EPFL

Thèse n° 8247

Physics-enhanced machine learning with symmetry-adapted and long-range representations

Présentée le 2 juillet 2021

Faculté des sciences et techniques de l'ingénieur
Laboratoire de science computationnelle et modélisation
Programme doctoral en science et génie des matériaux

pour l'obtention du grade de Docteur ès Sciences

par

Andrea GRISAFI

Acceptée sur proposition du jury

Prof. C. Hébert, présidente du jury
Prof. M. Ceriotti, directeur de thèse
Prof. B. Mennucci, rapporteuse
Prof. J. Hutter, rapporteur
Prof. N. Marzari, rapporteur

École polytechnique fédérale de Lausanne

2021

Tyger! Tyger! Burning bright
In the forests of the night:
What immortal hand or eye
Could frame thy fearful symmetry?
— William Blake

To my family, friends and colleagues.

Acknowledgements

I am truly grateful to my supervisor, Prof. Michele Ceriotti, for guiding me through a vibrant and exciting field of research, for his broad and constant support all throughout my PhD and for giving me the opportunity to take part of a stimulating academic environment, both in Switzerland and worldwide.
A special thank goes to Dr. David Wilkins, who gave me unbelievable support since the very beginning of my PhD and with whom I had the pleasure to work on several projects and to share countless funny moments. I also would like to thank all the previous and current members of COSMO, in particular, Andrea, Giulio, Edoardo, Michael, Natasha, Venkat, Kevin and Jigyasa for the great time spent together both in the office and outside. Among my collaborators at EPFL, a special thank goes to Alberto, with whom I greatly enjoyed doing science, as well as sharing joys and sorrows in our volleyball team. Not less importantly, I want to thank my italian crew in Lausanne, Federico, Matteo, Enrico and Francesco for keeping my mood up during the many pizzas and beers spent together.
A huge thank goes to my family in Lucca, my mother Antonella, my father Silvano and my brother Alberto, who put myself under the best conditions to cultivate a profound passion for science and always motivated me to undertake this experience. Last but not least, I am immensely grateful to Francesca, whose bright smiles, sincere love and inexorable support have enlightened my entire journey.

Lausanne, 21.01.2021 A. G.

Abstract

Theoretical and computational approaches to the study of materials and molecules have, over the last few decades, progressed at an exponential rate. Yet, the possibility of producing numerical predictions that are on par with experimental measurements is to date still hindered by a major computational barrier. In this context, machine-learning methods have emerged as an effective strategy to overcome this barrier by means of statistical approximations that rely only on the knowledge of the atomic coordinates of the system. The quality of these approximations strongly depends on the adoption of mathematical representations of the atomic structure that mirror the physical behaviour of the learning target. In this thesis, we make use of this general principle to tackle some particularly tricky aspects in the data-driven prediction of materials properties.

The first part addresses the problem of interpolating physical tensors, such as any quantity that follows a set of prescribed transformation rules under a three-dimensional rotation of the system. We derive mathematical representations of the atomic structures that satisfy the symmetry of spherical harmonics. This family of atomistic features can be used to efficiently regress the irreducible spherical decomposition of any Cartesian tensor. We benchmark the method on the optical series of water oligomers, the dielectric response of liquid water, as well as high-end polarizabilities of heterogeneous molecular datasets. Taking the crystal polymorphs of paracetamol as an example, we finally discuss the possibility of computing the Raman spectrum on top of predicted values of polarizabilities.

The second part of the thesis makes use of the symmetry-adapted representations previously introduced to address the challenging problem of learning and predicting scalar fields, such as the electronic charge density of a system. The main difficulty is associated with the decomposition of the field on a multi-centered non-orthogonal basis, which comes along with the derivation of a specifically designed regression algorithm. Making the electron density decomposition compatible with auxiliary basis sets commonly used in quantum-chemistry codes, we show the capability of the

method to perform highly transferable predictions for arbitrarily complex molecules, that scale linearly with the system size.
The last part of the thesis addresses the problem of incorporating a long-range description within state-of-the-art local machine-learning schemes. This is done by deriving a family of representations where a smooth Coulomb-like potential associated with the distribution of atoms is evaluated at the local scale. In particular, a suitable combination of long-range and local features makes it possible to design a learning framework that shows an asymptotic behaviour that allows us to capture repulsion, electrostatic, polarization and dispersion phenomena, on an equal footing. The method performance is tested on the binding energy of organic dimers, the mutual polarization between a water molecule and a metallic surface of lithium, and the dielectric response of peptidic chains.
By and large, this research study shows how a wise interplay between a totally agnostic learning method and a physically grounded approximation allows us to predict arbitrarily complex atomistic properties, paving the way to the accurate simulation of materials over time and length scales that are not accessible by first-principles methods.

Keywords: machine learning, atomic-scale representations, tensorial properties, electron densities, long-range interactions.

Riassunto

Durante gli ultimi decenni, gli approcci computazionali dedicati allo studio di materiali si sono sviluppati ad una velocità esponenziale. Tuttavia, la possibilità di ottenere predizioni numeriche che siano comparabili con le misurazioni sperimentali è ad oggi ancora ostacolata da una considerevole barriera computazionale. In questo contesto, i metodi di apprendimento automatico si sono rivelati una strategia efficace per superare questa barriera attraverso approssimazioni statistiche che si affidano all'esclusiva conoscenza delle coordinate atomiche del sistema. La qualità di queste approssimazioni dipende fortemente dall'adozione di rappresentazioni matematiche della struttura atomica che riflettono il comportamento fisico della proprietà in oggetto. In questo lavoro di tesi, questo principio generale viene impiegato per affrontare alcuni aspetti particolarmente complessi nell'apprendimento automatico delle proprietà dei materiali.

La prima parte della tesi tratta il problema dell'interpolazione di tensori fisici, ossia qualsiasi quantità che segue precise regole di trasformazione in seguito a una rotazione tridimensionale del sistema. Abbiamo derivato rappresentazioni matematiche della struttura atomica che soddisfano la simmetria delle armoniche sferiche. Questa famiglia di caratteristiche atomiche possono essere utilizzate per eseguire la regressione della decomposizione irriducibile di qualsiasi tensore Cartesiano. Il metodo viene messo alla prova sulla serie ottica di oligomeri d'acqua, sulla risposta dielettrica dell'acqua liquida e su polarizzabilità di alto livello di un insieme eterogeneo di molecole. Abbiamo infine discusso la possibilità di calcolare lo spettro Raman sulla base dei valori di polarizzabilità predetti.

La seconda parte della tesi ricorre alle rappresentazioni adattate alla simmetria precedentemente introdotte per affrontare il problema relativo alla predizione della densità elettronica di un sistema. La difficoltà principale è associata alla decomposizione del campo scalare su una base a molti centri, la quale porta con sé la derivazione di un algoritmo di regressione specifico. Rendendo la decomposizione della densità compatibile con le basi ausiliare comunemente usate nei codici di chimica quantistica, abbiamo dimostrato la capacità del metodo di realizzare predizioni altamente

trasferibili per molecole arbitrariamente complesse.
La terza e ultima parte della tesi tratta il problema di incorporare una descizione a lungo raggio all'interno di schemi locali di apprendimento. Abbiamo derivato una famiglia di rappresentazioni atomiche in cui un potenziale Coulombiano viene calcolato sulla scala locale. Una combinazione di caratteristiche locali e non-locali, in particolare, rende possibile uno schema di apprendimento che possiede un comportamento asintotico capace di predire effetti elettrostatici, di polarizzazione e di dispersione. Abbiamo dimostrato l'efficacia del metodo per le energie di legame di dimeri organici, la mutua polarizzazione tra una molecola d'acqua e una superficie metallica di litio, e la risposta dielettrica di catene proteiche.
Nel complesso, questo studio di ricerca mostra come una saggia combinazione di metodi agnostici di apprendimento e approssimazioni fisicamente fondate permetta di predire qualsiasi proprietà atomistica, aprendo la strada a simulazioni accurate dei materiali che hanno luogo su scale di tempo e di lunghezza non accessibili attraverso metodi ai primi principi.

Parole chiave: apprendimento automatico, rappresentazioni su scala atomica, proprietà tensoriali, densità elettroniche, interazioni a lungo raggio.

Contents

1 Introduction

Computational materials science is a constantly evolving field of research that targets the simulation of molecular and materials properties by borrowing theories and numerical methods from chemistry, physics and applied mathematics. Although constantly progressing, the theory underlying the prediction of physical quantities in solids, liquids and molecular systems has reached an advanced state. Yet, performing reliable and accurate simulations that can be compared with experimental measurements still presents a major computational barrier. This chapter introduces an overview of the machine-learning methods that can be used for overcoming this barrier by means of data-driven predictions. By focusing the attention on the level of physical insight that can be incorporated within the construction of mathematical representations of the atomic structure, it provides the general context and motivation for the research presented in this thesis.

1.1 The electronic structure problem

The major hurdle when addressing the atomic scale simulations of a material consists in dealing with its quantum nature. For any given instantaneous configuration of the atomic nuclei, the quantum-mechanical problem consists in solving the time-independent Schrödinger equation for a system of N electrons, i.e.,

$$\hat{H}_e \Psi_k = E_k \Psi_k \,, \tag{1.1}$$

with $\hat{H}_e$ the electronic Hamiltonian operator, while Ψ_k and E_k are the spectrum of electronic wave-functions and energies that satisfy the eigenvalue problem. Given its many-body character, solving this equation for an arbitrary number of electrons represents a particularly difficult task, and exact solutions can only be found for a handful of simple cases. For this reason, it comes as no surprise that the development

of approximated methods for solving Eq. (1.1) has driven the theoretical research in chemistry, molecular physics and materials science for half of a century [1].

1.1.1 Density functional theory

Among the possible methods, density functional theory (DFT) is by far the most widely adopted approach to compute the electronic properties of a system. The theory is grounded on the remarkable realization that knowing the electron density $n_e(\boldsymbol{r})$ of the system is sufficient to uniquely determine its ground-state properties [2]. This feature represents a tremendous computational advantage, because one can in principle entirely forget about the N-body wave-function Ψ and simply design a minimization algorithm of the ground-state electronic energy E_0 as a functional of the electron density $n_e(\boldsymbol{r})$. The major problem of DFT consists in deriving accurate approximations for this functional dependence, whose exact form is still unknown. In spite of the extensive theoretical investigation put in place by the scientific community, the search for accurate density functionals has proven to be exceedingly difficult, especially for the kinetic energy contribution to the electronic energy [3]. This fundamental issue of the theory has led to a reformulation of DFT as an effective single-particle theory, i.e., the Kohn-Sham DFT (KS-DFT) approach, where the notion of wave-function is restored within the definition of the electron density [4]. While KS-DFT has been broadly successful in predicting a wide variety of materials properties, the dimensionality of the problem depends, unlike to a pure DFT approach, on the number of electrons N of the system. This implies that, similarly to other wave-function based methods for solving the Schrödinger equation, common KS-DFT implementations present an unfavorable scaling of the computational cost with the system size ($\sim N^3$), thus hindering the application of the method to arbitrarily large and complex systems.

1.2 The problem of thermodynamic convergence

The calculation of electronic properties is not the only computational bottleneck in the simulation of materials. Most of measurable physical observables come in fact as a thermodynamic average over the instantaneous atomistic configurations of the system at a given temperature T. In this context, the most popular computational strategy to simulate the thermodynamics of a system is the *ab-initio molecular dynamics* (AIMD) method. In AIMD, the statistical ensemble of atomistic configurations is obtained simulating the dynamics of the atomic nuclei driven by the quantum forces that can be computed from the solution of the electronic-structure problem [5]. Whenever

the quantum nature of the nuclei can be neglected, this dynamics can be propagated over finite time-steps, of the order of 1 femtosecond, in accordance to Newton's laws. From a numerical point of view, the discrete character of the propagation forbids the simulation of any actual realistic trajectory [6]. Instead, specific algorithms are designed to obtain a simulated trajectory that satisfies the constraint of a constant phase-space distribution function (Liouville's theorem), guaranteeing a consistent calculation of the statistical averages [7]. Unfortunately, the amount of statistics that needs to be collected to convergence any thermodynamic property typically involves very long trajectories that can encompass a simulation time of several nanoseconds, therefore limiting the applicability of AIMD to relatively small systems (~100 atoms).

1.2.1 Thermodynamic integration

This problem is greatly worsened whenever one is interested in predicting the thermodynamic stability of the system in terms of a free-energy difference ΔF between two thermodynamic states A and B. In this circumstance, a *thermodynamic integration* (TI) needs to be performed along the path that adiabatically connects the initial state A to the final state B, so that any free energy difference can in principle be computed as follows,

$$\Delta F(A \rightarrow B) = \int_0^1 d\lambda \left\langle \frac{dU(\lambda)}{d\lambda} \right\rangle_\lambda . \tag{1.2}$$

Here, $U(\lambda)$ represents a suitable parametrization of the system potential energy as a function of the coupling parameter λ that modulates the transition between the two states, while $\langle\cdot\rangle_\lambda$ is the canonical ensemble average computed over the Boltzmann distribution defined by $U(\lambda)$. In practice, this implies that the path of adiabatic connection needs to be sampled by computing several thermodynamic averages, one for each λ-point of the quadrature grid chosen to perform the integral of Eq. (1.2). These kind of calculations are particularly demanding, for instance, when considering *ab initio* solvation free-energies, where the TI path involves a smooth embedding of a solute molecule from a fixed position in the gas-phase to a fixed position in the liquid solvent [8].

1.3 Machine learning at the atomic scale

According to the previous discussion, the extensive application of AIMD simulations to the calculation of quantum-level thermodynamic quantities is limited by the major computational burden associated with solving the Schrödinger equation

at each molecular dynamics step. Machine learning (ML) methods have recently emerged as an extremely successful strategy to sidestep the solution of the electronic structure problem, paving the way to the calculation of a number of electronic and thermodynamic properties that would be otherwise not accessible by other existing approaches [9–13].

The main idea behind the application of ML at the atomic scale is to provide a method to inexpensively interpolate the value of a given physical quantity over a representative set of reference data. The distinctive character of these data-driven interpolations is that they do not rely on any prior knowledge on the behaviour of the target property. Any atomic-scale ML model is in fact asked to find the hidden and arbitrarily complex relationship between the target property and the atomic coordinates of the system by means of a completely general regression framework. This aspect is in stark contrast with common fitting strategies of potential energy surfaces that make use of system-dependent functional forms for describing the interaction between the atoms of the material [14].

To produce accurate predictions within the vast chemical and conformational space spanned by the atomistic configurations, ML models typically require to map the spatial coordinates into a suitable structural representation of the system that lives in an arbitrarily complex *feature space* [15]. The interplay between the complexity of the structural representation and the complexity of the learning algorithm is the crucial aspect that determines the capability of ML methods to yield accurate predictions using a finite amount of training data. In particular, endowing the structural representation with a certain level of physical insight usually comes along with enhancing the learning power of the method, especially when the regression is carried out using linear functional forms. Conversely, highly non-linear learning algorithms such as deep neural networks (DNN) can often be used in conjunction with representations of the system built as a simple manipulation of the atomic coordinates, e.g., that are trivially related to pairwise interatomic distances [16–19].

The great learning power of DNN is due to the large number of parameters that control the correlation between the input structure and physical target. Although beneficial in many cases, this flexibility usually requires to limit the possibility of overfitting by training the model on a large amount of reference calculations [20]. Moreover, the high complexity of the DNN architecture makes it hard to rationalize the learning performance and suggest possible strategies to improve the prediction accuracy. The need for ML models that are both data-efficient and that can provide some physical insight on the underlying interpolation process has therefore motivated part of the scientific community to rely on a different learning paradigm.

1.3.1 Gaussian process regression

Gaussian process regression (GPR) represents a non-parametric statistical approximation theory that offers a more transparent and controllable learning framework than DNN. Within GPR, a property y is assumed to behave as a stochastic variable distributed according to a Gaussian probability of zero mean and covariance $k \equiv \langle y, y' \rangle$. Upon performing a set of (noisy) observations $\boldsymbol{y} \equiv \{y_n\}$ of the target property for a set of independent input structures $\{A_n\}$, the regression problem is then formulated on the question of finding the posterior probability distribution $P(y|\boldsymbol{y})$ associated with observing a value y for a new input A, conditioned on the observations $\boldsymbol{y}$ previously carried out [21]. From Bayes' theorem, $P(y|\boldsymbol{y})$ results to be in turn Gaussian distributed; its expectation value and variance are respecively given by

$$\big\langle y(A) \big\rangle_{P(y|\boldsymbol{y})} = \boldsymbol{k}^T(A) \cdot \big[\boldsymbol{K} + \eta^2 \mathbf{1}\big]^{-1} \cdot \boldsymbol{y} = \boldsymbol{k}^T(A) \cdot \boldsymbol{x}(\{y_n\}) \tag{1.3}$$

and

$$\langle y(A), y(A) \rangle_{P(y|\boldsymbol{y})} = k(A,A) - \boldsymbol{k}^T(A) \cdot \big[\boldsymbol{K} + \eta^{-1} \mathbf{1}\big] \cdot \boldsymbol{k}(A)\,, \tag{1.4}$$

with $\boldsymbol{k}(A) \equiv \{k(A, A_n)\}$ the vector of covariance functions between the new and reference observations, $\boldsymbol{K} \equiv \{k(A_n, A_{n'})\}$ the covariance matrix of the reference observations and η the intrinsic Gaussian noise of these observations. The whole point of the application of GPR to the prediction of physical quantities is to use the atomic coordinates of the system to provide a good prior approximation for the covariance function k, usually called *kernel*. According to Eq. (1.3), once a kernel approximation is computed the learning exercise can be performed at the cost of a simple matrix inversion, and the predicted value is obtained as a linear combination of the kernels $\boldsymbol{k}(A)$ weighted by the regression coefficients $\boldsymbol{x}$. In this respect, setting a value for the Gaussian noise η provides a lower-bound to the eigenvalues of the kernel matrix $\boldsymbol{K}$, which has the ultimate effect of regularizing the solution of the regression problem. Importantly, in the absence of reference values for the target quantity, the variance of Eq. (1.4) can be used to estimate the intrinsic uncertainty of the GPR predictions and assess the reliability of the learning exercise.

Relying on a matrix inversion, the cost of performing a GPR exercise scales cubically with the problem dimensionality N. Whenever this scaling becomes a computational bottleneck of the method, a dimensionality reduction can be adopted, which aims at projecting the problem in a smaller, sparse set $M < N$ that best represents the starting conformational space of dimension N. This procedure is known as the *subset of regressors* (SoR) approximation [22], and it is grounded on the assumption that a

given kernel matrix $\boldsymbol{K}_{NN}$ can be approximated as

$$\boldsymbol{K}_{NN} \approx \boldsymbol{K}_{NM}\boldsymbol{K}_{MM}^{-1}\boldsymbol{K}_{NM}^{T}, \tag{1.5}$$

with $\boldsymbol{K}_{NM}$ the kernel matrix that couples the reference and sparse points, while $\boldsymbol{K}_{MM}$ the full rank kernel matrix associated with the sparse set. From Eq. (1.5), an analogous derivation to that outlined in the previous paragraph can be carried out, which yields the following prediction formula:

$$\left\langle y(A)\right\rangle_{P(y|\boldsymbol{y})} = \boldsymbol{k}_{M}^{T}(A)\cdot\left[\boldsymbol{K}_{NM}^{T}\boldsymbol{K}_{NM}+\eta^{2}\boldsymbol{K}_{MM}\right]^{-1}\boldsymbol{K}_{NM}\boldsymbol{y}_{N} = \boldsymbol{k}_{M}^{T}(A)\cdot\boldsymbol{x}_{M}(\{y_n\})\,. \tag{1.6}$$

Note that the information brought by the *entire* training set is still included in the definition of the regression weights, and that the original GPR formula of Eq. (1.3) can be easily recovered for $M = N$.

Being a symmetric and positive-definite correlation function, a kernel can always be represented as an inner product defined in a possibly infinite dimensional space of functions (Mercer's theorem [21]). In practice, however, knowing the underlying space of functions is not needed to compute a kernel approximation, whose definition can solely rely on the notion of *similarity* between a pair of input structures [23, 24]. One of the most popular examples of this is the Gaussian kernel, i.e.,

$$k(A,A') = \exp\left(-\frac{\left\|\boldsymbol{A}-\boldsymbol{A}'\right\|^{2}}{2\sigma^{2}}\right), \tag{1.7}$$

with $\left\|\boldsymbol{A}-\boldsymbol{A}'\right\|$ measuring the Euclidean distance between the inputs on a given feature space, while the parameter σ reflects the standard deviation of the inputs similarity. Gaussian kernels have found applications in the prediction of numerous quantum-mechanical properties, going from electronic energies [25, 26] and densities [27, 28], density-functionals [29, 30], atomic forces [31], electric multipoles [32] and molecular polarizabilities [33], often outperforming DNN when training the model on a small amount of reference data. Their success is mainly due to their *infinite smoothness*, meaning that Eq. (1.7) is infinitely differentiatable in a mean-square sense [21]. The importance of smoothness in statistical approximations of physical quantities is well understood in terms of the fact that small variations of the atomic coordinates typically correspond to small variations of the value of the target property. As we will see in the following sections, this concept has also greatly inspired the construction of smooth structural representations of the system, that can either be used to compute kernel similarity measures as the one of Eq. (1.7), or enter the construction of linear regression models.

1.3.2 Linear regression models

It is particularly instructive to draw the connection between GPR and linear regression models. Here, and for the rest of the thesis, we will extensively rely on Dirac notation commonly used in quantum mechanics, so that bras $\langle\cdot|$ and kets $|\cdot\rangle$ are interpreted as abstract representations and brakets $\langle\cdot|\cdot\rangle$ are used to identify inner products. As detailed in Appendix A, this choice leaves us the freedom of adopting any arbitrary complete basis to compute an abstract structural representation $|A\rangle$ of the system A and to perform the integrals and summations that underlie the definition of inner products. Borrowing Dirac notation, a linear model to predict a property y for an input A is written as follows,

$$y(A) = \langle w|A\rangle \,, \tag{1.8}$$

where $\langle w|$ is an abstract vector for the regression weights we wish to learn. The usual approach for finding the weight vector is to minimize a quadratic loss function computed over N reference observations $\{y_n\}$, i.e.,

$$\ell(w) = \sum_{n=1}^{N} \big(y_n - \langle w|A_n\rangle\big)^2 + \lambda \langle w|w\rangle \,, \tag{1.9}$$

where the regularization parameter λ introduces a penalty for high-norm weight vectors, preventing from overfitting the model on the training data. $\ell(w)$ attains its maximum at

$$|w\rangle = \big(\hat{C} + \lambda \hat{I}\big)^{-1} \sum_{n=1}^{N} |A_n\rangle \, y_n \,, \tag{1.10}$$

with $\hat{C}$ the projector operator of the model covariance, i.e., $\hat{C} = \sum_{n=1}^{N} |A_n\rangle \langle A_n|$.

This construction in which one handles the representation $|A\rangle$ explicitly is often called the *primal* formulation. There is in fact another, complementary formulation, called the *dual*, which appears to be totally equivalent to GPR. In the dual formulation one does not handle the representation $|A\rangle$ explicitly but rather the similarity (kernel) between two inputs, $k(A, A')$. In this case, the loss function takes the form

$$\ell(\boldsymbol{x}) = \sum_{n=1}^{N} \big(y_n - \boldsymbol{k}^T(A_n)\boldsymbol{x}\big)^2 + \eta \, \boldsymbol{x}^T \boldsymbol{K} \boldsymbol{x}, \tag{1.11}$$

and minimizing with respect to $\boldsymbol{x}$ leads exactly to the GPR regression formula already reported in Eq. (1.3). Importantly, when computing the kernel as an inner product within a finite and known feature space, i.e., $k(A, A') = \big\langle A \big| A' \big\rangle$, the primal and dual

formulations are formally equivalent, and the choice of which to use is a purely practical question. Constructing a primal model requires inversion of the covariance matrix $\boldsymbol{C}$, while the dual requires inversion of the kernel matrix $\boldsymbol{K}$. If the feature space is larger than the training set then the dual approach is more convenient (*kernel trick*). Of course, the real utility of the kernel trick becomes apparent when the kernel is a complex, non-linear function for which the feature space is unknown and/or infinite-dimensional, as in Eq. (1.7). In these circumstances, working in the dual makes it possible to formulate regression as a linear problem, where reference configurations are used to define a basis for the target and all the complexity of the input space representation is contained in the definition of the kernel function.

1.4 Physics-enhanced machine learning

Thinking in terms of linear regression models, either in the primal or in the dual formulation, streamlines the connection between the learning target y and the structural representation $|A\rangle$ in terms of a one-to-one mapping $y \sim |A\rangle$. In this respect, it comes as no surprise that maximizing both the prediction accuracy and the data efficiency of the regression model comes along with the adoption of representations that mirror some general physical properties associated with the learning target.

1.4.1 The importance of locality

The statistical approximation of electronic energies is the prototypical example where the representation can be constructed to follow the physics of the learning target. In this case, a concept that has greatly inspired theoretical developments is that of *locality*. The local nature of the target reflects the fact that the response of a system to far-field perturbations is typically governed by screening phenomena that limit the spatial extent over which the effect of these perturbations can be propagated – a concept first introduced by Walter Kohn as the *nearsightedness of electronic matter* [34, 35]. Local representations are typically defined by means of finite spherical environments of a given cutoff radius r_c that are used to spatially limit the structural information around the atoms of the system. Crucially, this construction implies that the electronic energy is implicitly broken down in the sum of atomic contributions e_i, that, individually, encode the local many-body nature of the target [36]. Within a linear model, for instance, we would write

$$E(A) = \langle w|A\rangle = \langle w| \left(\sum_{i=1}^{N} |A_i\rangle \right) = \sum_{i=1}^{N} \langle w|A_i\rangle = \sum_{i=1}^{N} e_i \,, \tag{1.12}$$

with $|A_i\rangle$ indicating the abstract representation of the local environment of the atom i. This additivity property of the energy prediction carries a twofold advantage: on one hand, it enables a great transferability of the learning model across systems that share similar atomic environments within the selected cutoff radius r_c, in fact limiting the possibility of overfitting the model on the training configurations [10], and, on the other hand, it guarantees that the extensivity of target is automatically satisfied, so that the energy of two non-interacting systems A and B is given by the sum of the energies of the individual systems, i.e., $E(A+B) = E(A) + E(B)$.

1.4.2 The importance of symmetry

A similarly important aspect in the construction of efficient structural representations is that of *symmetry* [37]. Enforcing the expected physical symmetries of the target has the obvious advantage of letting the regression focus on the chemical and structural variability of the dataset, thus avoiding to waste a great amount of reference calculations in learning a piece of information that can be encoded *a priori* within the structural representation. The simplest example of this is the invariance of most electron-structure properties to permutation of identical atoms. Internal coordinates representations such as Coulomb matrices [38], where the atomic structure is mapped to a list of pairwise Coulomb interactions $Z_i Z_j / r_{ij}$ between the atoms, or similar bag-of-bonds (BoB) descriptors [39], for instance, suffer from the lack of permutational invariance and would require to average the representation over all the possible permutations of identical atomic pairs [40]. However, because of the unfavorable (exponential) scaling of the number of these permutations with the system size, the problem can be more effectively circumvented by adopting stochastic sorting algorithms of the feature vector components [41].

The need for representations that are naturally endowed with permutational symmetry has favoured the development of smooth functions of the atomic coordinates that are inherently defined to solely depend on the chemical species (H, C, O, ...) rather than on the identity of the atoms involved. *Symmetry functions* [36], for instance, are defined as a smooth representation of the internal coordinates of the system by collecting, for any given atom, the contributions coming for all the atomic neighbours of a given chemical species. A similar construction can be found in any field-based representation that is derived from the sum of species-dependent Gaussian functions centered on the atomic positions [27, 31, 42, 43]. For example, a generic atom-density representation would be constructed starting from the following real-space definition,

$$\rho_a(\boldsymbol{x}) = \sum_{j \in a} \exp\left(-\frac{|\boldsymbol{x} - \boldsymbol{r}_j|^2}{2\sigma^2}\right), \tag{1.13}$$

with a labeling the chemical species of the atoms j and σ defining the Gaussian width that modulates the spatial resolution of atomic field. It is worth noticing that the inherent smoothness of this class of field-derived representations is also important to limit the fluctuations of the machine-learning predictions in response to an arbitrary variation of the atomic coordinates. In fact, even though representations based on internal coordinates are also a smooth function of atomic positions, including *a posteriori* the permutational symmetry by sorting carries the major drawback of introducing derivative discontinuities that damage the function regularity needed to effectively predict the target quantity for arbitrarily distorted structures [44].

Of course, swapping the label of identical atoms is not the only symmetry relationship that can be attributed to the ground-state properties of a system. Assuming that no external field is introduced, any molecule or material lives in fact in a homogeneous and isotropic three-dimensional space; this implies that spatial symmetries such as rigid translations, rotations and reflections about a mirror plane, need to be considered when constructing any machine learning model. While using representations based on internal coordinates carries the advantage of having these symmetries naturally built-in, representations constructed as smooth functions of the atomic positions, such as the one of Eq. (1.13), typically require to explicitly introduce these symmetries by performing prescribed integral operations [45]. These are presented in details in the following section.

1.5 Atom density representations

To exemplify the inclusion of spatial symmetries within the machine-learning model, we consider here the case of atom density representations as derived in Ref. [45]. The rationale behind this construction is particularly relevant for our discussion, as it underpins the general concepts by which the original methods presented in this thesis are derived.

1.5.1 Translational symmetry

Let us start by considering a general atom density field as the one reported in Eq. (1.13). Borrowing Dirac notation (Appendix A), this can be defined as follows,

$$\langle a\boldsymbol{x}|\rho\rangle \equiv \sum_j \delta_{aa_j} g_\sigma(\boldsymbol{x}-\boldsymbol{r}_j)\,, \tag{1.14}$$

with g_σ a Gaussian function of width σ and $\langle a\boldsymbol{x}|$ indicating the species-dependent real space basis over which the abstract density state $|\rho\rangle$ of the atomic structure is projected. To introduce the invariance of the representation under translations, hence assuming that no external field is applied to the system, one can consider the two-body correlation function that arises upon performing the integral over all possible continuous translation operators $\hat{t}$ that are applied to the product of two atomic densities $|\rho\rangle$, i.e.,

$$\begin{aligned}\int d\hat{t}\,\langle a_1\boldsymbol{x}_1|\,\hat{t}\,|\rho\rangle\,\langle a_2\boldsymbol{x}_2|\,\hat{t}\,|\rho\rangle &= \sum_{ij}\delta_{a_1a_i}\delta_{a_2a_j}\int d\boldsymbol{t}\, g_{\sigma_1}(\boldsymbol{x}_1-\boldsymbol{r}_i+\boldsymbol{t})\,g_{\sigma_2}(\boldsymbol{x}_2-\boldsymbol{r}_j+\boldsymbol{t})\\ &= \sum_{ij}\delta_{a_1a_i}\delta_{a_2a_j}\,g_\sigma\left((\boldsymbol{x}_1-\boldsymbol{x}_2)-(\boldsymbol{r}_i-\boldsymbol{r}_j)\right)\\ &= \sum_{ij}\delta_{a_1a_i}\delta_{a_2a_j}\,g_\sigma\left(\boldsymbol{x}-\boldsymbol{r}_{ij}\right)\end{aligned} \tag{1.15}$$

where we defined $\boldsymbol{x}=\boldsymbol{x}_1-\boldsymbol{x}_2$ and $\boldsymbol{r}_{ij}=\boldsymbol{r}_i-\boldsymbol{r}_j$ in the last equality. Note that while the final pairwise Gaussian g_σ comes from the convolution properties of Gaussian functions, so that $\sigma=\sqrt{\sigma_1^2+\sigma_2^2}$, any pair of localized functions about the atomic positions would lead to a similar result. Unsurprisingly, introducing the invariance of the representation under translations therefore corresponds to fix the origin of the reference frame about the atoms of the system. In view of designing a local regression model for the prediction of physical quantities, one can then proceed to single out an atom-centered representation that relies on the following definition:

$$\langle a\boldsymbol{x}|\rho_i\rangle \equiv \sum_j \delta_{aa_j} g_\sigma(\boldsymbol{x}-\boldsymbol{r}_{ij})\,. \tag{1.16}$$

The actual localization of the representation about the atom i can be performed by applying to Eq. (1.16) a more or less smooth cutoff function of radius r_c. From now on, we will therefore refer to $|\rho_i\rangle$ as an abstract local density representation of the atomic structure centered about the atom i, thus always implying the inclusion of the translational symmetry.

1.5.2 2-body invariants

Unlike translational invariance, adapting the representation to the symmetries of the $O(3)$ group, namely three-dimensional rotations and reflections, is less trivial, and it strongly depends on the nature of the learning target. If one is interested in learning electronic energies, as well as any other scalar quantity, the rotational symmetry can be introduced by averaging the representation over all the possible three-dimensional rotations $\hat{R}$ that are defined by the triplet of Euler angles (α,β,γ) over $8\pi^2$. In doing so, we will make extensive use of spherical harmonics $\langle\hat{\boldsymbol{x}}|lm\rangle \equiv Y_m^l(\hat{\boldsymbol{x}})$ as a complete basis to represent angular correlations over the unit sphere, so that the actual regression features of the machine-learning model will be always identified by the spherical harmonics expansion coefficients of the symmetry-adapted structural representation. For example, performing the rotational average on Eq. (1.16) yields a rotationally invariant representation of the kind

$$\left\langle a\boldsymbol{x}\middle|\overline{\rho_i}\right\rangle = \frac{1}{8\pi^2}\int d\hat{R}\,\langle a\boldsymbol{x}|\,\hat{R}\,\middle|\rho_i\right\rangle = \langle\hat{\boldsymbol{x}}|00\rangle\left\langle ax00\middle|\overline{\rho_i}\right\rangle, \tag{1.17}$$

where $\left\langle ax00\middle|\overline{\rho_i}\right\rangle$ expresses the spherically symmetric components of the species-dependent density about the atom i. Note that, from here on, we will always use an overline notation to concisely indicate the result of the rotational average.

The isotropic density components $\left\langle ax00\middle|\overline{\rho_i}\right\rangle$ can directly be used in a regression to represent 2-body radial correlations. However, the real challenge in the interpolation of electronic energies, as well as of any other scalar physical observable, consists in representing the complicated many-body structural correlations that are encoded in the outcome of a quantum-mechanical calculation. For this reason, correlations of higher order must be introduced.

1.5.3 3-body invariants

Correlations beyond 2-body can be obtained by applying the rotational average on the tensor products of Eq. (1.16) with itself. At order $\nu+1$ in body-correlations, we can then generally write

$$\left\langle a_1\boldsymbol{x}_1;a_2\boldsymbol{x}_2;...;a_\nu\boldsymbol{x}_\nu\middle|\overline{\rho_i^{\otimes\nu}}\right\rangle = \frac{1}{8\pi^2}\int d\hat{R}\,\langle a_1\boldsymbol{x}_1|\,\hat{R}\,\middle|\rho_i\right\rangle\langle a_2\boldsymbol{x}_2|\,\hat{R}\,\middle|\rho_i\right\rangle...\langle a_\nu\boldsymbol{x}_\nu|\,\hat{R}\,\middle|\rho_i\right\rangle. \tag{1.18}$$

A particularly important case, illustrated in Fig. 1.1, is the one of $\nu=2$, corresponding to 3-body correlations. In this case, the rotational average of a pair of spherical

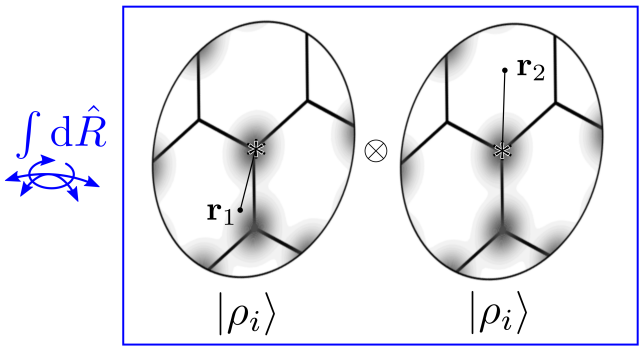


Figure 1.1 – 3-body correlations arise from the rotational average of a pair of smeared atomic densities sampled in a local environment of a given atom i.

harmonics components yields an expansion over Legendre polynomials P_l, which has the role of representing the angular correlation between two points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ located within a spherical environment of the central atom i:

$$\begin{aligned}
\left\langle a_1\boldsymbol{x}_1; a_2\boldsymbol{x}_2 \middle| \overline{\rho_i^{\otimes 2}} \right\rangle &= \frac{1}{8\pi^2}\int d\hat{R}\, \langle a_1\boldsymbol{x}_1|\hat{R}|\rho_i\rangle \langle a_2\boldsymbol{x}_2|\hat{R}|\rho_i\rangle \\
&= \frac{1}{8\pi^2}\sum_{l=0}^{\infty}\left[\sum_m \langle a_1 x_1 lm|\rho_i\rangle^\star \langle a_2 x_2 lm|\rho_i\rangle\right]\frac{8\pi^2}{2l+1}\sum_{m'}\langle\hat{\boldsymbol{x}}_1|lm'\rangle\langle lm'|\hat{\boldsymbol{x}}_2\rangle \\
&= \frac{1}{4\pi}\sum_{l=0}^{\infty}\left[\sum_m \langle a_1 x_1 lm|\rho_i\rangle^\star \langle a_2 x_2 lm|\rho_i\rangle\right] P_l(\hat{\boldsymbol{x}}_1\cdot\hat{\boldsymbol{x}}_2) \\
&= \frac{1}{4\pi}\sum_{l=0}^{\infty}\left\langle a_1 x_1 l; a_2 x_2 l \middle| \overline{\rho_i^{\otimes 2}} \right\rangle P_l(\hat{\boldsymbol{x}}_1\cdot\hat{\boldsymbol{x}}_2)\,,
\end{aligned} \tag{1.19}$$

where we used the spherical harmonics addition theorem,

$$P_l(\hat{\boldsymbol{x}}_1\cdot\hat{\boldsymbol{x}}_2) = \frac{4\pi}{2l+1}\sum_{m'}\langle\hat{\boldsymbol{x}}_1|lm'\rangle\langle lm'|\hat{\boldsymbol{x}}_2\rangle\,. \tag{1.20}$$

In the last equality of Eq. (1.19), the 3-body rotationally invariant coefficients that can be used as regression features are defined as follows,

$$\left\langle a_1 x_1 l; a_2 x_2 l \middle| \overline{\rho_i^{\otimes 2}} \right\rangle = \sum_m \langle a_1 x_1 lm|\rho_i\rangle^\star \langle a_2 x_2 lm|\rho_i\rangle\,, \tag{1.21}$$

with $\langle axlm|\rho_i\rangle$ the spherical harmonic projections of the atom-density field. The calculation of these projections typically requires to expand the representation over an orthogonal radial basis $\langle n|x\rangle \equiv R_n(x)$, so that, in practice, one has to deal with the discretized set of orthogonal projections $\langle anlm|\rho_i\rangle = \int_0^\infty dx x^2 \langle n|x\rangle\langle axlm|\rho_i\rangle$.

Analytical formulas for the calculation of $\langle anlm|\rho_i\rangle$, that have extensively been used to produce the results of this thesis, are detailed in Appendix B.

In practice, the spherical harmonics expansion has to be truncated at a certain cutoff value $l_{\max}$ that determines the accuracy by which angular correlations in real space are represented. In this regard, the local nature of the representation plays a crucial role: single-centered spherical harmonics expansions of arbitrarily extended three-dimensional fields are known to converge very slowly, so that their practical calculation becomes numerically affordable only when the field is spatially localized around the expansion center [46]. As a final remark, note that the structural features of Eq. (1.21) are, by construction, also invariant under inversion operations about the atomic center, so that the final representation is adapted to the symmetries of the $O(3)$ group.

1.5.4 4-body invariants

As a final example, we consider the case of 4-body invariants. From Eq. (1.18), these are generally defined as

$$\left\langle a_1\boldsymbol{x}_1; a_2\boldsymbol{x}_2; a_3\boldsymbol{x}_3 \middle| \overline{\rho_i^{\otimes 3}} \right\rangle = \frac{1}{8\pi^2}\int d\hat{R}\, \langle a_1\boldsymbol{x}_1|\,\hat{R}\,|\rho_i\rangle\, \langle a_2\boldsymbol{x}_2|\,\hat{R}\,|\rho_i\rangle\, \langle a_3\boldsymbol{x}_3|\,\hat{R}\,|\rho_i\rangle\,. \quad (1.22)$$

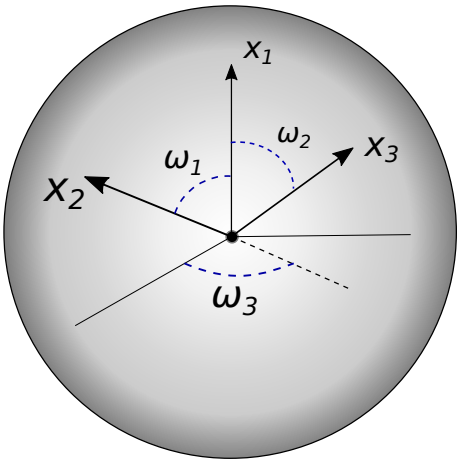


Figure 1.2 – 4-body angular correlations are uniquely identified by a triplet of angles.

Introducing an additional order in density-correlations implies that the effect of the rotational average cannot simply be represented through an expansion over Legendre polynomials. As illustrated in Fig. 1.2, given the triplet of versors $(\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2, \hat{\boldsymbol{x}}_3)$ that identify the position on the unit sphere of 3 points around the central atom, 4-body angular correlations are uniquely determined by a triplet of angles $(\omega_1, \omega_2, \omega_3)$ that define the relative orientation of the three versors. For instance, given $(\hat{\boldsymbol{e}}_x, \hat{\boldsymbol{e}}_y, \hat{\boldsymbol{e}}_z)$

the Cartesian unit vectors, any triplet $(\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2, \hat{\boldsymbol{x}}_3)$ that preserves the internal angles $(\omega_1, \omega_2, \omega_3)$ can be defined applying a rigid rotation to the following stencil,

$$\begin{cases} \hat{\boldsymbol{x}}_1 = \hat{\boldsymbol{e}}_z \\ \hat{\boldsymbol{x}}_2 = \sin\omega_1 \ \hat{\boldsymbol{e}}_x + \cos\omega_1 \ \hat{\boldsymbol{e}}_z \\ \hat{\boldsymbol{x}}_3 = \sin\omega_2 \ \left(\cos\omega_3 \ \hat{\boldsymbol{e}}_x + \sin\omega_3 \ \hat{\boldsymbol{e}}_y\right) + \cos\omega_2 \ \hat{\boldsymbol{e}}_z \, . \end{cases} \tag{1.23}$$

Thanks to this realization, the rotationally invariant expansion of Eq. (1.22) can be conveniently written as

$$\left\langle a_1 \boldsymbol{x}_1; a_2 \boldsymbol{x}_2; a_3 \boldsymbol{x}_3 \middle| \overline{\rho_i^{\otimes 3}} \right\rangle = \sum_{l_1 l_2 l_3} \left\langle a_1 x_1 l_1; a_2 x_2 l_2; a_3 x_3 l_3 \middle| \overline{\rho_i^{\otimes 3}} \right\rangle \langle \omega_1 \omega_2 \omega_3 | l_1 l_2 l_3 \rangle \, . \tag{1.24}$$

As in the description of molecular correlations in theories of dipolar fluids [47], the rotationally invariant angular functions $\langle \omega_1 \omega_2 \omega_3 | l_1 l_2 l_3 \rangle$ are defined by coupling a triplet of spherical harmonics via the Wigner-3J symbols:

$$\langle \omega_1 \omega_2 \omega_3 | l_1 l_2 l_3 \rangle = \sum_{m_1' m_2' m_3'} \begin{pmatrix} l_1 & l_2 & l_3 \\ m_1' & m_2' & m_3' \end{pmatrix} \langle \hat{\boldsymbol{x}}_1 | l_1 m_1' \rangle \langle \hat{\boldsymbol{x}}_2 | l_2 m_2' \rangle \langle \hat{\boldsymbol{x}}_3 | l_3 m_3' \rangle \, . \tag{1.25}$$

Similarly, the rotationally invariant expansion coefficients come from the coupling of a triplet of atom density spherical harmonics components:

$$\begin{aligned} \left\langle a_1 x_1 l_1; a_2 x_2 l_2; a_3 x_3 l_3 \middle| \overline{\rho_i^{\otimes 3}} \right\rangle = \sum_{m_1 m_2 m_3} & \begin{pmatrix} l_1 & l_2 & l_3 \\ m_1 & m_2 & m_3 \end{pmatrix} \\ & \langle a_1 x_1 l_1 m_1 | \rho_i \rangle \langle a_2 x_2 l_2 m_2 | \rho_i \rangle \langle a_3 x_3 l_3 m_3 | \rho_i \rangle \, . \end{aligned} \tag{1.26}$$

When compared with their 3-body counterparts, these coefficients are not necessarily invariant under inversion symmetry: applying an inversion operation $\hat{i}$ to the system would in fact bring a factor $(-1)^{l_1+l_2+l_3}$ that might or not cause a change of sign depending on the parity of the combination of angular momenta. This is coherent with the fact that the triplet of versors exemplified in Fig. 1.2 leaves total freedom on the *handedness* of the real space representation. If one is interested in learning physical quantities that are invariant under inversion of the atomic structure, such as the electronic energy of a system, then only the coefficients for which the combination of angular momenta $l_1 + l_2 + l_3$ is even need to be retained. Conversely, learning chiral properties of the system call for structural representations that change sign under an inversion operation, so that one should only retain the 4-body coefficients that realize an odd combination of angular momenta. For instance, the prediction of circular dichroism absorption spectra [48] is a clear example of a property that would require to adopt a chiral representation of the atomic structure [49].

1.5.5 Kernel trick and non-linearity

Upon the substitution $x \to n$, the discretized set of invariant features previously derived correspond to the local environment representations that underlie the *smooth overlap of atomic positions* (SOAP) method [42], namely the SOAP power spectrum and bispectrum. In the notation of Bartók *et al.* [42] and De *et al.* [23], these correspond to

$$p^{a_1 a_2}_{n_1 n_2 l} \equiv \left\langle a_1 n_1 l; a_2 n_2 l \middle| \overline{\rho^{\otimes 2}_i} \right\rangle , \quad b^{a_1 a_2 a_3}_{n_1 n_2 n_3 l_1 l_2 l_3} \equiv \left\langle a_1 n_1 l; a_2 n_2 l; a_2 n_3 l \middle| \overline{\rho^{\otimes 3}_i} \right\rangle . \quad (1.27)$$

In particular, the contraction of the 3-body representation over the feature space defined by the rotational invariant basis $\langle a_1 n_1 l; a_2 n_2 l|$ yields the popular SOAP kernel routinely used in machine-learning applications [50, 51]. For any pair of atomic environments i and j belonging to any pair of structures A and B, a SOAP kernel can then be computed as follows,

$$\begin{aligned} k(A_i, B_j) &= \left\langle \overline{\rho^{\otimes 2}_i}(A) \middle| \overline{\rho^{\otimes 2}_j}(B) \right\rangle \\ &= \sum_{a_1 a_2 n_1 n_2 l} \left\langle \overline{\rho^{\otimes 2}_i}(A) \middle| a_1 n_1 l; a_2 n_2 l \right\rangle \left\langle a_1 n_1 l; a_2 n_2 l \middle| \overline{\rho^{\otimes 2}_j}(B) \right\rangle . \end{aligned} \quad (1.28)$$

In this context, a crucial application of the kernel trick consists in elevating the SOAP kernel to an integer power ζ, which underlies non-linear representations that are built as the tensor product of the rotationally invariant 3-body expansion coefficients, i.e.,

$$\left| \left(\overline{\rho^{\otimes 2}_i} \right)^{\otimes \zeta} \right\rangle = \underbrace{\left| \overline{\rho^{\otimes 2}_i} \right\rangle \otimes \left| \overline{\rho^{\otimes 2}_i} \right\rangle \otimes \ldots \otimes \left| \overline{\rho^{\otimes 2}_i} \right\rangle}_{\zeta \text{ times}} . \quad (1.29)$$

For $\zeta = 2$, for instance, the kernel trick reads as follows,

$$k^{(\zeta=2)}(A_i, B_j) = \left\langle \left(\overline{\rho^{\otimes 2}_i} \right)^{\otimes 2}(A) \middle| \left(\overline{\rho^{\otimes 2}_j} \right)^{\otimes 2}(B) \right\rangle = \left\langle \overline{\rho^{\otimes 2}_i}(A) \middle| \overline{\rho^{\otimes 2}_j}(B) \right\rangle^2 = k^2(A_i, B_j) . \quad (1.30)$$

This choice has been widely exploited to improve the accuracy of machine-learning predictions [50, 52]. The reason for this improvement can be attributed to the higher order of atomic correlations that are introduced upon taking the tensor products of the 3-body representation with itself. However, because each of the individual representations is rotationally invariant, it is easy to show that this is only partially correct, as the angular information associated with correlations beyond 3-body would necessarily be missing [53].

