, and application within DFT," New J. Phys. **20**, 073040 (2018); arXiv:1803.00924.

[43]S. Baroni, S. de Gironcoli, A. Dal Corso, and P. Giannozzi, "Phonons and related crystal properties from density-functional perturbation theory," Rev. Mod. Phys. **73**, 515 (2001); arXiv:hep-ph/9702376.

[44]X. Andrade, S. Botti, M. A. L. Marques, and A. Rubio, "Time-dependent density functional theory scheme for efficient calculations of dynamic (hyper)polarizabilities," J. Chem. Phys. **126**, 184106 (2007); arXiv:cond-mat/0701632.

[45]Y. Shao, L. Andersson, L. Knijff, and C. Zhang, "Finite-field coupling via learning the charge response kernel," Electron. Struct. **4**, 014012 (2022).

[46]M. G. Zauchner, A. Horsfield, and J. Lischner, "Accelerating GW calculations through machine learned dielectric matrices," arXiv:2305.02990 [cond-mat] (2023).

[47]A. Grisafi, D. M. Wilkins, G. Csányi, and M. Ceriotti, "Symmetry-adapted machine learning for tensorial properties of atomistic systems," Phys. Rev. Lett. **120**, 36002 (2018); arXiv:1709.06757.

[48]Should strict charge conservation be required, this could be achieved by either applying a uniform *post-hoc* correction to the predicted density response, or by excluding $\lambda = 0$ auxiliary basis functions from the expansion in Eq. (1).

[49]V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter, and M. Scheffler, "*Ab initio* molecular simulations with numeric atom-centered orbitals," Comput. Phys. Commun. **180**, 2175 (2009).

[50]P. Hohenberg and W. Kohn, "Inhomogeneous electron gas," Phys. Rev. **136**, B864 (1964).

[51]W. Kohn and L. J. Sham, "Self-consistent equations including exchange and correlation effects," Phys. Rev. **140**, A1133 (1965).

[52]X. Ren, P. Rinke, V. Blum, J. Wieferink, A. Tkatchenko, A. Sanfilippo, K. Reuter, and M. Scheffler, "Resolution-of-identity approach to Hartree-Fock, hybrid density functionals, RPA, MP2 and GW with numeric atom-centered orbital basis functions," New J. Phys. **14**, 053020 (2012); arXiv:1201.0655.

[53]V. Kapil, M. Rossi, O. Marsalek, R. Petraglia, Y. Litman, T. Spura, B. Cheng, A. Cuzzocrea, R. H. Meißner, D. M. Wilkins, B. A. Helfrecht, P. Juda, S. P. Bienvenue, W. Fang, J. Kessler, I. Poltavsky, S. Vandenbrande, J. Wieme, C. Corminboeuf, T. D. Kühne, D. E. Manolopoulos, T. E. Markland, J. O. Richardson, A. Tkatchenko, G. A. Tribello, V. Van Speybroeck, and M. Ceriotti, "I-PI 2.0: A universal force engine for advanced molecular simulations," Comput. Phys. Commun. **236**, 214–223 (2019).

[54]The learning of the response to a field in the *y* direction also improves when using the modified kernels, since the applied field direction is no longer simply being inferred from the aligned geometries.

[55]The KS orbitals can be derived from a machine-learning model such as SALTED – see Ref. 28.

[56]J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple," Phys. Rev. Lett. **77**, 3865–3868 (1996).

[57]A. Tkatchenko, R. A. DiStasio, R. Car, and M. Scheffler, "Accurate and efficient method for many-body van der Waals interactions," Phys. Rev. Lett. **108**, 236402 (2012).

[58]A. Ambrosetti, A. M. Reilly, R. A. DiStasio, and A. Tkatchenko, "Long-range correlation energy calculated from coupled atomic response functions," J. Chem. Phys. **140**, 18A508 (2014).

[59]A. Grisafi, A. Bussy, and R. Vuilleumier, "Predicting the charge density response in metal electrodes," arXiv:2304.08966 [cond-mat] (2023).

19 July 2023 07:10:48