1.5.6 The importance of being many-body

Representations that are truly many-body need to be constructed following the recipe of Eq. (1.18). In this regard, one should consider that the actual calculation of many-body atom density representations is hindered by the exponential scaling of the number of combinations between the angular momentum components that need to be coupled when performing the rotational average. Remarkably, efficient recursive evaluation schemes have recently been proposed to overcome this obstacle [54], that could be used in the near future to entirely bypass the need for non-linear kernels similar to the one of Eq. (1.30). As a related aspect, one should also consider that increasing the body-order of structural correlations can be essential to guarantee the injective relationship between the physical target and the machine-learning representation, meaning that structures with different physical observables should always be associated with distinct atomistic representations. It was in fact a widespread belief that 3-body representations such as the one of Eq. (1.19) could always be used to distinguish any pair of different atomic environments, a belief that has only recently been proven wrong using specifically designed structural manifolds [55].

1.6 Research outline

This thesis is part of a general effort towards the derivation of machine-learning models of molecules and materials that can be used to predict a large variety of quantum-mechanical observables, including scalars, tensors and scalar fields, while also presenting a consistent description of long-range interactions. Following the line of thought introduced in the previous sections, the challenge of obtaining models that are at the same time highly transferable across systems of different nature and size, and that can yield accurate predictions using a relatively small amount of training data, is addressed by deriving mathematical representations of the atomic structure that satisfy some stringent physical principles. The thesis achievements are divided in three parts, each of which contains chapters that, individually, are adapted from the articles published by the candidate during his doctoral studies.

The first part addresses the problem of learning tensorial properties, such as polarizabilities, electronic multipoles and dielectric responses. When compared to scalar quantities, physical tensors carry the additional complexity of following prescribed transformation rules upon a three-dimensional rotation of the system, which ultimately requires one to design representations of the atomic structure that are *covariant*, rather than invariant, under rotations. Chapter 2 tackles this problem by deriving a class of symmetry-adapted representations and kernels that generalize

the SOAP construction to include spherical harmonics covariance properties [56]. Chapter 3 shows how this class of tensorial features can be used in practice to learn any physical tensor that is decomposed in its irreducible spherical components (ISCs), demonstrating the effectiveness of the method in predicting the optical response series of water oligomers and the electronic dielectric properties of liquid water [57]. In Chapter 4, these achievements are fully exploited within a learning model that is able to readily interpolate coupled-cluster-level molecular polarizabilities across a very heterogeneous dataset [58]. Finally, Chapter 5 shows how disposing of accurate polarizability predictions associated with an entire molecular dynamics trajectory enables the calculation of Raman spectra in different crystal polymorphs of paracetamol [59].

The second part of the thesis discusses the possibility of predicting electronic-structure properties by building on the symmetry-adapted regression method previously introduced. Chapter 6, in particular, introduces a learning framework that is able to regress electronic charge densities, as well as any three-dimensional scalar field that can be expanded on a multi-centered spherical harmonics basis [60]. In Chapter 7, the accuracy of the model is greatly increased by making the electron density calculations coherent with state-of-the-art resolution of the identity (RI) schemes commonly used in quantum-chemistry [61]. In both chapters, we will see how the local and symmetry-adapted nature of the learning model comes along with highly transferable predictions across broad chemical and conformational spaces, opening the door to inexpensive electron density calculations that scale linearly with the system size.

The last part of the thesis addresses the long-standing problem of overcoming the nearsighted nature of local machine-learning representations, which neglect, by construction, any long-range effect that occurs farther than the cutoff distance used to define the spatial extent of the atomic environments. In Chapter 8, this is done by deriving a method that incorporates the long-range information of the system via Coulomb-like potential representations, while still preserving a certain degree of transferability [43]. Finally, Chapter 9 reports an improved version of the method, where density and potential features are combined in a single multi-scale representation, that is flexible enough to learn both local and non-local effects on an equal footing, embracing Pauli repulsion, electrostatics, polarization and dispersion interactions [62]. Crucially, when applied to the regression of potential energy surfaces, the method proposed can also be put under rigorous formal correspondence with the multipole expansion of long-range interactions, in fact entirely bypassing the need of adopting arbitrary electrostatic baselines of the electronic energy, as well as any intermediate machine-learning model that targets the prediction of atomic partial-charges and multipoles.

Part I Tensorial properties

2 Symmetry-adapted representations

The importance of endowing structural representations with prescribed spatial symmetries is entirely manifested when considering the implications of learning tensorial properties, or, similarly, any quantity that is not invariant under a rigid rotation or reflection of the atomic structure, such as atomic forces, dipoles, multipoles and polarizabilities. This chapter defines the problem of tensor learning and provides the theoretical background that has led to the development of symmetry-adapted representations within linear and kernel-based regression models. Sections and figures are adapted the following book contribution:

A. Grisafi, D. M. Wilkins, M. J. Willatt and M. Ceriotti, "Atomic-Scale Representation and Statistical Learning of Tensorial Properties", in *Machine Learning in Chemistry*, Vol. 1326, edited by E. O. Pyzer-Knapp and T. Laino (American Chemical Society, Washington, DC, Jan.2019), pp. 1–21. Copyright © 2019 American Chemical Society. AG contributed to writing the manuscript and produced the figures for the examples reported.

2.1 Covariant transformations

Let us start by considering the prototypical case of a Cartesian tensor $y \equiv y_{\alpha\beta\ldots}$ of rank r, with the combination of indices $\{\alpha\beta\ldots\}$ running over a number of Cartesian components equal to 3^r. Given any arbitrary distorted atomic structure with no particular internal symmetry, we are interested in characterizing the transformations of the tensor under only three families of symmetry operations (viz., translations, rotations and reflections). Since these symmetry operations do not affect the internal geometry of an atomic structure, we can think equivalently in terms of active transformations, in which the system undergoes the symmetry operation and the reference frame remains fixed, or in terms of passive transformations, in which the reference frame undergoes the symmetry operation and the system remains fixed. In the following, we summarize

the symmetry operations by adopting an active picture and assume the system is not subjected to an external field.

Translations. Any physical property of an atomic structure A remains unchanged under a rigid translation $\hat{t}$ of atomic positions, that is,

$$y_{\alpha\beta\ldots}(\hat{t}A) = y_{\alpha\beta\ldots}(A)\ . \tag{2.1}$$

Rotations. Under the application of a rigid rotation $\hat{R}$ to an atomic structure A, we assume that each Cartesian component of the tensor undergoes a covariant linear transformation. Using Einstein notation for convenience, and representing by $\boldsymbol{R}$ the rotation matrix corresponding to $\hat{R}$, the rotated tensor is

$$y_{\alpha\beta\ldots}(\hat{R}A) = R_{\alpha\alpha'}R_{\beta\beta'} \times \ldots \times y_{\alpha'\beta'\ldots}(A)\ . \tag{2.2}$$

Reflections. Applying a reflection operator $\hat{Q}$ to an atomic structure A through any mirror plane leads to the following reflected tensor,

$$y_{\alpha\beta\ldots}(\hat{Q}A) = Q_{\alpha\alpha'}Q_{\beta\beta'} \times \ldots \times y_{\alpha'\beta'\ldots}(A)\ . \tag{2.3}$$

2.2 Covariant representations

In general terms, a primitive representation that mirrors a tensor of a given rank r could formally be built by considering

$$\left|A;\alpha\beta\ldots\right\rangle = |A\rangle \otimes |\alpha\rangle \otimes \left|\beta\right\rangle \otimes \ldots\,, \tag{2.4}$$

where $|A\rangle$ is an arbitrary description of the system, while $|\alpha\rangle$ represents a set of Cartesian axes which is rigidly attached to the system. When using this primitive representation in a linear regression model, the tensor component corresponding to $\alpha\beta\ldots$ would be

$$y_{\alpha\beta\ldots}(A) = \left\langle w\middle|A;\alpha\beta\ldots\right\rangle, \tag{2.5}$$

or

$$y_{\alpha\beta\ldots}(A) = \left\langle w_{\alpha\beta\ldots}\middle|A;\alpha\beta\ldots\right\rangle. \tag{2.6}$$

After minimizing the primal-space loss function of Eq. (1.9), however, the former possibility would lead to a model that predicts every component to be the same, while

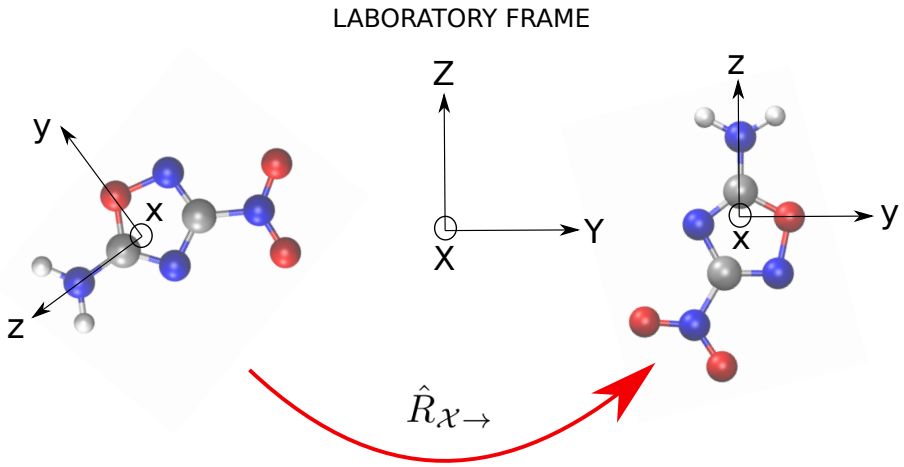


Figure 2.1 – Provided that one can define a local reference system, it is possible to learn tensorial properties by aligning the system into a fixed reference frame.

the latter would ignore the known correlations between different components.

To address these problems, one should adapt the primitive descriptor so that it fulfills each of the symmetries detailed in Eqs. (2.1-2.3). Since the Cartesian basis vectors are invariant under translations, Eq. 2.1 implies the core representation should itself be invariant under translations. Following the same procedure as in Sec. 1.5.1 one can construct a core representation that is invariant under translations by integrating an arbitrary representation over the translation operator $\hat{t}$. One can then proceed to consider the transformations under $SO(3)$ group operations, namely rotations. Eq. 2.2 implies that a covariant representation for $\left|A;\alpha\beta...\right\rangle$, namely $\left|\overline{A;\alpha\beta...}\right\rangle$, should satisfy the invariance relationship

$$\left[\hat{I}\otimes\hat{R}\otimes\hat{R}\otimes...\right]\left|\overline{\left(\hat{R}A\right);\alpha\beta...}\right\rangle=\left|\overline{A;\alpha\beta...}\right\rangle, \tag{2.7}$$

for any rotation $\hat{R}$. Starting from the primitive definition of Eq. 2.4, there are a variety of ways to enforce this invariance relationship. One possibility is to use

$$\left|\overline{A;\alpha\beta...}\right\rangle\equiv\left[\hat{I}\otimes\hat{R}_{A\rightarrow}\otimes\hat{R}_{A\rightarrow}\otimes...\right]\left|\left(\hat{R}_{A\rightarrow}A\right);\alpha\beta...\right\rangle \tag{2.8}$$

where the operator $\hat{R}_{A\rightarrow}$ is defined to rotate A into a specified orientation which is common to all the molecules of the dataset (Fig. 2.1).

This works under the assumption that it is always possible to define a unique (and therefore unambiguous) internal reference frame to rotate A into a specified orientation, which might be possible when the system involved has a particularly rigid internal structure. A more general strategy, which does not require any assumption on

the molecular geometry to be made, consists in considering the covariant integration over the operator $\hat{R}$ (Haar integration),

$$\left|\overline{A;\alpha\beta...}\right\rangle \equiv \int d\hat{R}\left[\hat{I}\otimes\hat{R}\otimes\hat{R}\otimes...\right]\left|\left(\hat{R}A\right);\alpha\beta...\right\rangle. \tag{2.9}$$

On the top of this definition, the requirement that a representation be covariant in $O(3)$, including the reflection symmetry of the tensor as in Eq. 2.3, means that improper rotations must be included, i.e., $\hat{S}=\hat{R}\times\{\hat{I},\hat{Q}\}$, with $\hat{Q}$ representing a reflection operator. This is done by a simple linear combination of the $SO(3)$ representation with its reflected counterpart with respect to any arbitrary mirror plane of the system; that is,

$$\left|\overline{A;\alpha\beta...}\right\rangle_{O(3)} = \left|\overline{A;\alpha\beta...}\right\rangle + \left[\hat{I}\otimes\hat{Q}\otimes\hat{Q}\otimes...\right]\left|\overline{\left(\hat{Q}A\right);\alpha\beta...}\right\rangle. \tag{2.10}$$

Upon this procedure, any other reflection operation is automatically included.

2.3 Covariant regression

Having shown how to build a symmetry-adapted representation of the system, let us see the implications of this procedure for linear regression. Using a symmetry-adapted representation in a linear regression model leads to the following solution for the regression weight,

$$|w\rangle = \sum_{n=1}^{N}\sum_{\alpha\beta...}\left(\hat{C}+\eta\hat{I}\right)^{-1}\left|\overline{A_n;\alpha\beta...}\right\rangle y_{\alpha\beta...}(n), \tag{2.11}$$

where the covariance is

$$\hat{C} = \sum_{n=1}^{N}\sum_{\alpha\beta...}\left|\overline{A_n;\alpha\beta...}\right\rangle\left\langle\overline{A_n;\alpha\beta...}\right|. \tag{2.12}$$

Note that the solution for the linear regression weight does not change when the training structures and corresponding tensors simultaneously undergo a symmetry operation that the representation has been adapted to. In other words, the same model results regardless of the arbitrary orientation of structures in the training set.

When moving to the dual, we find the kernel to be

$$k_{\alpha\beta...}^{\alpha'\beta'...}(A,B) = \left\langle\overline{B;\alpha'\beta'...}\middle|\overline{A;\alpha\beta...}\right\rangle. \tag{2.13}$$

From Eq. (2.9), this corresponds to considering the following Haar integration:

$$\int d\hat{R} \int d\hat{R}' \left\langle \hat{R}A \middle| \hat{R}'B \right\rangle (RR')_{\alpha\alpha'} (RR')_{\beta\beta'} \ldots \tag{2.14}$$

As stressed in the Introduction, performing the linear regression in the dual space should lead to a formally-equivalent model to that resulting from the primal formulation described above; yet, computing this kernel appears to be more complicated than Eq. (2.9) since it involves two integrations over rotations. If, however, we assume the core representation $|A\rangle$ undergoes a unitary transformation when the system is rotated, which is implied by the absence of an external field, the construction of a symmetry-adapted kernel reduces to performing a single integral over rotation:

$$k^{\alpha'\beta'\ldots}_{\alpha\beta\ldots}(A,B) = \int d\hat{R}\, \langle A|\, \hat{R}\, |B\rangle\, R_{\alpha\alpha'} R_{\beta\beta'} \ldots . \tag{2.15}$$

Note that upon defining a collective tensorial index $\{\alpha\beta\ldots\}$, a kernel matrix of size $3^r N \times 3^r N$ can be constructed by stacking together each of the $3^r \times 3^r$ vector-valued correlation functions. Then, a covariant tensorial prediction of the property of interest can be carried out according to the GPR prescription of Eq. 1.3. The symmetry-adapted kernel of Eq. (2.15) is just a generalization of the covariant kernels that have been introduced by De Vita and collaborators [31] in the context of learning the quantum atomic forces of a system.

It is instructive to compare the symmetry-adapted kernel definition of Eq. 2.15 to the kernel that one gets from the aligned descriptors of Eq. 2.8. In this case, building a kernel function on the top of this descriptor effectively means carrying out the structural comparison in a common reference frame where the two molecules are mutually aligned. One can then conveniently learn the tensor of interest component-by-component through a much simpler scalar regression framework. For the simple case of rank-1 tensors, for instance, we would get,

$$\boldsymbol{k}(A,B) = \langle B|\hat{R}_{B\to A}|A\rangle\, \boldsymbol{R}_{B\to A}, \tag{2.16}$$

where we have defined the best alignment operator as $\hat{R}_{B\to A} = \hat{R}_{B\to}\hat{R}^T_{A\to}$. This strategy has been successfully used in the learning of electronic multipoles of organic molecules [63] as well as for predicting optical response functions of water molecules in their liquid environments [33]. For the latter example, a representation of the best-alignment structural comparison is reported in Fig. 2.2.

This method for tensor learning has the clear drawback of relying on the definition of a rigid molecular geometry, for which an internal reference frame can be effectively used

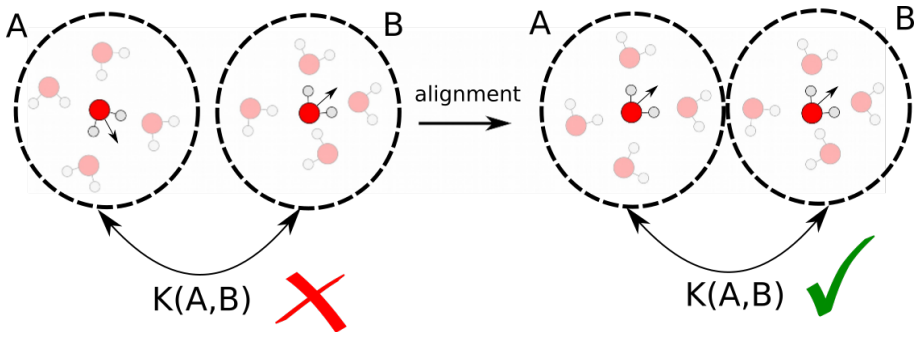


Figure 2.2 – Representation of the reciprocal alignment between water environments.

to perform the procedure of best alignment. Conversely, covariant kernel functions carries the great advantage to implicitly carry out both the structural comparison and the geometric alignment of two molecules simultaneously, neglecting any prior consideration about the internal structure of the molecule at hand.

2.4 Spherical representation

The family of symmetry-adapted descriptors previously introduced can be effectively used, in principle, to predict any Cartesian tensor of arbitrary rank. However, we should notice that having a tensor product for each additional Cartesian axis makes the cost of the regression scale unfavorably with the tensor rank, producing a global kernel matrix of dimension $(3^r)^2$. In fact, it is well established that a more natural representation of Cartesian tensors is given by their irreducible spherical components (ISCs) [64]. As described in Stone [64], the transformation matrix from Cartesian to spherical tensors can be found recursively, starting from the known transformation for rank-2 tensors.

Upon trivial manipulations, that might account for the non-symmetric nature of the tensor, each ISC transforms separately as spherical harmonics $|\lambda\mu\rangle$. Spherical harmonics form a complete basis set of the $SO(3)$ group. In particular, each λ-component of the tensor spans an orthogonal subspace of dimension $2\lambda+1$. For instance, the 9 components of a rank-2 tensor separate out into a term (proportional to the trace) that transforms like a scalar $|00\rangle$, three terms that transform like a vector $|1\mu\rangle$, and five terms that transform like $|2\mu\rangle$. When using a spherical representation, the kernel matrix is block diagonal, which greatly reduces the number of non-zero entries, and makes it possible to learn separately the different components. An additional advantage is that the possible symmetry of the tensor can be naturally incorporated by retaining only the spherical components λ that have the same parity as the tensor rank r. For instance, the $\lambda = 1$ component of a symmetric rank-2 tensor vanishes iden-

tically, meaning that only the 6 surviving elements of the tensor need to be considered when doing the regression. Especially for high rank tensors, this property means that the number of components can be cut down significantly.

In light of the discussion carried out for Cartesian tensors, it is straightforward to realize how a symmetry-adapted representation that transforms covariantly with spherical harmonics of order λ should look. Since each ISC is effectively a vector of dimension $2\lambda+1$, we can first write a primitive representation as

$$\left|A;\lambda\mu\right\rangle = |A\rangle \otimes \left|\lambda\mu\right\rangle \,, \tag{2.17}$$

where $\left|\lambda\mu\right\rangle$ is an angular momentum state of order λ, such that $\left\langle\hat{\boldsymbol{x}}\middle|\lambda\mu\right\rangle = Y^{\lambda}_{\mu}(\hat{\boldsymbol{x}})$. Its symmetry-adapted counterpart, which is covariant in $\hat{R}$, is

$$\left|\overline{A;\lambda\mu}\right\rangle = \int d\hat{R}\; \hat{R}\,|A\rangle \otimes \hat{R}\left|\lambda\mu\right\rangle \,. \tag{2.18}$$

Crucially, a tensorial kernel function built on the top of this representation would transform under rotations as the Wigner-D matrix of order λ, $D^{\lambda}_{\mu\mu'} = \langle\lambda\mu|\hat{R}|\lambda\mu'\rangle$:

$$k^{\lambda}_{\mu\mu'}(A,B) = \left\langle\overline{A;\lambda\mu}\middle|\overline{B;\lambda\mu'}\right\rangle = \int d\hat{R}\;\langle A|\,\hat{R}\,|B\rangle\, D^{\lambda}_{\mu\mu'}(\hat{R})\,, \tag{2.19}$$

a result that has been first introduced in Ref. [57] as the spherical tensor generalization of the covariant kernel prescription of Ref. [31]. Finally, since the parity of $\left|\lambda\mu\right\rangle$ with respect to the inversion operator $\hat{\imath}$ is determined by λ, a spherical tensor descriptor that is covariant in $O(3)$ can be obtained by considering

$$\left|\overline{A;\lambda\mu}\right\rangle_{O(3)} = \left|\overline{A;\lambda\mu}\right\rangle + (-1)^{\lambda}\;\left|\overline{\left(\hat{\imath}\,A\right)\lambda\mu}\right\rangle \,. \tag{2.20}$$

2.5 λ-SOAP representations

We now proceed to characterize the exact functional form of a symmetry-adapted representation of order λ which can be used to carry out a covariant prediction of any property that transforms as a spherical harmonic. By merging the general ideas previously discussed with the atom density construction already reported in the Sec. 1.5, we derive a family of λ-SOAP representations and kernels that recover the popular SOAP method of Bartók *et al.* [42] as the special $\lambda = 0$ limit.

2.5.1 2-body covariants

Let us start considering the definition of Eq. (2.18) for building an abstract spherical tensor representation of order λ. At the second order in structural correlations, replacing the abstract state $|A\rangle$ with the environmental state $\left|\rho_i\right\rangle$ in Eq. (2.18) reads

$$\left|\overline{\rho_i^{\otimes 1};\lambda\mu}\right\rangle = \frac{1}{8\pi^2}\int d\hat{R}\,\hat{R}\left|\rho_i\right\rangle \otimes \hat{R}\left|\lambda\mu\right\rangle . \tag{2.21}$$

When represented in real space, the previous representation carries a formal resemblance with the scalar 3-body representation of Eq. (1.19). This time, however, because one of the fields is given by $\left|\lambda\mu\right\rangle$, the effect of the rotational average is to single out the density expansion coefficients that transform covariantly with the spherical harmonics of order λ, i.e.,

$$\begin{aligned}\left\langle a\boldsymbol{x};\hat{\boldsymbol{x}}'\middle|\overline{\rho_i^{\otimes 1};\lambda\mu}\right\rangle &= \frac{1}{8\pi^2}\int d\hat{R}\ \langle a\boldsymbol{x}|\,\hat{R}\left|\rho_i\right\rangle\left\langle\hat{\boldsymbol{x}}'\middle|\,\hat{R}\left|\lambda\mu\right\rangle\right. \\ &= \frac{1}{4\pi}\left\langle ax\lambda\mu\middle|\rho_i\right\rangle P_\lambda(\hat{\boldsymbol{x}}\cdot\hat{\boldsymbol{x}}'),\end{aligned} \tag{2.22}$$

with P_λ the Legendre polynomial that expresses the angular correlation between the point $\boldsymbol{x}$ where the atom density is evaluated and the versor $\hat{\boldsymbol{x}}'$ that is used to evaluate the angular momentum ket $\left|\lambda\mu\right\rangle$. The covariance property of the real space representation is therefore entirely included in the λ spherical harmonics components of the local atom density, $\left\langle ax\lambda\mu\middle|\rho_i\right\rangle$, recovering the $\lambda = 0$ limit as $\left\langle ax00\middle|\rho_i\right\rangle$. Note that the covariance of the 2-body representation under inversion symmetry is naturally included in the definition of the coefficients $\left\langle ax\lambda\mu\middle|\rho_i\right\rangle$, so that the representation of Eq. (2.22) is automatically covariant within the $O(3)$ manifold.

2.5.2 3-body covariants

Extending the definition of Eq. (2.21), the inclusion of 3-body structural correlations within an abstract spherical tensor representation implies to deal with an additional tensor product with the environmental ket $\left|\rho_i\right\rangle$:

$$\left|\overline{\rho_i^{\otimes 2};\lambda\mu}\right\rangle = \frac{1}{8\pi^2}\int d\hat{R}\,\hat{R}\left|\rho_i\right\rangle \otimes \hat{R}\left|\rho_i\right\rangle \otimes \hat{R}\left|\lambda\mu\right\rangle . \tag{2.23}$$

Its real-space representation is illustrated in Fig. 2.3. In this case, the effect of the rotational average is more convoluted than the one previously reported, and it carries formal analogies with the scalar 4-body structural representation of Eq. (1.24). Following the same rationale outlined in Sec. 1.5.4, the correlation function in real space

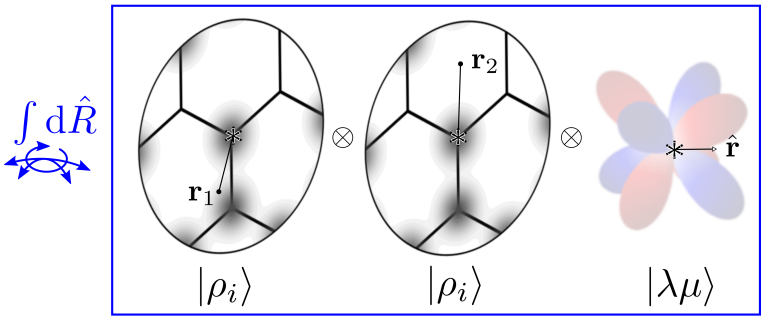


Figure 2.3 – Illustration of the real-space construction of 3-body density-correlations that are endowed with the rotational symmetry of spherical harmonics.

results as an expansion over rotational invariant functions of the kind $\langle\omega_1\omega_2\omega_3|l_1l_2\lambda\rangle$, with $(\omega_1,\omega_2,\omega_3)$ the triplet of angles that uniquely defines the reciprocal orientation of three versors $(\hat{\boldsymbol{x}}_1,\hat{\boldsymbol{x}}_2,\hat{\boldsymbol{x}}_3)$ on the unit sphere. In particular, we get

$$\begin{aligned}\left\langle a_1\boldsymbol{x}_1;a_2\boldsymbol{x}_2;\hat{\boldsymbol{x}}_3\middle|\overline{\rho_i^{\otimes 2};\lambda\mu}\right\rangle &= \frac{1}{8\pi^2}\int d\hat{R}\,\langle a_1\boldsymbol{x}_1|\,\hat{R}\,|\rho_i\rangle\,\langle a_2\boldsymbol{x}_2|\,\hat{R}\,|\rho_i\rangle\,\langle\hat{\boldsymbol{x}}_3|\,\hat{R}\,|\lambda\mu\rangle \\ &= \sum_{l_1l_2}\left\langle a_1x_1l_1;a_2x_2l_2\middle|\overline{\rho_i^{\otimes 2};\lambda\mu}\right\rangle\langle\omega_1\omega_2\omega_3|l_1l_2\lambda\rangle\,.\end{aligned} \tag{2.24}$$

As for the 2-body case, the 3-body expansion coefficients $\left\langle a_1x_1l_1;a_2x_2l_2\middle|\overline{\rho_i^{\otimes 2};\lambda\mu}\right\rangle$ play the role of expressing the covariance of the representation under rotations. Their exact definition comes from the combination of density angular momentum components that are compatible with the spherical tensor of order λ:

$$\left\langle a_1x_1l_1;a_2x_2l_2\middle|\overline{\rho_i^{\otimes 2};\lambda\mu}\right\rangle = \sum_{m_1m_2}\langle l_1m_1,l_2m_2|\lambda\mu\rangle\langle a_1x_1l_1m_1|\rho_i\rangle^\star\langle a_2x_2l_2m_2|\rho_i\rangle\,, \tag{2.25}$$

with $\langle l_1m_1,l_2m_2|\lambda\mu\rangle$ the Clebsch-Gordan (CG) coefficients that realize the combination of angular momenta. As in the quantum-theory of angular momentum, the pair of states $|l_1m_1\rangle$ and $|l_2m_2\rangle$ must satisfy the triangular relation $|l_1-l_2|\le\lambda\le|l_1+l_2|$. This makes clear that the $\lambda=0$ limit of Eq. (1.19) is immediately recovered by asking for the two angular momenta states to be the same, consistently with the properties of CG-coefficients $\langle l_1m_1,l_2m_2|00\rangle\sim\delta_{l_1l_2}\delta_{m_1m_2}$, i.e.,

$$\left\langle a_1x_1l_1;a_2x_2l_2\middle|\overline{\rho_i^{\otimes 2};00}\right\rangle = \delta_{l_1l_2}\sum_{m_1m_2}\delta_{m_1m_2}\langle a_1r_1l_1m_1|\rho_i\rangle^\star\langle a_2r_2l_2m_2|\rho_i\rangle\,. \tag{2.26}$$

Extending the discussion already carried out in Sec. 1.5.4, the application of an in-

version operator $\hat{\imath}$ to the atomic coordinates of the system implies that the covariant coefficients of Eq. (2.25) transform as

$$\left\langle a_1 x_1 l_1; a_2 x_2 l_2 \middle| \overline{\left(\hat{\imath}\rho_i\right)^{\otimes 2}; \lambda\mu} \right\rangle = (-1)^{l_1+l_1} \left\langle a_1 x_1 l_1; a_2 x_2 l_2 \middle| \overline{\rho_i^{\otimes 2}; \lambda\mu} \right\rangle . \tag{2.27}$$

As a result, representations that are $O(3)$-covariant with spherical tensors of order λ can be obtained by retaining only the components for which the combination $l_1 + l_2$ has the same parity as λ, i.e., for which $l_1 + l_2 + \lambda$ is even. Representations of this kind can straightforwardly be used to learn the ISCs of any symmetric Cartesian tensor, as well as of any other property that follows spherical harmonics transformations. The irreducible decomposition of asymmetric tensors, on the other hand, include ISCs that have an opposite parity under inversion symmetry, thus requiring to retain the only combinations of density angular momenta for which $l_1 + l_2 + \lambda$ is odd. This is the case, for instance, when asking to regress the response of the electronic energy to an applied magnetic field, as it comes from the definition of chemical shielding tensors that enter NMR solid-state spectroscopy [65]. Ultimately, this observation is consistent with the fact that the magnetic field behaves like a *pseudo-vector*, i.e., it preserves its direction under an inversion operation applied to the reference frame.

2.6 λ-SOAP kernels

The previous Section discusses the construction of representations that follow spherical harmonics transformations. The set of 3-body tensorial features reported in Eq. (2.25), in particular, underlie the definition of the λ-SOAP kernel first introduced in Ref. [57] starting from the prescription of Eq. (2.19). Considering the inner product of a pair of λ-SOAP representations, a λ-SOAP kernel can be computed as follows

$$\begin{aligned} k^{\lambda}_{\mu\mu'}(A_i, B_j) &= \left\langle \overline{\rho_i^{\otimes 2}(A); \lambda\mu} \middle| \overline{\rho_j^{\otimes 2}(B); \lambda\mu'} \right\rangle \\ &= \sum_{a_1 a_2 n_1 n_2 l_1 l_2} \left\langle \overline{\rho_i^{\otimes 2}(A); \lambda\mu} \middle| a_1 n_1 l_1; a_2 n_2 l_2 \right\rangle \left\langle a_1 n_1 l_1; a_2 n_2 l_2 \middle| \overline{\rho_j^{\otimes 2}(B); \lambda\mu'} \right\rangle , \end{aligned} \tag{2.28}$$

where we once again relied on the substitution $x \to n$ to discretize the radial degrees of freedom by expansion over an orthogonal basis (Appendix B).

When addressing the practical calculation of λ-SOAP representations, it is often more convenient to use real spherical harmonics [66], as this implies that the kernel of Eq. (2.28) is purely real. In fact, what one finds upon replacing $|\lambda\mu\rangle$ with a real spherical harmonic is that the components of Eq. (2.25) are either purely real or purely

imaginary, depending on whether the combination $l_1 + l_2 + \lambda$ is even or odd. For $\mu > 0$, for example, λ-SOAP coefficients computed using real spherical harmonics satisfy

$$\left\langle a_1 n_1 l_1; a_2 n_2 l_2 \middle| \overline{\rho_i^{\otimes 2}; \lambda\mu} \right\rangle^{\star} = (-1)^{l_1+l_2+\lambda} \left\langle a_1 n_1 l_1; a_2 n_2 l_2 \middle| \overline{\rho_i^{\otimes 2}; \lambda\mu} \right\rangle . \tag{2.29}$$

One can therefore discard all imaginary components to enforce the covariance of the representation under inversion.

2.6.1 Non-linearity

As already discussed in Sec 1.5.5, another crucial aspect to improve the regression performance is to incorporate non-linearities in the construction of the representation. For instance, tensor products of the scalar representation introduce higher body-order correlations, in a way that can be easily implemented in a kernel framework by raising the kernel to an integer power. When working with tensorial representations, however, one has to be careful to avoid breaking the covariant transformation properties of the feature vector. Taking products of $\left|\overline{\rho_i^{\otimes 2}; \lambda\mu}\right\rangle$ kets would require re-projecting the product onto the irreducible representations of the group, which would be as cumbersome as increasing the body-order ν. One obvious solution to this problem is to multiply the tensorial kernel of order λ by its scalar and rotationally invariant counterpart, which can then be raised to an integer power ζ without breaking the tensorial nature of the kernel. This procedure consists in considering

$$\boldsymbol{k}_{\zeta}^{\lambda}(A_i, B_j) = \boldsymbol{k}^{\lambda}(A_i, B_j) \left(k^0(A_i, B_j)\right)^{\zeta-1} , \tag{2.30}$$

which underlies a representation that is built as the following tensor product

$$\left|\overline{\rho_i^{\otimes 2}; \lambda\mu}\right\rangle \otimes \prod^{\zeta-1} \left|\overline{\rho_i^{\otimes 2}; 00}\right\rangle . \tag{2.31}$$

For $\zeta = 1$, one recovers the original tensorial kernel, while a non-linear behavior is introduced for $\zeta > 1$. A considerable improvement of the learning power is usually obtained when using $\zeta = 2$, while negligible further improvement is observed for $\zeta > 2$. These considerations also apply to the use of fully non-linear ML models like a neural network. To guarantee that the prediction of the model is consistent with the group covariances, the tensorial λ-SOAP features must enter the network at the last layer, and all the previous non-linear layers can only contribute to different linear combinations of the tensorial features. Similar ideas have already been implemented in the context of generalizing the construction of spherical convolutional neural networks [67].

3 Prediction of optical responses

This chapter provides a series of examples that demonstrate how symmetry-adapted representations and λ-SOAP kernels can be used for the prediction of optical response tensors of arbitrary rank, both for molecular and condensed-phase systems. An open-source implementation of the method can be found in the TENSOAP package [68]. Sections and figures are adapted from the following article:

A. Grisafi, D. M. Wilkins, G. Csányi and M. Ceriotti, "Symmetry-Adapted Machine Learning for Tensorial Properties of Atomistic Systems", *Physical Review Letters* 120, 036002 (2018). Copyright © 2018 by American Physical Society. All rights reserved. AG contributed to deriving and implementing the λ-SOAP method, collecting the results, running the reference calculations for the liquid water dataset, writing the manuscript and producing the figures for the examples reported.

3.1 Optical response series

The computational simulation of absorption and scattering spectra, such as infrared, Rayleigh, Raman, sum frequency generation (SFG) and second harmonic scattering (SHS), all require as ingredients the optical response tensors that modulate the susceptibility of the system to the electric field of the incoming radiation. The optical response series of a system is generally defined as the derivatives of the electronic energy U with an applied electric field $\boldsymbol{E}$:

$$T^{r}_{ijk\ldots} \equiv \frac{\partial^{r} U}{\partial E_i \partial E_j \partial E_k \ldots}, \tag{3.1}$$

with r the tensor rank and $ijk\ldots$ labeling the Cartesian components of the tensor. We aim to demonstrate that tensors as the ones of Eq. (3.1) can be efficiently regressed using the 3-body λ-SOAP kernel of Eq. (2.28). To show this, we consider the dipole

moment $\boldsymbol{\mu}$ ($r=1$), the polarizability $\boldsymbol{\alpha}$ ($r=2$) and first hyperpolarizability $\boldsymbol{\beta}$ ($r=3$) as prototypical examples of tensors of increasing rank.

3.2 Cartesian to spherical transformation

The simplest tensor is given by the dipole moment, whose Cartesian components can directly be mapped to the ones of $\lambda=1$ real spherical harmonics, i.e., $\{\mu_x,\mu_y,\mu_z\} = \{\mu_{-1},\mu_{+1},\mu_0\}$, and, as such, can straightforwardly be learned using a λ-SOAP representation of corresponding order. The interpolation of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is instead more involved, as it requires to first decompose them in the corresponding irreducibile spherical components (ISCs). Due to the symmetry with respect to permutations of Cartesian indices – which is implied by the definition of Eq. (3.1) – $\boldsymbol{\alpha}$ corresponds to an irreducible representation involving $\lambda=0$ and $\lambda=2$ spherical components only, while $\boldsymbol{\beta}$ is mapped to ISCs corresponding to $\lambda=1$ and $\lambda=3$. Note that, especially for $\boldsymbol{\beta}$, working in the spherical representation implies a massive simplification of the learning task, since the number of components to be regressed goes from 27 fully coupled Cartesian elements to just 10 spherical elements that are decoupled in 3 elements for $\lambda=1$ and 7 elements for $\lambda=3$. Explicit transformation matrices between symmetric Cartesian tensors and (complex) spherical tensors can be found in Ref. [64] both for r=2 and $r=3$. For $r=2$, for example, we have,

	$\lambda=0$			$\lambda=2$		
μ	0	-2	-1	0	$+1$	$+2$
xx (1)	$-1/\sqrt{3}$	$1/2$	0	$-1/\sqrt{6}$	0	$1/2$
xy (2)	0	$-i/2$	0	0	0	$i/2$
xz (2)	0	0	$1/2$	0	$-1/2$	0
yy (1)	$-1/\sqrt{3}$	$-1/2$	0	$-1/\sqrt{6}$	0	$-1/2$
yz (2)	0	0	$-i/2$	0	$-i/2$	0
zz (1)	$-1/\sqrt{3}$	0	0	$2/\sqrt{6}$	0	0

with the number in parenthesis representing the multiplicity M of the Cartesian component, i.e., the symmetry of the rank-2 matrix. To compute their inverses straightforwardly by taking the conjugate Hermitian, we chose the transformation matrices between the Cartesian and spherical tensor representations to be unitary. This is possible if the multiplicity M is taken into account by multiplying both the rows of the transformation matrix and the Cartesian tensor components by $\sqrt{M}$. Once the learning has been carried out, the predicted spherical tensors can be transformed back into the Cartesian representation and the resulting components are divided by $\sqrt{M}$ to obtain the predicted Cartesian tensor.

3.3 Definition of tensorial errors

In order to quantify the error when learning tensorial properties in a way that measures the capability of the method to capture their magnitude and geometric symmetries, we compute separately for each ISC the root mean square error (RMSE) in the prediction

$$\epsilon^\lambda = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \left\| \boldsymbol{T}^\lambda_{\text{pred}}(i) - \boldsymbol{T}^\lambda_{\text{ref}}(i) \right\|^2}, \tag{3.2}$$

as compared with the intrinsic variability of the property

$$\sigma^\lambda = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N} \left\| \boldsymbol{T}^\lambda(i)_{\text{ref}} - \langle \boldsymbol{T}^\lambda_{\text{ref}} \rangle \right\|^2}, \tag{3.3}$$

where N is the number of reference tensors used for testing the predictions and $\|\cdot\|$ indicates the Frobenius norm. This choice reflects the definition of the loss functions reported in Eqs. (1.9) and (1.11). Note that, for $\lambda > 0$, the tensor average $\langle \boldsymbol{T}^\lambda_{\text{ref}} \rangle$ is assumed to be statistically vanishing and we hence only compute it for $\lambda = 0$. In fact, only the scalar components of the tensor are expected to be normally distributed about an average value $\langle T^0_{\text{ref}} \rangle$. Importantly, when computed across the training data, this average can be used as a baseline value to facilitate the learning exercise by letting the regression to solely focus on the fluctuations of the property about the average. Therefore, while a kernel-based prediction for $\lambda > 0$ simply reads as

$$\boldsymbol{T}^\lambda_{\text{pred}}(A) = \sum_I \boldsymbol{k}^\lambda(A, A_I) \cdot \boldsymbol{x}^\lambda(A_I), \tag{3.4}$$

predictions for $\lambda = 0$ also require to add the scalar average of the tensor back:

$$T^0_{\text{pred}}(A) = \sum_I k^0(A, A_I) x^0(A_I) + \langle T^0_{\text{ref}} \rangle . \tag{3.5}$$

3.4 Water oligomers response series

As a first example, we consider a dataset made of 1000 flexible and arbitrarily oriented water molecules in vacuum, for which the optical response series ($\boldsymbol{\mu}$, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$) is computed from Eq. (3.1) using high-end quantum chemical methods. 3-body λ-SOAP kernels are computed centering the representation on the only oxygen atom, as the corresponding environment provides, for this simple system, a complete description of the molecular structure. Figure 3.1-a) shows the learning curves (the test error ϵ as a function of the

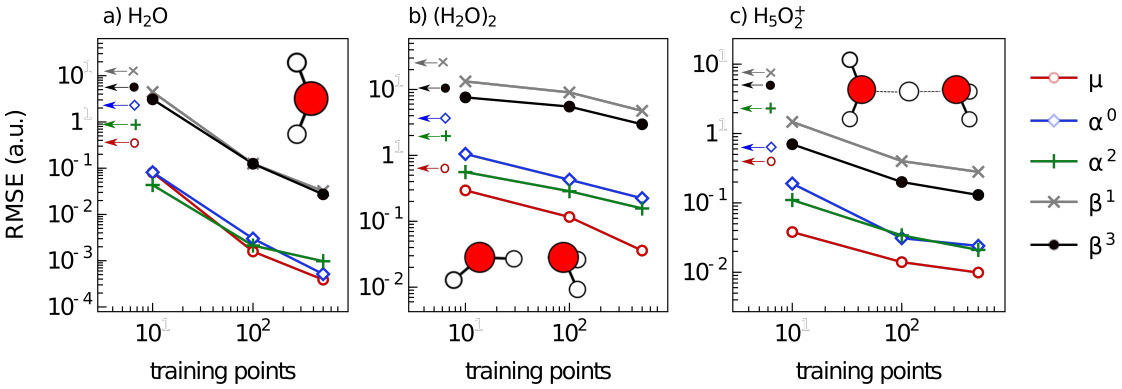


Figure 3.1 – Learning curves of the ISCs of dipole $\boldsymbol{\mu}$ ($\lambda = 1$), polarizability $\boldsymbol{\alpha}$ ($\lambda = 0, 2$) and hyperpolarizability $\boldsymbol{\beta}$ ($\lambda = 1, 3$) for water monomer (left), water dimer (center) and Zundel cation (right). For all cases the testing data set consists of 500 independent configurations. Arrows with symbols indicate the intrinsic standard deviation of the testing data set. λ-SOAP kernels have been computed with an environment cutoff of 4 Å for the monomer and $H_5O_2^+$, and 5 Å for the water dimer.

number of training structures) for all the ISCs of the optical series. Without explicitly using information on the orientation of water molecules, the λ-SOAP framework can easily achieve an error below 5% for all components using only 100 training points.

By following the same rationale discussed in Sec. 1.4.1, a natural approach to extend the λ-SOAP framework to complex molecules, and eventually to condensed phases, involves decomposing the global properties of the system into atom-centered terms. When working in the dual formulation, an atom-centered decomposition is equivalent to learning the system's properties using a single global kernel that is built as the sum of all possible local similarities between two configurations,

$$\boldsymbol{K}^{\lambda}(A,B) = \frac{1}{n_A n_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} \boldsymbol{k}^{\lambda}(A_i, B_j), \tag{3.6}$$

with A_i representing the i^{th} environment of the configuration A, while $\boldsymbol{k}^{\lambda}(A_i, B_j)$ is the tensorial kernel that compares the i^{th} local environment of the configuration A with the j^{th} local environment of the configuration B. Considering a water dimer as an example, we take the two oxygen atoms as centers of the representation (so that $n = 2$), and allow all of the surrounding atoms (H and O) to contribute to the smoothed atom density. From Eq. (3.6), a global tensor property of a dimer A is then predicted

as an average of individual monomer responses, e.g.,

$$\begin{aligned}\boldsymbol{\alpha}^{\lambda}(A) &= \sum_{B} \boldsymbol{K}^{\lambda}(A,B)\cdot\boldsymbol{x}^{\lambda}(B) = \sum_{B}\left[\frac{1}{2}\sum_{i\in O}\frac{1}{2}\sum_{j\in O}\boldsymbol{k}^{\lambda}(A_i,B_j)\right]\cdot\boldsymbol{x}^{\lambda}(B)\\ &= \frac{1}{2}\sum_{i\in O}\left[\sum_{B}\frac{1}{2}\sum_{j\in O}\boldsymbol{k}^{\lambda}(A_i,B_j)\cdot\boldsymbol{x}^{\lambda}(B)\right] = \frac{1}{2}\sum_{i\in O}\boldsymbol{\alpha}^{\lambda}(A_i)\,.\end{aligned} \tag{3.7}$$

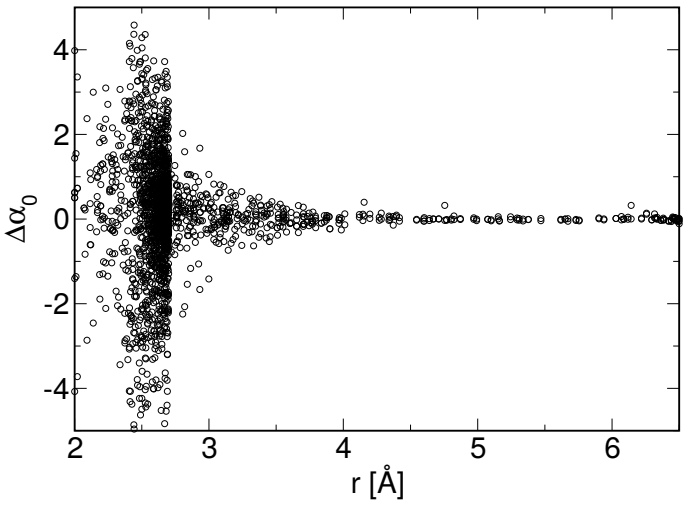


Figure 3.2 – Difference between the monomer contributions to the isotropic component of $\boldsymbol{\alpha}$ as obtained by λ-SOAP, and the value obtained by direct quantum chemical calculation of the individual monomer polarizabilities, as a function of the distance between the oxygen atoms.

With 500 training samples, both the isotropic and anisotropic components of the dimer polarizability can be learned with a RMSE below 10% of the intrinsic variance (Fig. 3.1-b)). As shown in the Fig. 3.2, when the two molecules are far apart the monomer polarizabilities predicted using Eq. (3.7) converge to the values computed separately for the two monomers. Thus, the discrepancy observed when the molecular separation is small can be seen as the two-body correction to the dielectric response function of individual monomers.

As the next step, we consider the case of the Zundel cation $H_2O_5^+$. Being both charged and chemically active, this system would be difficult to describe in terms of separate molecular contributions. Fig. 3.1-c) compares the learning curves for $\boldsymbol{\mu}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, obtained using a spherical cutoff of 4 Å around each oxygen atom. Note that although each environment encompasses the entire molecule, learning with atom-centered

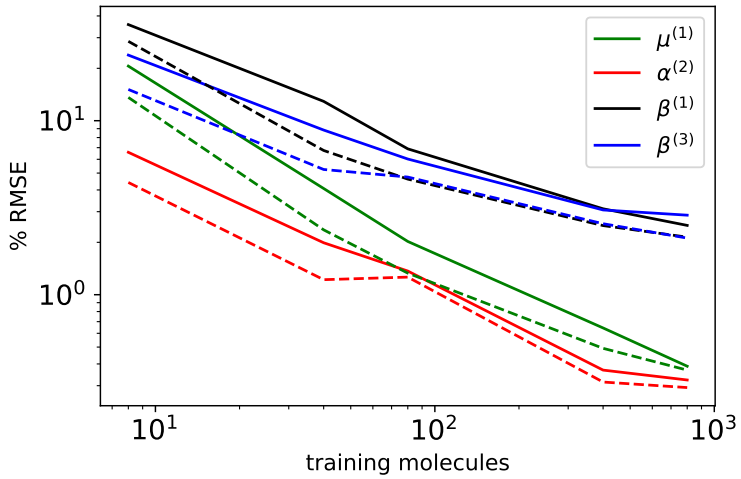


Figure 3.3 – Learning curves of the Zundel cation dielectric response series $\boldsymbol{\mu}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ as decomposed in their anisotropic ($\lambda > 0$) spherical tensor components. Full and dashed lines refer to predictions that are carried out with λ-SOAP kernel functions that are covariant in $SO(3)$ and $O(3)$ respectively.

environments implies enforcing the covariance condition at the level of O atoms, which better captures the physics of the problem. The errors for all components are well below 5% with 500 training samples, showing that λ-SOAP kernels are well suited to extend the method to systems which are intrinsically not separable into smaller molecular units.

All the previous results were obtained using λ-SOAP representations that are covariant within the $SO(3)$ manifold. In this regard, it is instructive to consider what happens if one introduces the covariance of the representation under inversion. The comparison between representations that are adapted in $SO(3)$ and $O(3)$ is reported in Fig. 3.3 taking the Zundel cation as an example. For all the anisotropic ISCs of the optical series, we observe a systematic improvement of the regression performance when endowing the kernel with inversion symmetry. This improvement is particularly pronounced for few training points, while it gets smaller for larger training set sizes, where the symmetry under inversion is eventually learned from data. This observation highlights the deep connection between symmetry-adapted structural representations and the efficiency of the machine-learning model in data-poor regimes.

3.5 Dielectric response of liquid water

In order to test the robustness and generality of the λ-SOAP approach, we consider the prediction of the dielectric response tensor $\boldsymbol{\varepsilon}$ of instantaneous configurations

of condensed phase water. Simulating the liquid bulk with 3D periodic boundary conditions, the electronic dielectric tensor can be computed as

$$\boldsymbol{\varepsilon} = \mathbf{1} + \frac{4\pi}{\Omega}\frac{d\boldsymbol{P}}{d\boldsymbol{E}}. \tag{3.8}$$

with Ω the cell volume and $\boldsymbol{P}$ the macroscopic polarization of the system across the cell. When making use of Eq. (3.8) to compute $\boldsymbol{\varepsilon}$ by finite differences, one should consider that the polarization $\boldsymbol{P}$ is ill-defined for periodic systems. In fact, according to the *modern theory of polarization* [69], the value of $\boldsymbol{P}$ in a periodic system is only defined up to a polarization quantum, so that only differences of polarization are physically meaningful. For this reason, we make use of the finite electric field method of Umari and Pasquareallo [70], as implemented in Quantum Espresso [71], which allows us to consistently estimate the derivative of Eq. (3.8) by computing the polarization of the system using the Berry phase approach [72]. Using this method, we compute $\boldsymbol{\varepsilon}$ for 1000 different snapshots of a 32-molecule path integral simulation [73] of room-temperature q-TIP4P/f water [74].

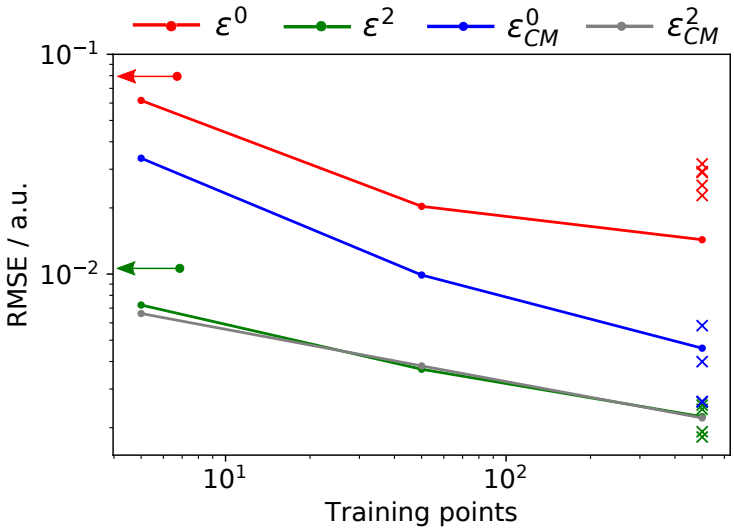


Figure 3.4 – Learning curves of the IST components of water dielectric response tensors $\boldsymbol{\varepsilon}$, through direct learning (red and green lines) and indirect learning going through the CM relation. The testing data set consists of 500 independent configurations. Arrows indicate the intrinsic standard deviation of the testing samples. Crosses show the predictions for 5 ice Ih structures using the ML model trained on liquid water.

Fig. 3.4 shows how an O-centered, $r_c = 4$ Å, λ-SOAP kernel allows us to learn directly both the isotropic and anisotropic components of $\boldsymbol{\varepsilon}$ with a RMSE well below 0.01 a.u. with just 500 training samples. Interestingly, the regression is much more effective if performed on the effective molecular polarizability as obtained from the Clausius-

Mossotti (CM) relation:

$$\boldsymbol{\alpha} = \frac{\Omega}{n}(\boldsymbol{\varepsilon} - \mathbf{1}) \cdot (\boldsymbol{\varepsilon} + \mathbf{2})^{-1} . \tag{3.9}$$

This underscores the importance of reducing the impact of non-local effects – which appear in the definition of $\boldsymbol{\varepsilon}$ through the volume and macroscopic field effects – when applying a machine-learning strategy that is based on an atom-centered decomposition. Indeed, a similar performance can be obtained by learning $\boldsymbol{\varepsilon}$ if r_c is increased to 5 Å, so that the macroscopic information is captured by the kernel (Fig. 3.5).

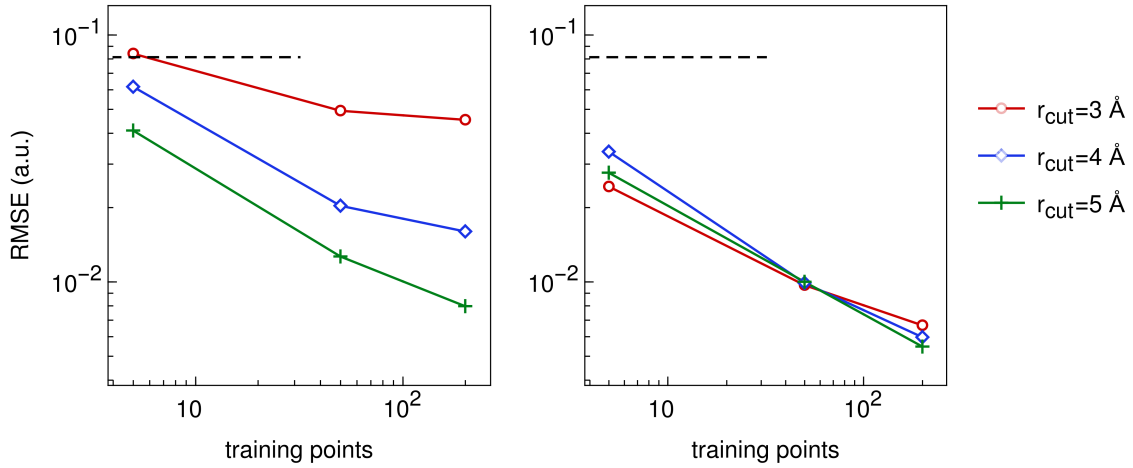


Figure 3.5 – Comparison between the learning of the isotropic component $\boldsymbol{\varepsilon}^0$ at three different environment cutoffs r_c. (*left*) direct learning. (*right*) indirect learning through CM molecular polarizability.

In Fig. 3.4, we also show the errors for predicting the dielectric constant of 5 proton-disordered configurations of ice Ih [75] using the model trained on liquid water. Direct predictions of $\boldsymbol{\varepsilon}$ are less accurate than what is seen for the liquid. When going through the local CM response, however, the accuracy becomes comparable, underscoring the transferability of the ML model, and the ease with which it can be applied to solids.

4 Prediction of accurate polarizabilities

Having shown the capability of λ-SOAP representations to learn tensorial properties in a symmetry-adapted fashion, we now test the accuracy of our predictions in a more challenging and useful scenario. In this chapter, we show how the λ-SOAP approach can be used for the regression of coupled-cluster-level polarizabilities across a very heterogeneous molecular dataset, yielding inexpensive tensorial predictions that present an accuracy comparable, if not larger, to the one of density functional theory. Sections and figures are adapted from the following article:

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", *Proceeding in the National Academy of Science* 116, 3401–3406 (2019). Copyright © 2019 National Academy of Sciences. AG contributed to re-implementing an improved version of the λ-SOAP method and to writing the manuscript.

4.1 Accurate polarizabilities from first principles

The dipole polarizability $\boldsymbol{\alpha}$ is a fundamental quantity of interest that underlies induction and dispersion interactions [76, 77], Raman and sum frequency generation (SFG) spectroscopy [78–81], and represents a key ingredient in the development of next-generation polarizable force fields [32, 82–84]. Beyond the toy examples reported in the previous Chapter, accurate and reliable polarizabilities can be quite difficult to compute [85]. This is primarily due to the fact that $\boldsymbol{\alpha}$ is a response property that is particularly sensitive to the quantum mechanical description of the underlying electronic structure. As such, non-trivial electron correlation effects and basis set incompleteness errors must be simultaneously accounted for. For these reasons, it is important to provide benchmark values for $\boldsymbol{\alpha}$ that go beyond the accuracy of relatively cheap quantum chemical methods such as density functional theory (DFT). In this regard, linear-response coupled-cluster theory [86–88] including single and double

excitations (LR-CCSD) has been shown to provide considerably more accurate and reliable predictions for the polarizability of a system when used in conjunction with a sufficiently large (diffuse) basis set [89–92]. However, such a prediction is accompanied by a substantially larger computational cost (scaling with the sixth power of the system size), which can become quite prohibitive even when treating molecules with as few as 10−15 atoms.

Machine-learning methods have already shown that an accuracy on par with (or even better than) DFT can be achieved in the prediction of many molecular properties [23, 93], and that DFT [94] or coupled-cluster [10] accuracy can be reached more easily when using a less accurate but more computationally efficient electronic structure method as a stepping stone. We aim to demonstrate that the application of a symmetry-adapted regression framework that makes use of λ-SOAP kernels can be used to predict $\boldsymbol{\alpha}$ with a similar level of accuracy. To do so, we present comprehensive, coupled-cluster level benchmarks for the polarizabilities of the $\sim 7,000$ small organic molecules contained in the QM7b database [95].

4.2 Electronic structure calculations

The QM7b database is made of 7,211 molecules (containing H, C, N, O, S, Cl atoms) and is based on a systematic enumeration of small organic compounds [38, 96]. Including a rich diversity of chemical groups, it represents a challenging test of the accuracy associated with DFT and quantum chemical methodologies. For this dataset, we computed DFT-based molecular polarizabilities by (numerical) differentiation of the molecular dipole moment $\boldsymbol{\mu}$, with respect to an external electric field $\mathbf{E}$, using the hybrid B3LYP [97]. Reference molecular polarizabilities were instead obtained using LR-CCSD [98]. To account for basis set incompleteness error, which can be even more important than higher-order electron correlation effects in an accurate and reliable determination of $\boldsymbol{\alpha}$ [90–92, 99], we employed the d-aug-cc-pVDZ basis set [100] for all calculations herein. To enable comparisons between molecules of different sizes, all error estimates are computed based on polarizabilites divided by the number of atoms n_i of each molecule. On the QM7b database, the popular B3LYP hybrid DFT functional predicts $\boldsymbol{\alpha}$ with a root mean square error (RMSE) of 0.404 a.u. with respect to the reference LR-CCSD values. These errors, which include both scalar and anisotropic contributions, are quite substantial and correspond to 18.3% of the intrinsic variability within the QM7b database, defined as the coupled-cluster standard deviation σ_{CCSD} of the full Cartesian tensor.

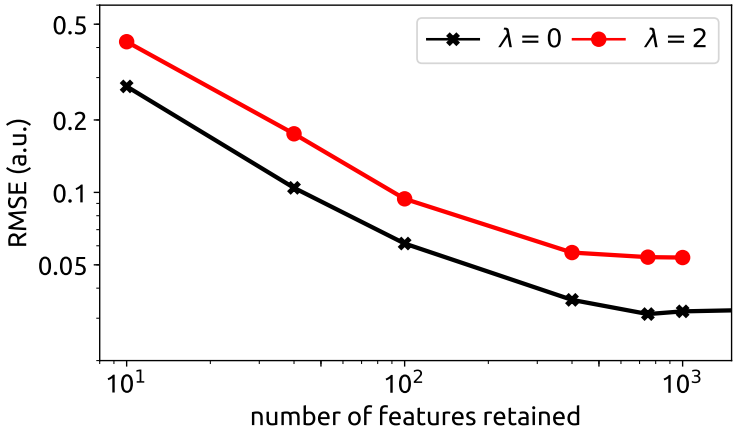


Figure 4.1 – Error in learning the $\lambda = 0$ and $\lambda = 2$ components of the per-atom polarizability for the QM7b dataset, with different percentages of the λ-SOAP features retained in calculating the kernels.

4.3 Learning model

3-body λ-SOAP representations were computed using an environment cutoff of r_c = 4Å, an angular cutoff of $l_{max} = 6$ and $n_{max} = 8$ radial functions for both the scalar ($\lambda = 0$) and tensorial ($\lambda = 2$) ISCs of $\boldsymbol{\alpha}$. Especially for $\lambda = 2$, this implies that the number of structural features, defined by the basis $\langle a_1 n_1 l_1; a_2 n_2 l_2|$ used to compute the λ-SOAP representation, is quite large, comprising several tens of thousands of components. To limit the feature-space size, we adopt the *farthest point sampling* (FPS) sorting algorithm, which allows us to retain the most diverse λ-SOAP features based on their reciprocal Euclidean distances [101]. Upon this procedure, we retain the most significant 400 features, amounting to ~2% of the 16,128 components in the original $\lambda = 0$ representation and ~0.7% of the 59,904 components in the original $\lambda = 2$ representation. As exemplified in Fig. 4.1, the rationale behind this choice is justified by the rapid drop of the prediction error with the number of features retained. In fact, even if only 10 components of the representation are kept, amounting to 0.06% for $\lambda = 0$ and 0.02% for $\lambda = 2$, then the error in predicting the polarizability is ~15%, which is comparable to the error incurred when using DFT to predict the CCSD polarizability. This massive reduction of the feature space size carries the obvious advantage of greatly speeding up the calculation of the λ-SOAP kernel when computing the inner product of Eq. (2.28). Finally, we adopt the prescription of Eq. (2.30) with $\zeta = 2$ to enhance the non-linear character of the tensorial kernels.

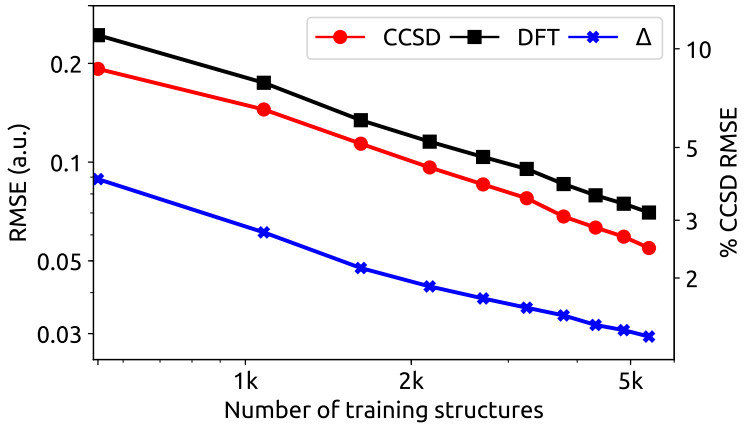


Figure 4.2 – Learning curves for the per-atom polarizabilities of the molecules in the QM7b database, calculated using either CCSD or DFT, as well as for the difference (Δ) between the two. The testing set consists of 1,811 molecules, and the right-hand axis shows the RMSE as a fraction of the intrinsic variability of the CCSD polarizability σ_{CCSD}.

4.4 Learning performance

The highly accurate reference CCSD calculations and the symmetry-adapted learning framework previously defined lay the foundation for a transferable model to predict molecular polarizabilities (AlphaML). We first test the regression performance by computing learning curves on both the DFT and CCSD polarizabilities. We used up to 5,400 structures for training, while predictions were tested on the remaining 1,811 structures. The structures were added to the training set starting from the most diverse configurations, according to the FPS algorithm. This procedure is representative of an efficient learning strategy that aims to obtain uniform accuracy with the minimum number of reference calculations [10]. We report ML errors in terms of the percentage of the intrinsic variability of the CCSD dataset ($\sigma_{\mathrm{CCSD}} = 2.216$ a.u. per atom), so as to provide a direct measure of the learning performance. As illustrated by the learning curves in Fig. 4.2, using up to 75% of the QM7b database for training yields a 2.5% RMSE with respect to σ_{CCSD} in predicting CCSD polarizabilities. To get a clearer idea of the accuracy associated with these ML-based predictions, one can compare these values against hybrid DFT. Using the same metric, the intrinsic error of DFT is 18% of σ_{CCSD} in the prediction of CCSD polarizabilities. This demonstrates that a ML-based model based on λ-SOAP kernels can yield polarizabilities with an accuracy that is approximately one order of magnitude greater than DFT. At the same time, the corresponding DFT polarizabilities can be learned with an error of 3.2% of σ_{CCSD}. As seen in other cases [10, 94], highly accurate quantum chemistry calculations are

smoother and slightly easier to learn than more approximate methods like DFT.

The AlphaML model can also be trained to evaluate the correction between different levels of theory, a correction commonly referred to as Δ-learning that is often found to result in much smaller error than learning the raw quantity itself [10, 94]. For instance, the use of DFT as a baseline to learn CCSD polarizabilities reduces the error by an additional factor of 2 relative to the direct learning of $\boldsymbol{\alpha}_{\mathrm{CCSD}}$ (see Fig. 4.2). Δ-learning therefore provides a way to further reduce the prediction error at the cost of performing a baseline DFT calculation.

4.5 Extrapolation to larger molecules

As already exemplified in Eq. (3.7), our definition of the kernel as an average of environmental kernels means that the polarizabilities predicted by AlphaML are given as a sum of predicted polarizabilities for each environment [10]. This feature allows one to predict $\boldsymbol{\alpha}$ for larger molecules. To test the behavior of the AlphaML in this extrapolative regime, we trained this model on the entire QM7b database, and then predicted the polarizabilities in a showcase dataset of 52 large molecules, which includes amino acids, nucleobases, drug molecules, carbohydrates, and 23 isomers of C_8H_n, as shown in the Figure below:

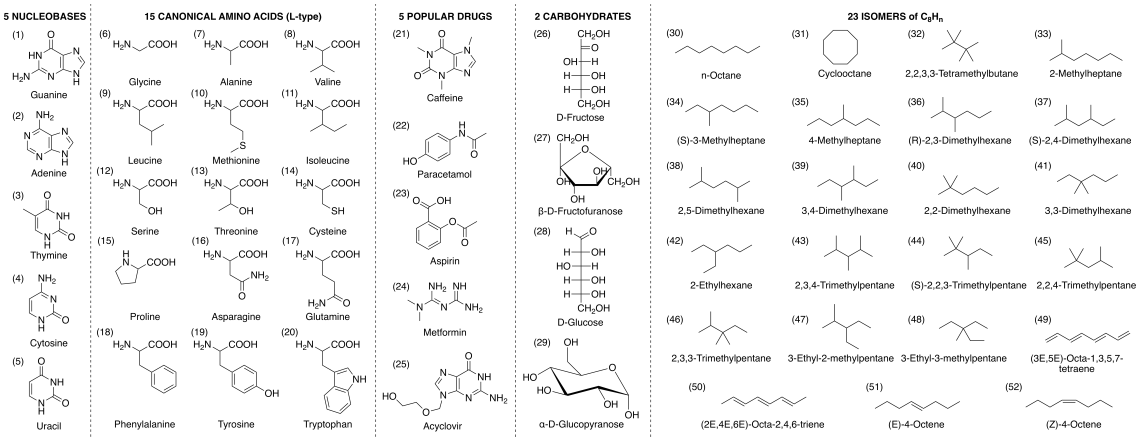


Figure 4.3 – Names and chemical structures of the 52 molecules included in the showcase dataset. The numbers refer to the position of each molecule in the dataset and are used for reference in the text and other figures.

Method	RMSE	RMSE($\lambda = 0$)	RMSE($\lambda = 2$)
CCSD/DFT	0.573	0.348	0.456
CCSD/ML	0.244	0.120	0.212
DFT/ML	0.302	0.143	0.266
Δ(CCSD-DFT)/ML	0.181	0.083	0.161

Table 4.1 – RMSE in the prediction of the per-atom polarizabilities of the 52 showcase molecules. CCSD/DFT denotes the discrepancy between CCSD and DFT values, while CCSD/ML and DFT/ML give the errors in predicting CCSD and DFT polarizabilities using AlphaML. Δ(CCSD-DFT)/ML gives the error in predicting the differences between the CCSD and DFT polarizabilities. All ML predictions are based on training on the full QM7b database. The total RMSE is expressed in a.u. per atom and broken down into the errors associated with the scalar ($\lambda = 0$) and tensorial ($\lambda = 2$) components of $\boldsymbol{\alpha}$.

In Table 4.1, we show the RMSE errors in predicting $\boldsymbol{\alpha}$ for the showcase molecules using AlphaML, as well as the error made when using DFT to approximate CCSD. Table 4.1 also breaks down the error into the $\lambda = 0$ and $\lambda = 2$ components of $\boldsymbol{\alpha}$; with an error in the anisotropic response comparable to that in the trace, this demonstrates that AlphaML learns both components with similar efficiency. As seen in the previous section, we again note that using the AlphaML model to predict CCSD polarizabilities is more accurate than simply using DFT. However, the use of DFT as the baseline in the Δ-learning sense leads to a further reduction of $\sim 20-30\%$ in the error. While AlphaML predicts CCSD polarizabilities of the showcase molecules with better-than-DFT accuracy, we observe a substantial decrease in accuracy, which is to be expected when the model is extrapolated to the larger molecules in the showcase dataset.

We can investigate the performance of AlphaML in more detail by analyzing the errors of individual molecules in the showcase dataset. Fig. 4.4 shows that the errors are actually very small for most molecules. Large errors occur predominantly for highly-polarizable compounds, particularly those that show a large degree of conjugation, such as long-chain alkenes and the purine nucleobases. For these systems, the underlying electronic structure is characterized by a high degree of delocalization, which requires larger cutoffs and more complex reference molecules to ensure accurate predictions. The ML predictions for the tensorial component of the polarizability, $\boldsymbol{\alpha}^{(2)}$, tend to be slightly less accurate than the DFT reference, except for the highly-polarizable alkenes, for which AlphaML dramatically outperforms DFT. Sulfur-containing structures, which are poorly represented in QM7b, also exhibit comparatively large errors.

The large discrepancy between DFT, CCSD, and AlphaML observed for alkenes (like octatetraene) reflects the non-local and collective nature of the underlying physics in

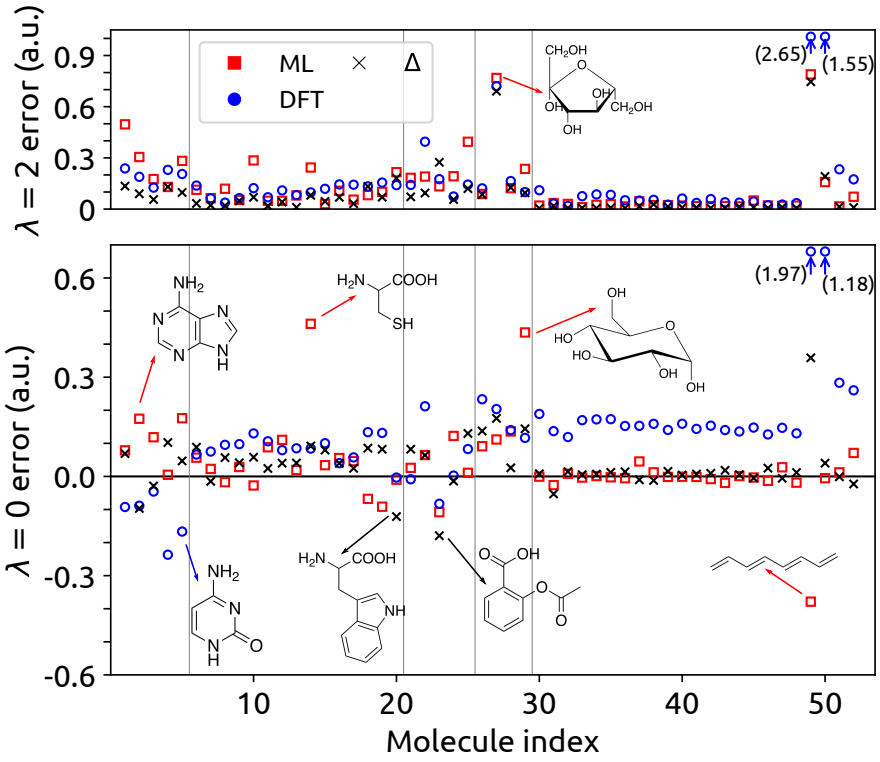


Figure 4.4 – RMSE made in approximating the $\lambda = 0$ (bottom panel) and $\lambda = 2$ (top panel) components of the per-atom polarizability in the showcase dataset. The x-axis corresponds to the numerical indices provided in the showcase molecule key in the SI, and the vertical lines show the partitioning of the dataset into the different groups outlined in the same figure. Red squares show the ML error, blue circles the error made in using DFT to approximate CCSD, and black crosses the error made when Δ-learning the CCSD correction with respect to DFT.

these systems, as well as the inherent local structure of the AlphaML model. For DFT and CCSD, the narrowing HOMO-LUMO gaps in conjugated hydrocarbons leads to near-metallic states which are known to exhibit strong multi-reference character [105]. As such, these systems represent a significant challenge for electronic structure methods (like DFT and CCSD) that are not explicitly based on a multi-reference wavefunction. In practice, this leads to divergent polarizabilities [106, 107], and methods like CCSD are no longer reliable as the source of reference quantum chemical data for machine learning. A machine-learning framework like AlphaML, which relies on local atomic environments to represent structures, tacitly disregards any collective (non-local) behavior that extends beyond the range of the local domains and the size of the molecules included in the training set.

As shown in Fig. 4.5, the per-carbon polarizabilities predicted by AlphaML there-

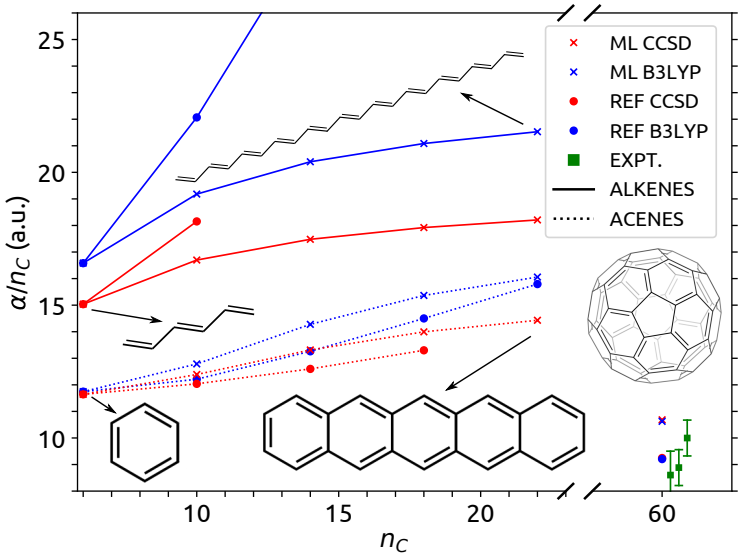


Figure 4.5 – Polarizability per carbon atom (α/n_C) vs. number of carbons (n_C) for the series of s-*trans* alkenes (from C_6H_8 to $C_{22}H_{24}$) and acenes (from benzene to pentacene), as well as fullerene (C_{60}). The reference CCSD results for anthracene and tetracene were taken from Ref. [102], and that for C_{60} from Ref. [103]. The green squares (and error bars) indicate the experimental measurements for C_{60} [104]. Results are provided from DFT and CCSD calculations, as well as the corresponding AlphaML models.

fore saturate to a constant value for the s-*trans* alkenes and acenes that are larger than those included in the QM7b dataset, i.e., hexatriene and benzene, respectively. Although this is a limitation when trying to learn collective and non-local physics, the local structure of AlphaML is also instrumental for obtaining the accurate and transferable predictions that we demonstrated on the showcase dataset. Even when it comes to challenging, conjugated systems with a vanishing HOMO-LUMO gap, the predictions of AlphaML are stable and completely avoid the unphysical and divergent predictions of more costly (but far from reference) quantum mechanical methods like DFT and CCSD. For molecules with a sizable gap (like C_{60}), the non-locality is less pathological and AlphaML performs remarkably well. For this prototypical nanotechnological system, machine-learning predictions are within 10% of DFT and CCSD results, and within the range of experimental values, despite the extrapolation to a system size that one order of magnitude larger than the molecules in the training set.

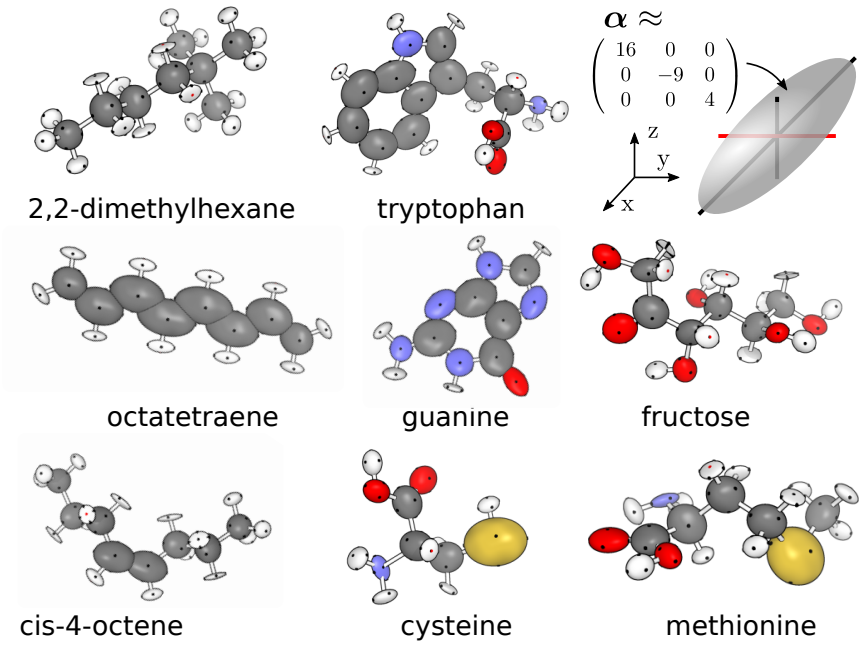


Figure 4.6 – Predicted atomic contributions to the total CCSD polarizability tensor for a selection of showcase molecules. The ellipsoids are aligned along the principal axes of $\boldsymbol{\alpha}_i$, and their extent is proportional to the square root of the corresponding eigenvalue. The principal axes are shown, and are colored based on whether the corresponding eigenvalues are positive (black) or negative (red). See the figure key above (which is not drawn to scale) for additional details.

4.6 Atomic polarizabilities

The atom-centered structure of AlphaML provides a natural additive decomposition of $\boldsymbol{\alpha}$ into a sum of local atomic terms, $\boldsymbol{\alpha}_i$, which can be used to better understand how different functional groups contribute to the molecular polarizability. Unlike other methods for decomposing the polarizability, such as an atoms-in-molecules scheme [108] or a self-consistent decomposition [109], the approach used in this section does not require any additional calculations on top of the molecular polarizability, as the atom-centered polarizabilities are obtained as a byproduct of the local nature of the λ-SOAP scheme. When interpreting the $\boldsymbol{\alpha}_i$, one should keep in mind that each term corresponds to the contribution from the *entire* atom-centered environment, and the way that the polarizability is split between neighboring atoms is entirely inductive, reflecting the interplay between data, structure (as represented by the kernels), and regression, rather than explicit physicochemical considerations. For instance, a few atoms within the showcase dataset (in particular several H environments) have $\boldsymbol{\alpha}_i$ with negative eigenvalues, which reflects the fact that they reduce the dielectric

response of the functional group to which they belong.

With this in mind, one can recognize physically-meaningful features in the magnitude and anisotropy of the $\boldsymbol{\alpha}_i$. Fig. 4.6 depicts eight representative examples. Comparing saturated and unsaturated hydrocarbons, e.g., 2,3-dimethylhexane, cis-4-octene and octatetraene, one sees that AlphaML predicts the contribution from the unsaturated carbon atoms to be large and very anisotropic, which is consistent with the higher degree of electron delocalization along conjugated molecules. Similarly large and anisotropic contributions are associated with aromatic systems, as seen in guanine and the indole ring of tryptophan. Oxygen atoms are associated with a very anisotropic $\boldsymbol{\alpha}_i$; a large fraction of the polarizability of OH and COOH groups is assigned to the environments centered around nearby H and C atoms, but O atoms systematically contribute a further anisotropic term, oriented perpendicularly to the highly-polarizable lone pairs (see for instance fructose as well as the carboxyl group in the amino acids). The sulfur-centered environments in cysteine and methionine have the largest contribution to the total polarizability in the showcase set, and exhibit a strongly anisotropic response. All of these examples suggest that AlphaML can utilize relatively local structural information to determine an atom-centered decomposition of $\boldsymbol{\alpha}$ that encodes non-trivial quantum mechanical contributions from each functional group (or moiety) contained within a given molecule. It is this ability to predict such an environment-dependent decomposition of $\boldsymbol{\alpha}$ that underlies the observed better-than-DFT performance of AlphaML when faced with the often insurmountable challenge of transferability to a sector of chemical compound space which contains molecules that are quite distinct and notably larger than those included in the training set.

5 Prediction of Raman spectra

One of the major applications of a machine-learning model that can yield accurate polarizabilities of a system is the computational simulation of Raman spectra. This chapter addresses the calculation of the Raman spectra of paracetamol in its molecular and crystal forms, as obtained from the predicted polarizability time series of a simulated molecular dynamics trajectory. In doing so, we also adopt an uncertainty estimation procedure that allows us to propagate the error made on the polarizabilities to the predicted Raman intensities. Sections and figures are adapted from the following article:

N. Raimbault, A. Grisafi, M. Ceriotti and M. Rossi, "Using Gaussian process regression to simulate the vibrational Raman spectra of molecular crystals", *New Journal of Physics* 21, 105001 (2019). Copyright © 2019 Institute of Physics. AG contributed to performing the polarizability predictions using the λ-SOAP method and to writing the manuscript.

5.1 Simulation of vibrational Raman spectra

Vibrational Raman spectra are widely used to monitor phase transitions, as well as for the identification of global and local structural patterns [110–112]. These kind of spectra represent the perfect example of a physical observable that requires the knowledge of the response of the system to electric field perturbations. In particular, any technique used to simulate this property requires the calculation of several instances of the polarizability tensor (in molecules) or the dielectric susceptibility (in crystals). For simplicity, we will refer to the polarizability tensor $\boldsymbol{\alpha}$ all throughout, but one should keep in mind that for solids the quantity of interest is rather the dielectric tensor of the system. As discussed in Ref. [59] and others [113, 114], the vibrational Raman spectrum can be calculated using several approximations, the simplest of which is the harmonic approximation. Here, we focus on a linear-response time-correlation

formalism, which is suited to take into account the anharmonicity of the potential energy surface. In this case, the Raman intensity can be obtained from the Fourier transform of the static polarizability autocorrelation function at thermodynamic equilibrium [115]. In particular, the so-called *powder spectrum intensity* is given by a combination of isotropic and anisotropic contributions as $I(\omega) = I_{\text{iso}}(\omega) + \frac{7}{3} I_{\text{aniso}}(\omega)$, with

$$\begin{aligned} I_{\text{iso}}(\omega) &= \frac{n}{2\pi} \int_{-\infty}^{+\infty} dt e^{-i\omega t} \langle \bar{\alpha}(0) \bar{\alpha}(t) \rangle \\ I_{\text{aniso}}(\omega) &= \frac{n}{2\pi} \int_{-\infty}^{+\infty} dt e^{-i\omega t} \frac{1}{10} \langle \text{Tr}\left[\tilde{\boldsymbol{\alpha}}(0) \cdot \tilde{\boldsymbol{\alpha}}(t) \right] \rangle \,, \end{aligned} \tag{5.1}$$

where n is the number of atoms in the system, the brackets $\langle \cdot \rangle$ denote an ensemble average and Tr is the trace. $\bar{\alpha}$ and $\tilde{\boldsymbol{\alpha}}$ are the Cartesian isotropic and anisotropic parts of the polarizability tensor, defined as $\bar{\alpha} = (\alpha_{xx} + \alpha_{yy} + \alpha_{zz})/3$ and $\tilde{\boldsymbol{\alpha}} = \boldsymbol{\alpha} - \bar{\alpha}\mathbf{1}$, respectively.

Computing anharmonic vibrational Raman spectra as in Eq. (5.1) can be a powerful tool to identify structural fingerprints in molecular crystals [116, 117]. Within this formalism, it is necessary to calculate *ab initio* molecular dynamics trajectories and compute $\boldsymbol{\alpha}$ for subsequent atomic configuration, employing, for instance, *density-functional perturbation theory* (DFPT) [118–121]. These calculations are computationally demanding, not only because of the tens of thousands of force evaluations that need to be performed to provide sufficient statistical sampling, but also because each DFPT calculation is typically four times more expensive than a force evaluation [117]. Furthermore, while there are several empirical potentials available that can be used to simulate the dynamics of molecular crystals [122], empirical models of the $\boldsymbol{\alpha}$ are rare and often poorly transferable [123]. The possibility of exploiting a machine-learning model that is able to inexpensively predict accurate polarizability tensors at each step of the molecular dynamics simulation is therefore particularly attractive. A similar strategy have been adopted, for instance, in the context of computing the Raman intensity of liquid water by means of a DNN architecture that is suitable to predict the polarizability of the system in terms of effective molecular contributions [124].

5.2 Dataset generation and model definition

In this study, we do not address the problem of obtaining forces; instead, we only predict the Raman intensities associated with precomputed *ab initio* trajectories. We test our method on a single paracetamol molecule in vacuum, as well as on the first and second crystal polymorphs of paracetamol, as represented in Fig. 5.1. The *ab initio* calculations were performed using the FHI-aims package [125] with light basis

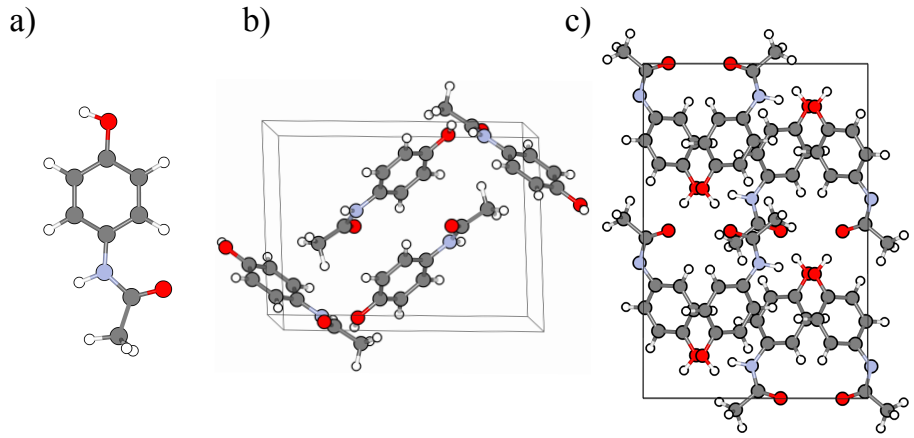


Figure 5.1 – The systems considered in this work: (a) Isolated Paracetamol molecule. (b) Paracetamol crystal form I (monoclinic). (c) Paracetamol crystal form II (orthorhombic).

set settings for all atomic species. AIMD trajectories were obtained using the PBE functional with many-body dispersion corrections [126, 127], employing a time step of 0.5 fs. The polarizability tensors were instead obtained every 1 fs, extending the DFPT calculations already carried out in Refs. [59, 116]. For each system, we ran 20 picoseconds of simulation in the NVT ensemble at 300 K, which is mainly used to train and validate the machine-learning model, and 15 picoseconds in the NVE ensemble, which is instead used to test the accuracy of the Raman spectra predictions. The construction of a symmetry-adapted GPR (SA-GPR) model that is suited to learn $\boldsymbol{\alpha}$ in a covariant fashion mirrors the discussion already carried out in Chapter 4. In particular, 3-body λ-SOAP kernels with $\zeta = 2$ and a cutoff of $r_c = 4$ Å are used to learn the ISCs of $\boldsymbol{\alpha}$ in terms of a $\lambda = 0$ component, proportional to the trace $\bar{\alpha}$, and a $\lambda = 2$ component that is linearly related to the anistropic Cartesian tensor $\tilde{\boldsymbol{\alpha}}$.

5.3 Covariant vs. component-wise regression

Taking the paracetamol molecule as an example, it is instructive to compare the performance of SA-GPR against a standard GPR model that measures the structural similarity between configurations by means of the best-alignment prescription already discussed in Chapter 2. In the latter case, the reciprocal alignment of the molecules is performed adopting the Kabsch algorithm [128], which is known to work particularly well for relatively rigid molecules. Upon this procedure, we place each molecule in a box of $6 \times 4 \times 2.5$ Å and build a smooth representation of the atomic structure as a species-dependent density field $\rho_a(\boldsymbol{x})$ sampled on a uniform three-dimensional grid of spacing $\delta = 0.5$ Å. Similarly to the SOAP construction, the density field comes from

the sum of Gaussian functions centered on the atomic positions:

$$\rho_a(\boldsymbol{x}) = \sum_{i \in a} \exp\left(-\frac{|\boldsymbol{x} - \boldsymbol{r}_i|^2}{2\sigma^2}\right), \tag{5.2}$$

with the Gaussian width chosen as $\sigma = 0.5$ Å and a labels the chemical species. A 2D slice of $\rho_a(\boldsymbol{x})$ is represented in Fig. 5.2. For each structure A, the actual feature

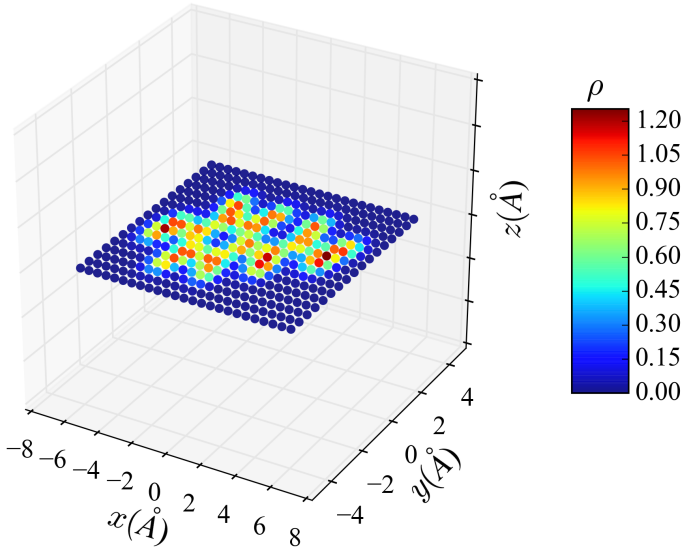


Figure 5.2 – 2D view of the smooth density field of a paracetamol molecule. Blue (red) points indicate a low (high) density.

vector is constructed from the point-by-point concatenation of the species-dependent densities of Eq. (5.2), i.e., $\boldsymbol{u}(A) \equiv \{\rho_a(\boldsymbol{x}; A)\}$. Finally, the structural similarity between any pair of molecules A and B is measured using a Gaussian kernel analogous to the one already introduced in Eq. (1.7), i.e.,

$$k(A,B) = \exp\left(-\frac{\|\boldsymbol{u}(A) - \boldsymbol{u}(B)\|^2}{2d^2}\right), \tag{5.3}$$

where the Euclidean distance $\|\boldsymbol{u}(A) - \boldsymbol{u}(B)\|$ is computed through the point-by-point difference of the feature vectors on the 3D-grid. The adimensional hyperparameter that modulates the kernel similarity is optimized as $d = 10$.

Upon the best-alignment procedure, the kernel of Eq. (5.3) can be used to learn the polarizability tensor $\boldsymbol{\alpha}$ component by component. The learning performance is tested on a FPS sub-selection of the NVT trajectory that comprises 2000 molecular configurations. In particular, 500 randomly selected molecules, out of the total of 2000,

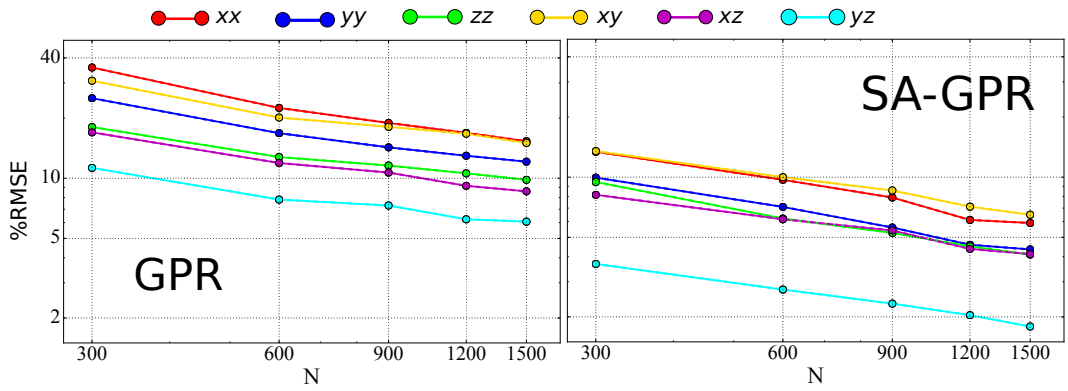


Figure 5.3 – % RMSE associated with the prediction of each of the six distinct components of the molecular polarizability tensor as a function of the number training configurations N. (*left*) Component-wise GPR results using a Gaussian kernel. (*right*) SA-GPR results with λ-SOAP kernels.

are retained for testing the accuracy of the $\boldsymbol{\alpha}$-predictions, while the other 1500 are used for training. Fig. 5.3 compares the learning performance of the component-wise regression model against the SA-GPR framework. The learning of all the Cartesian polarizability components follow a similar slope, but they are predicted with different accuracy due to the strong anisotropy of the dielectric response of the paracetamol molecule at the given orientation. Because of the π-conjugation of the system in the molecular plane, in particular, the system appears much more polarizable along the x-axis rather than along other directions, making it harder to regress the corresponding variations across the dataset. The α_{xx} component presents the largest error, going down to 17% of the intrinsic variation with 1500 training points. The best learning performance is instead obtained for the α_{yz} component, where the prediction error can be brought down to about 6% RMSE. When using the SA-GPR model, the possibility to learn the irreducible spherical components of $\boldsymbol{\alpha}$ in a covariant fashion yields predictions that are systematically more accurate than the GPR ones. In fact, all the Cartesian components show an accuracy that is more than doubled at any training set size, with an error on the α_{xx} and α_{yz} components of 6% and 2% RMSE respectively. These results underscore the importance of adopting an atom-centered symmetry-adapted approach even when the system presents enough structural rigidity to enable the practical application of a global kernel similarity measure. From here on, we will hence only report results associated with the SA-GPR model.

5.4 Uncertainty estimation and error propagation

Beyond a direct comparison with the quantum-mechanical reference calculations, it is important to dispose of a method to estimate the learning uncertainty associated with the predicted polarizability components. In fact, one would like to propagate the expected error that incurs in the prediction of $\boldsymbol{\alpha}$ to the actual Raman intensity, in order to obtain a quantitative measure of the reliability of the predicted spectra. In the particular case of GPR, the uncertainty estimate can be computed *a priori* from the GPR intrinsic variance of Eq. (1.4). This strategy is however not very practical because of its computational expense, so that other kind of methods such as *bootstrapping* or *subsampling* can rather be used to estimate the prediction errors [129]. Moreover, propagating the error to the Raman spectrum would be difficult to carry out on top of the GPR intrinsic variance.

In this work, N_{RS} subselections of the training dataset are considered to generate an ensemble of predictions for the polarizability. From these, N_{RS} Raman spectra are computed by Fourier transforming the time series of each model in the ensemble. Finally, the average and the standard deviation of the predicted spectra over the N_{RS} subselections would give the final Raman spectrum prediction and the propagated estimated error respectively. The downside of this approach is that this model works under the assumption that the training data correspond to independently distributed samples. This is of course not true in general, so that one needs to correct the model to take into account for the underlying correlations. Following Ref. [129], a maximum likelihood recipe can be adopted to linearly scale the variance of the predictions by a constant factor ν^2. The calibration of this scaling factor is carried out by computing the actual prediction errors of the polarizabilites over a suitably selected validation set N_{val}, for which the reference polarizabilities are known, and then considering

$$\nu^2 = \frac{1}{N_{\mathrm{val}}} \sum_{j=1}^{N_{\mathrm{val}}} \frac{\left\| \boldsymbol{\alpha}_{\mathrm{pred}}(j) - \boldsymbol{\alpha}_{\mathrm{ref}}(j) \right\|^2}{\sigma^2(j)}, \tag{5.4}$$

where $\sigma^2(j)$ are the variances of the predicted polarizabilities. Once the value of ν has been determined, each polarizability prediction of a given training model k can be updated as follows:

$$\boldsymbol{\alpha}'_k = \bar{\boldsymbol{\alpha}} + \nu(\boldsymbol{\alpha}_k - \bar{\boldsymbol{\alpha}}), \tag{5.5}$$

where $\bar{\boldsymbol{\alpha}}$ is the predicted polarizability averaged over the N_{RS} models. This scaling procedure guarantees that the variance of the models is consistent with the outcome of the likelihood maximization. By computing the Raman spectrum for each scaled

model k, the propagated uncertainty estimation associated with the spectra will automatically take into account the calibration of the variance.

5.5 Raman spectrum of a paracetamol molecule

By making use of the subsampling strategy previously described, we first test the quality of our predictions on the paracetamol molecule using a committee model made of 16 different training sets. In particular, each training set is obtained by a random subselection of 2000 configurations over a total of 2500 FPS configurations extracted from the NVT ensemble. Upon computing the (calibrated) predictions of the polarizability tensor over the full NVE trajectory, the Raman intensity for each member of the committee model was computed as in Eq. (5.1), and the average prediction and estimated error were computed as described in the previous Section. Fig. 5.4 shows excellent results of the predicted spectrum for the entire range of frequencies, with error estimates that are in fact negligibly small. This is in agreement with the high accuracy already reported in Fig. 5.3 when testing the actual prediction error on the individual Cartesian components of $\boldsymbol{\alpha}$.

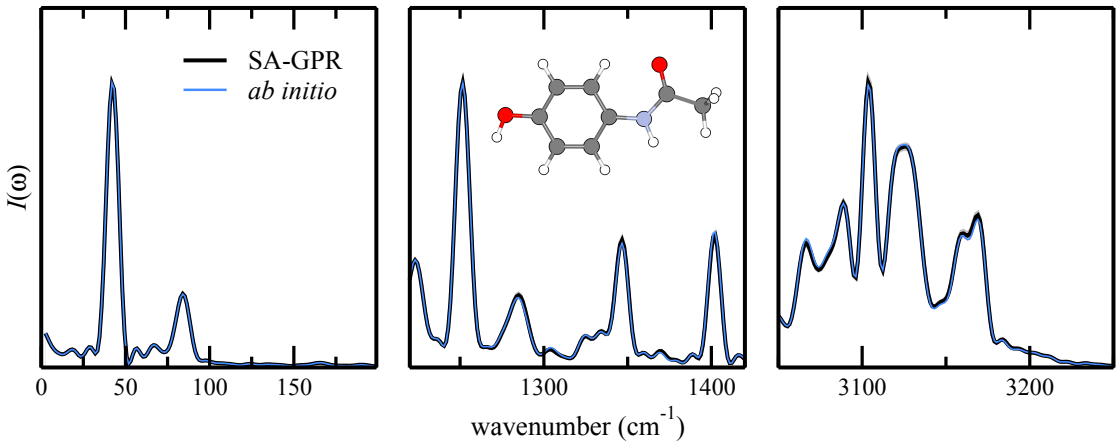


Figure 5.4 – (*black line*) Raman spectrum prediction of the paracetamol molecule in vacuum. (*blue line*) reference *ab initio* Raman spectrum. (*shaded area*) uncertainty estimate.

5.6 Raman spectrum of a paracetamol crystal

We now consider the more challenging case of predicting the spectrum of the first crystal polymorph of paracetamol, as represented in Fig. 5.1-(b). The training set is built by considering a random selection of 2500 configurations extracted from a NVT trajectory, so that to uniformly sample the underlying canonical distribution.

For this example, we rely on a two-step procedure. We first consider the sum of the polarizability predictions associated with the individual monomers of the molecular crystal within the unit cell, i.e., $\boldsymbol{\alpha}^{\mathrm{mol}} = \sum_{I=1}^{4} \boldsymbol{\alpha}_I$. Then, $\boldsymbol{\alpha}^{\mathrm{mol}}$ is used as a baseline value for the prediction of $\boldsymbol{\alpha}$. In doing so, the regression framework is mainly asked to learn the variations of the polarizability tensor associated with the intermolecular interactions between the monomers. As shown by the learning curves in Fig. 5.5, centering $\boldsymbol{\alpha}$ about the sum of the molecular polarizabilities has the effect of greatly improving the accaracy of model, especially when using few training data.

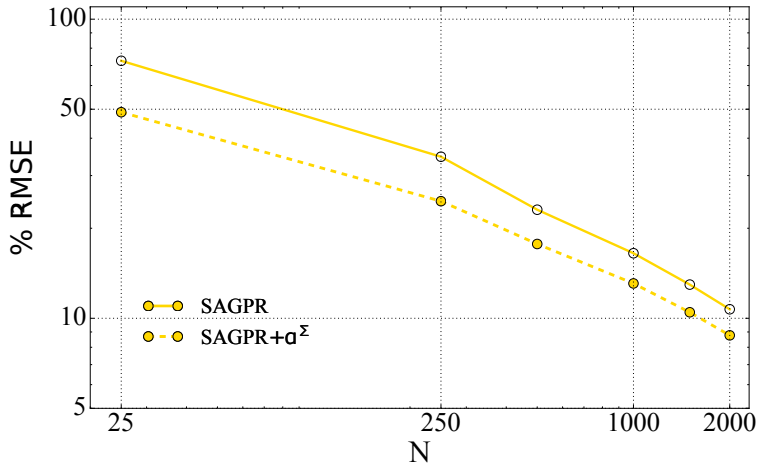


Figure 5.5 – Learning curves for the full polarizability Cartesian tensor of paracetamol crystal form I. Predictions are tested within the NVT ensemble. (*full lines*) Plain SA-GPR predictions. (*dashed lines*) SA-GPR predictions obtained using the predicted $\boldsymbol{\alpha}^{\mathrm{mol}}$ as a baseline.

To perform the Raman predictions, we once again define a committee model made of 16 random subselections, each of which contains 80% of the training set. Using the baseline strategy previously introduced, the polarizabilities of the full NVE trajectory are predicted and the associated Raman spectra computed. Figure 5.6 shows the prediction results and the corresponding estimated error. In this case, we find that the estimated variance computed over the committee members has to be increased by roughly an order of magnitude, i.e., $\nu^2 = 10.9$ in Eq. (5.4), meaning that the 16 subselections are strongly correlated to each other. One can observe that the excellent agreement between the reference and predicted spectrum at low frequencies is consistent with a negligible estimated error, while larger discrepancies and error bars can be observed in the more challenging high-frequency domain.

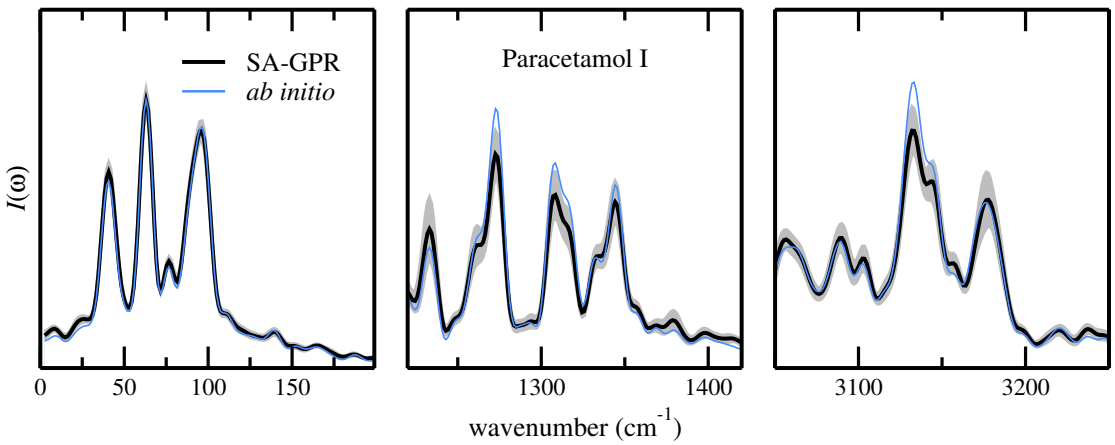


Figure 5.6 – (*black line*) Raman spectrum prediction of paracetamol crystal form I. (*shaded area*) error estimate. (*blue line*) Reference *ab initio* Raman spectrum.

5.7 Extrapolation to other crystal polymorphs

Thanks to the already discussed additive nature of the λ-SOAP predictions, one can think of predicting the polarizability of the crystal form II (Fig. 5.1-(c)) with the model trained on form I only. Since different polymorphic forms are mainly distinguished by the different intermolecular interactions, major difficulties in this extrapolation procedure are expected to be associated with the low-frequency (intermolecular) modes of the molecular crystal. To put this idea to the test, we trained SA-GPR on form I using the same committee model as before, and made predictions for a NVE trajectory of form II.

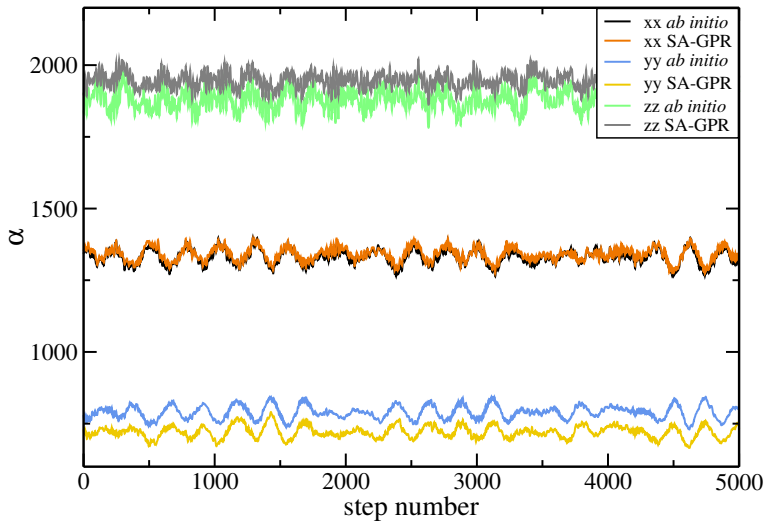


Figure 5.7 – Comparison between DFPT and SA-GPR time series for three Cartesian components of $\boldsymbol{\alpha}$.

We find that the error in the prediction of the polarizability tensor is mostly associated to an offset in the time series of some of the Cartesian components (Fig. 5.7). Interestingly, because the Raman intensity comes from the Fourier transform of the time series, these offsets do not have a substantial impact on the predicted spectrum. As shown in Fig. 5.8, the general lineshape is in agreement with the reference *ab initio* spectrum, even though the error in the intensities is overall larger than for the direct prediction of the first crystal polymorph. As expected, high frequencies are this time better described and that errors are more pronounced at low frequencies. This suggests that the model can accurately reproduce changes in polarizability associated with intramolecular vibrations, while it is less effective in predicting low-frequency components that are specific to the molecular packing of form II.

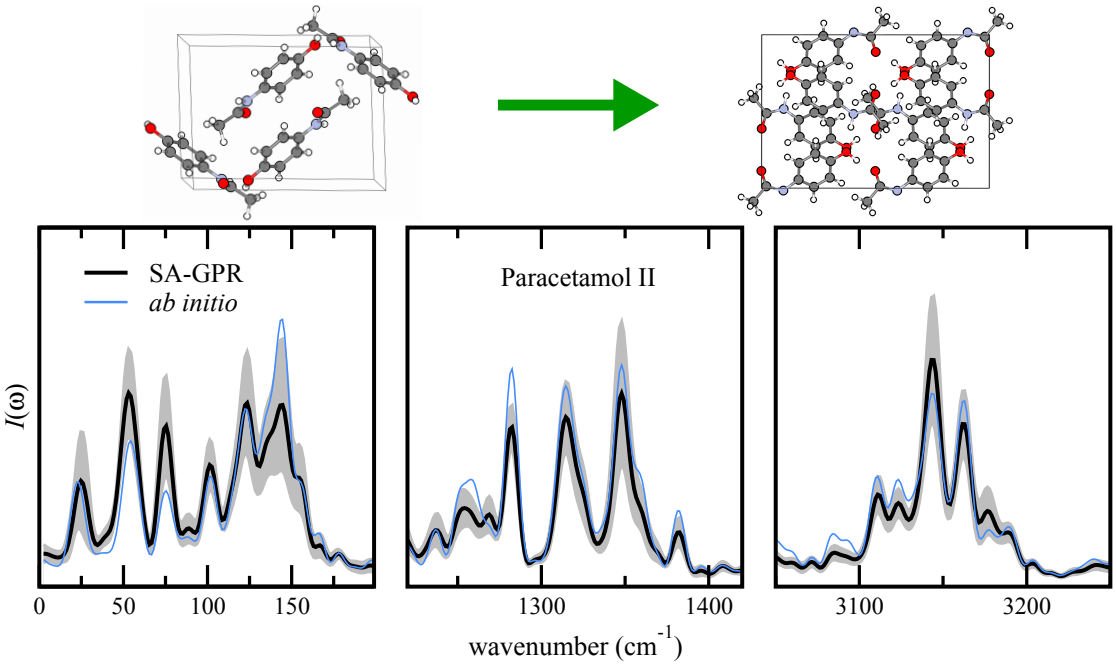


Figure 5.8 – (*black line*) Raman spectrum prediction of paracetamol crystal form II upon learning the polarizability on form I. (*shaded area*) error estimate. (*blue line*) Reference *ab initio* Raman spectrum.

Electronic charge densities Part II

6 Machine learning of electron densities

In the first part of this thesis we derived a class of symmetry-adapted representations of the atomic structure and showed how they can be used to efficiently regress electronic response tensors. In this chapter, we show that the same class of representations can also be used within a learning framework that is specifically designed to regress three-dimensional scalar fields. Taking the electron density of a system as an example, we underscore the importance of adopting a multi-centered basis to expand the scalar field in order to predict each electron-density component in a data-efficient and highly transferable fashion. Sections and figures are adapted from the following article:

A. Grisafi, A. Fabrizio, B. Meyer, D. M. Wilkins, C. Corminboeuf and M. Ceriotti, "Transferable Machine-Learning Model of the Electron Density", *ACS Central Science* 5, 57–64 (2019). Copyright © 2019 American Chemical Society. AG contributed to deriving and implementing the symmetry-adapted scalar-field regression framework, to carry out the electron-density predictions, to produce the figures and to writing the manuscript.

6.1 Electronic charge density

The electron density $n_e(\boldsymbol{r})$ is a fundamental property of atoms, molecules and condensed phases of matter. It is generally defined as the integral of the electronic probability distribution over $N-1$ degrees of freedom:

$$n_e(\boldsymbol{r}) = \int d\boldsymbol{r}_1 \int d\boldsymbol{r}_2 \ldots \int d\boldsymbol{r}_{N-1} \, |\Psi_N(\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N)|^2 \,, \tag{6.1}$$

with Ψ_N the many-body wave-function for a system of N electrons. $n_e(\boldsymbol{r})$ can be measured directly by high-resolution electron diffraction [130, 131] and transmission electron microscopy [132], and can be analyzed to identify covalent and non-covalent patterns [133–137]. Based on density-functional theory (DFT), in the framework of the

first Hohenberg-Kohn theorem [4], knowledge of $n_e(\boldsymbol{r})$ gives access, in principle, to any ground-state property. Especially for large systems, however, the computation of $n_e(\boldsymbol{r})$ requires considerable effort, involving the solution of an electronic structure problem with a more or less approximate level of theory. Sidestepping these calculations and directly accessing the ground-state electron density for a given configuration of atoms would have broad implications, including real-time visualization of chemical fingerprints, acceleration of DFT calculations by providing an estimate of the self-consistent $n_e(\boldsymbol{r})$ that is closer to the functional minimum, and the analysis of X-ray crystallographic experiments.

The first machine-learning model that has been proposed to bypass the quantum-mechanical calculation of $n_e(\boldsymbol{r})$ consists in making use of the Hohenberg-Kohn mapping to learn the electron density on a basis of plane-waves [27, 28]. Although successful, the choice of adopting a global basis set to decompose the density field carries the downside of limiting the transferability of the model to relatively small and rigid systems. Another approach, that can instead achieve a certain degree of transferability between different systems, exploits a representation of the atomic environment that can be used to directly learn and predict the electronic density on a three-dimensional grid around the molecule [138, 139]. While free of any basis-set decomposition error, representing the density field on a large number of grid points, rather than on a small set of basis set coefficients, has the disadvantage of dramatically increasing the computational effort. In this study, we show how to interpolate the electron density of a system by combining a local multi-centered basis set to represent $n_e(\boldsymbol{r})$ with the symmetry-adapted structural representations already adopted in the context of tensors learning. In the process, we derive a completely general regression framework that makes use of λ-SOAP kernels to predict any three-dimensional scalar field in a strictly linear-scaling and highly transferable manner.

6.2 Multi-centered spherical harmonics expansion

The problem of decomposing the electronic charge density of a system on atom-centered contributions has long been known in the context of determining $n_e(\boldsymbol{r})$ from experimental X-ray diffraction data [140–144]. One of the most widely used methods is the multipole model proposed by Stewart [145] and by Hansen and Coppens [146], which models the valence charge density with both spherical and anisotropic components that are specific of the nature of the molecular fragments involved.

To tackle the problem of decomposing the density field in such a way that it can be effectively regressed through λ-SOAP kernels, we adopt an approach that is similar

in spirit to the aforementioned multipolar expansion, but that does not rely on any prior knowledge of discrete molecular fragments. Instead, we adopt a multi-centered basis set decomposition analogous to the one commonly used in quantum chemistry for representing the molecular orbitals of a system [147]. In particular, we expand $n_e(\boldsymbol{r})$ over a non-orthogonal basis made of radial functions and spherical harmonics centered on the atomic positions. For a system of n_{at} atoms, we write

$$n_e(\boldsymbol{r}) = \sum_{i=1}^{n_{\text{at}}} \sum_{\lambda=0}^{\lambda_{\text{cut}}(a_i)} \sum_{n=1}^{n_{\text{cut}}(a_i,\lambda)} \sum_{\mu=-\lambda}^{\lambda} c^i_{n\lambda\mu}\, R^{a_i}_{\lambda n}(|\boldsymbol{r}-\boldsymbol{r}_i|)\, Y_{\lambda\mu}\left(\frac{\boldsymbol{r}-\boldsymbol{r}_i}{|\boldsymbol{r}-\boldsymbol{r}_i|}\right), \tag{6.2}$$

where $c^i_{n\lambda\mu}$ are the non-orthogonal coefficients that realize the multi-centered expansion. Here, the angular cutoff λ_{cut} depends on the species a_i of the atom i, while the number n_{cut} and type of radial functions $R^{a_i}_{\lambda n}$ depend both on the species a_i and on the angular momentum λ. Note that Eq. (6.2) realizes a decomposition of $n_e(\boldsymbol{r})$ in effective atomic contributions without making use of any prior knowledge about the molecular conformation and identity. As such, it provides the optimal representation of the target that can be regressed via a machine-learning model that relies on the definition of local atomic environments.

6.3 Regression of three-dimensional scalar fields

In order to learn the density decomposition of Eq. (6.2) through a local and symmetry adapted representation of the atomic coordinates, we now make the assumption that the expansion coefficients $c^i_{n\lambda\mu}$ arise from the outcome of a sparse SA-GPR prediction. In particular, we assume that the covariance of order λ of each coefficient under rotation and inversion operations is expressed by λ-SOAP representations of corresponding order. In so doing, each set of coefficients with different atomic species a, radial channels n and angular orders λ are approximated as follows,

$$c^i_{n\lambda\mu} \approx \sum_{j=1}^{M} \delta_{a_i a_j} \sum_{\mu'=-\lambda}^{\lambda} k^{\lambda}_{\mu\mu'}(A_i, A_j)\, x^j_{n\lambda\mu'}, \tag{6.3}$$

with $x^j_{n\lambda\mu'}$ the set of (covariant) regression weights we wish to learn and $k^{\lambda}_{\mu\mu'}$ a λ-SOAP kernel. Here, j runs over a set of M atomic environments that are selected to best represent the local structural and chemical diversity across the dataset, according to the sparse GPR framework already outlined in Sec. 1.3.1. Note that since the kernel $k^{\lambda}_{\mu\mu'}$ does not depend on the type of radial functions $R^{a_i}_{\lambda n}$, the deltas $\delta_{a_i a_j}$ are introduced to only couple those environments that are centered on atoms that belong to the same chemical species.

The determination of the regression weights $x^{j}_{n\lambda\mu'}$ follows the minimization of a loss function that describes the collective error in representing the real-space density field of N training molecules. Using the compact notation $\phi^{a_i}_{n\lambda\mu} \equiv R^{a_i}_{n\lambda} Y_{\lambda\mu}$ to indicate the multi-centered basis functions, we have

$$\ell(\boldsymbol{x}_M) = \sum_{A=1}^{N} \int d\boldsymbol{r} \left| n_e(A;\boldsymbol{r}) - \sum_{i\in A} \sum_{n\lambda\mu} c^{i}_{n\lambda\mu}(\boldsymbol{x}_M)\, \phi^{a_i}_{n\lambda\mu}(\boldsymbol{r}-\boldsymbol{r}_i) \right|^2 + \eta\, \boldsymbol{x}_M^T \boldsymbol{K}_{MM} \boldsymbol{x}_M \quad (6.4)$$

with η accounting for the intrinsic noise of the training densities. The subscript M labels the fact that the dimension of the weight vector $\boldsymbol{x}_M$ and kernel matrix $\boldsymbol{K}_{MM}$ is determined by the number of sparse atomic environments M, times the number of basis functions associated with the corresponding chemical species, i.e., $\{jn\lambda\mu\}$. Upon substituting the prediction ansatz of Eq. (6.3) into Eq. (6.4), minimization of the loss function with respect to $\boldsymbol{x}_M$ yields the following regression formula:

$$\boldsymbol{x}_M = \left(\boldsymbol{K}_{NM}^T \boldsymbol{S}_{NN} \boldsymbol{K}_{NM} + \eta \boldsymbol{K}_{MM}\right)^{-1} \boldsymbol{K}_{NM}^T \boldsymbol{w}_N \,. \quad (6.5)$$

Here, the vector $\boldsymbol{w}_N$ contains the projections $\langle\phi|n_e\rangle$ of the N reference densities on the basis functions,

$$w^{i}_{n\lambda\mu} = \int d\boldsymbol{r}\, n_e(\boldsymbol{r})\, \phi^{a_i}_{n\lambda\mu}(\boldsymbol{r}-\boldsymbol{r}_i)\,, \quad (6.6)$$

while $\boldsymbol{S}_{NN}$ is the block-diagonal matrix containing the spatial overlap $\langle\phi|\phi'\rangle$ between the basis functions of each training molecule,

$$S^{i'n'\lambda'\mu'}_{in\lambda\mu} = \int d\boldsymbol{r}\, \phi^{a_{i'}}_{n'\lambda'\mu'}(\boldsymbol{r}-\boldsymbol{r}_{i'})\, \phi^{a_i}_{n\lambda\mu}(\boldsymbol{r}-\boldsymbol{r}_i)\,. \quad (6.7)$$

Finally, the rectangular matrix $\boldsymbol{K}_{NM}$ contains the symmetry-adapted kernels that couple the atomic environments of the N training molecules with the ones of the representative sparse set M. If these were chosen to correspond to the ones of the N training densities, then $\boldsymbol{K}_{NM} = \boldsymbol{K}_{MM} = \boldsymbol{K}_{NN}$, and Eq. (6.5) simplifies to

$$\boldsymbol{x}_N = \left(\boldsymbol{K}_{NN} \boldsymbol{S}_{NN} + \eta \mathbf{1}_{NN}\right)^{-1} \boldsymbol{w}_N \,. \quad (6.8)$$

In fact, Eq. (6.5) expresses a sparse approximation to the Equation above that allows us to massively cut down the kernel dimensionality by projecting the problem on a reduced set M of most representative atomic environments.

6.4 The curse of non-orthogonality

From the previous discussion, once the regression weights $\boldsymbol{x}_M$ are determined according to Eq. (6.5), the covariant prediction of the density expansion coefficients $c^i_{n\lambda\mu}$ is obtained as in Eq. (6.3). At this point, one could ask why not addressing the regression of the different families of expansion coefficients separately, rather than using the full density $n_e(\boldsymbol{r})$ as a learning target. In fact, one could simply obtain the training expansion coefficients as $\boldsymbol{c} = \boldsymbol{S}^{-1}\boldsymbol{w}$ and design an independent covariant regression for each set of coefficients that belong to different atomic species a, angular orders λ and radial channels n. The downside of this approach is that the error made in the prediction of each family $\{a\lambda n\}$ would sum up in an uncontrollable way when reconstructing the density field. This is because the non-orthogonal nature of the basis functions $\phi^{a_i}_{n\lambda\mu}$ implies that the different families of coefficients are necessarily correlated to each other, so that the regression must deal with *all* the density components at once.

If the basis functions were orthogonal, then $\boldsymbol{S} = \boldsymbol{1}$, and the problem of learning $n_e(\boldsymbol{r})$ could be conveniently recast into the problem of learning each set of (uncorrelated) orthogonal projections $\boldsymbol{c} = \boldsymbol{w}$ separately. This is exactly what is done in Ref. [27], where the electron density is expanded on a plane-wave basis $e^{i\boldsymbol{k}\cdot\boldsymbol{r}}$ and a separate regression problem is solved for each distinct set of Fourier components $\hat{n}_e(\boldsymbol{k})$. Although this framework is undoubtedly more convenient than having to deal with all the set of density components simultaneously, the fact that each Fourier component depends on the density of the entire system carries the major drawback of hindering the transferability of the method across highly heterogeneous datasets. Conversely, the atom-centered nature of the basis functions $\phi^{a_i}_{n\lambda\mu}$ can be exploited to transfer the information encoded in the density projections $\boldsymbol{w}$ across atomic environments that share a similar nature – the same principle that underlies the construction of local machine-learning representations.

A further downside brought by the non-orthogonality of the basis functions is the fact that the λ-SOAP kernel neglects, by construction, the statistical correlations between different families $\{a\lambda n\}$ of coefficients. As already discussed in Sec. 1.3.1, any kernel is in fact interpreted as a prior for the statistical correlations of the quantity we wish to predict, i.e., the expansion coefficients $c^i_{n\lambda\mu}$ in this case. As such, it should therefore be able to couple different pairs of atomic species (a, a'), radial channels (n, n') and angular orders (λ, λ'). Building covariant kernel functions that satisfy these criteria is in principle possible, but it would also imply dealing with a formulation that is way more convoluted and expensive to compute with respect to λ-SOAP.

6.4.1 The Löwdin approach

The aforementioned issues associated with the use of a non-orthogonal basis to represent $n_e(\boldsymbol{r})$ would be entirely resolved if one used as learning targets the density projections that arise from the following orthogonalization procedure: $\tilde{\boldsymbol{w}} = \boldsymbol{S}^{-1/2}\boldsymbol{w}$. This kind of orthogonal transformation is known in quantum chemistry as Löwdin orthogonalization [148], and it is typically adopted to work in the basis of *atomic natural orbitals* (ANOs) [149]. In contrast to other hierarchical orthogonalization algorithms, such as Gram-Schmidt, orthogonalizing the basis functions using $\boldsymbol{S}^{-1/2}$ carries the advantage of preserving both the covariant and local nature of the density projections $\tilde{\boldsymbol{w}}$ about the atoms of the system. While this may sound as the optimal scenario to deal with, one should however notice that, upon the orthogonalization procedure, the projections $\tilde{\boldsymbol{w}}$ not only include information about the density field, but also encode the transformation of the basis functions that is induced by the $\boldsymbol{S}^{-1/2}$ operator. As a result, the regression would carry the additional burden of describing the system-dependent variations of the basis functions across the dataset, which would make the learning of $n_e(\boldsymbol{r})$ both more challenging and data hungry. For this reason, while we do not disregard the Löwdin approach as a valuable strategy to tackle the problem of density-learning, from here on we will only refer to the non-orthogonal regression approach previously derived.

6.5 Dataset generation and error definition

We test our density-learning model on the valence electron density of 1000 molecules of ethene (C_2H_4), ethane (C_2H_6), butadiene (C_4H_6) and butane (C_4H_{10}), computed at the DFT/PBE/SBKJC-LFK level [150, 151] with SBKJC effective core potentials [152]. The basis functions chosen for the density expansion correspond to *Gaussian type orbitals* (GTOs), routinely used in quantum chemistry codes. This choice allows us to compute analytically the spatial overlap $\boldsymbol{S}$ between basis functions by relying on well-known transformations between spherical and Cartesian Gaussian functions [153]. The projections of the reference DFT densities on GTOs are instead computed numerically as follows,

$$w^i_{n\lambda\mu} = \int d\boldsymbol{s}\; s^2 R^{a_i}_{\lambda n}(s)\; Y^{\star}_{\lambda\mu}(\hat{\mathbf{s}})\; n_e(\boldsymbol{r}_i + \boldsymbol{s})\,, \tag{6.9}$$

with $\mathbf{s} = \boldsymbol{r} - \boldsymbol{r}_i$. In particular, the integration of the spherical harmonics on the unit sphere is performed using the Lebedev quadrature with 2030 points [154], while the radial integral is computed with an equispaced radial mesh of 200 points spanning a distance of 6 Å from the central atom i.

Upon reconstructing the real space density on uniform Cartesian grids of 0.1 Bohr of spacing, the percentage error $\varepsilon(\%)$ incurred in the approximation of $n_e(\boldsymbol{r})$ is measured all throughout as the integrated absolute difference with respect to the DFT density, as a fraction of the number of electrons N_e:

$$\varepsilon_\rho(\%) = 100 \times \frac{1}{N_\mathrm{e}} \sum_k \left| n_e(\boldsymbol{r}_k) - n_e^{\mathrm{DFT}}(\boldsymbol{r}_k) \right| \tag{6.10}$$

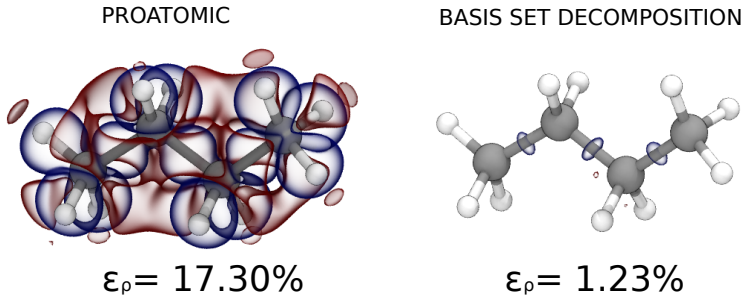
6.6 Basis set optimization

When using a multi-centred expansion, one can easily encounter ill-conditioning issues consisting in a lowering of the rank of the overlap matrix $\boldsymbol{S}$. The standard strategy to avoid these problems involves the contraction of the radial functions $R_n(r)$ into a smaller set of optimized functions $R'_n(r)$. For each chemical species a and angular order λ, this contraction reads as the following linear combination,

$$R'_{a\lambda n}(r) = \sum_{n'} C_{nn'} R_{a\lambda n'}(r)\,, \tag{6.11}$$

with $C_{nn'}$ the rectangular matrix that realize the basis set contraction. In our case, for each atomic species (C and H), we considered angular functions up to $\lambda_\mathrm{cut} = 3$ and 12 primitive radial functions that are contracted down to a total of 4 optimized functions. The contraction is optimized in such a way that both the percentage density error $\varepsilon(\%)$ associated with the solution of the linear problem $\boldsymbol{c} = \boldsymbol{S}^{-1}\boldsymbol{w}$ and the condition-number ω of the overlap matrix $\boldsymbol{S}$ are simultaneously minimized [155]. In particular, we found that a good compromise for the optimization of the contraction matrix $\boldsymbol{C}_{12\times4}$ consists in using $\varepsilon(\%) + 0.1\log_{10}(\omega)/N_\mathrm{e}$ as a minimization target.

The table included in Figure 6.1 reports the basis set errors $\varepsilon_\rho(\%)$ averaged over each of the individual C_2 and C_4 datasets, as compared with the superposition of the isolated atomic densities, i.e., the *protoatomic* density. Clearly, the optimized multi-centered expansion of Eq. (6.2) yields a representation of $n_e(\boldsymbol{r})$ that is about 20 times more accurate than the protoatomic density, obtaining a mean absolute error that is ∼1% of the electronic charge. A visual representation of this comparison is also shown in Figure 6.1 for a given configuration of butane, where it is apparent that the small basis set error is concentrated in the C–C bond regions.



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

Figure 6.1 – Density errors at different level of representation: (*left*) superposition of isolated atomic densities, (*right*) optimized basis set. Red and blue isosurfaces refer to an error of ± 0.005 Bohr^{-3} respectively. The Table reports the proatomic and basis set decomposition errors averaged over the dataset of C_2 and C_4 molecules.

6.7 Angular spectrum of the valence density

From the optimal basis set decomposition previously discussed, it is instructive to single out the contributions to $n_e(\boldsymbol{r})$ carried by each angular order λ, i.e.,

$$n_e^\lambda(\boldsymbol{r}) = \sum_{i=1}^{n_{\text{at}}} \sum_{n\in\{a_i,\lambda\}} \sum_{\mu=-\lambda}^{\lambda} c^i_{n\lambda\mu}\phi^{a_i}_{n\lambda\mu}(\boldsymbol{r}-\boldsymbol{r}_i) \tag{6.12}$$

As exemplified in Fig. 6.2, while the isotropic $\lambda = 0$ functions determine the general shape of the density, the $\lambda = 1$ functions primarily describe the gradient of electronegativity in the region close to C–H bonds. Furthermore, the $\lambda = 2$ functions describe the charge modulation associated with the C–C bonds along the main chain as well as the π-cloud along the conjugated backbone, while the $\lambda = 3$ functions act as a further modulation that captures the non-trivial anisotropy. The Figure also shows the λ-spectrum of the valence charge density computed as the collective contribution to the electron-density variability carried by each angular order λ and atomic type a, i.e.,

$$\sigma_\lambda(a) = \sqrt{\left\langle \sum_{i\in a} \sum_{n\in\{a,\lambda\}} \left|\boldsymbol{c}^i_{\lambda n} - \langle \boldsymbol{c}^i_{\lambda n}\rangle\right|^2 \right\rangle} \tag{6.13}$$

where the average $\langle\cdot\rangle$ is computed on the entire the dataset. After having subtracted

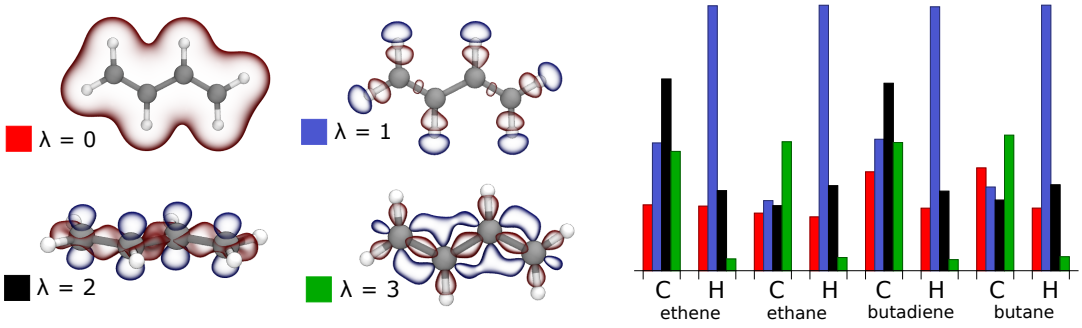


Figure 6.2 – (*left*) representation of the angular momentum decomposition of the electron density. Red and blue isosurfaces refer to ± 0.01 Bohr^{-3} respectively. (*right*) angular momentum spectrum $\sigma_\lambda(a)$ in arbitrary units of the valence electron density of C_2 and C_4 datasets. The isotropic contributions $\lambda = 0$ express the collective variations with respect to the dataset's mean value, while the mean is statistically zero for $\lambda > 0$.

the mean spherical contribution of pure $\lambda = 0$ character, the $\lambda = 1$ components largely dominate the charge density variability associated with hydrogen atoms. As previously demonstrated [156], functions with $\lambda = 2$ symmetry also carry a substantial contribution, particularly for the carbon atoms of alkenes, while $\lambda = 3$ functions appear to be dominant for carbon atoms of alkanes and almost irrelevant for hydrogen atoms in all the four molecules. Note that in comparison to an atom-centered expansion of the wave-function Ψ, the choice of using a larger basis set is justified by the greater complexity in describing an electron density field rather than the $N_e/2$ occupied molecular orbitals defined as the eigen-solutions of an effective single-particle Hamiltonian. In fact, the squaring of Ψ that yields $n_e(\boldsymbol{r})$ introduces components with up to twice the maximum λ used to expand the wave-function.

6.8 Learning performance

Having analyzed the variability of the electron density when expanded over an optimized multi-centered basis, we now proceed to test the learning performance associated with the symmetry-adapted sparse-GPR formulation of Eq. (6.5). Following the choice of basis functions previously described, we generated λ-SOAP kernels up to $\lambda = 3$ using a cutoff of $r_c = 4.5$ Å and a non-linearity degree of $\zeta = 2$. For all the spherical orders λ, the feature space is reduced down to the 500 principal components obtained from the diagonalization of the covariance matrix defined in space of data points. The number M of reference environments has been fixed to the 1500 most diverse, FPS-selected, environments contained in each dataset. Learning curves are then

obtained by varying the number of training molecules up to 800 randomly selected configurations out of the total of 1000. The remaining 200 molecules for each of these random selections are used to estimate the error in the density prediction.

The difficulty of the learning exercise largely depends on the structural flexibility of the molecules. Small, rigid systems such as ethene and ethane require little training, and could be equivalently learned through a machine-learning framework based on a pairwise comparison of aligned molecules. Butadiene data, containing both *cis* and *trans* conformers, as well as distorted configurations approaching the isomerization transition-state, poses a more significant challenge, due to an extended conjugated system that makes the electronic structure very sensitive to small molecular deformations. The case of butane is also particularly challenging because of the broad spectrum of intramolecular non-covalent interactions spanned by the many different conformers contained in the dataset. Being fully flexible, this kind of system is expected to benefit most from a ML scheme that can adapt its kernel similarity measure to different orientations of molecular sub-units.

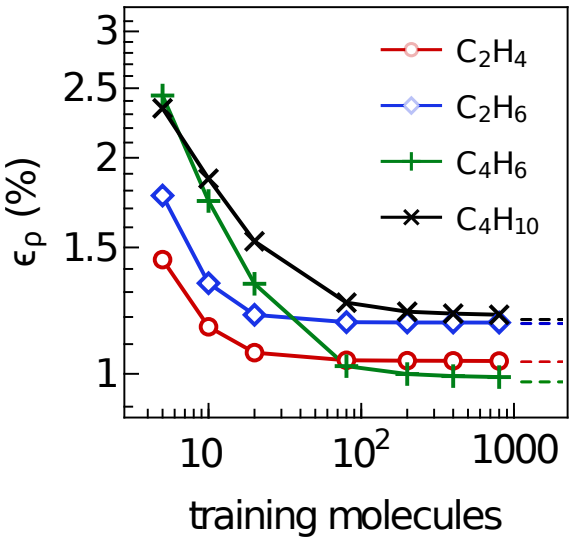


Figure 6.3 – Learning curves for C_2 and C_4 molecules expressed as the % mean absolute error of the predicted SA-GPR densities as a function of the number of training molecules.

Fig. 6.3 shows the performance of the method in terms of prediction accuracy of the electron density as a function of the number of training molecules. The prediction errors of ethene and ethane saturate to the limit set by the basis set representation, which is around 1% for all molecules, with as few as 10 training points. As expected, given the greater flexibility, learning the charge density of butadiene and butane is more challenging, requiring the inclusion of more than 100 training structures in order

to approach the basis set limit. Overall, the prediction accuracy that we can possibly achieve using this dataset is therefore limited by the error incurred in the basis set decomposition of the scalar field.

6.9 Indirect energy prediction

When obtaining the reference densities from a density-functional calculation, the regression framework previously discussed can in principle be used to indirectly compute the energy of the system by feeding the functional back with the predicted $n_e(\boldsymbol{r})$. As a benchmark for this application, we evaluated the PBE exchange-correlation functional $E_{\mathrm{XC}}[n_e]$ used for the reference quantum-mechanical calculations. Depending on the gradient of the density, this quantity is very sensitive to small density variations, especially those localized around the atomic nuclei.

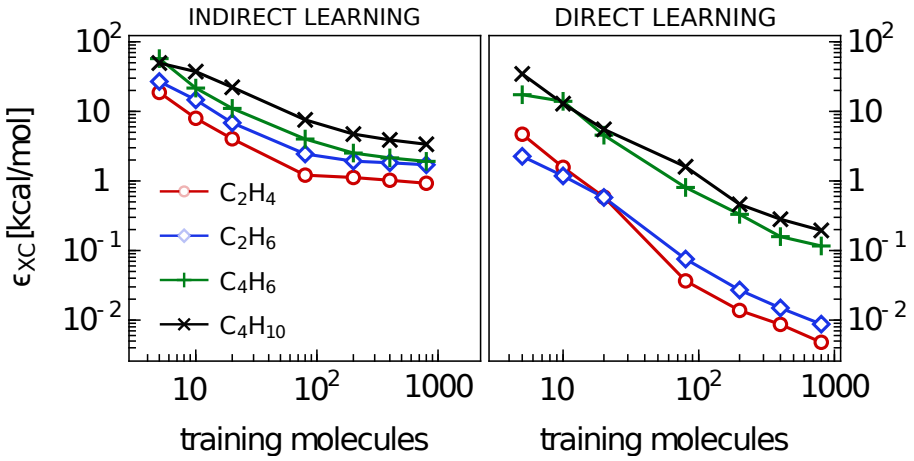


Figure 6.4 – Root mean square errors of the PBE exchange-correlation energies indirectly predicted from the SA-GPR densities and directly predicted via a scalar SOAP kernel, as a function of the number of training molecules. Dashed lines refer to the error carried by the basis set representation.

Fig. 6.4 shows the root mean square error for the exchange-correlation energies $\varepsilon_{\mathrm{XC}}$. Using the full set of 800 training molecules, we reach a RMSE of 0.9 and 1.7 kcal/mol for ethene and ethane, 1.9 kcal/mol for butadiene and 3.5 kcal/mol for butane, basically matching the basis set limit. It is clear that the ML scheme has the potential to reach higher accuracy with a small number of reference configurations, but a significant reduction of the basis set error is necessary to reach chemical accuracy (roughly 1 kcal/mol RMSE) in the prediction of E_{XC}. At the same time, it is not obvious that computing E_{XC} indirectly, by first predicting $n_e(\boldsymbol{r})$, is the most effective strategy to obtain a machine-learning model of DFT energetics. As shown in the Figure, applying

a direct, scalar regression based on conventional SOAP kernels to learn the relationship between the molecular structure and E_{XC} leads to vastly superior performance while requiring a much simpler machine-learning model.

6.10 Linear-scaling extrapolation

While incremental improvements of the underlying density representation framework are desirable to use the predicted density as the basis of DFT calculations, we can already demonstrate the potential of our SA-GPR scheme in terms of transferability of the model. From the prediction formula of Eq. (6.3), it is clear that no assumption is made about the identity of the molecule for which the electron density is predicted. Practically speaking, the regression weights $x^{j}_{n\lambda\mu}$ are associated with representative environments that could be taken from any kind of compound, not necessarily the same as that for which the density is being predicted. As long as the training set is capable of describing different chemical environments, and contains local configurations similar to the ones of our prediction target, accurate densities can be obtained simply by computing the kernels between the atomic environments of an arbitrarily large molecule and the subset of representative environments M. The cost of this prediction is proportional to the number of environments, making this method of evaluating the electron density strictly linear scaling in the size of the target molecule.

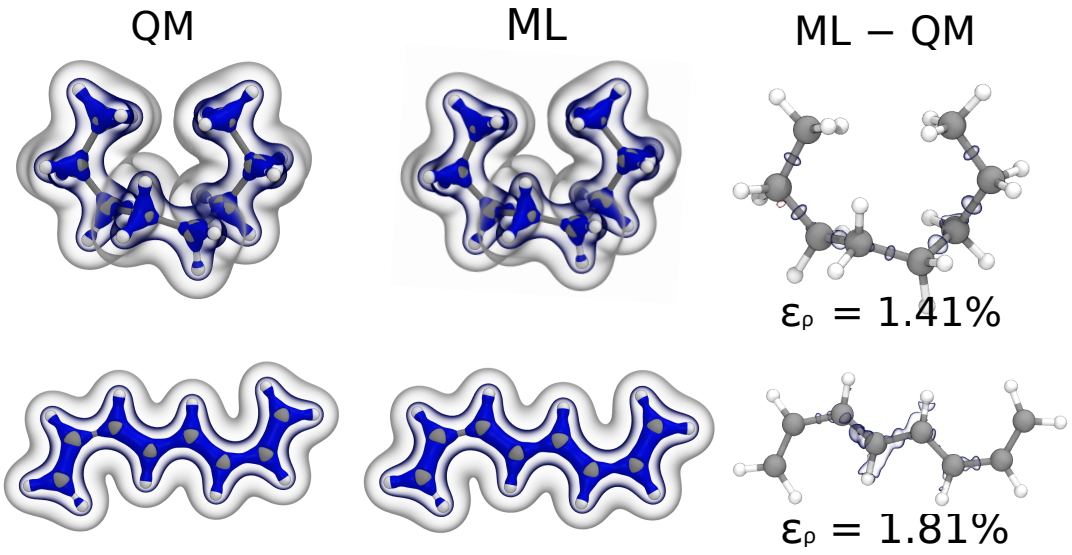


Figure 6.5 – Extrapolation results for the valence electron density of one octane (*top*) and one octatetraene (*bottom*) conformer. (*left*) DFT density isosurface at 0.25, 0.1, 0.01 Bohr^{-3}, (*middle*) SA-GPR prediction isosurface at 0.25, 0.1, 0.01 Bohr^{-3}, (*right*) machine-learning error, red and blue isosurfaces refer to ± 0.005 Bohr^{-3} respectively.

As a proof of concept of this extrapolation procedure, we use environments and training information from the butadiene and butane configurations already discussed to predict the electron density of similar molecules that are twice as large, namely octatetraene (C_8H_{10}) and octane (C_8H_{18}), respectively. For both octane and octatetraene, the extrapolation is carried out on a challenging dataset made of the 100 FPS structures extracted from the 300K replica of a long replica-exchange MD run. When learning on the full dataset of butadiene and butane, we obtain a low density mean absolute error of 1.8% for octatetraene and of 1.4% for octane. As shown in Fig. 6.5 for two representative configurations, the linear-scaling predictions accurately reproduce the structure of the electron density for both octane and octatetraene. Because of the high sensitivity of the electronic π-cloud to the molecular identity and configuration, major difficulties arise in predicting the electron density of octatetraene, particularly in the middle regions, for which no analogous examples are contained in the butadiene training dataset.

It is important to stress that the transferability of the method is due to the fact that, on a local scale, the larger molecules are similar to those used for training. Therefore, the prediction is effectively an *interpolation* in the space of local environments. This is emphasized by the observation that the optimal extrapolation accuracy is obtained using a λ-SOAP cutoff of $r_{\text{cut}} = 3$ Å, versus a value of $r_{\text{cut}} = 4.5$ Å that was optimal for same-molecule predictions. On a scale larger than 3 Å, the C_8 environments differ substantially from those in the corresponding C_4 compound, which negatively affects the transferability of the model. Ideally, as the training dataset is extended to include larger and larger molecules, this locality constraint can be relaxed until no substantial difference can be appreciated between the prediction accuracy of the interpolated and extrapolated density.

7 Density learning with quantum-chemical accuracy

In the previous chapter, we have seen how to construct a highly transferable regression model of the electron density of a system that makes use of symmetry-adapted λ-SOAP representations. As demonstrated by the examples provided, the model is not limited by the learning accuracy, but rather by the quality of the basis set decomposition of the scalar field. In this chapter, we show how to solve this issue by adopting density-fitted auxiliary basis functions that are routinely used in quantum-chemical applications. Crucially, this methodological advancement allows us to push the accuracy of the electron-density predictions up to the level of state-of-the-art all-electron calculations. An open source implementation of the method can be found in the SALTED package [157]. Sections and figures are largely based on the following article:

A. Fabrizio, A. Grisafi, B. Meyer, M. Ceriotti and C. Corminboeuf, "Electron density learning of non-covalent systems", *Chemical Science* 10, 9424 (2019). Copyright © 2019 The Royal Society of Chemistry. AG contributed to enhance the density-learning implementation by making it both faster and memory-saving, produced the density-learning results and contributed to writing the manuscript.

7.1 Electron density in single-particle theories

When solving the Schrödinger equation using an effective single-particle theory, such as the Kohn-Sham DFT approach, the electron density $n_e(\boldsymbol{r})$ of a system can be simply computed as a sum over the single-electron probability distributions associated with the occupied *molecular orbitals* (MOs), $\psi_k(\boldsymbol{r})$. In finite molecular systems, each MO is typically expanded over a basis of *atomic orbitals* (AOs), $\chi_\nu(\boldsymbol{r})$, which resemble the Slater-type functions that solve the Schrödinger equation for the isolated hydrogen atom. In particular, the coefficients $\tilde{c}_\nu^k$ of this expansion represent the variational parameters of the quantum-chemical calculation that are used to minimize the many-body electronic energy. As a result, the variationally optimized $n_e(\boldsymbol{r})$ is written as an

expansion over pairs of one-center atomic orbitals,

$$\begin{aligned} n_e(\boldsymbol{r}) &= \sum_{k\in\text{occ.}}^{\text{MOs}} \left|\psi_k(\boldsymbol{r})\right|^2 = \sum_{k\in\text{occ.}}^{\text{MOs}} \left|\sum_{\nu}^{\text{AOs}} \tilde{c}_\nu^k \chi_\nu(\boldsymbol{r})\right|^2 \\ &= \sum_{\nu\nu'}^{\text{AOs}} \left(\sum_{k\in\text{occ.}}^{\text{MOs}} \tilde{c}_\nu^k \tilde{c}_{\nu'}^{k\star}\right) \chi_\nu(\boldsymbol{r}) \chi_{\nu'}^\star(\boldsymbol{r}) = \sum_{\nu\nu'}^{\text{AOs}} D_{\nu\nu'} \chi_\nu(\boldsymbol{r}) \chi_{\nu'}^\star(\boldsymbol{r}) \,, \end{aligned} \tag{7.1}$$

where the expansion weights $D_{\nu\nu'}$ are defined as the elements of the one-electron *reduced density matrix*. From Eq. (7.1), all the electronic energy contributions that depend on the electron density of the system can be computed from knowledge of $D_{\nu\nu'}$. The Hartree energy, in particular, is a universal functional of $n_e(\boldsymbol{r})$ that can be computed as the following two-center four-electron integral:

$$\begin{aligned} E_\text{H}[n_e] &= \frac{1}{2} \int d\boldsymbol{r} \int d\boldsymbol{r}' \, n_e(\boldsymbol{r}) \frac{1}{|\boldsymbol{r}-\boldsymbol{r}'|} n_e(\boldsymbol{r}') \\ &= \frac{1}{2} \sum_{\nu_1\nu_2}^{\text{AOs}} \sum_{\nu_3\nu_4}^{\text{AOs}} D_{\nu_1\nu_2} D_{\nu_3\nu_4}^\star \int d\boldsymbol{r} \int d\boldsymbol{r}' \, \chi_{\nu_1}(\boldsymbol{r}) \chi_{\nu_2}^\star(\boldsymbol{r}) \frac{1}{|\boldsymbol{r}-\boldsymbol{r}'|} \chi_{\nu_3}^\star(\boldsymbol{r}') \chi_{\nu_4}(\boldsymbol{r}') \,. \end{aligned} \tag{7.2}$$

Because of the slow decay of the Coulomb operator, the calculation of the integral above presents an unfavorable scaling with the system size and it therefore represents a computational bottleneck of any effective single-particle approximation of the electronic wave-function.

7.2 Resolution of the identity auxiliary basis

To alleviate the cost of computing the Hartree energy, specialized basis sets have been designed to represent $n_e(\boldsymbol{r})$ as a linear expansion that is formally equivalent to the one introduced in Eq. (6.2), i.e.,

$$n_e(\boldsymbol{r}) \approx \sum_k c_k \phi_k(\boldsymbol{r}) \,, \tag{7.3}$$

with $\phi_k(\boldsymbol{r})$ some optimized auxiliary functions and $k \equiv \{inl\mu\}$ a compact index for the basis set labels. This approximation to the *ab initio* density of Eq. (7.1) is known as the *resolution of the identity* (RI) approximation, and it has long been used to sidestep the four-electron integral of Eq. (7.2) [158–164]. The determination of the RI coefficients c_k, in particular, comes from the minimization of the error Δ_H incurred in representing the Hartree energy using the approximation of Eq. (7.3), i.e.,

$$\Delta_\text{H}[\boldsymbol{c}] = \frac{1}{2} \int d\boldsymbol{r} \int d\boldsymbol{r}' \left(\sum_k c_k \phi_k(\boldsymbol{r}) - n_e(\boldsymbol{r})\right) \frac{1}{|\boldsymbol{r}-\boldsymbol{r}'|} \left(\sum_k c_k \phi_k(\boldsymbol{r}') - n_e(\boldsymbol{r}')\right) , \tag{7.4}$$

which yields the optimal density-fitted coefficients as a linear transformation of the one-electron reduced density matrix:

$$c_k = \sum_{\nu\nu'} d_k^{\nu\nu'} D_{\nu\nu'} \,. \tag{7.5}$$

Here, the RI transformation matrix $d_k^{\nu\nu'}$ is defined by the multiplication of the inverse of the RI-Coulomb matrix $\boldsymbol{J}$ with the three-electron integral that expresses the Coulomb-coupling of a pair of AOs with the RI-auxiliary basis functions:

$$d_k^{\nu\nu'} = \sum_{k'} \left[\boldsymbol{J}^{-1}\right]_{kk'} \int d\boldsymbol{r} \int d\boldsymbol{r}' \chi_\nu(\boldsymbol{r}) \chi_{\nu'}(\boldsymbol{r}) \frac{1}{|\boldsymbol{r}-\boldsymbol{r}'|} \phi_{k'}(\boldsymbol{r}') \,, \tag{7.6}$$

with the RI-Coulomb matrix defined as

$$J_{kk'} = \int d\boldsymbol{r} \int d\boldsymbol{r}' \phi_k(\boldsymbol{r}) \frac{1}{|\boldsymbol{r}-\boldsymbol{r}'|} \phi_{k'}(\boldsymbol{r}') \,. \tag{7.7}$$

Upon the procedure previously described, a quantum-chemical RI-calculation can be carried out at the price of making an error in the Hartree energy equal to Eq. (7.4). This possibility not only allows us to speed up the generation of the reference calculations, but it also provides an accurate and well-conditioned basis for the regression of $n_e(\boldsymbol{r})$ that can in principle be used to predict state-of-the-art electron densities.

7.3 Mean spherical baseline

When compared with the density decomposition adopted in the previous Chapter, the RI-auxiliary basis carries the further advantage of describing the full, rather than pseudo-valence, $n_e(\boldsymbol{r})$. The direct application of the scalar-field regression framework discussed in Sec. 6.3 to the all-electron density of a molecule would imply that a great portion of the learning effort is spent on capturing the core-electron density peaks close to the atomic nuclei. Given that the behaviour of these peaks is mostly determined by the nuclear charge [165], they encode a piece of information that is merely constant across the dataset. Therefore, it is convenient to set a baseline value to the reference densities and let the regression focuses on the sole chemically driven fluctuations of $n_e(\boldsymbol{r})$. Here, the baseline is chosen such as considering, for each atomic type a and radial function n, the average of the spherical components of the density over the training set, i.e., $\bar{c}_{an00}$. These average components build up an effective density field $\bar{n}_e(\boldsymbol{r})$ given by the superposition of spherically symmetric contributions. One can then build a sparse vector $\bar{\boldsymbol{c}}$ that has the spherical components $\bar{c}_{an00}$ as the only non-zero entries, and use it to compute the basis set projections of

$\bar{n}_e(\boldsymbol{r})$ as $\bar{\boldsymbol{w}} = \boldsymbol{S}\bar{\boldsymbol{c}}$. The resulting vector of baselined density projections $\Delta\boldsymbol{w} = \boldsymbol{w} - \bar{\boldsymbol{w}}$ can then be used as the actual learning target. Once the learning is carried out as in Eq. (6.5), the interpolated RI-density can finally be obtained by adding the mean spherical density components $\bar{c}_{an00}$ back to the predicted differences of expansion coefficients $\Delta\boldsymbol{c}$. Note that the recipe just discussed represents the scalar field analog of the baseline strategy already seen in the context of predicting the isotropic ($\lambda = 0$) ISC of the polarizability tensor.

7.4 Bio-fragment dataset

In this study, we consider a dataset of molecular dimers selected from the side-chain side-chain interaction (SSI) subset of the BioFragment Database (BFDb) [166]. The original set is made of 3558 dimers formed by amino-acids side-chain fragments taken from 47 different protein structures. Dimers with more than 25 atoms as well as those containing sulfur atoms were not considered. While the total number of sulfur-containing structures is too small to enable the machine-learning model to accurately capture its rich chemistry, the inclusion of the larger systems does not increase dramatically the chemical diversity of the dataset. The final dataset contains a total of 2291 dimers.

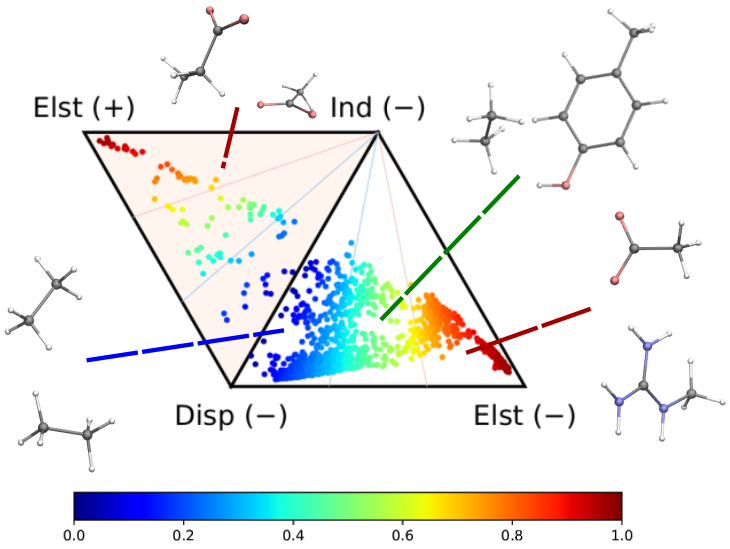


Figure 7.1 – Ternary diagram representation of the symmetry-adapted perturbation theory (SAPT) attractive components of the dimer interaction energies for the 2291 systems considered in this work. The values are taken from Ref. [166].

As shown in Fig. 7.1, the complete set of 2291 dimers spans a large variety of dominant interaction types, ranging from purely dispersion dominated complexes (in blue)

to mixed-influence (green and yellow) to hydrogen-bonded and charged systems (red). We retain the same classification criteria as in the original database to attribute the nature of the dominant interaction. For each dimer, the reference all-electron density has been computed at the DFT/ωB97X-D level using the JK-fit cc-pVQZ basis set [163], henceforth RI-cc-pVQZ, where the resolution of the identitiy approximation is adopted for both the Coulomb and exchange potential (JK). This implies that auxiliary functions up to $\lambda = 5$ are included for the C, N and O atoms, while auxiliary functions up to $\lambda = 4$ are used for hydrogens.

7.5 Learning results

λ-SOAP kernels are once again generated with a radial cutoff of r_c=4 Å and a non-linearity degree of $\zeta = 2$. The training set for the density-learning model was chosen by randomly picking 2000 dimers out of a total of 2291. The remaining 291 were used to test the accuracy of the predictions. Given the tremendous number of possible atomic environments (~40 000) associated with such a chemically diverse database, a subset of M representative environments was selected via FPS to reduce the dimensionality of the regression problem. To assess the consequences of this dimensionality reduction, the learning exercise was performed on three different sizes $M = \{100, 500, 1000\}$. The collective error made in the density-predictions is measured as the cumulative integrated mean absolute difference expressed as a fraction of the total number of electrons included in the test set, i.e.,

$$\varepsilon(\%) = 100 \times \frac{1}{\sum_A N_e(A)} \sum_A \int d\boldsymbol{r} \left| n_e^{\mathrm{ML}}(A;\boldsymbol{r}) - n_e^{\mathrm{QM}}(A;\boldsymbol{r}) \right| . \tag{7.8}$$

Figure 7.2 summarizes the performance of the machine learning algorithm, expressed in terms of the mean absolute difference between the predicted and *ab-initio* densities reconstructed on the RI auxiliary-basis. As shown in the first panel of Fig. 7.2, 100 training dimers were sufficient to reach saturation of the density error around 0.5% for M=100. This result already outperforms the level of accuracy reached in our previous work, which is remarkable given the large chemical diversity of the dataset and the consideration of all-electron densities. Learning curves obtained with M=500 and M=1000 show steeper slopes, approaching saturation at about 2000 training dimers with errors that were reduced to ~0.2-0.3% of the cumulative electronic charge. The predicted full-electron densities are hence five times more accurate than the valence-only predictions (~1%) reported in Ref. [60]. The second panel of Fig. 7.2 reports a more detailed analysis of the M=1000 learning curve as a function of the nature of the dominant interaction between the monomers. Specifically, stronger non-local

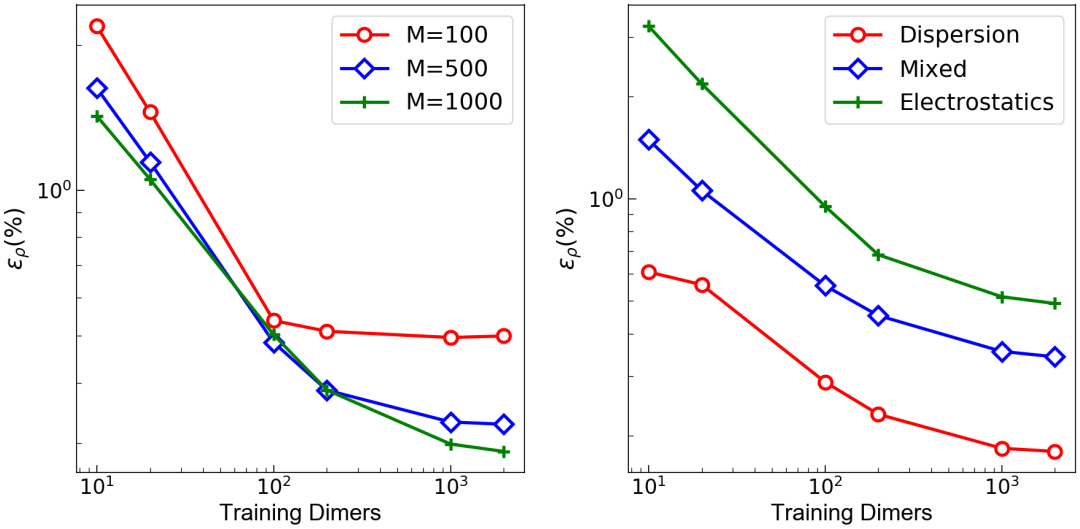


Figure 7.2 – Learning curves. (*left*) weighted mean absolute percentage error (ϵ_ρ(%)) of the predicted SA-GPR densities as a function of the number of training dimers. The weights correspond to the number of electrons in each dimer and the normalization is defined by the total number of electrons. Color code reflects the number of reference environments. (*right*) ϵ_ρ(%) of the predicted SA-GPR densities (M=1000) divided per dominant contribution to the interaction energy according to Ref. [166].

character in the interaction yields a larger error. This is especially prevalent for dimers dominated by electrostatic interactions, which are characterized by errors that are twice as large as those found in other regimes. The origin of this slow convergence arises from two factors. First, only about 20% of the dimers are dominantly bound by electrostatics. The priority of the regression model is thus to minimize the error on the other classes. Second, there is a fundamental dichotomy between the local nature of our symmetry-adapted learning scheme and the long-range nature of the interactions. In this respect, a global ML representation of the density field would be more suitable, but this would imply renouncing the scalability and transferability of the model.

7.6 Density-derived interaction indexes

The fundamental advantage of setting the electron density as the machine-learning target is the broad spectrum of chemical properties that are directly derivable from $n_e(\boldsymbol{r})$. For instance, the predicted charge densities are the key ingredient in density-dependent scalar fields aimed at visualizing and characterizing interactions between atoms and molecules in real space [137]. Routinely used examples include the quan-

tum theory of atoms in molecules (QTAIM) [167, 168], the density overlap region indicator (DORI) [136], and the non-covalent interaction (NCI) index [135, 169]. Figure 7.3 shows an example of the DORI indicator for representative dimers. Compared to the rather featureless $n_e(\boldsymbol{r})$, DORI reveals fine details of electronic structure, which constitute a more sensitive probe for the quality of the machine-learning predictions. In particular, it reveals density overlaps (or clashes) associated with bonding and non-covalent regions on equal footing through the behavior of the local wave-vector, $\nabla n_e(\boldsymbol{r})/n_e(\boldsymbol{r})$ [170–172].

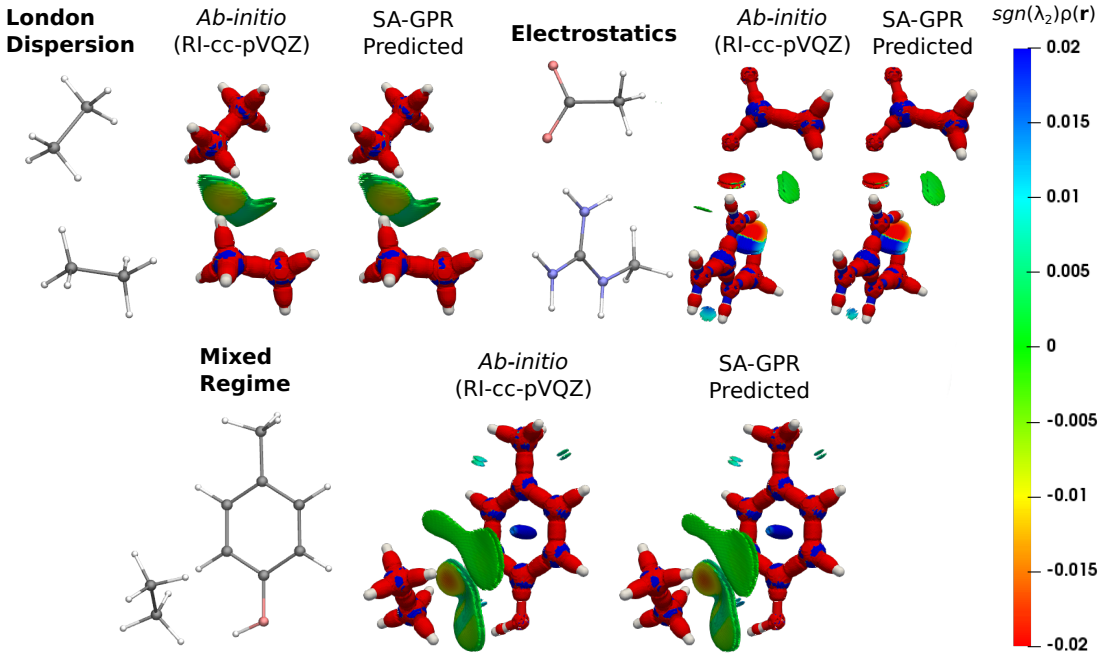


Figure 7.3 – DORI maps of representative dimers for each type of dominant interaction (DORI isovalue: 0.9). Isosurfaces are color-coded [135] with $sgn(\lambda_2)n_e(\boldsymbol{r})$ in the range from attractive -0.02 a.u. (red) to repulsive 0.02 a.u. (blue). In particular, $sgn(\lambda_2)n_e(\boldsymbol{r}) < 0$ characterizes covalent bonds or strongly attractive NCIs (e.g., H-bonds); $sgn(\lambda_2)n_e(\boldsymbol{r}) \sim 0$ indicates weak attractive interactions (van der Waals); $sgn(\lambda_2)n_e(\boldsymbol{r}) > 0$ repulsive NCIs (e.g., steric clashes).

As shown in the Figure, the intra- and intermolecular DORI domains obtained with the SA-GPR densities are indistinguishable from those in the *ab initio* maps. This performance is especially impressive for the density clashes associated with low density values, as is typical for the non-covalent domains. All the features are well captured by the predicted densities ranging from large and delocalized basins typical of the van der Waals complexes (in green) to the compact and directional domains typical of electrostatic interactions, to intramolecular steric clashes, e.g., phenol,

mixed regime. Overall, these results illustrate that the residual 0.2% error does not significantly affect the density amplitude in the valence and intermolecular regions that are accurately described by the SA-GPR model. The highest amplitude errors are concentrated near the nuclei in the region dominated by the core-density fluctuations.

7.7 Electrostatic potential

The versatility of the machine-learning prediction is further illustrated by using the predicted densities to compute the molecular electrostatic potential (ESP) for the same representative dimers (Figure 7.4). ESP maps based on predicted densities agree quantitatively with the *ab initio* reference and correctly attribute the sign and magnitude of the electrostatic potential in all regions of space. Importantly, the accuracy of the ESP magnitude remains largely independent of the dominant interaction type. This is especially relevant for charged dimers (electrostatics) as it demonstrates that despite slower convergence of the learning curve for this category, the achieved accuracy of the model is sufficient to describe the key features of the electrostatic potential.

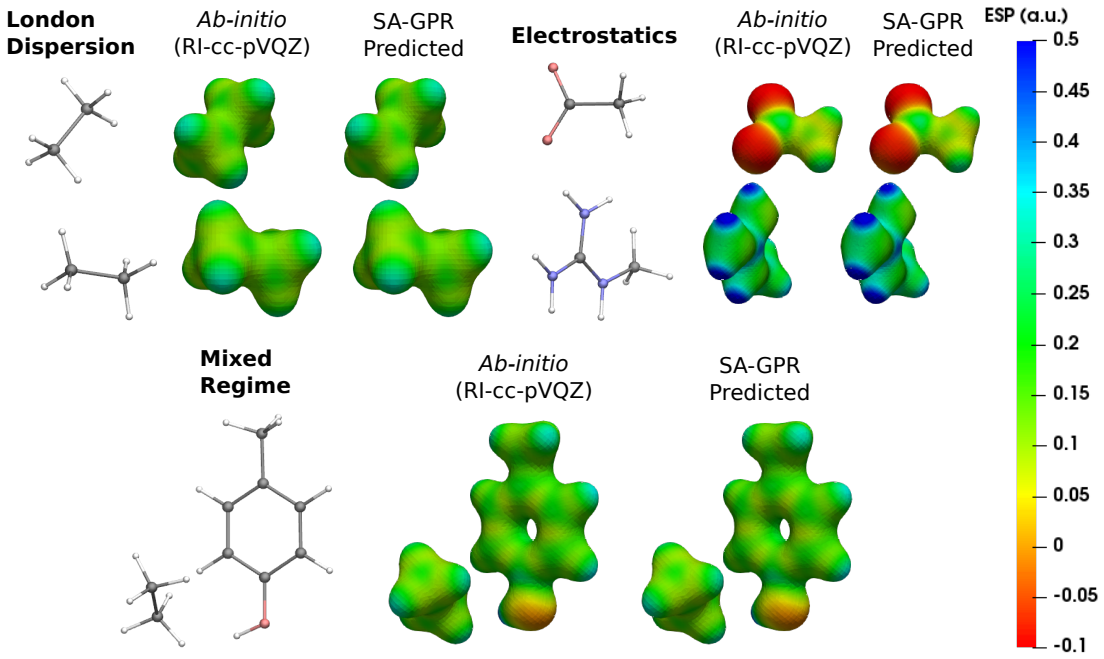


Figure 7.4 – Electrostatic potential (ESP) maps of representative dimers for each type of dominant interaction (density isovalue: 0.05 e^- $Bohr^{-3}$). ESP potential is given in Hartree atomic units (a.u.).

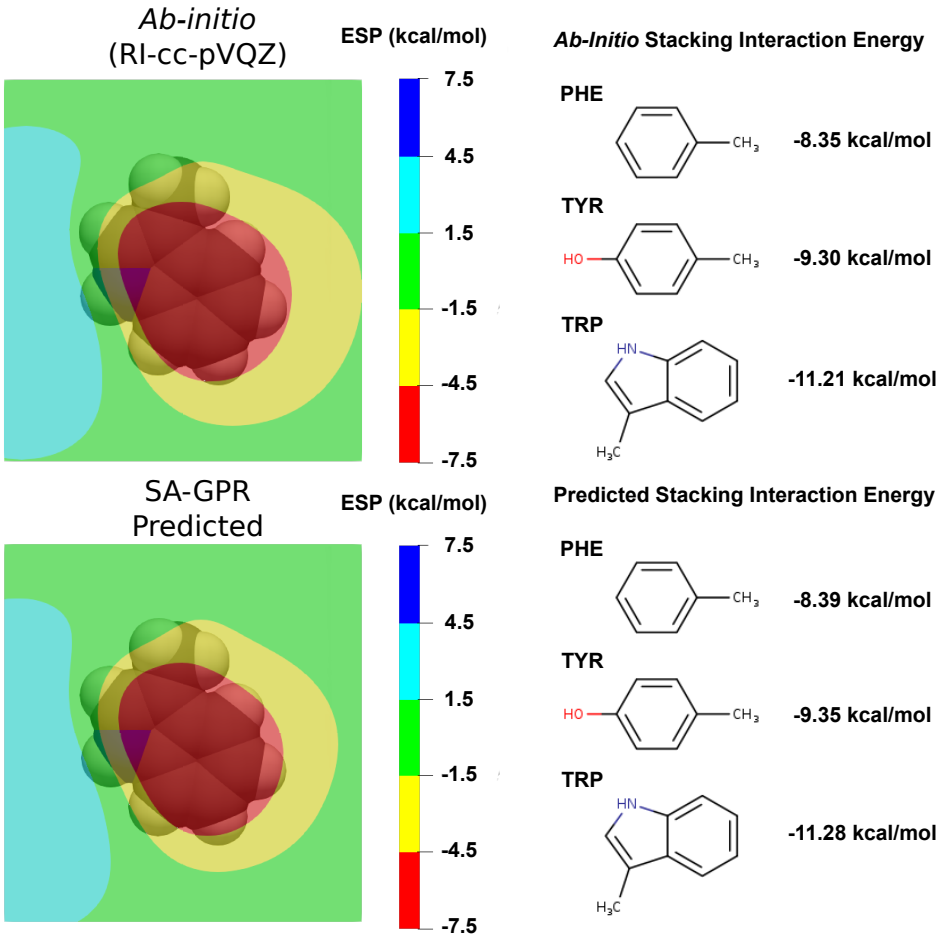


Figure 7.5 – (*left*) Electrostatic potential maps 3.25 Å above the plane of the tryptophan (TRP) side-chain. The van der Waals volume of TRP is represented in transparency. (*Right*) Stacking interaction energies of TRP with the phenylalanine (PHE), tyrosin (TYR) and tryptophan (TRP) side-chains computed as detailed in Ref. [173].

The most widespread applications of ESP maps exploit qualitative information, e.g., identification of the molecular regions most prone to electrophilic/nucleophilic attack, but the electrostatic potentials can be related to quantitative properties such as the degree of acidity of hydrogen bonds and the magnitude of binding energies [173–176]. As a concrete example related to structure-based drug design, we used a recent model that estimates the strength of the stacking interactions between heterocycles and aromatic amino acid side-chains directly from the ESP maps [173]. This model derives the stacking energies of drug-like heterocycles from the maximum and mean value of their ESP within a surface delimited by molecular van der Waals volume (at 3.25 Å above the molecular plane). Following this procedure, we used the ESP derived from the ML predicted densities to compute the binding energies between a representative

heterocycle included in our dataset, the tryptophan side-chain, and the three aromatic amino acid side-chains (Figure 7.5).

Comparison between *ab initio* and ML predicted stacking interaction energies shows that the deviations in the ESP maps lead to minor errors on the order of 0.05 kcal/mol. The largest deviations in the ESP would appear further away from the molecule, beyond the region exploited for the computation of the energy descriptors, i.e., the sum of the atomic van der Waals radii. This behavior can be understood in relation to the propagation of the error made in the density predictions $\delta n_e(\boldsymbol{r})$ to the Fourier components of the electrostatic potential $\delta\hat{V}(\boldsymbol{k}) = 4\pi\delta\hat{n}_e(\boldsymbol{k})/k^2$. In particular, the error associated with the slow-varying components of $n_e(\boldsymbol{r})$ is greatly amplified as $k \to 0$, yielding predictions of $V(\boldsymbol{r})$ that show larger errors in the smooth far-field regions.

7.8 Electrostatic energy

Integration of the electrostatic potential with the all-electron density field gives access to the *ab-initio* electrostatic energy of any given molecule. This is defined as follows

$$U_{\text{ele}}[n_e] = \frac{1}{2}\sum_{i\,j\neq i}\frac{Z_i Z_j}{|\boldsymbol{R}_j - \boldsymbol{R}_i|} + \frac{1}{2}\int d\boldsymbol{r}\int d\boldsymbol{r}'\,\frac{n_e(\boldsymbol{r})n_e(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|} - \sum_i Z_i\int d\boldsymbol{r}\,\frac{n_e(\boldsymbol{r})}{|\boldsymbol{r}-\boldsymbol{R}_i|}. \quad (7.9)$$

The first term corresponds to the (exact) nuclear repulsion energy, the second term to the Hartree energy U_{H} already encountered, while the third term to the electron-nucleus interaction energy U_{en}.

7.8.1 Basis set error

The computation of U_{H} from the RI electron density follows the density-fitting strategy already seen. Since the RI approximation has been derived to explicitly minimize the error on electron repulsion integrals, the error associated with the calculation of the RI Hartree energy (Eq. (7.2)) is negligibly small when compared with its *ab-initio* counterpart, i.e., $\sim 10^{-6}$ kcal/mol. Conversely, the RI construction is not optimized to give minimal error on the electron-nucleus energy U_{en}. As shown in Appendix D, the calculation of this term is simple enough to be carried out analytically. The resulting RI basis set error is enormous, of the order of ~1 kcal/mol per electron. While this error is way too large for any reasonable application, the systematic nature of this basis set driven error is such that it perfectly cancels out when considering electrostatic energy differences, rather than absolute energies. In our particular case, the ultimate goal

consists in using $n_e(\boldsymbol{r})$ to predict the electrostatic interaction between the monomers (which we will label as A and B) included in each molecular dimer, i.e., $U_{\text{int}} = U_{\text{dimer}} - U_{\text{A}} - U_{\text{B}}$. In this scenario, the error incurred in the RI representation of U_{en} does not affect the interaction energy, implying that our predictions of U_{int} can be directly compared with the *ab initio* electrostatic interaction energies.

7.8.2 Prediction error

When computing U_{ele} on top of the predicted $n_e(\boldsymbol{r})$, another crucial error cancellation occurs between the individual predictions of U_{H} and U_{en} . In fact, the two (opposite) terms screen each other out, guaranteeing that the final prediction error δU_{ele} is greatly attenuated with respect to the error associated with the individual predictions δU_{H} and δU_{en}. Formally, the effect of the reciprocal screening on the ML prediction can be understood by considering that, at the first order in the density error $\delta n_e(\boldsymbol{r})$, the error made in the total electrostatic energy reads as follows,

$$\begin{aligned}
\delta U_{\text{ele}}[n_e] &= \delta U_{\text{H}}[n_e] + \delta U_{\text{en}}[n_e] \\
&\approx \int d\boldsymbol{r} \int d\boldsymbol{r}' \, \frac{\delta n_e(\boldsymbol{r}) n_e^0(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|} - \sum_{i=1}^{N} Z_i \int d\boldsymbol{r} \, \frac{\delta n_e(\boldsymbol{r})}{|\boldsymbol{r}-\boldsymbol{R}_i|} \\
&= \int d\boldsymbol{r} \, \delta n_e(\boldsymbol{r}) \left[\int d\boldsymbol{r}' \left(\frac{n_e^0(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|} - \sum_{i=1}^{N} Z_i \, \frac{\delta(\boldsymbol{r}'-\boldsymbol{r})}{|\boldsymbol{r}'-\boldsymbol{R}_i|} \right) \right] \\
&= \int d\boldsymbol{r} \, \delta n_e(\boldsymbol{r}) V^0(\boldsymbol{r}) \,,
\end{aligned} \tag{7.10}$$

with $n_e^0(\boldsymbol{r})$ and $V^0(\boldsymbol{r})$ the *ab initio* electron density and electrostatic potential, respectively.

To effectively predict the electrostatic interaction energies within our dataset, a global learning exercise that contains both the 2000 selected dimers and the corresponding non-interacting monomers has been performed. This procedure has the advantage that the predictions for the dimers and the monomers included in the test set is carried out by using the same set of regression weights $\boldsymbol{x}_M$. As a consequence, the discrepancies in the prediction accuracy between dimers and monomers are supposed to be linearly comparable in the definition of the SOAP representation. Learning curves for the density-based predictions of U_{ele} are reported in Fig. 7.6. As shown in the first panel of the Figure, as few as 10 training structures are enough to bring the error made on the individual predictions of the monomers and dimers well below the intrinsic variability of U_{ele} within the test set. At this training set size, the resulting electrostatic interaction energy shows a large error compensation between monomers and dimers

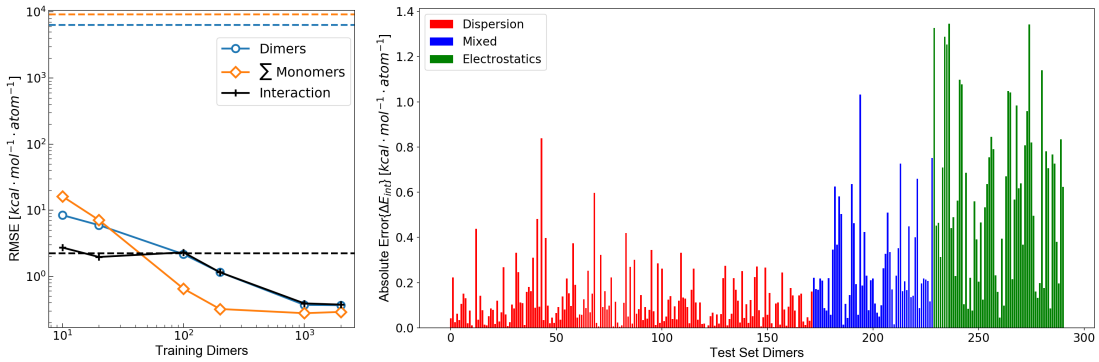


Figure 7.6 – *(left)* Root mean square error (RMSE) of the electrostatic energy for the dimers, the sum of the monomers and their difference (interaction) as a function of the number of dimers in the training set. Dashed lines correspond to the standard deviation of the target property within the test set. *(right)* Absolute error of the interaction electrostatic energy U_{int} for each dimer in the test set divided per dominant interaction type.

predictions, with an error that is roughly equivalent to the intrinsic variability of U_{int} within the test set. Increasing the number of training structures results in a quick saturation of the interaction energy error at about 0.3 kcal/mol per atom. The absolute distribution of errors across the test set is reported in the second panel of Fig. 7.6. Overall, we find that, especially for the electrostatic-driven configurations, the density-based prediction of the molecular electrostatic interaction between the monomers is well above the chemical accuracy of ~1 kcal/mol. These results underscore, once again, that the indirect calculation of the system's energetics through the prediction of $n_e(\boldsymbol{r})$ represents a very challenging task, which would be better tackled by a learning model that is specifically constructed to predict U_{int} directly.

7.9 Density extrapolation on polypeptides

If, on the one hand, our ML model requires further optimization to deliver predictions of $n_e(\boldsymbol{r})$ that enable the calculation of chemically-accurate interaction energies, on the other hand, its inherent transferability can be exploited to provide access to density information of large macromolecules, at the sole price of training the model on a sufficiently heterogeneous and chemically diverse dataset. The predictive power of this extrapolation procedure is demonstrated by using the machine-learning model exclusively trained on the 2291 BFDb dimers to predict the electron density of 8 polypeptides taken from the Protein DataBank (PDB) [177]. The performance of the ML model for each macromolecules, labelled by their PBD ID, is reported in Figure 7.7.

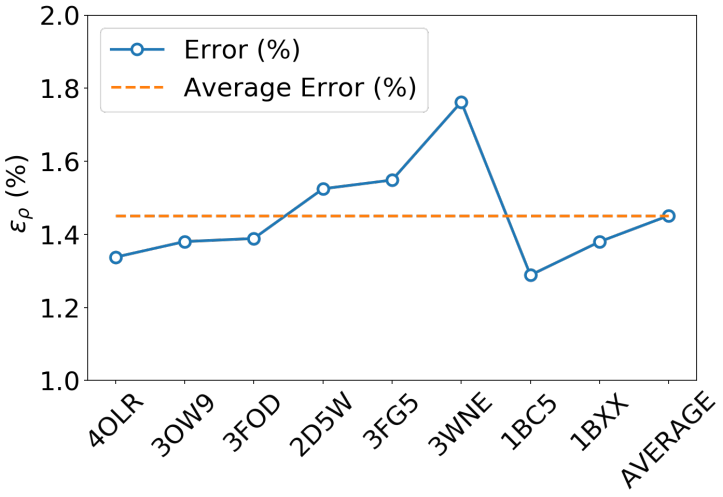


Figure 7.7 – Weighted mean absolute percentage error (ϵ_ρ(%)) of the predicted densities extrapolated for 8 biologically relevant peptides (protein databank ID).

Overall, the predictions lead to a low average error of only 1.5% for the 8 polypeptides, which is in line with the highest density errors obtained on the BFDb test set. Relevantly, the largest discrepancies are obtained for 3WNE, which is the only cyclopeptide of the set. The origin of these differences can be understood by performing a more detailed analysis on a representative polypeptide, the leu-enkephalin (4OLR). The errors in this percentage range do not affect the density-based properties, such as the spatial analysis of the intramolecular interactions with scalar fields (Figure 7.8 top right panel). Yet, most of the discrepancies occur along the amino acid backbone (Figure 7.8 lower panels), which is especially sensitive for the more strained 3WNE cyclopeptide. Although similar chemical environments were included in the training set, the error is determined by the lack of an explicit peptide bond motif and cyclopeptides in the training set. While this limitation could be addressed by *ad hoc* modification of the training set, the overall performance of the machine-learning model is rather exceptional as it provides in only a few minutes electron densities of DFT quality for large and complex molecular systems.

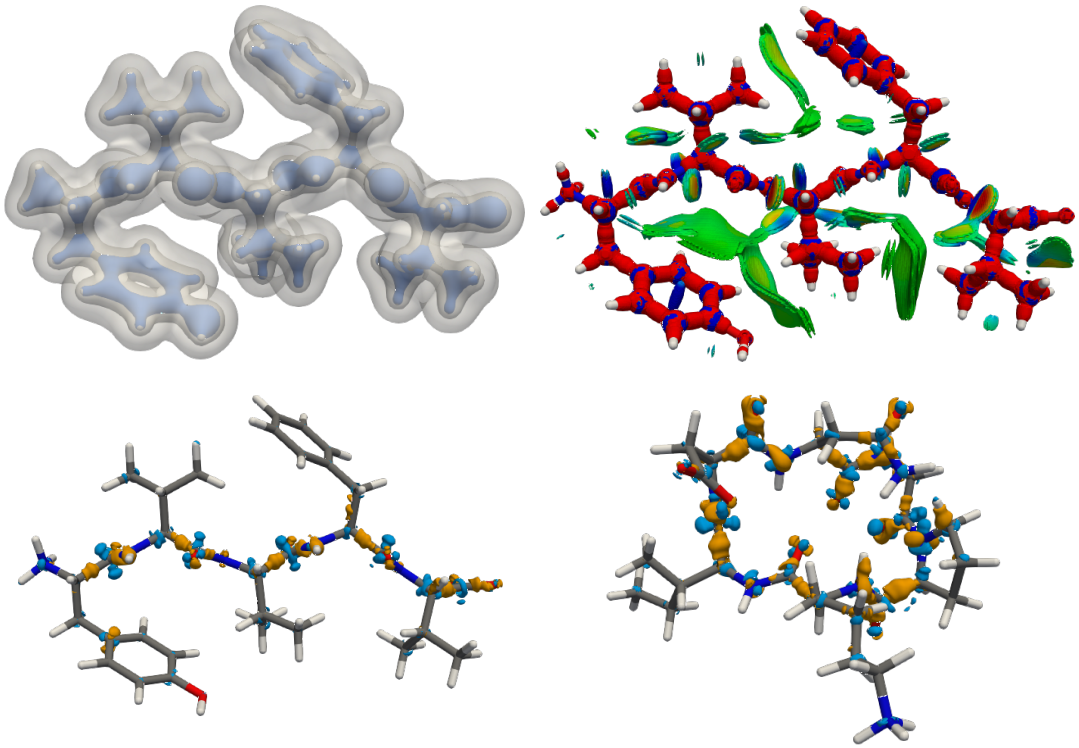


Figure 7.8 – *(top left)* predicted electron density of enkephalin (PBD ID: 4OLR) at three isovalues: 0.25, 0.1, and 0.01 e^- $Bohr^{-3}$. *(top right)* DORI map of enkephalin (DORI isovalue: 0.9) colored by $sgn(\lambda_2)n_e(\boldsymbol{r})$ in the range from -0.02 a.u. (red) to 0.02 a.u. (blue) *(lower left)* density difference between predicted and *ab-initio* electron density (isovalues ± 0.01e^- $Bohr^{-3}$). *(lower right)* density difference between predicted and *ab-initio* electron density of 3WNE (isovalues ± 0.01e^- $Bohr^{-3}$).

Part III Long-range interactions

8 Incorporating long-range physics in atomic-scale machine learning

During the previous chapters, we underscored the importance of three-dimensional symmetries in constructing structural representations of the system and demonstrated their effectiveness in predicting a broad variety of physical observables, including scalars, tensors and scalar fields. All throughout the dissertation, we reported examples that show how the inherent local nature of the machine-learning model plays a crucial role for transferring the information learned in a neighbourhood of the system's atoms across chemically heterogeneous datasets of varying structural complexity. The assumption that lies underneath these results consists in neglecting the effect of long-range and non-local phenomena, such as those related to electrostatics, dispersion and quantum coherence. While this assumption is in many cases valid to a first approximation, there are several contexts in molecular and material science where long-range effects play a leading role in the determining the system's properties. In this chapter, we show how to tackle this problem by deriving a representation of the system where the non-local structural information is evaluated at the local scale, thus preserving the atom-centered and additive nature of the predictions. By doing so, we provide a conceptual framework to incorporate long-range physics into atomistic machine learning, that entirely bypasses the *ad hoc* prescriptions commonly adopted to circumvent the nearsightedness of local representation. Sections and figures are adapted from the following article:

A. Grisafi and M. Ceriotti, "Incorporating long-range physics in atomic-scale machine learning", *The Journal of Chemical Physics* 151, 204105 (2019), with the permission of AIP Publishing. AG contributed to deriving and implementing the LODE method, to produce the results and figures reported and to writing the manuscript.

8.1 Long-range effects in materials science

Long-range phenomena are ubiquitous in materials science [178]. The prototypical example is given by electrolyte solutions, where the pathologically slow decay of Coulomb interactions, $\sim 1/r$, between the ions of the system has a macroscopic influence on the instantaneous properties of the material. At thermodynamic equilibrium, the Coulomb potential is exponentially screened by the statistical distribution of ions, so that, in practice, the spatial extent of correlations in the liquid is determined by the characteristic screening length λ of the electrolyte. This screening effect is associated with an average electrostatic potential that decays monotonically in the regime of low ionic concentrations (Debye-Hückel theory [179]), and that appears as a dumped oscillation beyond a critical concentration value c_0 [180, 181]. In contrast to the screening properties of dilute electrolytes, where the Debye screening length λ_{D} is predicted to decrease with the ionic concentration ($\lambda_{\mathrm{D}} \propto c^{-1/2}$), values of $c > c_0$ are associated with screening lengths λ that *increase* with the electrolyte concentration [182]. This behaviour has been experimentally observed in a large variety of electrolyte solutions for concentrations $c \gtrsim 0.5$M, finding that the electrostatic screening can take place over a length scale that can reach up to ~ 120 times the expected value of λ_{D} [183]. When considering polar solvents like water, this scenario is made even more complicated by the fact that the ionic electric field introduces a long-range dipolar correlation between the solvent molecules [184–187], which have been experimentally found to occur over nanometric scales [188]. Finally, while the thermodynamic properties of pure liquid water can be well reproduced by means of a local (short-range) model, the properties of vapour phases can dependent substantially on the inclusion of a long-range description [189].

Of course, water and ionic systems do not represent the only scenario where long-range effects can be manifested. In the context of electrochemical simulations, for instance, the spontaneous [190, 191], or externally induced [192–195] polarization of an electrodic interface between a metallic surface and a liquid medium is yet another example where non-local effects play a crucial role. Moreover, while electrostatic and polarization phenomena are both examples of interactions that can be represented by means of classical physics, there are a series of long-range effects that have an intrinsic quantum nature. These include dispersion interactions [196], $\sim 1/r^6$, responsible for the stabilization of molecular crystals and biomolecules [197, 198], quantum delocalization effects that are associated with large polarizabilities [58] and nanoscopic charge-transfer phenomena [199], as well as geometric properties of the wave-function that determine the anomalous quantum Hall conductivity [200] and the behaviour of topological insulators [201].

8.2 Machine learning of long-range phenomena

According to the previous discussion, long-range effects occur in many physical contexts and can influence the statistical correlations between the atoms of the system over several nanometers. In these circumstances, the local nature of the atomic environments commonly used in machine-learning approximations, such as symmetry functions [202], SOAP [42], SNAP [203], MTP [204], ACE [205] and NICE [54], reflects a fundamental limit to the accuracy of the regression model.

When it comes to electronic energies, the problem of including long-range effects can be tackled by explicitly separating the local quantum many-body contribution to the total energy from a more or less parametrized reference term that includes the long-range properties of the system (Δ-learning). This can be done either by using inexpensive methods, e.g., Hartree-Fock and density functional tight binding, as a baseline for the learning exercise [94, 206, 207], or by directly subtracting the classical electrostatic contribution to the total energy [25, 208]. Δ-learning approaches have also found applications in the description of the cohesive energy in homogeneous and isotropic bulk systems, where the long-range dispersion tails of the interatomic potential are introduced by means of van der Waals corrections [11, 209].

A different approach to tackle the problem involves a two-step process, where a parallel regression is performed to predict, in turn, the ingredients that enter the calculation of the long-range interaction. This is the case, for instance, of the partial charges and atomic multipoles that determine the long-range electrostatics of the system [32, 63, 210–213]. More sophisticated models rely instead on a charge equilibration scheme that makes use of predicted atomic electronegativities to compute the partial charges of the system by minimizing a quadratic functional form of the electrostatic energy [214–216].

Beyond electronic energies, the breakdown of a local machine learning model is particularly pronounced when dealing with intrinsically non-local quantities like the dielectric response of a condensed-phase medium [57]. This non-locality has to do both with the effect of the far-field electrostatics [217], and to the topological geometric nature of the macroscopic polarization of an infinitely extended (periodic) material [69]. In this case, the problem can possibly be bypassed by adopting specific physical prescriptions. Examples of this have already been shown in Sec. 3.5, where the dielectric tensor $\boldsymbol{\varepsilon}$ of liquid water is learned indirectly by building a model for an effective molecular polarizability that is mapped to $\boldsymbol{\varepsilon}$ through the Clausius-Mossotti relationship [217]. In the context of reproducing the autocorrelation function of the macroscopic polarization of liquid water, another strategy has instead been adopted,

where the selected learning targets are the positions of the Wannier centers [218] that are used to recast the electron density of the system into a set of point-charges [219]. Finally, when addressing the prediction of the dipole moment in molecular systems, combining a tensorial representation, such as $\lambda = 1$ SOAP, with a model that explicitly breaks down the dipole in its classical atomic contributions, i.e., $\boldsymbol{\mu} = \sum_i q_i \boldsymbol{r}_i$, has demonstrated to be essential to capture the non-local character of the dipole in zwitterionic molecular chains [220].

By and large, the learning models previously described tackle the problem of including long-range phenomena by making use of an *ad hoc* definition of the electrostatic energy, or dielectric response, in terms of local atomic quantities. Although successful, these kind of approaches have the downside of being very system dependent and, as such, hardly transferable across systems that have a different nature. Capturing long-range effects without any prior assumption on the nature of the learning target is a difficult task to accomplish with the methods currently available. Most of the approaches that have explicitly attempted to do so, such as Coulomb kernels [38], many-body tensor representations [221], or multi-scale wavelet invariants [222], are built upon a global representation of the system rather than on an additive atom-centred model. In what follows, we propose a simple, yet elegant, solution to this problem, where the non-local character of the target property is incorporated through an atomic Coulomb-like potential field evaluated at the local scale.

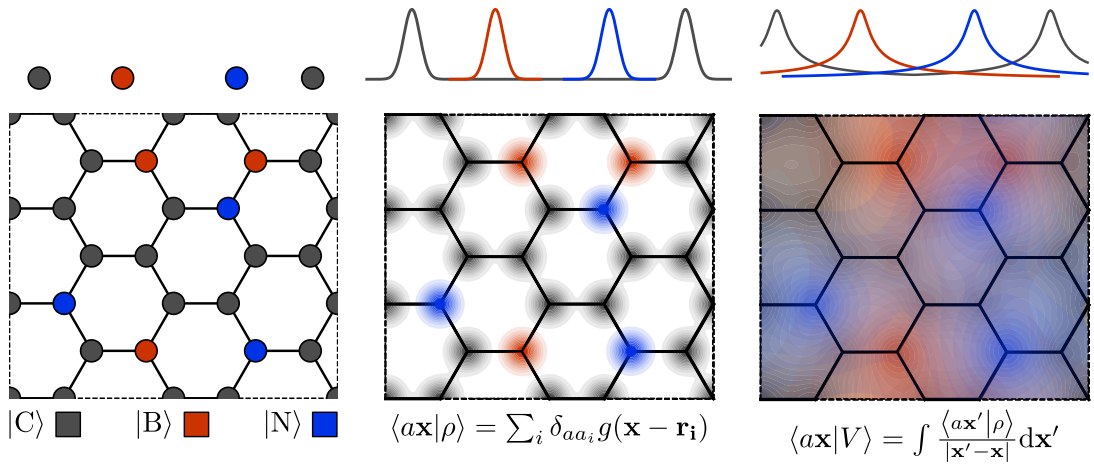


Figure 8.1 – Schematic representation of the construction of potential-field representation for $p = 1$. Top: 1D atomic chain; bottom: hypothetical 2D system. (*left*) Atoms are represented by their position in Cartesian coordinates.(*middle*) the structure is represented as an atom-density field; each element is associated with a separate channel, represented by color coding. (*right*) the atomic potential field generated by the decorated atom density.

8.3 Long-distance equivariant representations

Let us start from the formal definition of a global atom-density representation, as already reported in Eq. (1.13), i.e., $\langle a\boldsymbol{x}|\rho\rangle = \sum_i g_\sigma(\boldsymbol{x}-\boldsymbol{r}_i)\delta_{a_i a}$. We now introduce a generic potential-field representation through the following integral transformation applied to the density field:

$$\langle a\boldsymbol{x}|V^{(p)}\rangle = \int d\boldsymbol{x}' \frac{\langle a\boldsymbol{x}'|\rho\rangle}{|\boldsymbol{x}'-\boldsymbol{x}|^p}. \tag{8.1}$$

The rationale behind this transformation is that, whereas $\langle a\boldsymbol{x}|\rho\rangle$ contains information only about the atoms in the vicinity of $\boldsymbol{x}$, the non-local nature of the potential fields $\langle a\boldsymbol{x}|V^{(p)}\rangle$ can be used to encode information about *all* atoms of the system, with an asymptotic dependence on the position of the i-th atom that follows an algebraic decay $|\boldsymbol{x}-\boldsymbol{r}_i|^{-p}$. A graphical representation of this construction is reported in Fig. 8.1 for the special case of $p=1$. The physical significance of the potential field $|V^{(p)}\rangle$ becomes obvious when considering specific interatomic interactions. For instance, if we replaced the atom-density field with an actual charge density, the $p=1$ case would correspond to the electrostatic potential of the system. Similarly, under a different interpretation of the atom-density field, the $p=6$ case recalls the asymptotic limit of the exchange-correlation energy per particle [223] that underlies the definition of dispersion interactions [196].

Given the translational invariant nature of the non-local kernel $|\boldsymbol{x}'-\boldsymbol{x}|^{-p}$, imposing the translational symmetry to Eq. (8.1) follows a derivation similar to the one reported in Sec. 1.5.1. In particular, one can build a two-body correlation function that comes from the translational convolution of the tensor product between $|V^{(p)}\rangle$ and an atom density representation $|\rho\rangle$, i.e.,

$$\begin{aligned}
&\langle a_1\boldsymbol{x}_1; a_2\boldsymbol{x}_2|\rho\otimes V^{(p)}\rangle_{\hat{t}} = \int d\hat{t}\,\langle a_1\boldsymbol{x}_1|\,\hat{t}\,|\rho\rangle\,\langle a_2\boldsymbol{x}_2|\,\hat{t}\,|V^{(p)}\rangle \\
&= \sum_{ij}\delta_{a_i a_1}\delta_{a_j a_2}\int d\boldsymbol{t}\, g_{\sigma_1}(\boldsymbol{x}_1-\boldsymbol{r}_i+\boldsymbol{t})\int d\boldsymbol{x}_3 \frac{g_{\sigma_2}(\boldsymbol{x}_3-\boldsymbol{r}_j)}{|\boldsymbol{x}_2-\boldsymbol{x}_3+\boldsymbol{t}|^p} \\
&= \sum_{ij}\delta_{a_i a_1}\delta_{a_j a_2}\int d\boldsymbol{s}\,\frac{1}{|\boldsymbol{x}_2-\boldsymbol{s}|^p}\int d\boldsymbol{t}\, g_{\sigma_1}(\boldsymbol{x}_1-\boldsymbol{r}_i+\boldsymbol{t})\, g_{\sigma_2}(\boldsymbol{s}-\boldsymbol{r}_j+\boldsymbol{t}) \\
&= \sum_{ij}\delta_{a_i a_1}\delta_{a_j a_2}\int d\boldsymbol{s}\,\frac{g_\sigma((\boldsymbol{x}_1-\boldsymbol{s})-(\boldsymbol{r}_i-\boldsymbol{r}_j))}{|\boldsymbol{x}_2-\boldsymbol{s}|^p} \\
&= \sum_{ij}\delta_{a_i a_1}\delta_{a_j a_2}\int d\boldsymbol{x}'\,\frac{g_\sigma(\boldsymbol{x}'-\boldsymbol{r}_{ij})}{|\boldsymbol{x}'-\boldsymbol{x}|^p},
\end{aligned} \tag{8.2}$$

where we used the twofold change of variables $\boldsymbol{s} \to \boldsymbol{x}_3 + \boldsymbol{t}$ and $\boldsymbol{x}' \to \boldsymbol{x}_1 - \boldsymbol{s}$, and we set $\boldsymbol{r}_{ij} = \boldsymbol{r}_i - \boldsymbol{r}_j$ and $\boldsymbol{x} = \boldsymbol{x}_1 - \boldsymbol{x}_2$ in the last equality. Note that, once again, the width σ of the Gaussian function $g_\sigma(\boldsymbol{x}' - \boldsymbol{r}_{ij})$ is defined from the Gaussian convolution properties as $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$. Proceeding as in Sec. 1.5.1, from Eq. (8.2) we can then single out an atom-centered representation defined as follows:

$$\left\langle a\boldsymbol{x} \middle| V_i^{(p)} \right\rangle = \int d\boldsymbol{x}' \frac{\sum_j \delta_{a_j a} g_\sigma(\boldsymbol{x}' - \boldsymbol{r}_{ij})}{|\boldsymbol{x}' - \boldsymbol{x}|^p} = \int d\boldsymbol{x}' \frac{\langle a\boldsymbol{x}' | \rho_i \rangle}{|\boldsymbol{x}' - \boldsymbol{x}|^p}. \tag{8.3}$$

Similarly to the atom-density case, the potential-field information can be localized in a neighbourhood of the central atom i using a cutoff function of radius r_c. Especially for small values of p, however, the integral of Eq. (8.3) introduces a substantially non-local behavior in the definition of the potential field, which makes the atom-centered representation aware of the structural variations that occur over the *entire* system. In other words, thanks to the algebraic decay of the potential tails, the representation $\left\langle a\boldsymbol{x} \middle| V_i^{(p)} \right\rangle$ can in principle depend on the positions of atoms that are located arbitrarily far from $\boldsymbol{r}_i$, in fact beyond the selected cutoff distance r_c. In this regard, the construction previously introduced lies in between a global and local description of the system, as it incorporates the far-field information while still retaining the atom-centered and additive nature characteristic of transferable learning models.

From of Eq. (8.3), the derivation of potential-field representations of arbitrary body-order ν that are adapted to rotational and inversion symmetry follows the very same discussion reported in Sec. 1.5 for the atom-density case. We will refer from now on to the resulting class of atomistic representations as the *long-distance equivariant* (LODE) framework. In the following, we will focus on the important case of $p = 1$.

8.3.1 2-body LODE and points-charge limit

2-body ($\nu = 1$) spherical invariants come from performing the rotational average on Eq. (8.3). In this case, it is particularly instructive to take the δ-Dirac limit corresponding to $\sigma \to 0$ in the definition of the atom-density field, i.e.,

$$\langle a\boldsymbol{x} | V_i^{\otimes 1} \rangle \to \sum_j \delta_{a_j a_1} \frac{1}{|\boldsymbol{x} - \boldsymbol{r}_{ij}|}, \tag{8.4}$$

which yields the 2-body invariant features

$$\left\langle ax00 \middle| \overline{V_i^{\otimes 1}} \right\rangle \to \sum_j \delta_{a_j a} \min\left[\frac{1}{x}, \frac{1}{r_{ij}}\right]. \tag{8.5}$$

Clearly, Eq. (8.5) simply sums up pairwise Coulomb interaction terms of the kind $1/r_{ij}$ over all atoms *outside* the region over which the potential-field is computed. Ignoring the contribution from the atoms within the cutoff, that can be better characterized by other atomic structure representations, a linear model built on these features is hence equivalent to a fixed point-charge electrostatic model. In particular, upon promoting the valence of the central atom i in the feature space, the linear regression weights associated with each pair of chemical species a_i and a_j can be interpreted as the product of atomic charges q_{a_i} and q_{a_j}. While this construction is very revealing, it is clear that its descriptive power carries the same limitations of the fixed points-charge models routinely used in atomistic simulations [83].

8.3.2 3-body LODE features

Following the same derivation of Sec. 1.5.3, increasing the order of structural correlations up to 3-body ($\nu = 2$) yields the following rotational invariant features:

$$\left\langle a_1 x_1 l; a_2 x_2 l \middle| \overline{V_i^{\otimes 2}} \right\rangle = \sum_m \langle a_1 x_1 l m | V_i \rangle^\star \langle a_2 x_2 l m | V_i \rangle \,, \tag{8.6}$$

with $\langle a x l m | V_i \rangle$ the spherical harmonics components of the potential field. Its real-space counterpart underlies a construction that is schematically represented in Fig. 8.2. The extension to higher body-orders $\nu > 2$ and/or to rotationally covariant representations of a given spherical-tensor order $\lambda > 0$ is straightforward based on the analogous density-based counterparts. Note that it is also possible to compute representations that combine different values of p, and even $p = 0$, corresponding to the atom-density field. These possibilities are investigated in more details in the next chapter.

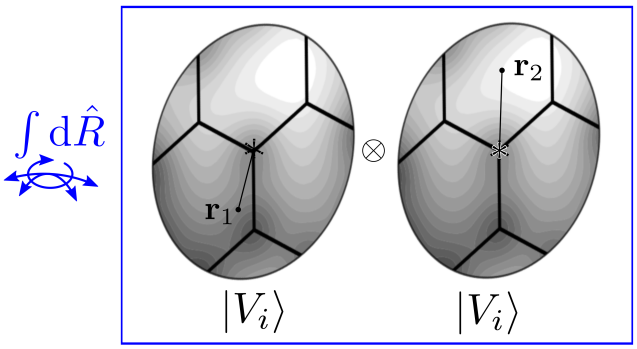


Figure 8.2 – 3-body LODE correlations arise from the rotational average of a pair of smeared Coulomb potentials sampled in a local environment of a given atom i.

8.4 Calculation of potential harmonics projections

The calculation of the LODE features of Eq. (8.6) requires to first compute the spherical harmonics projections of the potential field, $\langle axlm|V_i\rangle$. As for the density case, it is convenient to expand the radial degrees of freedom over orthogonal radial functions $R_n(r)$, so that in practice one is required to compute coefficients of the kind $\langle anlm|V_i\rangle$. As presented in details in Appendix C, for molecules and clusters the spherical harmonics projections $\langle anlm|V_i\rangle$ can be computed conveniently in real space, by numerical integration on appropriate atom-centred grids. For a bulk (infinite) systems, described by a periodically-repeated supercell, the slow decay of the Coulomb potential would instead make the real-space computation prohibitive. This is exactly the same problem one faces when evaluating electrostatic interactions in the condensed phase, and fortunately it has long been solved, e.g., with the many techniques based on the use of a plane-waves auxiliary basis [224]. Consider in particular the plane-wave definition as $\langle \boldsymbol{x}|\boldsymbol{k}\rangle \equiv e^{i\boldsymbol{k}\cdot\boldsymbol{x}}$, with $\boldsymbol{k}$ representing the wave-vectors that are compatible with the simulation box. Starting from a smooth, Gaussian atom density, means that in practice one needs only a manageable number of plane waves. In particular, the width σ of the Gaussian density determines the minimum wavelength that should be introduced in the the plane-wave expansion, so that $\boldsymbol{k}$-vectors only need to be generated within a sphere of radius $k_{\max}$ of the order of $2\pi/\sigma$. In order to evaluate the orthogonal potential projections, it is then enough to include the identity resolution $\sum_{\boldsymbol{k}} |\boldsymbol{k}\rangle\langle\boldsymbol{k}|$ within the definition of the coefficients $\langle anlm|V_i\rangle$, i.e.,

$$\begin{aligned}\langle anlm|V_i\rangle &= \sum_{\boldsymbol{k}\neq\boldsymbol{0}} \langle nlm|\boldsymbol{k}\rangle\langle a\boldsymbol{k}|V_i\rangle \\ &= \sum_{\boldsymbol{k}\neq\boldsymbol{0}} \left[4\pi\, i^l I_{nl}(k)\, Y_{lm}^{\star}(\hat{\boldsymbol{k}})\right] \left[\frac{1}{\Omega}\left(\sum_j \delta_{a_j a} e^{-i\boldsymbol{k}\cdot\boldsymbol{r}_{ij}}\right)\frac{4\pi}{k^2} e^{-\frac{k^2\sigma^2}{2}}\right]\end{aligned} \tag{8.7}$$

where Ω the volume of the simulation box, and k and $\hat{\boldsymbol{k}}$ are, respectively, the modulus and direction of the wavevector. As detailed in Appendix C, the bracket $\langle nlm|\boldsymbol{k}\rangle$ realizes the expansion in plane waves of the local environment basis, and can be computed analytically once and for all if the radial functions are taken to be Gaussian type orbitals. Conversely, $\langle a\boldsymbol{k}|V_i\rangle$ represent the Fourier components of the potential generated by the Gaussian density of type a for the entire system, and can be readily computed analytically [5]. As a result, the geometric local nature of the representation of Eq. (8.7) is formally factorized from its system-dependent global character. In fact, Eq. (8.7) could also be used to compute efficiently the coefficients of the atom-density expansion that enter, for instance, the SOAP framework. Note that the $\boldsymbol{k}$=$\boldsymbol{0}$ component can be safely excluded from the sum of Eq. (8.7): this is in fact equivalent to the application of a charge-neutrality constraint while solving the regression problem.

8.5 Random gas of point charges

We begin testing the LODE representations by considering a toy system made of randomly distributed point-charges in a cubic box that is infinitely repeated in the three dimensions using periodic boundary conditions. The number of positive charges is equal to the number of negative charges, so that the system is overall neutral. To limit the amplitude of energy fluctuations, we discard configurations in which two charges are closer together than 2.5 Å. Following these prescriptions, we generate a total of 2000 configurations, each of which contains 64 atoms in cubic boxes spanning a broad range of densities, with side lengths between 12 and 20 Å. For each of these configurations, we compute the electrostatic energy using the Ewald method, as implemented in LAMMPS [225]. Fig. 8.3 compares the learning performance obtained using a 3-body SOAP ($\nu = 2$) representation with different cutoffs, to the one obtained using 2-body ($\nu = 1$) and 3-body ($\nu = 2$) LODE features. A Gaussian width of σ=1.0 Å has been used to construct the atom-density that enters both representations.

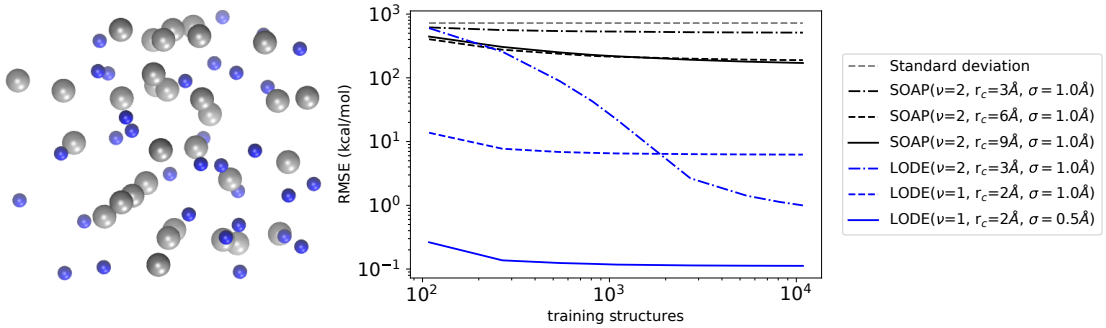


Figure 8.3 – (*left*) snapshot of a random gas configuration of sodium chloride. (*right*) Learning curves in kcal/mol for the electrostatic energy of an idealized random gas of point charges. The model is trained on 1500 randomly selected configurations and tested on other 500 independent configurations. (*black full and dashed lines*) Local SOAP results at environment cutoffs of 3, 6 and 9 Å. (*blue lines*) LODE($\nu = 1$) results at an environment cutoff of 2 Å and Gaussian smearing of 0.5 and 1.0 Å, and LODE($\nu = 2$) results with a cutoff of 3 Å and 1.0 Å.

The figure clearly demonstrates the inefficiency of a local model when attempting to learn a property that is dominated by long-range effects. Given that the training set contains few configurations with atoms closer than 3 Å, the model with r_c=3 Å is almost completely ineffective. Even increasing the cutoff up to 9 Å, a SOAP model barely reaches an accuracy of about 20% RMSE when using the maximum number of training structures, corresponding to an error larger than 100 kcal/mol. On the other hand, upon promoting the species of the central atom to the feature space, a linear model built using the 2-body LODE ($\nu = 1$) representation yields an error that is one

order of magnitude smaller, ∼ 10 kcal/mol, by using a handful of training points. This is not surprising, since the functional form of Eq. (8.5) is formally equivalent to the fixed point-charges interaction as the Gaussian smearing of the atomic density tends to zero ($\sigma \to 0$). In fact, chemically accurate predictions with ∼ 0.1 kcal/mol RMSE can be obtained under halving the Gaussian width down to $\sigma = 0.5$ Å. A 3-body LODE ($\nu = 2$) model, although initially less effective, possesses sufficient descriptive power to reach, and then overcome, the accuracy of the linear $\nu = 1, \sigma = 1$ Å model. This simple example highlights the fundamental difficulty in incorporating long-range physics with a conventional local structure representation, and demonstrates that the LODE features can, on their own, be used as a very efficient description to predict the electrostatic energy of a system of fixed point charges.

8.6 Binding curves of charged dimers

We now consider a more realistic scenario, namely the problem of predicting the binding curves of a dataset of organic molecular dimers that carry an electric charge. We extract 661 different dimers containing H, C, N and O atoms from the BioFragment Database (BFDb) [166], where at least one of the two monomers in each dimer configuration has a net charge. This choice ensures that we focus the exercise on a problem for which permanent electrostatic interactions play a prominent role. Contrary to the toy system previously discussed, however, one cannot expect that a fixed point-charge model would suffice to predict the binding curves. The dataset contains a multitude of chemical moieties, including neutral polar fragments, highly polarizable groups, and provides a realistic assessment of how well a LODE model can perform in practice. For each of the 661 dimers, we consider 13 configurations where the reciprocal distance between the two monomers, defined as the distance between their geometric centers, spans an interval that can go from a minimum of ∼3 Å to a maximum of ∼8 Å. For each of these configurations, unrelaxed binding curves are computed at the DFT/B3LYP level using the FHI-aims quantum-chemistry package [125]. The training dataset is defined by considering the binding curves of the first 600 dimers out of the total of 661, while predictions are tested on the remaining 61. We also include the isolated monomers in the training set, so that the ML model has knowledge of the dissociation limit, and compute a few additional reference energies at larger separations, which are however not used for training. SOAP and LODE representations are defined within spherical environments of $r_c = 3.0$ Å, while the Gaussian width of the density field is chosen to be σ=0.3 and 1.0 Å respectively.

Before carrying out the learning exercise, the reference DFT energies are baselined with respect to the monomer energies, so that the model only has to reproduce the

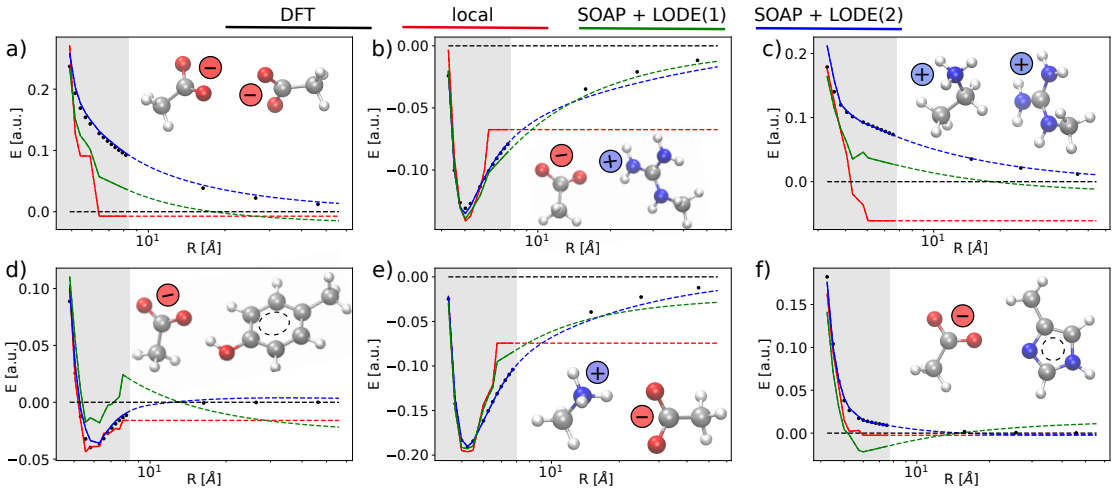


Figure 8.4 – Comparison of reference and predicted binding curves of six molecular dimers. (*black dots*) DFT reference calculations, (*red lines*) local SOAP predictions, (*green lines*) combined SOAP and LODE(1) predictions, (*blue lines*) combined SOAP and LODE(2) predictions. Full lines and shaded background represent the range of distances that is comparable to the geometries included in the training set. Dashed lines refer to predictions carried out in an extrapolative (long-range) regime. Panels (a,c) correspond to repulsive charge-charge interactions, panels (b,e) correspond to attractive charge-charge interactions, panel (d) corresponds to an attractive charge-dipole interaction and panel (f) to repulsive charge-dipole interaction.

interaction energies between the two molecules. Upon this baselining, we find that optimal SOAP performances correspond to a RMSE ∼20%, whereas a suitable combination between SOAP and LODE($\nu = 2$) allows us to bring the error down to ∼4%. This substantial improvement can be justified by the large discrepancy between the SOAP and SOAP+LODE accuracy in representing the interaction between the monomers at intermediate and large distance. To clarify the issue further, we plot in Fig. 8.4 the predicted binding curves of 6 test dimers, against the reference DFT calculations. We observe that a SOAP-based local description is overall able to capture the short-range interactions with good accuracy. However, it becomes less and less effective as the distance between the monomers increases, to the point of being completely blind to changes in interatomic distances when the environments cutoff distance is overcome. Note that the performance of the local model at small separations is degraded substantially by the inclusion of fully dissociated dimers in the training set, because the representation cannot distinguish these configurations from those barely beyond the cutoff distance, that correspond to a non-zero value of the binding curve. The SOAP+LODE multiscale description, in contrast, can recognize the changes in separation between the monomers, leading to a smooth asymptotic behavior of the predicted binding curve. Although a linear model incorporating LODE($\nu = 1$) allows

us to halve the error made by SOAP down to ~10%, it is not sufficiently expressive to achieve predictive accuracy - particularly for binding curves that involve neutral monomers that do not have a $1/r$ asymptotic behavior. This limitation can be addressed, on one side, by increasing the body-order of the LODE descriptor to $\nu = 2$, and, on the other side, by considering non-linear kernels of the form $\left(k_{ij}^{\mathrm{SOAP}(2)} + k_{ij}^{\mathrm{LODE}(2)}\right)^2$. The resulting model is able to accurately predict the binding curves in the entire domain of distances, demonstrating its transferability across a vast spectrum of different chemical species and intermolecular configurations. This is particularly remarkable, as the SOAP(ν =2)$\oplus$LODE($\nu = 2$) model does not only predict accurately systems that are dominated by monopole electrostatics (Fig. 8.4 -(a,b,c,e)), but also systems in which only one of the molecules is charged, and so interactions involve polarization as well as charge-dipole electrostatics (Fig. 8.4 -(d,f)).

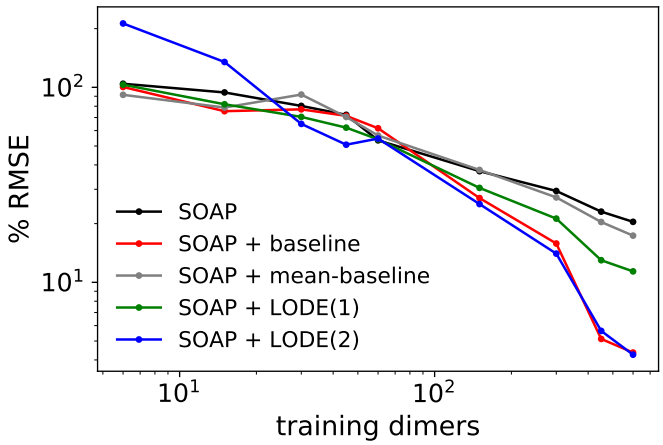


Figure 8.5 – Learning curves for the binding energies of the molecular dimers. The model is trained on a maximum of 600 binding trajectories and predictions are tested on an additional 61 independent binding trajectories. (*black line*) SOAP results obtained with r_{cut} =3 Å and σ =0.3 Å. (*red line*) SOAP results obtained with r_{cut} =3 Å and σ =0.3 Å upon baselining the binding curves with the electrostatic interaction energies coming from the Mulliken partial charges of the isolated monomers. (*gray line*) SOAP results obtained with r_{cut} =3 Å and σ =0.3 Å upon baselining the binding curves with the electrostatic interaction energies coming from the fixed partial charges of the isolated monomers obtained by averaging the Mulliken charges over the entire dataset. (*green line*) SOAP+LODE(1) multiscale results obtained with r_{cut} =3 Å and σ =1.0 Å for the LODE representation. (*blue line*) SOAP+LODE(2) multiscale results obtained with r_{cut} =3 Å and σ =1.0 Å for the LODE representation.

In order to verify how a LODE-based model compares with a treatment of electrostatics based on machine-learning atomic partial charges, we performed a Mulliken population analysis to compute the partial charges of each isolated monomer included in

the dataset, and used these partial charges to compute the classical electrostatic interaction between the monomers. Asymptotically, this term exactly represents the permanent electrostatic interaction between the monomers and it can therefore be used as an excellent baseline value for the binding curves. As shown in Fig. 8.5, we find that, using this baselining, a local SOAP description can be used to reach a similar learning accuracy (~4% RMSE) to the one obtained through the SOAP($\nu = 2$)$\oplus$LODE($\nu = 2$) multiscale approach. This result highlights the capability of the LODE($\nu = 2$) representation to describe environment-dependent electrostatic effects beyond a model of fixed point charges. In fact, the performance of a SOAP model baselined on the electrostatic energy of fixed atomic charges equal to the mean of Mulliken partial charges performs only marginally better than a purely local model.

8.7 Dielectric response of liquid water

As a final example, we revisit the problem already discussed in Sec. 3.5 of constructing a model of the electronic dielectric tensor $\boldsymbol{\varepsilon}$ of liquid water. In that context, we argued that a local model was inefficient in learning the dielectric response because of its non-local nature, and showed that using the Clausius-Mossotti relationship to map $\boldsymbol{\varepsilon}$ to effective molecular polarizabilities was greatly improving the model. Here, LODE learning performances are only tested for the isotropic component of the tensor $\varepsilon_0 = \mathrm{Tr}[\boldsymbol{\varepsilon}]$, which was shown to be most sensitive to the collective nature of the physics of dielectrics. Similarly to the case of the BFDb, we use a non-linear kernel that combines a SOAP representations computed using an optimal Gaussian width of σ=0.3 Å, and LODE($\nu = 2$) features constructed starting from a Gaussian density of σ=1.0 Å. Figure 8.6 reports results obtained when learning on 800 randomly selected structures and predicting on other 200 independent configurations.

When relying upon a local description of r_c=3 Å, LODE features perform much better than a local description. In this case, however, we observe a substantial improvement of the performance of SOAP when increasing the size of the local environments, eventually overcoming the accuracy of a LODE-based model with a radial cutoff of r_c=6 Å. This might be a consequence of a less pronounced contribution of long-range tails, or of the fact that a cutoff of 6 Å encompasses the entirety of the supercell, and therefore effectively provides a complete description of the input space of this specific dataset. Optimal ML predictions can be obtained when combining the fine-grained local description of SOAP at r_c=3 Å with the coarse-grained and non-local description of LODE at the same cutoff. This behaviour highlights the multiscale character of $\boldsymbol{\varepsilon}$, meaning that both the local many-body information and the long-range electrostatic effects need to be considered to get accurate predictions. It is

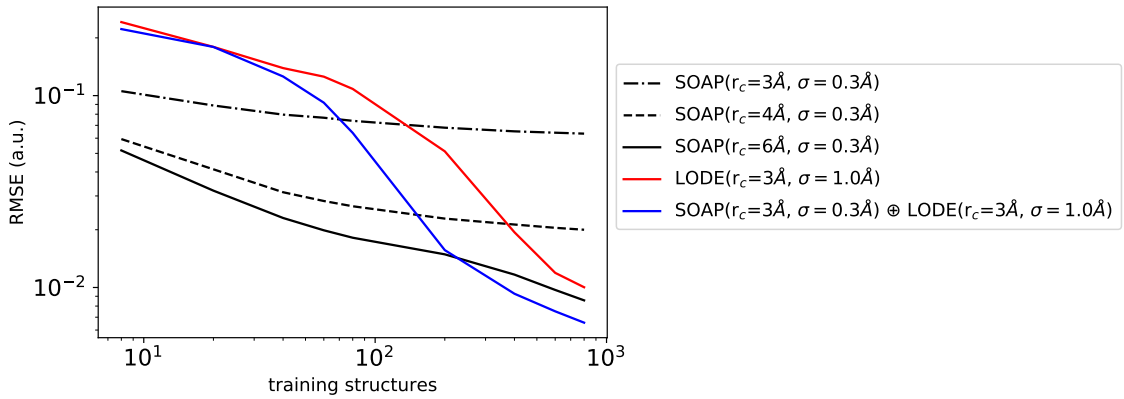


Figure 8.6 – Learning curves for the isotropic component of the electronic dielectric tensor $\boldsymbol{\varepsilon}$ of liquid water. The model is trained on up to 800 randomly selected configurations and tested on other 200 independent configurations. (*black full and dashed lines*) SOAP results with r_c= 3, 4 and 6 Å. (*red line*) LODE results with r_c=3 Å. (*blue line*) combined results of SOAP and LODE, both using r_c=3 Å.

also important to stress that a combination of SOAP and LODE is not only beneficial in terms of learning performance. The formal similarity with methods to evaluate empirical potentials separating long-range and short-range interactions suggests that by choosing judiciously the local cutoff and the density smearing it might be possible to evaluate SOAP⊕LODE at a lower cost than a SOAP model with a very large cutoff distance.

9 Multi-scale equivariant representations with consistent electrostatics

The LODE framework introduced in the previous chapter represents a novel paradigm in the regression of physical quantities that attempts to incorporate a consistent description of long-range effects within state-of-the-art local machine-learning approximations. In this chapter, this concept is fully generalized thanks to a symmetry-adapted combination of atom-density and potential-field representations of the system that allows us to treat a broad spectrum of short-range and long-range phenomena, such as Pauli repulsion, dispersion, polarization and electrostatic effects, on an equal footing. The derived class of multi-scale equivariant features not only shows a greater accuracy than a pure potential-based method, but it is also suitable to map the electrostatic energy of the system to a functional form that resembles the classical multipolar description of long-range interactions, realizing the optimal balance between physics-based and data-driven models. An open-source implementation of the method can be found in the TENSOAP package [68]. Sections and figures are largely based on the following article:

A. Grisafi, J. Nigam and M. Ceriotti, "Multi-scale approach for the prediction of atomic scale properties", *Chemical Science* 12, 2078-2090, (2021). Copyright © 2021 The Royal Society of Chemistry. AG contributed to deriving and implementing the multiscale LODE method, to produce results and figures associated with the water-carbon dioxide and organic dimers examples reported and to writing the manuscript.

9.1 Multi-scale equivariant representations

During the discussion carried out in the previous chapter, we have shown that a multi-scale learning model that can treat both local and non-local effects can be obtained by combining structural descriptions of short-range interatomic correlations, equivalent to SOAP [42], with long-distance equivariants (LODE) features [43]. Here, we introduce

a more explicit multi-scale approach, that couples $|\rho_i\rangle$ and $|V_i\rangle$ terms in a unified representation that is adapted to the symmetries of the $O(3)$ group. In this regard, consider the following tensor product that is averaged over all the possible improper rotations $\hat{S} = \hat{i}^k \hat{R}$, with $\hat{R}$ rotation operators and $\hat{i}^k$ inversion operators, as

$$\int d\hat{S}\, \underbrace{\hat{S}|\rho_i\rangle \otimes \ldots \hat{S}|\rho_i\rangle}_{\nu \text{ times}} \otimes \underbrace{\hat{S}|V_i\rangle \otimes \ldots \hat{S}|V_i\rangle}_{\nu' \text{ times}} \otimes \hat{S}|\lambda\mu\rangle \otimes \hat{S}|\sigma\rangle \,, \tag{9.1}$$

Within this construction, the ket $|\lambda\mu\rangle$ has the role of making the resulting features transform as a Y^μ_λ spherical harmonics, paving the way to the regression of tensorial properties, while $|\sigma\rangle$ indicates the parity of the structural representation under inversion, leaving the freedom to treat both polar ($\sigma = 1$) and pseudo-tensors ($\sigma = -1$). Clearly, Eq. (9.1) corresponds to the λ-SOAP framework for $\nu' = 0$, while it represents the tensorial extension of LODE features for $\nu = 0$. In what follows, we will restrict the discussion to the hybrid density-potential case $\nu = 1$ and $\nu' = 1$, i.e.,

$$\left|\overline{\rho_i;\, V_i;\, \lambda\mu;\, \sigma}\right\rangle = \int d\hat{S}\, \hat{S}|\rho_i\rangle \otimes \hat{S}|V_i\rangle \otimes \hat{S}|\lambda\mu\rangle \otimes \hat{S}|\sigma\rangle \,, \tag{9.2}$$

where we used the short-hand notation $\left|\overline{\rho_i;\, V_i;\, \lambda\mu;\, \sigma}\right\rangle$ to indicate that the representation follows the symmetries of the $O(3)$ manifold. As sketched in the figure below, its real-space representation underlies 3-body structural correlations where one point of the density-field is coupled with one point of the potential-field about the central atom, together with a spherical harmonic rigidly attached to the frame of reference.

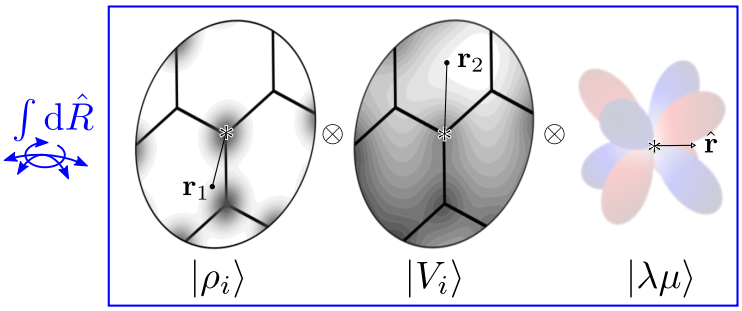


Figure 9.1 – A schematic real-space representation of Eq. (9.1) for $\nu = 1$ and $\nu' = 1$. Different representations of an atomic environment are combined as tensor products, i.e., evaluated at different points, and averaged over all possible rotations of the system. Including also a set of spherical harmonics makes it possible to build ML models endowed with an equivariant behavior.

9.2 Linear models for electrostatic interactions

Linear models are notoriously useful to reveal the physical meaning of a structural representation. For example, they can be used to draw the connection between short-range density correlations ($\nu' = 0$) and the body-order expansion of interatomic potentials [45, 53, 205, 226], as well as to relate the LODE($\nu = 0, \nu' = 1$) features to a fixed point-charge electrostatic model [43]. In this section, we use this idea to show that even though neither the atom density $\left|\rho_i\right\rangle$ nor the associated potential field $|V_i\rangle$ correspond to actual physical quantities, the multi-scale combination of the two entails formal similarities with the physics of long-range interactions. In particular, we show that the simplest multi-scale LODE($\nu = 1, \nu' = 1$) defined in Eq. (9.2) can be put under formal correspondence with a multipolar expansion for the electrostatic energy of a system [76]. This connection is demonstrated both analytically and with numerical benchmarks.

9.2.1 Analytical connection with the multipole expansion

Consider the scalar ($\lambda = 0$) and polar ($\sigma = 1$) limits within the multi-scale abstract representation of Eq. (9.2), i.e., $\left|\overline{\rho_i; V_i}\right\rangle$. This can be used to build a linear regression model for the electronic energy U of a system A, that is approximated through the following atom-centered decomposition,

$$\begin{aligned} U(A) &\approx \sum_{i=1}^{N} U_i(A) = \sum_{i=1}^{N} \left\langle w \middle| \overline{\rho_i(A); V_i(A)} \right\rangle \\ &= \sum_{i=1}^{N} \sum_{a_1 a_2} \sum_{l=0}^{l_{\max}} \int_0^{r_c} dx_1\, x_1^2 \int_0^{r_c} dx_2\, x_2^2 \, \langle w | a_1 x_1 l; a_2 x_2 l \rangle \left\langle a_1 x_1 l; a_2 x_2 l \middle| \overline{\rho_i(A); V_i(A)} \right\rangle . \end{aligned} \tag{9.3}$$

In the last equality, the linear problem is projected on the rotationally invariant basis $\langle a_1 x_1 l; a_2 x_2 l|$ already encountered to represent 3-body structural correlations, so that $\langle w | a_1 x_1 l; a_2 x_2 l \rangle$ are the regression weights to be determined, while the actual multi-scale representation of the system is given by:

$$\left\langle a_1 x_1 l; a_2 x_2 l \middle| \overline{\rho_i(A); V_i(A)} \right\rangle = \sum_m \left\langle a_1 x_1 l m \middle| \rho_i(A) \right\rangle \langle a_2 x_2 l m | V_i(A) \rangle^\star . \tag{9.4}$$

We aim to prove that the functional form of Eq. (9.3) can be used to model rigorously a multipolar expansion of the long-range contributions to U.

To see this, let us start by separating the near-field from the far-field potential in the

definition of the atom-centered potential field $|V_i\rangle$, that is,

$$\langle a\boldsymbol{x}|V_i\rangle = \left\langle a\boldsymbol{x}\middle|V_i^{<}\right\rangle + \left\langle a\boldsymbol{x}\middle|V_i^{>}\right\rangle = \int d\boldsymbol{x}' \frac{\left\langle a\boldsymbol{x}'\middle|\rho_i^{<}\right\rangle}{|\boldsymbol{x}-\boldsymbol{x}'|} + \int d\boldsymbol{x}' \frac{\left\langle a\boldsymbol{x}'\middle|\rho_i^{>}\right\rangle}{|\boldsymbol{x}-\boldsymbol{x}'|}, \tag{9.5}$$

where $\rho_i^{<}$ and $\rho_i^{>}$ are the atomic densities located inside and outside the i-th spherical environment of radius r_c. While the near-field term contributes to a piece of information that is similar to that included in $\left|\rho_i\right\rangle$, the far-field contribution determines the effect of the density beyond r_c. Upon this splitting, we can in turn partition each atom-centred contribution to the energy prediction in range separated terms, i.e., $U_i = U_i^{<} + U_i^{>}$. Focusing in particular on the long-range contribution, we have:

$$U_i^{>} = \sum_{a_1 a_2} \sum_{l=0}^{l_{\max}} \int_0^{r_c} dx_1\, x_1^2 \int_0^{r_c} dx_2\, x_2^2 \langle w|a_1 x_1 l; a_2 x_2 l\rangle \sum_m \left\langle a_1 x_1 lm\middle|\rho_i^{<}\right\rangle \left\langle a_2 x_2 lm\middle|V_i^{>}\right\rangle^{\star}. \tag{9.6}$$

Using the Laplace expansion of the Coulomb operator, the spherical harmonics components of the far-field potential can be explicitly written as follows:

$$\left\langle a_2 x_2 lm\middle|V_i^{>}\right\rangle^{\star} = \frac{4\pi}{2l+1} \int_{r_c^+}^{\infty} dx x^2 \left\langle \rho_i^{>}\middle|a_2 x lm\right\rangle \frac{x_2^l}{x^{l+1}}. \tag{9.7}$$

Finally, plugging this expression into Eq. (9.6), one sees that the long-range contribution to the energy prediction can be formally rewritten as

$$\begin{aligned} U_i^{>} &= \sum_{a_1 a_2} \sum_{l=0}^{l_{\max}} \int_{r_c^+}^{\infty} dx \frac{x^2}{x^{l+1}} \sum_m \left\langle \rho_i^{>}\middle|a_2 x lm\right\rangle \left\langle a_1 a_2; lm\middle|M_i^{<}(w)\right\rangle \\ &= \sum_{a_1 a_2} \int_{|\boldsymbol{x}|>r_c} d\boldsymbol{x} \left\langle \rho_i^{>}\middle|a_2\boldsymbol{x}\right\rangle \left[\sum_{l=0}^{l_{\max}} \frac{1}{x^{l+1}} \sum_m \langle\hat{\boldsymbol{x}}|lm\rangle \left\langle a_1 a_2; lm\middle|M_i^{<}(w)\right\rangle\right] \\ &= \sum_{a_1 a_2} \int_{|\boldsymbol{x}|>r_c} d\boldsymbol{x} \left\langle \rho_i^{>}\middle|a_2\boldsymbol{x}\right\rangle \left\langle a_1 a_2; \boldsymbol{x}\middle|V_i^{<}(w)\right\rangle. \end{aligned} \tag{9.8}$$

Eq. (9.8) shares a striking resemblance with the expression for the interaction of a far-field charge density $\rho_i^{>}$ with the multipole expansion (represented in square brackets) of the electrostatic potential $V_i^{<}$ generated by a localized charge distribution [227]. To write this expression, we relied on a definition of the atom-centered spherical multipoles $M_i^{<}$ that reads as a non-linear transformation of the inner cutoff density distribution $\rho_i^{<}$ through the fitting parameters w:

$$\left\langle a_1 a_2; lm\middle|M_i^{<}(w)\right\rangle = \frac{4\pi}{2l+1} \int_0^{r_c} dx_2\, x_2^2 x_2^l \int_0^{r_c} dx_1\, x_1^2 \langle w|a_1 x_1 l; a_2 x_2 l\rangle \left\langle a_1 x_1 lm\middle|\rho_i^{<}\right\rangle.$$

(9.9)

Crucially, however, ρ_i and V_i are not physical quantities, but are just a representation of the spatial arrangement of atoms. In fact, atoms in the far-field respond in a way that depends only on their chemical nature, while the local multipoles are modulated in a highly flexible, non-trivial fashion by the distribution of atoms in the local environment. The form of Eq. (9.9) also hints at how changing the representation would affect this derivation. Increasing the density order ν would bring a more flexible, higher-body-order dependence of the local multipoles on the distribution of atoms in the vicinity of the central atom i, while increasing ν' would bring a more complicated dependency on the distribution of atoms in the far-field, leading to a linear regression limit that does not match formally the electrostatic multipole expansion.

As we shall see in what follows, the formal equivalence previously outlined underpins the ability of multi-scale LODE features to model accurately several kinds of interactions. For instance, a fixed point-charge interaction model can be obtained under truncating the expansion at $l_{\max} = 0$ and taking the limits of vanishing cutoff radius ($r_c \to 0$) and δ-like atom-density distributions ($\sigma \to 0$), i.e.,

$$\begin{aligned}
U_i^{>} &= \sum_{a_1 a_2} \int_{r_c^+}^{\infty} dx \, \frac{x^2}{x} \left\langle \rho_i^{>} \middle| a_2 x 00 \right\rangle \left\langle a_1 a_2; 00 \middle| M_i^{<}(w) \right\rangle \\
&\xrightarrow{r_c \to 0} \sum_{a_2} \int_{0+}^{\infty} dx \, \frac{x^2}{x} \left\langle \rho_i^{>} \middle| a_2 x 00 \right\rangle \left\langle a_j a_2; 00 \middle| M_i^{<}(w) \right\rangle \\
&\xrightarrow{\sigma \to 0} \sum_{a_2} \int_{0+}^{\infty} dx \, \frac{x^2}{x} \left(\sum_{j \neq i} \delta_{a_j a_2} \frac{\delta(x - r_{ij})}{x^2} \right) \left\langle a_i a_2; 00 \middle| M_i^{<}(w) \right\rangle \\
&= \sum_{j \neq i} \frac{\left\langle a_i a_j; 00 \middle| M_i^{<}(w) \right\rangle}{r_{ij}},
\end{aligned} \tag{9.10}$$

where we considered that the only atom inside the cutoff is the central atom i and we made use of the properties of Dirac-δ distribution functions. If one interprets $\left\langle a_i a_j; 00 \middle| M_i^{<}(w) \right\rangle$ as the product of the partial charges of the two species q_{a_i}, and q_{a_j}, this form is equivalent to a simple Coulomb interaction between fixed point-charges. In this regard, while relaxing the cutoff constraint would make the atomic charges dependent on the chemical environment, hence matching a flexible point-charge model, including multipoles for $l > 0$ makes it possible to represent the structural anisotropy of the electrostatic interaction.

9.2.2 A toy model for multipolar interactions

To show this, we analyze the performance of the method in representing the long-range interaction between a H_2O and a CO_2 molecule. We build a dataset made of 33 non-degenerate reciprocal orientations between the two molecules, and learn the corresponding interaction within the far-field regime defined by an intermolecular distance that goes from 6.5 to 9 Å between the molecular center of mass. The regression is performed using a LODE(1,1) model built using a cutoff r_c = 3 Å, so that we can unambiguously interpret our results in terms of the long-range energy formula reported in Eq. (9.8). We then extrapolate the predicted interaction profile in the asymptotic regime of $R > 9$ Å, verifying how the model converges towards the dissociated limit of $R \to \infty$, which is also included in the training set.

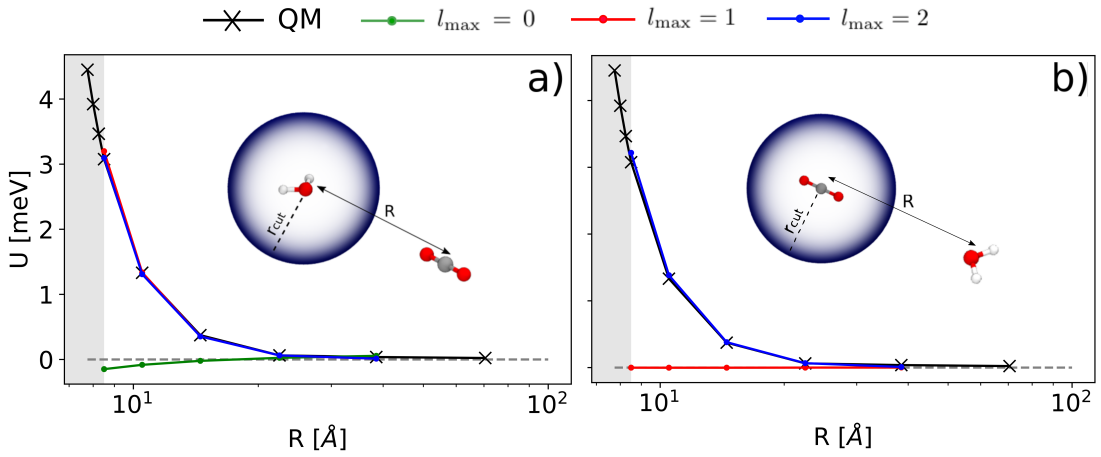


Figure 9.2 – Extrapolated interaction profiles for a given configuration of H_2O and CO_2 at different angular cutoff values l_{max}. Left and right panels show the results of the asymptotic extrapolation when centring the representation on the oxygen atom of H_2O and the carbon atom of CO_2 respectively.

According to our construction, the cutoff value l_{max} chosen to define the angular resolution of the representation determines the number of multipoles that are included within the expansion of Eq. (9.7). In Fig 9.2 we report the results of the extrapolation for a highly symmetric reciprocal orientation at increasing angular cutoffs l_{max}. We also compare different choices for the possible atomic centres that contribute to the energy prediction: in panel (a) we express the energy in terms of a single environment centred on the oxygen atom of the H_2O molecule; in panel (b) we use a single environment centred on the carbon atom of CO_2. As one would expect from a classical interpretation of the long-range energy, the binding profile for the selected test configuration is solely determined by the interaction between the dipole moment of the water molecule and the quadrupole moment of CO_2, while the interaction term

associated with the quadrupole moment of H_2O vanishes for symmetry reasons. This is reflected in the sharp transition of the prediction accuracy when crossing a critical angular cutoff $l_{\max}$. When centring the local environment on the water molecule (Fig. 9.2-a)), truncating the expansion at $l_{\max} = 1$ is enough to reproduce the interaction between the dipolar potential of water and the atom-density distribution of the CO_2 molecule. Conversely, when centring the representation on carbon dioxide (Fig. 9.2-b)), the H_2O atom-density in the far-field can only interact with a CO_2 potential that is quadrupolar in nature, requiring an angular cutoff of at least $l_{\max} = 2$. Fig. 9.3 reports what happens when centering the LODE(1,1) representation on all the

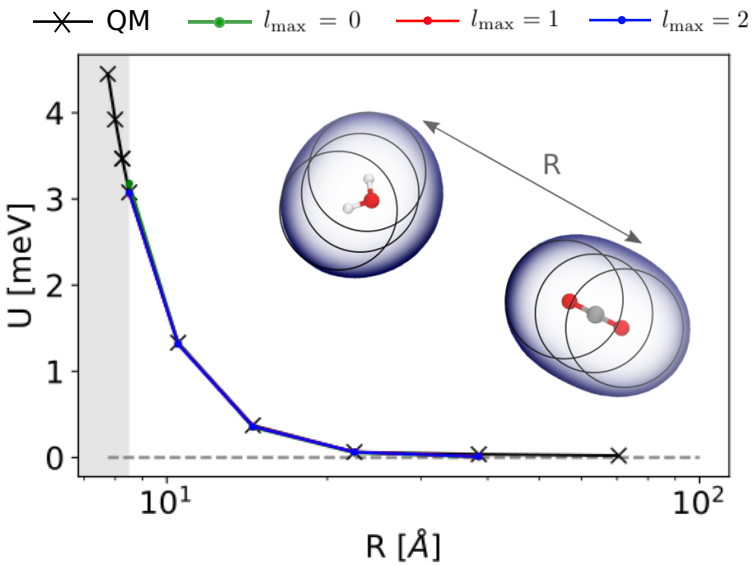


Figure 9.3 – Extrapolated interaction profiles for a highly symmetric configuration of H_2O and CO_2 at different angular cutoff values $l_{\max}$. Results are obtained centring the representation on the all the atoms of the system.

atoms of the system. Interestingly, for this particular highly symmetric configuration, using an angular cutoff of $l_{\max} = 0$ suffices to obtain an accurate asymptotic profile, underlying a model that can be interpreted through the fixed point-charge limit derived in Eq. (9.10). Overall, these results remark the distinction between a learning model for the electrostatic energy that relies on the definition of *molecular* multipoles (Fig. 9.2), and a model that is instead based on *atomic* multipoles (Fig. 9.3). In this regard, it is worth stressing that in contrast to other learning models that explicitly target the prediction of the actual *ab initio* atomic multipoles [32], our data-driven definition of atomic (or molecular) multipoles only needs to be interpreted in the sense of providing a formal connection with a functional form for the long-range electrostatic energy that resembles the one of multipole interactions.

To prove the relationship between a fixed point-charge model and the $l_{\max} = 0$ truncation of the LODE(1,1) representation, we fitted the partial charges associated with the H, C and O atomic species by numerically minimizing the discrepancy between the training interaction energies and the interatomic Coulomb energies $\sum_{j>i} q_{a_i} q_{a_j} / r_{ij}$. This minimization yields optimal charges that correspond to q_{H} =0.24e, q_{C} =0.96e and $q_0 = -0.49e$, guaranteeing the global eletroneutrality of the system. The final error is

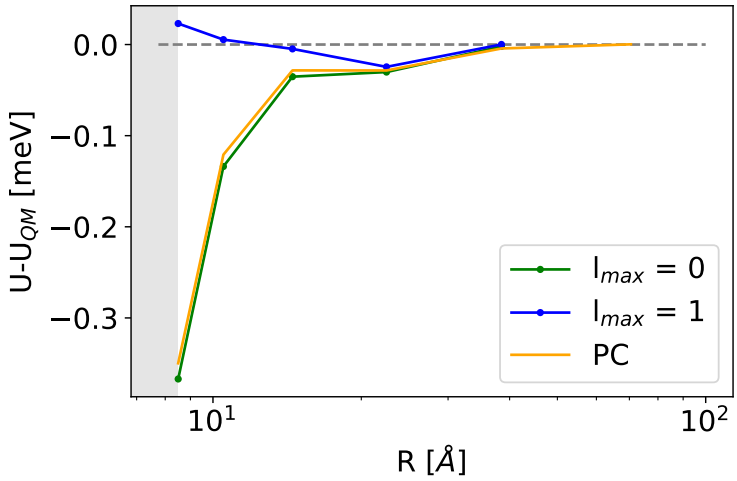


Figure 9.4 – Asymptotic prediction errors of a representative CO_2/H_2O configuration obtained at different levels of theory. Green and blue lines correspond to $l_{\max} = 0$ and $l_{\max} = 1$ LODE(1,1) models while the orange line refers to a fixed point-charges model.

reported in Figure 9.4 for a representative configuration of the dataset that does not have any particular symmetry. The aforementioned relationship is apparent by the almost perfect agreement between the two models. As shown in the figure, increasing the order of the expansion has the beneficial effect to go beyond a fixed point-charge model, which is shown to improve the accuracy of the prediction particularly in the intermediate distance range. For this simple toy problem, in particular, truncating the atom-centered expansion at $l_{\max} = 1$ allows us to achieve almost perfect predictions.

9.3 Beyond electrostatics

The discussion carried out during the previous section reflects the capability of the multi-scale LODE representation to regress electrostatic energies in a physically consistent fashion. However, the data-driven nature of the method implies that its applicability is not limited to the prediction of electrostatic properties. In this section we present three examples to demonstrate that even in their simplest form, this family of multi-scale features is suitable to address the complexity of challenging, real-life atom-

istic modelling problems, and physics well beyond that of permanent electrostatics. For the sake of comparing our approach with a local machine-learning scheme, the LODE($\nu = 1, \nu' = 1$) results will be reported all throughout against the ones obtained using the SOAP($\nu = 2$) method of Ref. [42].

9.3.1 Binding energies of organic dimers

We start by testing the ability of multi-scale LODE to describe different kinds of molecular interactions. To this end, we consider the interaction energy between the 2291 pairs of organic molecules already used in the context of charge-density learning [60]. For each dimer configuration, binding curves are generated by considering 12 rigid displacements in steps of 0.25 Å along the direction that joins the geometric centres of the two molecules. Then, unrelaxed binding energies are computed at the DFT/PBE0 level using the Tkatchenko-Scheffler self-consistent van der Waals method [228] as implemented in the FHI-aims package[125]. For each binding trajectory, we also include in the training set the dissociated limit of vanishing interaction energy, where the two monomers are infinitely far apart. The dataset so generated includes all the possible spectrum of interactions, spanning pure dispersion, induced polarization and permanent electrostatics. In order to better rationalize the learning capability of such a large variety of molecular interactions, we choose to partition the molecules in the dataset in three independent classes, namely, 1) molecules carrying a net charge, 2) neutral molecules that contain heteroatoms (N, O), and can therefore exhibit a substantial polarity 3) neutral molecules containing only C and H, that are considered apolar and interacting mostly through dispersive interactions. Considering all the possible combinations of these kinds of molecules partitions the dimers into six classes, i.e., 184 charged-charged (CC), 267 charged-polar (CP), 210 charged-apolar (CA), 161 polar-polar (PP), 418 polar-apolar (PA) and 1051 apolar-apolar (AA) interactions. For each of the six classes, several, randomly selected binding curves are held out of the training set, to test the accuracy of our predictions. The remaining curves are used to fit one separate linear model for each class, using either local SOAP features or multi-scale LODE($\nu = 1, \nu' = 1$) features using a cutoff of $r_c = 3$ Å. In order to also assess the reliability of our predictions, we use a calibrated committee estimator [129] for the model uncertainty, which allows us to determine error bars for the binding curves. 8 random subselections of 80% of the total number of training configurations were are considered to construct the committee model. The internal validation set is then defined by selecting the training structures that are absent from at least 25% of the committee members.

Figure 9.5 shows characteristic interaction profiles for the six different classes of

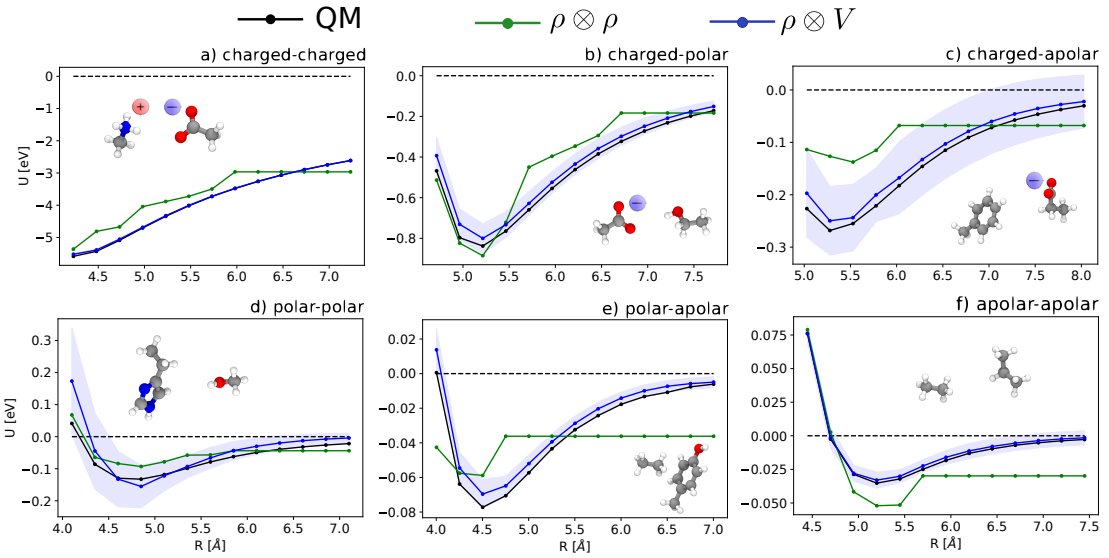


Figure 9.5 – Median-error binding curves (in eV) for six different classes of intermolecular interactions. (*black lines*) quantum-mechanical calculations. (*green lines*) local $\rho \otimes \rho$ predictions. (*blue lines*) multiscale $\rho \otimes V$ predictions.

molecular pairs. The configurations reported are those that exhibit median integrated errors within the test set of each class. The root mean square errors associated with the predictions over the entire test sets of each class are listed in Table 9.1.

			RMSE/eV		
class	n_{train}	STD/eV	$\rho \otimes \rho$	$\rho \otimes V$	$V \otimes V$
CC	100	1.86	0.72	0.049	0.058
CP	200	0.379	0.25	0.074	0.092
CA	150	0.083	0.056	0.041	0.034
PP	100	0.131	0.10	0.062	0.125
PA	350	0.046	0.032	0.013	0.021
AA	950	0.063	0.026	0.004	0.006

Table 9.1 – Prediction performance expressed in terms of the RMSE over all the points of the binding curves, for the six classes of interactions and $\rho \otimes \rho$, $\rho \otimes V$ and $V \otimes V$ models. For each class we also indicate the number of training samples, and the characteristic energy scale, expressed in terms of the standard deviation of the energies in the test set.

The results clearly show that while SOAP(2) is limited by the nearsightedness of the local environments, the LODE(1,1) multi-scale model is able to predict both the short and the long-range behaviour of the binding profiles on an equal footing. What is particularly remarkable is the fact that a simple, linear model can capture accurately

different kinds of interactions, that occur on wildly different energy scales and asymptotic behavior: the typical binding energy of charged dimers is of the order of several eV, and has a $1/r$ tail, while the typical interaction energy of two apolar molecules is of the order of a few tens of meV, and decays roughly as $1/r^6$. A LODE($\nu' = 2$) model also allows us to predict the binding curves beyond the 3 Å cutoff, but usually yields 50-100% larger errors than those observed with LODE(1,1). The multi-scale nature of LODE($\nu = 1, \nu' = 1$) yields a better balance of short and long-range descriptions, and is sufficiently flexible to be adapted to the description of systems that are not dominated by permanent electrostatics, even though interactions between charged fragments are considerably easier to learn, in comparison to the others. We also observe that the uncertainty model works reliably, as the predicted curves always fall within the estimated error bar. Larger uncertainties are found for interaction classes that have few representative samples in the training set, such as those associated with polar-polar molecular pairs (Fig 9.5-d)).

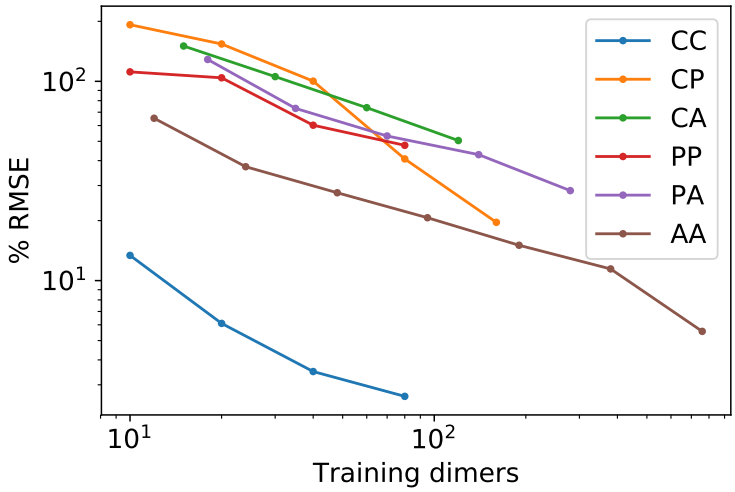


Figure 9.6 – Learning curves for the 6 classes of molecular interactions computed using the LODE(1,1) representation. The curves indicate that all interactions can be learned with comparable efficiency and that the accuracy of the model is limited by the small number of available reference structures. Interactions between charged molecules, that have a formal connection with the form of the multi-scale features, can be learned effectively with a small number of training samples.

The learning curves, plotted in Figure 9.6, provide insights into the performance of LODE(1,1) for different kinds of interactions. CC dimers are learned with excellent relative accuracy – which is unsurprising given the formal connection with the multipole expansion. All other classes of interactions yield a relative accuracy for a given training set size which is an order of magnitude worse (with the exception of AA interactions, whose learning performance is intermediate). However, learning curves show no

sign of saturation [229], reflecting the fact that multi-scale features have sufficient flexibility to provide accurate predictions, but that the lack of a natural connection to the underlying physics would require a larger train set size. This is consistent with the considerations we made in the previous section based on the simple H_2O/CO_2 example.

9.3.2 Induced polarization on a metal surface

The previous example proves that linear LODE(1,1) models capture a wide class of molecular interactions, ranging from pure dispersion to permanent electrostatics. Beyond molecular systems, however, a large number of phenomena occur in solid state physics that are driven by long-range effects, and involve more subtle, self-consistent interactions between far-away atoms. A particularly relevant example is represented by the induced macroscopic polarization that a metallic material undergoes in response to an external electric field, which underlies fundamentally and technologically important phenomena for surface science and nanostructures [230–232]. Physics-based modelling of these kinds of systems usually exploits the fact that, for a perfectly-conductive surface, the interaction is equivalent to that between the polar molecule and the mirror image, relative to the surface plane, of its charge distribution, with an additional inversion of polarity [233]. It would not appear at all obvious that our atom-centred framework, which does not include an explicit response of the far-field atom density to the local data-driven multipole, can capture the physics of a phenomenon associated with the polarization of electrons that are delocalized over the entire extension of the metallic solid.

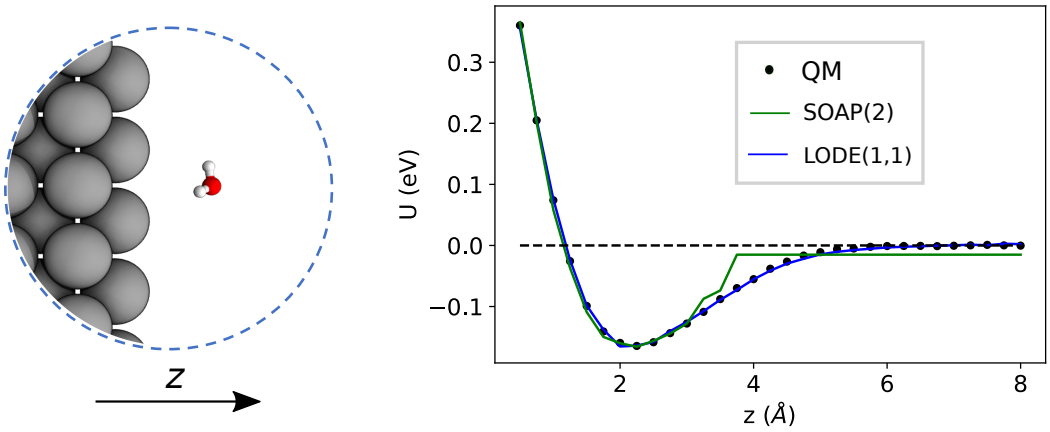


Figure 9.7 – Predicted binding curve of a test water-lithium configuration. (*black dots*) reference DFT calculations. (*green line*) SOAP(2) predictions. (*blue line*) LODE(1,1) predictions.

To benchmark the performance of multi-scale LODE in this challenging scenario we consider the interaction of a slab of *bcc* lithium with a water molecule that is located at various distances from the (100)-surface. We start by selecting 81 water molecule configurations, differing in their internal geometry or in their spatial orientation relative to the surface. For each of these configurations, 31 rigid displacements are performed along the (100)-direction, spanning a range of distances between 0.5 Å and 8 Å from the lithium surface. Using this dataset we compute unrelaxed binding energies at the DFT/PBE level using the FHI-aims package[125]. We converge the slab size along the periodic xy-plane, minimizing the self-interaction between the periodic images of the water molecule, resulting in a 5×5 unit cell repetitions and a k-points sampling of $4 \times 4 \times 1$ Å^{-1}. We set the slab extension along the non-periodic z-direction so that the Fermi energy is converged within 10 meV, resulting in a total of 13 layers. To remove the spurious interactions along the z-axis, we set a large vacuum space of roughly 80 Å in conjunction with a correction suitable to screen the dipolar potential [234]. Following these prescriptions, we obtain attractive potential profiles for all molecular geometries and orientation, consistently with the interaction between the dipolar field of the water molecule and the induced metal polarization.

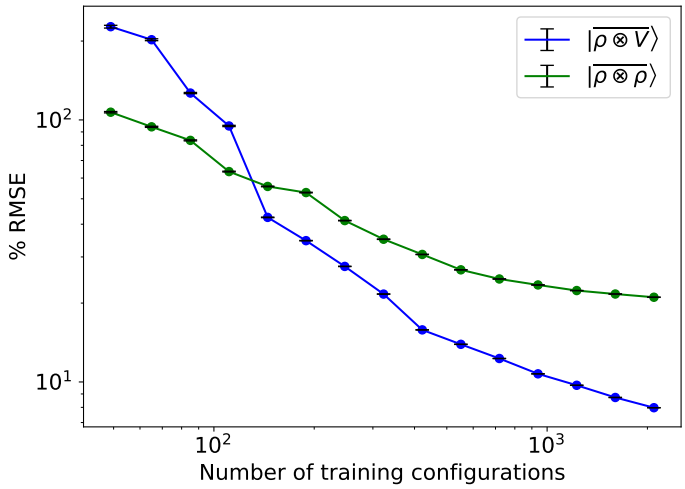


Figure 9.8 – Learning curves for the binding energy of water-lithium slab interaction. The energies of 75 slab-water molecule trajectories were learnt with SOAP (*green*) and LODE(1,1) (*blue*). Error in predictions on the 6 test trajectories is shown here.

For this example, we construct $|\rho_i\rangle$ and $|V_i\rangle$ representations within spherical environments of $r_c = 4$ Å with a Gaussian-density width of $\sigma = 0.3$ Å. In this case, computing the potential projections in reciprocal space would be very expensive because of the large number of plane-waves that arise from having to deal with a huge vacuum space along z. For this reason, we rely on a plain Ewald method to break down the calculation of $\langle anlm|V_i\rangle$ in a short-range, screened contribution computable in real space, and a

long-range, smooth contribution computable in reciprocal space (see Appendix C.3 for more details). The regression model is trained on 75 lithium-water binding curves while the remaining 6 are used for testing the accuracy of our predictions. Figure 9.7 shows a comparison between a local SOAP model and a multi-scale LODE(1,1) model in learning the interaction energy of the metal slab and the water molecule for one representative test trajectory. We observe that SOAP is able to capture the short-range interactions but becomes increasingly ineffective as the water molecule moves outside the atomic environment, leading to an overall error of about 19 RMSE%. This is in sharp contrast to the performance of the multi-scale representation, which can capture both the effects of electrostatic induction at a large distance and the Pauli-like repulsion at short range with the same level of accuracy, halving the prediction error to about 9%. A comparison between the SOAP(2) and LODE(1,1) learning curves is reported in Fig. 9.8.

To further investigate what aspects of the physics of the molecule-surface interaction can be captured by the model, we perform a Mulliken population analysis on the reference DFT calculations, to extract the polarization vector of the water molecule in response to the interaction with the metal, i.e., $\boldsymbol{P}^W = \boldsymbol{\mu}^W - \boldsymbol{\mu}_0^W$, where $\boldsymbol{\mu}^W$ and $\boldsymbol{\mu}_0^W$ are the dipole moment of the water molecule in the lithium-slab system and in vacuum respectively. Physically, the polarization $\boldsymbol{P}^W$ involves the response of water's electrons to the rearrangement of the electronic charge in the surface triggered by the dipolar field, and so it involves explicitly a back-reaction. Furthermore, the polarization shows both a (usually larger) component along the z-axis, and a tangential component in the xy-plane. To account for the vectorial nature of $\boldsymbol{P}^W$, we take advantage of the tensorial extension of the multiscale model reported in Eq. (9.2). To single out the long-range nature of the polarization interaction, we restrict the regression of $\boldsymbol{P}^W$ to water configurations that are more than 4.5 Å far from the surface. Our dataset contains 1215 such configurations, out of which we randomly select 1000 for training, while the remaining 215 are retained for testing.

Results are shown in Figure 9.9. Given that the training set contains no structures within the local descriptor cutoff, it comes as no surprise that a pure density-based tensor model entirely fails to learn the long-range polarization induced on the water molecule. Making use of the tensorial extension of the multiscale model of Eq. (9.2), in contrast, allows us to effectively learn the polarization vector $\boldsymbol{P}^W$, showing an error that decreases to ~20 %RMSE at the maximum training set size available. This example provides a compelling demonstration of the ability of LODE(1,1) to embrace effects that go well-beyond permanent electrostatics.

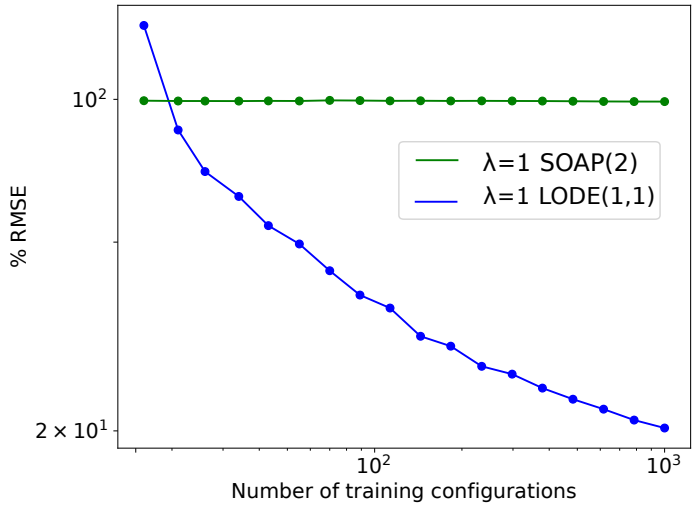


Figure 9.9 – Learning curves for the induced polarization of the water molecule due to interaction with image charges in the metal slab, computed only for separations greater than 4.5Å. The error is computed as a fraction of the intrinsic variability of the test set of 215 configurations. Contrary to the local model (green), a linear LODE(1,1) model (blue) can learn this self-consistent polarization, with no significant reduction of the learning rate up to 1000 training configurations.

9.3.3 Response functions of oligopeptides

As a final example, we consider the challenging task of predicting the polarizabilty of a dataset of poly-aminoacids. Dielectric response functions are strongly affected by long-range correlations, because of the cooperative nature of the underlying physical mechanism. Poor transferability of local models between structures of different sizes has been observed for molecular dipole moments [220], polarizability [58], and the electronic dielectric constant of bulk water [57]. For this purpose, we use a training set composed of 27428 conformers of single aminoacids and 370 dipeptides, testing the predictions of the model on a smaller test set containing 30 dipeptides, 20 tripeptides, 16 tetrapeptides and 10 pentapeptide configurations. Reference polarizability calculations are carried out with the Gaussian16 quantum-chemistry code using the double-hybrid DFT functional PWPB95-D3 and the aug-cc-pVDZ basis set [235]. We compute the multi-scale LODE(1,1) features and their local counterparts using a Gaussian width of σ =0.3 Å and a spherical environment cutoff of r_c =4 Å. This data set is interesting, because it combines large structural variability with tens of thousands of distorted aminoacid configurations with longer-range interactions described by a few hundred dipeptide conformers. We consider three models: a linear $\left|\overline{\rho \otimes V}\right\rangle$ multi-scale model; a square kernel model, that is equivalent to using a quadratic functional of the SOAP features, $\left|\left[\overline{\rho_i^{\otimes 2}}\right]^{\otimes 2}\right\rangle$, which partially incorporates 4 and 5-body

correlations and enhance the many-body character of the representation at the local scale [55]; a weighted combination of the two.

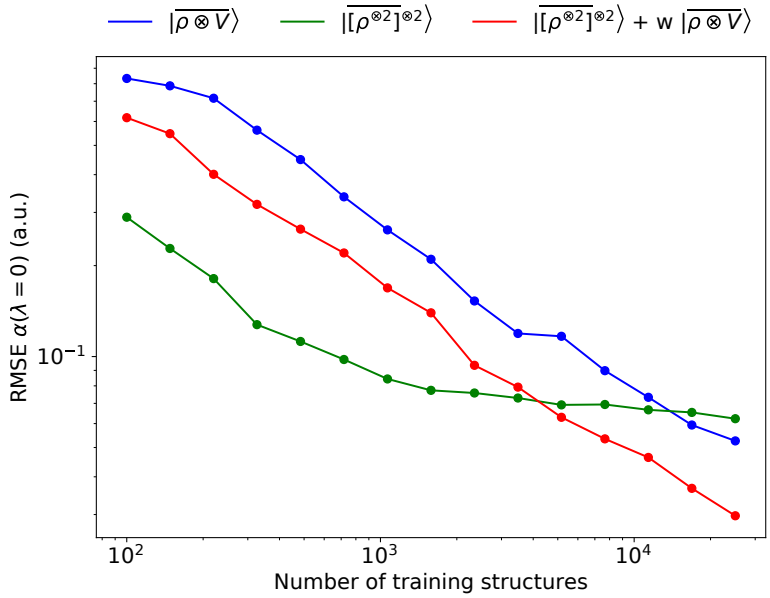


Figure 9.10 – Learning curves for the $\lambda = 0$ component of the polarizability tensor of a database of polypeptide conformers. (*green curve*) non-linear kernel model which underlies a density-based $\left|\left[\overline{\rho_i^{\otimes 2}}\right]^{\otimes 2}\right\rangle$ representation. (*blue curve*) linear kernel model based on $\left|\overline{\rho \otimes V}\right\rangle$. (*red curve*) optimal linear combination of the two.

The learning curves for the trace ($\lambda = 0$) of the polarizability tensor, shown in Fig. 9.10, are very revealing of the behavior of these three models. The $\left|\left[\overline{\rho_i^{\otimes 2}}\right]^{\otimes 2}\right\rangle$ model, which disregards any non-local behavior beyond the atomic environment, is initially very efficient, but saturates to an error of 0.06a.u.. In contrast, equipped with non-local information, the LODE(1,1) representation reduces the error of prediction to 0.05a.u., but is initially much less effective. This is not due to the lack of higher-order local density correlations: a linear SOAP model performs well, despite showing saturation due to its local nature. We interpret the lackluster performance of the LODE model in the data-poor regime as an indication of the dominant role played by short-range effects in this diverse dataset, which can be learned more effectively by a nearsighted kernel, similarly to what observed in Refs.[10, 236, 237]. Inspired by those works, we build a tunable kernel model based on a weighted sum of the local and the LODE kernels, that can be optimized to reflect the relative importance of the different ranges. We optimize the weight by cross-validation at the largest train size, obtaining a reduction of 50% of the test error, down to 0.028a.u.

An analysis of the test error which separates the contributions from oligopeptides

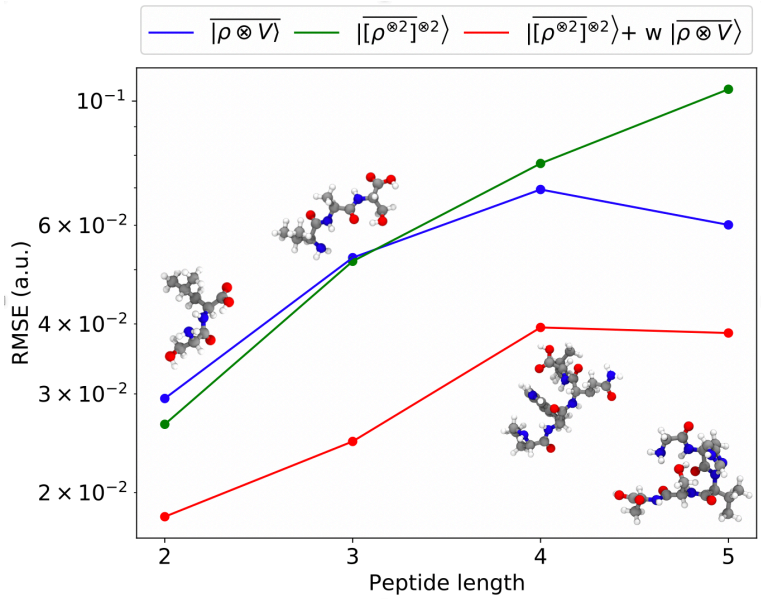


Figure 9.11 – Absolute RMSE in learning the $\lambda = 0$ spherical tensor of polarizability of polypeptides as a function of the peptide length. The model was trained on 27428 single-amino acids and 370 dipeptides. The error was computed on 30 dipeptides, 20 tripeptides, 16 tetrapeptides and 10 pentapeptides respectively.

of different length, shown in Fig. 9.11, is consistent with this interpretation of the learning curves. All models show an error that increases with the size of the molecule, because there are interactions that are just not described at the smaller train set size. However, the purely local model shows by far the worst extrapolative performance, while multi-scale models – in particular the one combining a non-linear local kernel and LODE features – show both a smaller overall error, and a saturation of the error for tetra and penta-peptides. This example illustrates the different approaches to achieve a multi-scale description of atomic-scale systems: the LODE(1,1) features offer simplicity and physical interpretability, while a multi-kernel model makes it possible to optimize in a data-driven manner the balance between local and long-ranged correlations.

10 Conclusions and perspectives

The work presented in this thesis demonstrates how the statistical approximation of atomic-scale properties can benefit from the adoption of mathematical representations of the atomic structure that are grounded in physical principles. We tackled three classes of fundamental problems in atomistic machine learning: i) deriving a symmetry-adapted structural representation that is suitable to learn physical tensors of arbitrary rank, ii) designing a regression algorithm able to predict scalar fields, such as the electron density of a system, in a linear-scaling and highly transferable fashion, iii) incorporating a long-range description within a local featurization of the system, that, by construction, encodes the structural information within finite spherical environments centered about the atomic positions.

We first derived a family of λ-SOAP features and kernels that follows the same transformation rules as spherical harmonics, and showed how to use them within a GPR framework that is naturally adapted to the symmetries of the $O(3)$ group. Deriving representations that follow different covariance relationships, such as those of discrete translational symmetries or point-group symmetries, is straightforward following the rationale of the construction outlined. Thanks to a prior decomposition over irreducible spherical components, we were able to recast the problem of interpolating tensorial properties into multiple independent regression tasks, each of which can be performed using the proper λ-SOAP kernel. When compared with its Cartesian counterpart [31, 32], tackling the problem of tensor-learning in the space of ISCs carries a massive advantage, as it allows us, on the one hand, to reduce the dimensionality of the problem, and, on the other hand, to exploit the possible symmetry of the tensor by discarding those ISCs that have a different parity than the tensor rank. The proposed SA-GPR method finds a natural application in the regression of dielectric response properties, and, as a consequence, it paves the way to the inexpensive calculation of any derived scattering and absorption spectra. As a relevant example, we focused the

attention on the accurate prediction of the polarizability of a system, showing that an accuracy greater than, or equal to, DFT can be achieved upon training the model on high-end coupled-cluster calculations. We then investigated the possibility of computing the vibrational Raman spectrum of a molecular crystal of paracetamol as a post-processing to the polarizability prediction, obtaining remarkable extrapolations of the line-shapes and absorption intensities across different crystal polymorphs.

The SA-GPR method has already been applied by other researchers to the prediction of the dipole moment of a system [220], as well as to compute the infrared and Raman spectrum of liquid water [238]. An important future application of the method will involve the computational simulation of sum-frequency generation (SFG) spectra, where a combination of both dipole and polarizability predictions is needed to compute the scattering intensity associated with the non-linear optical response of atomistic surfaces, e.g., electrochemical interfaces [239]. An additional field of application, not covered during our dissertation, is the one associated with the use of tensorial features that have an opposite parity under inversion symmetry. These kind of applications include, for example, the prediction of the pseudo-scalars and vectors that define the chiral response in circular dichroism spectroscopies [49], the asymmetric chemical shielding tensors that determine the peaks shape and position in solid-state NMR experiments [65], and the frequency-dependent polarizabilities that enter an *ab initio* calculation of the van der Waals dispersion coefficients [240]. Another application with immense potential is related to the calculation of the building blocks of electronic-structure methods, such as the matrix elements of single-particle Hamiltonians, $\langle a|\,\hat{H}\,|b\rangle$, written in an atom-centered, spherical harmonics basis. In this case, the problem carries a twofold complication: i) each matrix element transforms under rotations as a product of two spherical harmonics, according to the Slater-Koster rules [241], ii) the prediction of off-centered matrix elements call for a structural representation that is centered about atomic pairs, rather than on individual atoms. A tentative solution has already been proposed within a deep neural network context, which however requires to learn the rotational symmetry of the Hamiltonian via a massive dataset augmentation [242].

As a non-trivial application of the SA-GPR framework, we showed how to regress electronic charge density fields represented on a basis of spherical harmonics centered on the atomic positions. The challenge of dealing with a non-orthogonal basis has been tackled by deriving a regression algorithm that predicts each individual family of atomic density coefficients while still having the entire scalar field as the learning target. The computational burden carried by the coupling of λ-SOAP kernels with the overlap matrix between basis functions is compensated by the acquired transferability of the learning model across similar chemical environments. We showed

that this transferability opens the door to the atom-wise prediction of the electron density of molecules that are much larger than those used to train the model, such as hydrocarbons and peptides, relying on the same "divide-and-conquer" principle that underlies linear-scaling quantum-chemistry approaches [243]. Importantly, while the prediction of $n_e(\boldsymbol{r})$ allows us to access quantities such as electrostatic potentials and non-covalent interaction indexes, we found that using the electron density to indirectly predict the energy of the system yields poor performance when compared to a direct interpolation of U. In fact, the non-linear functional dependence of the electronic energy from the electron density greatly amplifies the error incurred in the statistical approximation of $n_e(\boldsymbol{r})$, making it hard to apply an indirect $n_e(\boldsymbol{r}) \rightarrow U$ approach beyond relatively simple structural manifolds [27].

The training effort carried out for the bio-fragment molecular dataset introduced in Chapter 7 has recently been exploited to predict the electron density of a full protein [244]. Although remarkable, this example underscores the importance of enlarging even further the spectrum of environments to be used as a representative basis of the structural and chemical diversity spanned by the local atomic configurations. In this regard, the current implementation of the method presents a technical bottleneck related to the prohibitive computational cost associated with the inversion of large regression matrices. This issue could in principle be bypassed by direct numerical minimization of the loss function of the problem, or implementing the alternative Löwdin approach already discussed in Chapter 6. Being generally applicable to the regression of any scalar field, the method has also been applied by other researchers to the prediction of the on-top pair density tha can be used to visualize electronic correlations [245]. A further development will involve the prediction of the electron density in the condensed-phase, which carries the additional complexity of treating the overlap between basis functions that belong to different periodic images of the unit cell. Crucially, this technical advancement could give access to accurate X-ray intensities that enter the determination of the atomic structure in crystallographic scattering experiments [156]. In addition, a possible application would concern the prediction of the spatially-resolved density of states of a material at the Fermi energy, which is the fundamental ingredient for the first-principles calculation of scanning tunneling microscopy (STM) images [246].

The desired transferability of the machine-learning model, which is implied by the local nature of the structural representation, forbids that long-range and/or non-local effects can be accurately captured. During this thesis, we encountered a clear manifestation of this problem when predicting the dielectric tensor of liquid water, as well as when trying to predict the polarizability of highly conjugated molecules. In the last part of the thesis, we proposed a solution by constructing a Coulomb-like

potential field generated by a smooth density representation of the atomic positions, that can be evaluated in a finite local environment of the system's atoms. In doing so, we derived a *long-distance equivariant* (LODE) representation that retains the additive and atom-centered nature of the learning model, while still presenting a description of long-range effects brought by the algebraic tails, $\sim 1/r$, of the potential field. We demonstrated that such a model can accurately capture the long-range nature of electrostatic interactions and describe the non-local character of the dielectric response of liquid water. We then showed that a suitable combination between density and potential features allows us to build a multi-scale version of LODE that is flexible enough to represent arbitrary interactions, including the polarization of a metallic surface and the dielectric response of peptidic chains. Importantly, such a combined density-potential representation presents also some analytical limits that allows us to rationalize the asymptotic prediction of the binding energy between two systems in terms of well-established multipolar interaction terms.

The main difficulty associated with the description of long-range phenomena is to find the balance between a functional form that is flexible enough to describe arbitrary interactions, and one that maps naturally onto the physics of the problem. On the one hand, a too general structural representation is prone to overfitting and requires enormous amounts of training data, as it appears when increasing by brute force the cutoff of a local featurization [10, 43, 236], or when adopting a global representation of the system [221]. On the other hand, pushing too far the physical consistency of the regression model makes the learning effort very system-specific and limits its broad applicability across diverse structural and chemical patterns. In the future, drawing the fine line between these two limits will be essential to assess up to which extent we can effectively use LODE features within a highly transferable machine-learning model. If successful, its application could potentially solve a broad class of problems in the data-driven simulation of materials that are to date still hindered by a lack of a long-range description. For example, a current open problem in the simulation of electrochemical interfaces consists in including a sufficiently large portion of electrolyte solution that is able to perfectly screen the surface charge collected at the electrode [195, 247]. In fact, while one can adopt suitable continuum models to represent the solution [248], the explicit simulation of such a system is way too expensive to carry out by first-principles and it would necessarily require to perform a physically consistent extrapolation of the metal-electrolyte interaction that extends well beyond the length scales that can be spanned by the reference calculations.

In perspective, the presented achievements will be integrated within efficient and reliable statistical approximation programs that can be interfaced with state-of-the-art

molecular dynamics engines. The possibility of performing high-end simulations at a cost that is at least an order of magnitude smaller than the one of first-principles methods will lead to the accurate modeling of systems over time and length scales that cannot be reached by standard *ab initio* approaches. To this end, the great challenge of enlarging either the simulation box or the molecular size without disregarding the importance of long-range phenomena will be addressed by the inclusion of LODE-derived features as a generally transferable correction term that is added to a local, many-body representation of the atomic structure. A related important aspect will consists in assessing the reliability of the data-driven predictions within those size regimes that forbid any direct comparison with quantum-level calculations. While suitable error estimations can be implemented for this purpose, disposing of a tool to readily access electronic-structure properties beyond energies and atomic forces will also be of tremendous help to pinpoint the deficiencies of the statistical approximation. For example, accessing scalar fields such as the electron density will enable one to assess up to which extent the machine-learning model is able to reproduce the overall electro-neutrality of the system, as well as to highlight any unphysical partitioning of the electronic charge over the molecular space. When it comes to the time domain, the possibility of simulating sufficiently long trajectories that meet thermodynamic convergence criteria will finally lay the groundwork for a direct comparison of computer simulations with experimental results. First and foremost, this comparison will be made through absorption and scattering spectra, for which the inexpensive prediction of electronic response tensors will be an essential cornerstone.

A Dirac notation for structural representations

A mathematical representation of the atomic structure can be interpreted as a unique characterization of the system's state that lives in an abstract Hilbert space. For this reason, we find convenient to rely on the Dirac notation routinely used in quantum mechanics. Using this notation, the abstract state of a generic structure A is indicated as a ket $|A\rangle$. The representation of this state on a given complete basis X is then interpreted as the projection of $|A\rangle$ on the bra $\langle X|$, which is indicated by the braket $\langle X|A\rangle = \langle A|X\rangle^{\star}$. This construction leaves us the freedom of introducing an arbitrary transformation of the abstract state $|A\rangle$, without necessarily specify the kind of basis chosen to represent the state of the system. For example, one could generically refer to a rotation of the system as $\hat{R}|A\rangle$, with $\hat{R}$ a rotation operator. Conversely, one could consider to apply the rotation operator to the basis as $\langle X|\hat{R}$, leaving the state $|A\rangle$ unchanged. When considering the braket $\langle X|\hat{R}|A\rangle$, the two distinct pathways can be interpreted in terms of a passive rotation of the reference frame, when the operator $\hat{R}$ is applied to the bra, or to an active rotation of the system, when the operator $\hat{R}$ is applied to the ket.

Another advantage of Dirac notation is apparent when considering the kernel defined by the inner product $k(A,B) = \langle A|B\rangle$. This definition has in fact a standalone meaning as a distance (similarity) measure between A and B, regardless from the kind of representation adopted. In this case, including a generic operator within the braket, e.g., $k(A,B) = \langle A|\hat{R}|B\rangle$, would conveniently indicate that the similarity between A and B is measured either by comparing $|A\rangle$ with respect to a rotation of $\langle B|$, or viceversa. Note that a representation of the kernel on a given complete basis X can be obtained upon including the resolution of the identity within the braket:

$$k(A,B) = \langle A|B\rangle = \langle A|\left(\sum_X |X\rangle\langle X|\right)|B\rangle = \sum_X \langle A|X\rangle\langle X|B\rangle\ . \tag{A.1}$$

To give a practical example of the use of this notation in the construction of structural representations, consider the abstract state $|A\rangle \equiv \big|\rho; V\big\rangle$, with ρ and V a generic density and potential field derived from the atomic coordinates. As in the standard Dirac notation, the semicolon stands for the fact that the representation lives in the tensor product of $\big|\rho\big\rangle$ and $|V\rangle$, i.e., $\big|\rho; V\big\rangle = \big|\rho\big\rangle \otimes |V\rangle$. Projecting such an abstract description of the system on a given basis would read, for instance,

$$\big\langle \boldsymbol{x}; \boldsymbol{k}\big|\rho; V\big\rangle = (\langle \boldsymbol{x}| \otimes \langle \boldsymbol{k}|)\left(\big|\rho\big\rangle \otimes |V\rangle\right) = \big\langle \boldsymbol{x}\big|\rho\big\rangle \langle \boldsymbol{k}|V\rangle \,, \tag{A.2}$$

with $\langle \boldsymbol{x}|$ and $\langle \boldsymbol{k}|$ indicating the basis of positions and momenta that are used to represent the density and potential fields, $\big\langle \boldsymbol{x}\big|\rho\big\rangle \equiv \rho(\boldsymbol{x})$ and $\langle \boldsymbol{k}|V\rangle \equiv V(\boldsymbol{k})$, respectively. As a result, the kernel that is derived from this representation is written as

$$\begin{aligned} k(A,B) &= \big\langle \rho(A); V(A)\big|\rho(B); V(B)\big\rangle \\ &= \big\langle \rho(A); V(A)\big|\left(\int d\boldsymbol{x}\, |\boldsymbol{x}\rangle \langle \boldsymbol{x}| \otimes \sum_{\boldsymbol{k}} |\boldsymbol{k}\rangle \langle \boldsymbol{k}|\right)\big|\rho(B); V(B)\big\rangle \\ &= \int d\boldsymbol{x} \sum_{\boldsymbol{k}} \big\langle \rho(A); V(A)\big|\boldsymbol{x}; \boldsymbol{k}\big\rangle \big\langle \boldsymbol{x}; \boldsymbol{k}\big|\rho(B); V(B)\big\rangle \,. \end{aligned} \tag{A.3}$$

Note that this notation also allows us to transparently perform any change of variable. For example, expressing the density field in reciprocal space from its real-space counterpart reads

$$\big\langle \boldsymbol{k}\big|\rho\big\rangle = \langle \boldsymbol{k}|\left(\int d\boldsymbol{x}\, |\boldsymbol{x}\rangle \langle \boldsymbol{x}|\right)\big|\rho\big\rangle = \int d\boldsymbol{x}\, \langle \boldsymbol{k}|\boldsymbol{x}\rangle \big\langle \boldsymbol{x}\big|\rho\big\rangle \,, \tag{A.4}$$

where the integral underlies a Fourier transform and $\langle \boldsymbol{k}|\boldsymbol{x}\rangle \equiv e^{-i\boldsymbol{k}\cdot\boldsymbol{x}}$ is a plane wave.

B Calculation of SOAP coefficients

Consider to compute the orthogonal projections $\langle anlm|\rho_i\rangle$ associated with the atom-centered expansion of the atom-density field, where n stands for a discrete set of orthogonal radial functions $R_n(x)$ that are defined within the spherical cutoff r_c. From the real-space definition of the density field $\langle a\boldsymbol{x}|\rho_i\rangle$ as a superposition of Gaussian functions centered on the atomic positions (Eq. (1.13)), the spherical harmonics projection can be carried out analytically [249], leading to

$$\langle anlm|\rho_i\rangle = \sum_{j\in a} \langle lm|\hat{\mathbf{r}}_{ij}\rangle \exp\left\{-\frac{|\mathbf{r}_i-\mathbf{r}_j|^2}{2\sigma^2}\right\} \int_0^\infty \mathrm{d}x\, x^2\, \langle n|x\rangle \exp\left\{-\frac{x^2}{2\sigma^2}\right\} \imath_l\left(\frac{x r_{ij}}{\sigma^2}\right), \tag{B.1}$$

where the sum over j runs over the neighboring atoms of a given type a, and $\imath_l$ represents a modified spherical Bessel function of the first kind. Under suitable choices of the functions $\langle n|x\rangle \equiv R_n(r)$, the radial integration can be carried out analytically, too. One possibility is to start with non-orthogonal Gaussian type functions, $\tilde{R}_k(x)$, reminiscent of the Gaussian-type orbitals commonly used in quantum chemistry:

$$\tilde{R}_k(x) = \mathcal{N}_k\, x^k \exp\left\{-\frac{1}{2}\left(\frac{x}{\sigma_k}\right)^2\right\}, \tag{B.2}$$

where $\mathcal{N}_k$ is a normalization factor, such that $\int_0^\infty dr\, x^2 \tilde{R}_k^2(x) = 1$. The set of Gaussian widths $\{\sigma_k\}$ can be chosen to uniformly span the radial interval $[0, r_c]$. In our case, we consider $\sigma_k = r_c \max(\sqrt{k}, 1)/n_{\max}$. The explicit formula of the primitive radial integrals

is

$$\int_0^\infty \mathrm{d}x\, x^2\, \tilde{R}_k(x)\, e^{-\frac{x^2}{2\sigma^2}} \imath_l\left(\frac{x r_i}{\sigma^2}\right) =$$
$$\mathcal{N}_k\, 2^{-\frac{1}{2}(1+l-k)} \left(\frac{1}{\sigma^2}+\frac{1}{\sigma_k^2}\right)^{-\frac{3+l+k}{2}} \frac{\Gamma(\frac{3+l+k}{2})}{\Gamma(\frac{3}{2}+l)} \left(\frac{r_i}{\sigma^2}\right)^l {}_1F_1\left(\frac{3+l+k}{2}, \frac{3}{2}+l;\, \frac{1}{2}\frac{\sigma_k^2\, r_i^2}{\sigma^4+\sigma_k^2\sigma^2}\right), \tag{B.3}$$

where Γ is the Gamma function, while ${}_1F_1$ is the confluent hypergeometric function of the first kind. These integrals can be finally orthogonalized by applying the Löwdin orthogonalization matrix $\boldsymbol{S}^{-1/2}$, with $\boldsymbol{S}$ the overlap between primitive functions, i.e., $S_{kk'} = \int_0^\infty dx\, x^2 \tilde{R}_k(x) \tilde{R}_{k'}(x)$, which can also be computed analytically. A representation of the orthogonal radial functions up to $n_{\max} = 8$ is reported in Fig. B.1.

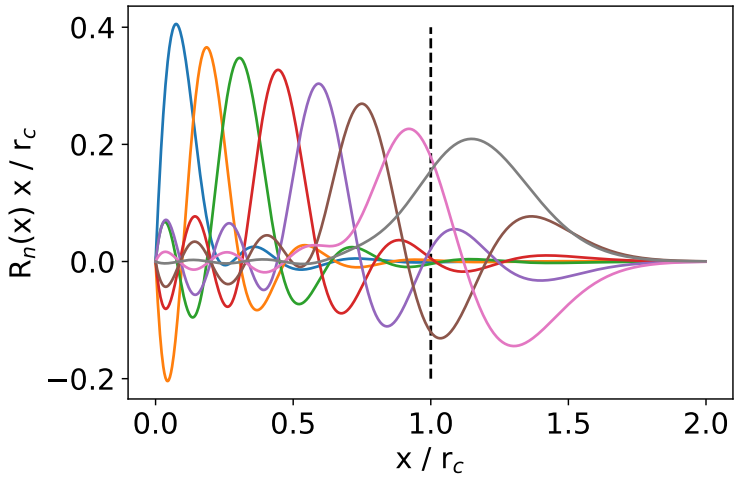


Figure B.1 – Representation of 8 orthogonal radial functions, multiplied by x, built to evenly span the radial interval $[0, r_c]$. Distances are reported in units of r_c.

C Calculation of LODE coefficients

We report below the details of the real and reciprocal space calculation of the potential harmonic coefficients, $\langle anlm|V_i\rangle$, that enter the construction of the LODE framework.

C.1 Direct-space formulation for finite systems

For a given atom-type a, the electrostatic potential generated by a localized Gaussian density distribution is [5]

$$\langle a\boldsymbol{x}|V\rangle = \sum_j \delta_{a_j a} \frac{1}{|\boldsymbol{x}-\boldsymbol{r}_j|} \mathrm{erf}\left(\frac{|\boldsymbol{x}-\boldsymbol{r}_j|}{\sqrt{2}\sigma}\right), \tag{C.1}$$

with σ the Gaussian width and erf$(\cdot)$ an error function. Upon centering the field about an atom i, the spherical harmonic projections can be worked out analytically as

$$\langle axlm|V_i\rangle = \frac{4\pi}{2l+1} \sum_j \delta_{a_j a} \exp\left\{-\frac{1}{2}\left(\frac{r_{ij}}{\sigma}\right)^2\right\} Y_{lm}^{\star}(\hat{\boldsymbol{r}}_{ij}) \times$$
$$\left[\frac{1}{x^{l+1}} \int_0^x dx'\, x'^{2+l} \exp\left\{-\frac{1}{2}\left(\frac{x'}{\sigma}\right)^2\right\} \iota_l\left(\frac{x' r_{ij}}{\sigma^2}\right) + x^l \int_x^\infty dx'\, x'^{1-l} \exp\left\{-\frac{1}{2}\left(\frac{x'}{\sigma}\right)^2\right\} \iota_l\left(\frac{x' r_{ij}}{\sigma^2}\right)\right] \tag{C.2}$$

with $\boldsymbol{r}_{ij} = \boldsymbol{r}_j - \boldsymbol{r}_i$ and ι_l a modified spherical Bessel functions of the first kind. To produce the results presented in this thesis, however, we directly sample Eq. (C.1) on atom-centered spherical grids and perform the projections on the basis functions by numerical integration. In particular, a Gauss-Legendre quadrature of 50 points and a Lebedev quadrature [154] of 146 points were used to compute the radial and spherical projections, respectively.

C.2 Reciprocal-space formulation for periodic systems

Given Ω the cell volume, the Fourier transform of an a-type Gaussian density is

$$\langle a\boldsymbol{k}|\rho\rangle = \frac{1}{\Omega}\int d\boldsymbol{x}\,\langle \boldsymbol{k}|\boldsymbol{x}\rangle\,\langle a\boldsymbol{x}|\rho\rangle = \frac{1}{\Omega}\left(\sum_j \delta_{a_j a} e^{-i\boldsymbol{k}\cdot\boldsymbol{r}_j}\right) e^{-\frac{k^2\sigma^2}{2}}\,. \tag{C.3}$$

Then, given the plane-waves solution of the Poisson equation,

$$\nabla^2 V(\boldsymbol{r}) = -4\pi\rho(\boldsymbol{r}) \;\longrightarrow\; -k^2 V(\boldsymbol{k}) = -4\pi\rho(\boldsymbol{k})\,, \tag{C.4}$$

we can write the potential in real space as

$$\langle a\boldsymbol{x}|V\rangle = \frac{1}{\Omega}\sum_{\boldsymbol{k}}\left(\sum_j \delta_{a_j a} e^{-i\boldsymbol{k}\cdot\boldsymbol{r}_j}\right)\frac{4\pi}{k^2} e^{-\frac{k^2\sigma^2}{2}} e^{i\boldsymbol{k}\cdot\boldsymbol{x}}\,. \tag{C.5}$$

Consider now the spherical harmonics expansion of the plane wave:

$$e^{i\boldsymbol{k}\cdot\boldsymbol{x}} = 4\pi\sum_{l=0}^{\infty}\sum_{|m|\le l} i^l\, j_l(kr)\, Y_{lm}^{\star}(\hat{\boldsymbol{k}})\, Y_{lm}(\hat{\boldsymbol{x}})\,, \tag{C.6}$$

with j_l a spherical Bessel function. Upon centering the field on an atom i, which brings an additional phase factor $e^{i\boldsymbol{k}\cdot\boldsymbol{r}_i}$, the expansion above allows us to single out the spherical harmonic components of the potential, obtaining

$$\langle arlm|V_i\rangle = \frac{16\pi^2}{\Omega}\sum_{\boldsymbol{k}}\left(\sum_j \delta_{a_j a} e^{-i\boldsymbol{k}\cdot\boldsymbol{r}_{ij}}\right)\frac{e^{-\frac{k^2\sigma^2}{2}}}{k^2} i^l\, j_l(kr)\, Y_{lm}^{\star}(\hat{\boldsymbol{k}})\,. \tag{C.7}$$

Finally, we can project on GTO-like primitive radial functions to get the discretized set of LODE coefficients:

$$\langle anlm|V_i\rangle = \frac{16\pi^2}{\Omega}\sum_{\boldsymbol{k}}\left(\sum_j \delta_{a_j a} e^{-i\boldsymbol{k}\cdot\boldsymbol{r}_{ij}}\right)\frac{e^{-\frac{k^2\sigma^2}{2}}}{k^2} i^l\, I_{nl}(k)\; Y_{lm}^{\star}(\hat{\boldsymbol{k}})\,. \tag{C.8}$$

The radial integral can be computed analytically as

$$\begin{aligned} I_{nl}(k) &= \frac{1}{\mathcal{N}_n}\int_0^{\infty} dr\; r^{2+n}\,\exp\left\{-\frac{1}{2}\left(\frac{r}{\sigma_n}\right)^2\right\} j_l(kr) = \\ &= \frac{1}{\mathcal{N}_n}\sqrt{\pi}\,2^{\frac{1}{2}(n-l-1)}\;k^l\,\sigma_n^{3+n+l}\;\frac{\Gamma\left(\frac{1}{2}(3+n+l)\right)}{\Gamma\left(\frac{3}{2}+l\right)}\,{}_1F_1\left(\frac{1}{2}(3+n+l),\frac{3}{2}+l,-\frac{1}{2}(k\sigma_n)^2\right) \end{aligned} \tag{C.9}$$

with $\{\sigma_n\}$ the set of GTOs widths chosen to have equally spaced peaks within the cutoff radius r_c and $\mathcal{N}_n$ a normalization factors. Note that the convergence of this integral is guaranteed by the superexponential decay of GTOs. The choice of primitive radial functions and the orthogonalization of the radial projections follows the same procedure discussed in Appendix B.

In compact Dirac notation, Eq. (C.8) can be simply written as

$$\langle anlm|V_i\rangle = \sum_{\boldsymbol{k}} \langle nlm|\boldsymbol{k}\rangle \langle a\boldsymbol{k}|V_i\rangle \,, \tag{C.10}$$

with

$$\langle nlm|\boldsymbol{k}\rangle = 4\pi\, i^l\, I_{nl}(k)\, Y^{\star}_{lm}(\hat{\boldsymbol{k}}) \tag{C.11}$$

the scattering partial-wave coefficients of pure geometric nature, and

$$\langle a\boldsymbol{k}|V_i\rangle = \frac{1}{\Omega}\left(\sum_j \delta_{a_j a} e^{-i\boldsymbol{k}\cdot\boldsymbol{r}_{ij}}\right)\frac{4\pi}{k^2} e^{-\frac{k^2\sigma^2}{2}} \tag{C.12}$$

the Fourier components of the species-dependent potential field. One can make a few remarks on these expressions:

- The information about the system is entirely included in the combination of the complex phase factors that depend on the atomic positions. All the other terms can be computed only once for each atomic configuration if the box size is allowed to vary, and only once for each ensemble of atomic configurations if the box size is kept fixed.
- Assuming a statistically uniform distribution of atoms in the supercell, the calculation of the system-dependent phase factors scales quadratically with the number of atoms N. To alleviate the cost of this operation, one should in principle adopt suitable fast Fourier transform (FFT) algorithms that reduce the cost from N^2 to $N\log N$ [250].
- Given the real nature of the potential field, the sum over the $\boldsymbol{k}$-vectors can be restricted on a semi-sphere in reciprocal space, e.g., $k_x > 0$ for orthorombic cells, with the sphere radius $k_{\max} = 2\pi/\lambda_{\min}$ defined by the minimum wavelength introduced in the calculation, $\lambda_{\min} \sim \sigma$. Then, given the parity of the spherical harmonics under inversion of the direction of $\boldsymbol{k}$, which brings a $(-1)^l$ phase, one can replace the complex exponential by $2\cos\left(\boldsymbol{k}\cdot\boldsymbol{r}_{ij}\right)$ and $-2i\sin\left(\boldsymbol{k}\cdot\boldsymbol{r}_{ij}\right)$ for even and odd values of l, respectively.

C.3 Plain Ewald method

To afford the calculation of the spherical harmonic projections associated with a periodic potential generated by an arbitrarily sharp atom-density distribution, we rely on a plain implementation of the Ewald method. For each different atomic species a, we introduce a smooth Gaussian density $\tilde{\rho}_{\sigma'}$, with $\sigma' > \sigma$, that is able to perfectly screen the original density field ρ_σ, i.e.,

$$\langle a\boldsymbol{x}|\rho_{\mathrm{S}}\rangle = \langle a\boldsymbol{x}|\rho_\sigma\rangle - \langle a\boldsymbol{x}|\tilde{\rho}_{\sigma'}\rangle\,, \tag{C.13}$$

such that,

$$\int d\boldsymbol{x}\,\langle a\boldsymbol{x}|\rho_{\mathrm{S}}\rangle = 0\,. \tag{C.14}$$

The resulting potential is short-ranged, meaning that it decays as fast as the compensating Gaussian density $\tilde{\rho}_{\sigma'}$; as such, its spherical harmonics projections can be computed in real space using the same implementation discussed in Sec. C.1. The remaining long-ranged (unscreened) potential generated by $\tilde{\rho}_{\sigma'}$ is smooth enough to be represented via a manageable number in plane-waves; as such, its spherical harmonics projections can be computed in reciprocal space using the same implementation discussed in Sec. C.2. Finally, the two contributions can be added together to obtain the desired potential field projections associated with the arbitrarily sharp Gaussian density ρ_σ.

D Electron-nucleus interaction on a density-fitted basis

Consider to compute the interaction of the electron density, represented on a linear density-fitted basis, with the nuclear potential of a system of N atoms. In atomic units,

$$\begin{aligned} U_{\text{ex}}[n_e] &= -\sum_{i=1}^{N} Z_i \int d\mathbf{r}\, \frac{n_e(\mathbf{r})}{|\mathbf{r}-\mathbf{r}_i|} \\ &\approx -\sum_{i=1}^{N} Z_i \sum_{j=1}^{N} \sum_{nlm} c^j_{nlm} \int d\mathbf{r}\, \frac{R_n(|\boldsymbol{r}-\mathbf{r}_j|)\, Y_{lm}(\widehat{\boldsymbol{r}-\boldsymbol{r}_j})}{|\mathbf{r}-\mathbf{r}_i|} \\ &= -\sum_{i=1}^{N} Z_i \sum_{j=1}^{N} \sum_{nlm} c^j_{nlm} \int d\mathbf{r}\, \frac{R_n(r)\, Y_{lm}(\theta,\phi)}{|\mathbf{r}-\mathbf{r}_{ij}|}, \end{aligned} \tag{D.1}$$

where we adopted the change the of variable $\mathbf{r} \to \mathbf{r}-\mathbf{r}_j$ and set $\mathbf{r}_{ij} = \mathbf{r}_i - \mathbf{r}_j$ in the last equality. Assuming the space is isotropic, we now introduce a rotation of the reference frame in such a way that the direction of $\mathbf{r}_{ij}$ is aligned with the $\hat{\mathbf{z}}$-axis. Upon this transformation, the spherical harmonics undergo a passive rotation that is expressed by the proper Wigner-D matrix:

$$U_{\text{ex}}[n_e] = -\sum_{i=1}^{N} Z_i \sum_{j=1}^{N} \sum_{nlm} c^j_{nlm} \sum_{m'} D^{l\dagger}_{mm'}(\hat{\mathbf{r}}_{ij} \to \hat{\mathbf{z}}) \int d\mathbf{r}\, \frac{R_n(r)\, Y_{lm'}(\theta,\phi)}{|\mathbf{r}-\mathbf{r}_{ij}|}, \tag{D.2}$$

where now $\mathbf{r}_{ij} \parallel \hat{\mathbf{z}}$ within the integral. This alignment allows us to rewrite

$$|\mathbf{r}-\mathbf{r}_{ij}| = \sqrt{r^2 + r_{ij}^2 - 2 r\, r_{ij} \cos\theta}\,. \tag{D.3}$$

The spherical harmonics within the integral can now be averaged over the azimuthal angle ϕ. Upon this procedure, only the $m' = 0$ components survive and we are left

with the following expression,

$$U_{\text{ex}}[n_e] = -\sum_{i=1}^{N} Z_i \sum_{j=1}^{N} \sum_{nl} \left(\sum_{m} c^{j}_{nlm} Y_{lm}(\hat{\mathbf{r}}_{ij}) \right) \\ 2\pi \int_0^{\infty} dr\, r^2 R_n(r) \int_{-1}^{1} d\cos\theta \frac{P_l(\cos\theta)}{\sqrt{r^2 + r_{ij}^2 - 2r r_{ij}\cos\theta}}, \tag{D.4}$$

where we used the identity $D^{l}_{0m}(\hat{\mathbf{r}}_{ij} \rightarrow \hat{\mathbf{z}}) = \sqrt{\frac{4\pi}{2l+1}} Y_{lm}(\hat{\mathbf{r}}_{ij})$. At this point, we introduce the Laplace expansion of the Coulomb potential, i.e.,

$$\begin{cases} \frac{1}{\sqrt{r^2+r_{ij}^2-2r r_{ij}\cos\theta}} = \sum_l \frac{r^l}{r_{ij}^{l+1}} P_l(\cos\theta) & \text{for } r < r_{ij} \\ \frac{1}{\sqrt{r^2+r_{ij}^2-2r r_{ij}\cos\theta}} = \sum_l \frac{r_{ij}^l}{r^{l+1}} P_l(\cos\theta) & \text{for } r > r_{ij} \end{cases} \tag{D.5}$$

Plugging into Eq. (D.4), the orthogonality of Legendre polynomials can finally be exploited to obtain

$$U_{\text{ex}}[n_e] = -\sum_{i=1}^{N} Z_i \sum_{j=1}^{N} \sum_{nl} \left(\sum_{m} c^{j}_{nlm} Y_{lm}(\hat{\mathbf{r}}_{ij}) \right) \\ \frac{4\pi}{2l+1} \left(\frac{1}{r_{ij}^{l+1}} \int_0^{r_{ij}} dr\, r^{2+l} R_n(r) + r_{ij}^{l} \int_{r_{ij}}^{\infty} dr\, r^{1-l} R_n(r) \right). \tag{D.6}$$

Note that if $i = j$ then $r_{ij} = 0$ and $Y_{lm}(\hat{\mathbf{r}}_{ij}) = \frac{1}{\sqrt{4\pi}} \delta_{l0}\delta_{m0}$, which implies that only the isotropic density components contribute to the electron-nucleus interaction. Under this limit, the previous formula simplifies to

$$-\sum_{i}^{N} Z_i \sum_{n} c^{i}_{n00} \sqrt{4\pi} \int_0^{\infty} dr\, r R_n(r). \tag{D.7}$$

As a final remark, note that all the radial integrals can be computed analytically if the density is expanded on a basis of GTOs.

Bibliography

[1] C. Coulson and R. McWeeny, *Coulson's valence*, Oxford Chemistry Series (Oxford University Press, 1979).

[2] P. Hohenberg and W. Kohn, "Inhomogeneous electron gas", Phys. Rev. **136**, B864–B871 (1964).

[3] K. Burke, "Deriving approximate functionals with asymptotics", Faraday Discuss., - (2020).

[4] R. G. Parr and W. Yang, *Density-functional theory of atoms and molecules*, 1. iss. as ... paperback, International Series of Monographs on Chemistry 16 (Oxford Univ. Press [u.a.], New York, NY, 1994).

[5] D. Marx and J. Hutter, "Ab initio molecular dynamics: theory and implementation", in Modern Methods and Algorithms of Quantum Chemistry, Vol. 1, edited by J. Grotendorst (2000), pp. 301–449.

[6] D. Frenkel and B. Smit, *Understanding Molecular Simulation*, Second (Academic Press, London, 2002).

[7] M. Tuckerman, *Statistical Mechanics and Molecular Simulations* (Oxford University Press, 2008).

[8] A. Ben-Naim, *Solvation thermodynamics* (Springer US, 2013).

[9] J. Behler, "First principles neural network potentials for reactive simulations of large molecular and condensed systems", Angewandte Chemie International Edition **56**, 12828–12840 (2017).

[10] A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi, and M. Ceriotti, "Machine learning unifies the modeling of materials and molecules", Sci. Adv. **3**, e1701816 (2017).

[11] M. Veit, S. K. Jain, S. Bonakala, I. Rudra, D. Hohl, and G. Csányi, "Equation of State of Fluid Methane from First Principles with Machine Learning Potentials", J. Chem. Theory Comput. **15**, 2574–2586 (2019).

[12]B. Cheng, E. A. Engel, J. Behler, C. Dellago, and M. Ceriotti, "Ab initio thermodynamics of liquid and solid water", Proc. Natl. Acad. Sci. U. S. A. **116**, 1110–1115 (2019).

[13]B. Cheng, G. Mazzola, C. J. Pickard, and M. Ceriotti, "Evidence for supercritical behaviour of high-pressure liquid hydrogen", Nature **585**, 217–220 (2020).

[14]F. Libbi, N. Bonini, and N. Marzari, "Thermomechanical properties of honeycomb lattices from internal-coordinates potentials: the case of graphene and hexagonal boron nitride", 2D Materials **8**, 015026 (2020).

[15]C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)* (Springer, 2011).

[16]B. Jiang and H. Guo, "Permutation invariant polynomial neural network approach to fitting potential energy surfaces", The Journal of Chemical Physics **139**, 054112 (2013).

[17]K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, "Quantum-chemical insights from deep tensor neural networks", Nat. Commun. **8**, 13890 (2017).

[18]K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, "SchNet – A deep learning architecture for molecules and materials", J. Chem. Phys. **148**, 241722 (2018).

[19]L. Zhang, J. Han, H. Wang, R. Car, and W. E, "Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics", Phys. Rev. Lett. **120**, 143001 (2018).

[20]K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. von Lilienfeld, A. Tkatchenko, and K.-R. Müller, "Assessment and validation of machine learning methods for predicting molecular atomization energies", Journal of Chemical Theory and Computation **9**, 3404–3419 (2013).

[21]C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)* (The MIT Press, 2005).

[22]J. Quiñonero-Candela and C. E. Rasmussen, "A unifying view of sparse approximate gaussian process regression", J. Mach. Learn. Res. **6**, 1939–1959 (2005).

[23]S. De, A. P. Bartók, G. Csányi, and M. Ceriotti, "Comparing molecules and solids across structural and alchemical space", Phys. Chem. Chem. Phys. **18**, 13754–13769 (2016).

[24]S. De, F. Musil, T. Ingram, C. Baldauf, and M. Ceriotti, "Mapping and classifying molecules from a high-throughput structural database", J. Cheminformatics **9**, 1–14 (2017).

[25]A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, "Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons", Phys. Rev. Lett. **104**, 136403 (2010).

[26]C. Scherer, R. Scheid, D. Andrienko, and T. Bereau, "Kernel-Based Machine Learning for Efficient Simulations of Molecular Liquids", J. Chem. Theory Comput. **16**, 3194–3204 (2020).

[27]F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, K. Burke, and K. R. Müller, "Bypassing the Kohn-Sham equations with machine learning", Nat. Commun. **8**, 872 (2017).

[28]M. Bogojeski, L. Vogt-Maranto, M. E. Tuckerman, K.-R. Müller, and K. Burke, "Quantum chemical accuracy from density functional approximations via machine learning", Nat. Commun. **11**, 5223 (2020).

[29]J. C. Snyder, M. Rupp, K. Hansen, K.-R. Müller, and K. Burke, "Finding Density Functionals with Machine Learning", Phys. Rev. Lett. **108**, 253002 (2012).

[30]R. Meyer, M. Weichselbaum, and A. W. Hauser, "Machine learning approaches toward orbital-free density functional theory: simultaneous training on the kinetic energy density functional and its functional derivative", Journal of Chemical Theory and Computation **16**, 5685–5694 (2020).

[31]A. Glielmo, P. Sollich, and A. De Vita, "Accurate interatomic force fields via machine learning with covariant kernels", Phys. Rev. B **95**, 214302 (2017).

[32]T. Bereau, R. A. DiStasio, A. Tkatchenko, and O. A. von Lilienfeld, "Non-covalent interactions across organic and biological subsets of chemical space: Physics-based potentials parametrized from machine learning", The Journal of Chemical Physics **148**, 241706 (2018).

[33]C. Liang, G. Tocci, D. M. Wilkins, A. Grisafi, S. Roke, and M. Ceriotti, "Solvent fluctuations and nuclear quantum effects modulate the molecular hyperpolarizability of water", Phys. Rev. B **96**, 041407 (2017).

[34]W. Kohn, "Density Functional and Density Matrix Method Scaling Linearly with the Number of Atoms", Phys. Rev. Lett. **76**, 3168–3171 (1996).

[35]E. Prodan and W. Kohn, "Nearsightedness of electronic matter", Proc. Natl. Acad. Sci. **102**, 11635–11638 (2005).

[36]J. Behler and M. Parrinello, "Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces", Phys. Rev. Lett. **98**, 146401 (2007).

[37]M. A. Collins and D. F. Parsons, "Implications of rotation–inversion–permutation invariance for analytic molecular potential energy surfaces", The Journal of Chemical Physics **99**, 6756–6772 (1993).

[38]M. Rupp, A. Tkatchenko, K.-R. Müller, and A. von Lilienfeld, "Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning", Phys. Rev. Lett. **108**, 058301 (2012).

[39]K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. Von Lilienfeld, K. R. Müller, and A. Tkatchenko, "Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space", J. Phys. Chem. Lett. **6**, 2326–2331 (2015).

[40]Z. Xie and J. M. Bowman, "Permutationally Invariant Polynomial Basis for Molecular Energy Surface Fitting via Monomial Symmetrization", J. Chem. Theory Comput. **6**, 26–34 (2010).

[41]G. Montavon, K. Hansen, S. Fazli, M. Rupp, F. Biegler, A. Ziehe, A. Tkatchenko, A. V. Lilienfeld, and K.-R. Müller, "Learning invariant representations of molecules for atomization energy prediction", in *Advances in neural information processing systems 25*, edited by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Curran Associates, Inc., 2012), pp. 440–448.

[42]A. P. Bartók, R. Kondor, and G. Csányi, "On representing chemical environments", Phys. Rev. B **87**, 184115 (2013).

[43]A. Grisafi and M. Ceriotti, "Incorporating long-range physics in atomic-scale machine learning", J. Chem. Phys. **151**, 204105 (2019).

[44]O. Çaylak, A. von Lilienfeld, and B. Baumeier, "Wasserstein metric for improved quantum machine learning with adjacency matrix representations", Mach. Learn.: Sci. Technol., 10.1088/2632-2153/aba048 (2020).

[45]M. J. Willatt, F. Musil, and M. Ceriotti, "Atom-density representations for machine learning", J. Chem. Phys. **150**, 154110 (2019).

[46]K. Cahill, *Physical mathematics* (Cambridge University Press, 2013).

[47]J. Hansen and I. McDonald, *Theory of simple liquids* (Elsevier Science, 2006).

[48]R. H. Meißner, J. Schneider, P. Schiffels, and L. Colombi Ciacchi, "Computational prediction of circular dichroism spectra and quantification of helicity loss upon peptide adsorption on silica", Langmuir **30**, PMID: 24627945, 3487–3494 (2014).

[49]M. Chen, T. Wu, K. Xiao, T. Zhao, Y. Zhou, Q. Zhang, and J. Aires-de-Sousa, "Machine learning to predict the specific optical rotations of chiral fluorinated molecules", Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy **223**, 117289 (2019).

[50]V. L. Deringer and G. Csányi, "Machine learning based interatomic potential for amorphous carbon", Phys. Rev. B **95**, 094203 (2017).

[51]V. L. Deringer, M. A. Caro, and G. Csányi, "A general-purpose machine-learning force field for bulk and nanostructured phosphorus", Nat. Commun. **11** (2020).

[52]F. Musil, S. De, J. Yang, J. E. Campbell, G. M. Day, and M. Ceriotti, "Machine learning for the structure-energy-property landscapes of molecular crystals", Chem. Sci. **9**, 1289–1300 (2018).

[53]A. Glielmo, C. Zeni, and A. De Vita, "Efficient nonparametric n -body force fields from machine learning", Phys. Rev. B **97**, 184307 (2018).

[54]J. Nigam, S. Pozdnyakov, and M. Ceriotti, "Recursive evaluation and iterative contraction of N -body equivariant features", J. Chem. Phys. **153**, 121101 (2020).

[55]S. N. Pozdnyakov, M. J. Willatt, A. P. Bartók, C. Ortner, G. Csányi, and M. Ceriotti, "Incompleteness of atomic structure representations", Phys. Rev. Lett. **125**, 166001 (2020).

[56]A. Grisafi, D. M. Wilkins, M. J. Willatt, and M. Ceriotti, "Atomic-Scale Representation and Statistical Learning of Tensorial Properties", in *Machine Learning in Chemistry*, Vol. 1326, edited by E. O. Pyzer-Knapp and T. Laino (American Chemical Society, Washington, DC, Jan. 2019), pp. 1–21.

[57]A. Grisafi, D. M. Wilkins, G. Csányi, and M. Ceriotti, "Symmetry-Adapted Machine Learning for Tensorial Properties of Atomistic Systems", Phys. Rev. Lett. **120**, 036002 (2018).

[58]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).

[59]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).

[60]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).

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

[62]A. Grisafi, J. Nigam, and M. Ceriotti, "Multi-scale approach for the prediction of atomic scale properties", Chem. Sci. **12**, 2078–2090 (2021).

[63]T. Bereau, D. Andrienko, and O. A. Von Lilienfeld, "Transferable Atomic Multipole Machine Learning Models for Small Organic Molecules", J. Chem. Theory Comput. **11**, 3225–3233 (2015).

[64]A. J. Stone, "Transformation between cartesian and spherical tensors", Mol. Phys. **29**, 1461–1471 (1975).

[65]J. C. Facelli, "Calculations of chemical shieldings: theory and applications", Concepts in Magnetic Resonance Part A **20A**, 42–69 (2004).

[66]M. A. Blanco, M. Flórez, and M. Bermejo, "Evaluation of the rotation matrices in the basis of real spherical harmonics", J. Mol. Struct. **419**, 19 (1997).

[67]R. Kondor and S. Trivedi, "On the generalization of equivariance and convolution in neural networks to the action of compact groups", arXiv:1802.03690 (2018).

[68]D. Wilkins, A. Grisafi, A. Anelli, G. Fraux, J. Nigam, E. Baldi, L. Folkmann, and M. Ceriotti, *Tensoap: program for doing symmetry-adapted regression of tensorial properties*, https://github.com/dilkins/TENSOAP, 2020.

[69]R. Resta, "Electrical polarization and orbital magnetization: the modern theories", Journal of Physics: Condensed Matter **22**, 123201 (2010).

[70]P. Umari and A. Pasquarello, "Ab initio molecular dynamics in a finite homogeneous electric field", Phys. Rev. Lett. **89**, 157602 (2002).

[71]P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, A. D. Corso, S. de Gironcoli, S. Fabris, G. Fratesi, R. Gebauer, U. Gerstmann, C. Gougoussis, A. Kokalj, M. Lazzeri, L. Martin-Samos, N. Marzari, F. Mauri, R. Mazzarello, S. Paolini, A. Pasquarello, L. Paulatto, C. Sbraccia, S. Scandolo, G. Sclauzero, A. P. Seitsonen, A. Smogunov, P. Umari, and R. M. Wentzcovitch, "QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials", J. Phys. Condens. Matter **21**, 395502–395519 (2009).

[72]N. A. Spaldin, "A beginner's guide to the modern theory of polarization", Journal of Solid State Chemistry **195**, Polar Inorganic Materials: Design Strategies and Functional Properties, 2–10 (2012).

[73]M. Ceriotti, J. More, and D. E. Manolopoulos, "I-PI: A Python interface for ab initio path integral molecular dynamics simulations", Comput. Phys. Commun. **185**, 1019–1026 (2014).

[74]S. Habershon, T. E. Markland, and D. E. Manolopoulos, "Competing quantum effects in the dynamics of a flexible water model.", J. Chem. Phys. **131**, 24501 (2009).

[75]J. A. Hayward and J. R. Reimers, "Unit cells for the simulation of hexagonal ice", J. Chem. Phys. **106**, 1518–1529 (1997).

[76]A. Stone, *The theory of intermolecular forces*, International Series of Monographs on Chemistry (Clarendon Press, 1997).

[77]J. Hermann, R. A. DiStasio, and A. Tkatchenko, "First-principles models for van der waals interactions in molecules and materials: concepts, theory, and applications", Chemical Reviews **117**, 4714–4758 (2017).

[78]Y. R. Shen, "Surface properties probed by second harmonic and sum-frequency generation", Nature **337**, 519 (1989).

[79]S. Luber, M. Iannuzzi, and J. Hutter, "Raman spectra from ab initio molecular dynamics and its application to liquid s-methyloxirane", J. Chem. Phys. **141**, 094503 (2014).

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

[81]G. R. Medders and F. Paesani, "Dissecting the molecular structure of the air/water interface from quantum simulations of the sum-frequency generation spectrum", Chem. Phys. Lett. **138**, 11 (2016).

[82]G. S. Fanourgakis and S. S. Xantheas, "Development of transferable interaction potentials for water. v. extension of the flexible, polarizable, thole-type model potential (TTM3-F, v. 3.0) to describe the vibrational spectra of water clusters and liquid water", J. Chem. Phys. **128**, 074506 (2008).

[83]J. W. Ponder, C. Wu, P. Ren, V. S. Pande, J. D. Chodera, M. J. Schnieders, I. Haque, D. L. Mobley, D. S. Lambrecht, R. A. DiStasio, M. Head-Gordon, G. N. I. Clark, M. E. Johnson, and T. Head-Gordon, "Current status of the AMOEBA polarizable force field.", J Phys Chem B **114**, 2549–2564 (2010).

[84]G. R. Medders, V. Babin, and F. Paesani, "Development of a "first-principles"water potential with flexible monomers. III. Liquid phase properties", J. Chem. Theory Comput. **10**, 2906–2910 (2014).

[85]D. Hait and M. Head-Gordon, "How accurate are static polarizability predictions from density functional theory? an assessment over 132 species at equilibrium geometry", Phys. Chem. Chem. Phys. **20**, 19800–19810 (2018).

[86]H. J. Monkhorst, "Calculation of properties with the coupled-cluster method", Int. J. Quantum Chem. **12**, 421–432 (1977).

[87]H. Koch and P. Jørgensen, "Coupled cluster response functions", The Journal of Chemical Physics **93**, 3333–3344 (1990).

[88]O. Christiansen, P. Jørgensen, and C. Hättig, "Response functions from fourier component variational perturbation theory applied to a time-averaged quasienergy", Int. J. Quantum Chem. **68**, 1–52 (1998).

[89]O. Christiansen, J. Gauss, and J. F. Stanton, "Frequency-dependent polarizabilities and first hyperpolarizabilities of CO and H_2O from coupled cluster calculations", Chemical Physics Letters **305**, 147–155 (1999).

[90]J. R. Hammond, W. A. de Jong, and K. Kowalski, "Coupled-cluster dynamic polarizabilities including triple excitations", J. Chem. Phys. **128**, 224102 (2008).

[91]J. R. Hammond, N. Govind, K. Kowalski, J. Autschbach, and S. S. Xantheas, "Accurate dipole polarizabilities for water clusters n=2-12 at the coupled-cluster level of theory and benchmarking of various density functionals", J. Chem. Phys. **131**, 214103:1–9 (2009).

[92]K. U. Lao, J. Jia, R. Maitra, and R. A. DiStasio Jr., "On the geometric dependence of the molecular dipole polarizability in water: a benchmark study of higher-order electron correlation, basis set incompleteness error, core electron effects, and zero-point vibrational contributions", J. Chem. Phys. **149**, 204303:1–17 (2018).

[93]F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley, and A. von Lilienfeld, "Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error", J. Chem. Theory Comput. **13**, 5255–5264 (2017).

[94]R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. Von Lilienfeld, "Big data meets quantum chemistry approximations: The Δ-machine learning approach", J. Chem. Theory Comput. **11**, 2087–2096 (2015).

[95]G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K. R. Müller, and O. Anatole Von Lilienfeld, "Machine learning of molecular electronic properties in chemical compound space", New J. Phys. **15**, 095003 (2013).

[96]L. C. Blum and J.-L. Reymond, "970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13", J. Am. Chem. Soc. **131**, 8732–8733 (2009).

[97]A. D. Becke, "Density-functional thermochemistry. III. The role of exact exchange", J. Chem. Phys. **98**, 5648 (1993).

[98]Y. Yang, K.-U. Lao, D. M. Wilkins, A. Grisafi, and M. Ceriotti, "Quantum mechanical static dipole polarizabilities in the QM7b and AlphaML showcase databases", Sci Data **6**, 152 (2019).

[99]O. Christiansen, C. Hättig, and J. Gauss, "Polarizabilities of co, n2, hf, ne, bh, and ch+ from ab initio calculations: systematic studies of electron correlation, basis set errors, and vibrational contributions", J. Chem. Phys. **109**, 4745–4757 (1998).

[100]D. E. Woon and T. H. Dunning Jr., "Gaussian basis sets for use in correlated molecular calculations. iv. calculation of static electrical response properties", J. Chem. Phys. **100**, 2975–2988 (1994).

[101] G. Imbalzano, A. Anelli, D. Giofré, S. Klees, J. Behler, and M. Ceriotti, "Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials", J. Chem. Phys. **148**, 241730 (2018).

[102] M. Huzak and M. S. Deleuze, "Benchmark theoretical study of the electric polarizabilities of naphthalene, anthracene, and tetracene", J. Chem. Phys. **138**, 024319 (2013).

[103] K. Kowalski, J. R. Hammond, W. A. de Jong, and A. J. Sadlej, "Coupled cluster calculations for static and dynamic polarizabilities of C_{60}", J. Chem. Phys. **129**, 226101 (2008).

[104] D. S. Sabirov, "Polarizability as a landmark property for fullerene chemistry and materials science", RSC Adv. **4**, 44996 (2014).

[105] E. Voloshina and B. Paulus, "First multireference correlation treatment of bulk metals", J. Chem. Theory Comput. **10**, 1698 (2014).

[106] S. M. Smith, A. N. Markevitch, D. A. Romanov, X. Li, R. J. Levis, and H. B. Schlegel, "Static and dynamic polarizabilities of conjugated molecules and their cations", J. Phys. Chem. A **108**, 11063 (2004).

[107] M. Grüning, O. V. Gritsenko, and E. J. Baerends, "Exchange potential from the common energy denominator approximation for the kohn–sham green's function: application to (hyper)polarizabilities of molecular chains", J. Chem. Phys. **116**, 6435 (2002).

[108] K. E. Laidig and R. F. W. Bader, "Properties of atoms in molecules: atomic polarizabilities", J. Chem. Phys. **93**, 7213 (1990).

[109] J. Applequist, J. R. Carl, and K.-K. Fung, "Atom dipole interaction model for molecular polarizability. application to polyatomic molecules and determination of atom polarizabilities", J. Am. Chem. Soc. **94**, 2952 (1972).

[110] K. V. Agrawal, S. Shimizu, L. W. Drahushuk, D. Kilcoyne, and M. S. Strano, "Observation of extreme phase transition temperatures of water confined inside isolated carbon nanotubes", Nature Nanotech. **12**, 267–273 (2016).

[111] Z. Heiner, I. Zeise, R. Elbaum, and J. Kneipp, "Insight into plant cell wall chemistry and structure by combination of multiphoton microscopy with raman imaging", J. Biophotonics **11**, e201700164 (2018).

[112] B. Monserrat, N. D. Drummond, P. Dalladay-Simpson, R. T. Howie, P. López Ríos, E. Gregoryanz, C. J. Pickard, and R. J. Needs, "Structure and metallicity of phase v of hydrogen", Phys. Rev. Lett. **120**, 255701 (2018).

[113] A. Putrino and M. Parrinello, "Anharmonic raman spectra in high-pressure ice from ab initio simulations", Phys. Rev. Lett. **88**, 176401 (2002).

[114]M. Pagliai, C. Cavazzoni, G. Cardini, G. Erbacci, M. Parrinello, and V. Schettino, "Anharmonic infrared and raman spectra in car–parrinello molecular dynamics simulations", J. Chem. Phys. **128**, 224514 (2008).

[115]S. Mukamel, *Principles of nonlinear optical spectroscopy*, Oxford series in optical and imaging sciences (Oxford University Press, 1995).

[116]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).

[117]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).

[118]J. Gerratt and I. M. Mills, "Force constants and dipole-moment derivatives of molecules from perturbed hartree–fock calculations. i", J. Chem. Phys. **49**, 1719–1729 (1968).

[119]X. Gonze and C. Lee, "Dynamical matrices, born effective charges, dielectric permittivity tensors, and interatomic force constants from density-functional perturbation theory", Phys. Rev. B **55**, 10355–10368 (1997).

[120]A. Putrino, D. Sebastiani, and M. Parrinello, "Generalized variational density functional perturbation theory", J. Chem. Phys. **113**, 7102 (2000).

[121]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).

[122]A. M. Reilly, R. I. Cooper, C. S. Adjiman, S. Bhattacharya, A. D. Boese, J. G. Brandenburg, P. J. Bygrave, R. Bylsma, J. E. Campbell, R. Car, D. H. Case, R. Chadha, J. C. Cole, K. Cosburn, H. M. Cuppen, F. Curtis, G. M. Day, R. A. DiStasio Jr, A. Dzyabchenko, B. P. van Eijck, D. M. Elking, J. A. van den Ende, J. C. Facelli, M. B. Ferraro, L. Fusti-Molnar, C. A. Gatsiou, T. S. Gee, R. de Gelder, L. M. Ghiringhelli, H. Goto, S. Grimme, R. Guo, D. W. M. Hofmann, J. Hoja, R. K. Hylton, L. Iuzzolino, W. Jankiewicz, D. T. de Jong, J. Kendrick, N. J. J. de Klerk, H. Y. Ko, L. N. Kuleshova, X. Li, S. Lohani, F. J. J. Leusen, A. M. Lund, J. Lv, Y. Ma, N. Marom, A. E. Masunov, P. McCabe, D. P. McMahon, H. Meekes, M. P. Metz, A. J. Misquitta, S. Mohamed, B. Monserrat, R. J. Needs, M. A. Neumann, J. Nyman, S. Obata, H. Oberhofer, A. R. Oganov, A. M. Orendt, G. I. Pagola, C. C. Pantelides, C. J. Pickard, R. Podeszwa, L. S. Price, S. L. Price, A. Pulido, M. G. Read, K. Reuter, E. Schneider, C. Schober, G. P. Shields, P. Singh, I. J. Sugden, K. Szalewicz, C. R. Taylor, A. Tkatchenko, M. E. Tuckerman, F. Vacarro, M. Vasileiadis, A. Vazquez-Mayagoitia, L. Vogt, Y. Wang, R. E. Watson, G. A. de Wijs, J. Yang, Q. Zhu, C. R. Groom, and IUCr, "Report on the sixth

blind test of organic crystal structure prediction methods", Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials **72**, 439–459 (2016).

[123]J. A. Lemkul, J. Huang, B. Roux, and A. D. MacKerell Jr., "An Empirical Polarizable Force Field Based on the Classical Drude Oscillator Model: Development History and Recent Applications", Chem. Rev. **116**, 4983–5013 (2016).

[124]G. M. Sommers, M. F. Calegari Andrade, L. Zhang, H. Wang, and R. Car, "Raman spectrum and polarizability of liquid water from deep neural networks", Phys. Chem. Chem. Phys. **22**, 10592–10602 (2020).

[125]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**, 2175–2196 (2009).

[126]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).

[127]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).

[128]W. Kabsch, "A solution for the best rotation to relate two sets of vectors", Acta Crystallogr. Sect. A **32**, 922–923 (1976).

[129]F. Musil, M. J. Willatt, M. A. Langovoy, and M. Ceriotti, "Fast and Accurate Uncertainty Estimation in Chemical Machine Learning", J. Chem. Theory Comput. **15**, 906–915 (2019).

[130]T. S. Koritsanszky and P. Coppens, "Chemical Applications of X-ray Charge-Density Analysis", Chem. Rev. **101**, 1583–1628 (2001).

[131]C. Gatti and P. Macchi, eds., *Modern charge-density analysis*, 1st (Springer Netherlands, 2012).

[132]J. C. Meyer, S. Kurasch, H. J. Park, V. Skakalova, D. Künzel, A. Groß, A. Chuvilin, G. Algara-Siller, S. Roth, T. Iwasaki, U. Starke, J. H. Smet, and U. Kaiser, "Experimental analysis of charge redistribution due to chemical bonding by high-resolution transmission electron microscopy", Nat. Mater. **10**, 209–215 (2011).

[133]P. Coppens, "Charge Densities Come of Age", en, Angew. Chem., Int. Ed. **44**, 6810–6811 (2005).

[134]A. D. Becke and K. E. Edgecombe, "A simple measure of electron localization in atomic and molecular systems", J. Chem. Phys. **92**, 5397–5403 (1990).

[135]E. R. Johnson, S. Keinan, P. Mori-Sánchez, J. Contreras-García, A. J. Cohen, and W. Yang, "Revealing Noncovalent Interactions", en, J. Am. Chem. Soc. **132**, 6498–6506 (2010).

[136]P. de Silva and C. Corminboeuf, "Simultaneous Visualization of Covalent and Noncovalent Interactions Using Regions of Density Overlap", en, J. Chem. Theory Comput. **10**, 3745–3756 (2014).

[137]E. Pastorczak and C. Corminboeuf, "Perspective: Found in translation: Quantum chemical tools for grasping non-covalent interactions", J. Chem. Phys. **146**, 120901 (2017).

[138]J. M. Alred, K. V. Bets, Y. Xie, and B. I. Yakobson, "Machine learning electron density in sulfur crosslinked carbon nanotubes", Composites Science and Technology **166**, 3–9 (2018).

[139]A. Chandrasekaran, D. Kamal, R. Batra, C. Kim, L. Chen, and R. Ramprasad, "Solving the electronic structure problem with machine learning", npj Comput Mater **5**, 22 (2019).

[140]P. M. Dominiak, A. Volkov, A. P. Dominiak, K. N. Jarzembska, and P. Coppens, "Combining crystallographic information and an aspherical-atom data bank in the evaluation of the electrostatic interaction energy in an enzyme–substrate complex: influenza neuraminidase inhibition", Acta Crystallogr., Sect. D **65**, 485–499 (2009).

[141]V. Pichon-Pesme, C. Jelsch, B. Guillot, and C. Lecomte, "A comparison between experimental and theoretical aspherical-atom scattering factors for charge-density refinement of large molecules", Acta Crystallogr., Sect. A **60**, 204–208 (2004).

[142]B. Guillot, C. Jelsch, A. Podjarny, and C. Lecomte, "Charge-density analysis of a protein structure at subatomic resolution: the human aldose reductase case", Acta Crystallogr., Sect. D **64**, 567–588 (2008).

[143]N.-E. Ghermani, N. Bouhmaida, and C. Lecomte, "Modelling electrostatic potential from experimentally determined charge densities. I. Spherical-atom approximation", Acta Crystallogr., Sect. A **49**, 781–789 (1993).

[144]N. Bouhmaida, N.-E. Ghermani, C. Lecomte, and A. Thalal, "Modelling electrostatic potential from experimentally determined charge densities. II. Total potential", Acta Crystallogr., Sect. A **53**, 556–563 (1997).

[145]R. F. Stewart, "Electron population analysis with rigid pseudoatoms", Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr. **32**, 565–574 (1976).

[146]N. K. Hansen and P. Coppens, "Testing aspherical atom refinements on small-molecule data sets", Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr. **34**, 909–921 (1978).

[147] T. Helgaker, P. Jørgensen, and J. Olsen, *Molecular Electronic-Structure Theory* (John Wiley & Sons, Ltd, Chichester, UK, 2000).

[148] P.-O. Löwdin, "On the Non-Orthogonality Problem Connected with the Use of Atomic Wave Functions in the Theory of Molecules and Crystals", J. Chem. Phys. **18**, 365–375 (1950).

[149] P.-O. Löwdin and H. Shull, "Natural orbitals in the quantum theory of two-electron systems", Phys. Rev. **101**, 1730–1739 (1956).

[150] J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized Gradient Approximation made simple", Phys. Rev. Lett. **77**, 3865 (1996).

[151] N. P. Labello, A. M. Ferreira, and H. A. Kurtz, "An augmented effective core potential basis set for the calculation of molecular polarizabilities", J. Comput. Chem. **26**, 1464–1471 (2005).

[152] W. J. Stevens, H. Basch, and M. Krauss, "Compact effective potentials and efficient shared-exponent basis sets for the first- and second-row atoms", J. Chem. Phys. **81**, 6026–6033 (1984).

[153] H. B. Schlegel and M. J. Frisch, "Transformation between cartesian and pure spherical harmonic gaussians", Int. J. Quantum Chem. **54**, 83–87 (1995).

[154] V. I. Lebedev, "Quadratures on a sphere", USSR Comput. Maths Math. Phys. **16**, 10–24 (1976).

[155] J. VandeVondele and J. Hutter, "Gaussian basis sets for accurate calculations on molecular systems in gas and condensed phases", J. Chem. Phys. **127**, 114105 (2007).

[156] V. Pichon-Pesme, C. Lecomte, and H. Lachekar, "On Building a Data Bank of Transferable Experimental Electron Density Parameters Applicable to Polypeptides", The Journal of Physical Chemistry **99**, 6242–6250 (1995).

[157] A. Grisafi, A. Fabrizio, A. Lewis, M. Rossi, C. Corminboeuf, and M. Ceriotti, *Salted: program for doing symmetry-adapted learning of three-dimensional electron densities*, https://github.com/andreagrisafi/SALTED, 2021.

[158] J. L. Whitten, "Coulombic potential energy integrals and approximations", J. Chem. Phys. **58**, 4496–4501 (1973).

[159] B. I. Dunlap, J. W. D. Connolly, and J. R. Sabin, "On the applicability of LCAO-Xα methods to molecules containing transition metal atoms: The nickel atom and nickel hydride", Int. J. Quantum Chem. Symp. **11**, 81–87 (1977).

[160] "Use of approximate integrals in ab initio theory. An application in MP2 energy calculations", Chem. Phys. Lett. **208**, 359–363 (1993).

[161]A. P. Rendell and T. J. Lee, "Coupled-cluster theory employing approximate integrals: An approach to avoid the input/output and storage bottlenecks", J. Chem. Phys. **101**, 400–408 (1994).

[162]K. Eichkorn, O. Treutler, H. Öhm, M. Häser, and R. Ahlrichs, "Auxiliary basis sets to approximate Coulomb potentials", Chem. Phys. Lett. **240**, 283–290 (1995).

[163]F. Weigend, "A fully direct RI-HF algorithm: Implementation, optimised auxiliary basis sets, demonstration of accuracy and efficiency", Phys. Chem. Chem. Phys. **4**, 4285–4291 (2002).

[164]H.-J. Werner, F. R. Manby, and P. J. Knowles, "Fast linear scaling second-order Møller-Plesset perturbation theory (MP2) using local and density fitting approximations", J. Chem. Phys. **118**, 8149–8160 (2003).

[165]T. Kato, "On the eigenfunctions of many-particle systems in quantum mechanics", Communications on Pure and Applied Mathematics **10**, 151–177 (1957).

[166]L. A. Burns, J. C. Faver, Z. Zheng, M. S. Marshall, D. G. A. Smith, K. Vanommeslaeghe, A. D. MacKerell, K. M. Merz, and C. D. Sherrill, "The BioFragment Database (BFDb): An open-data platform for computational chemistry analysis of noncovalent interactions", The Journal of Chemical Physics **147**, 161727 (2017).

[167]R. F. W. Bader, "A quantum theory of molecular structure and its applications", Chem. Rev. **91**, 893–928 (1991).

[168]R. Bader, *The Quantum Theory of Atoms in Molecules*, edited by C. F. Matta and R. J. Boyd (Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2007).

[169]J. Contreras-García, E. R. Johnson, S. Keinan, R. Chaudret, J.-P. Piquemal, D. N. Beratan, and W. Yang, "NCIPLOT: A Program for Plotting Noncovalent Interaction Regions", J. Chem. Theory Comput. **7**, 625–632 (2011).

[170]A. NAGY and N. H. MARCH, "Ratio of density gradient to electron density as a local wavenumber to characterize the ground state of spherical atoms", Mol. Phys. **90**, 271–276 (1997).

[171]H. J. Bohórquez and R. J. Boyd, "On the local representation of the electronic momentum operator in atomic systems", J. Chem. Phys. **129**, 024110 (2008).

[172]Á. Nagy and S. Liu, "Local wave-vector, Shannon and Fisher information", Phys. Lett. A **372**, 1654–1656 (2008).

[173]A. N. Bootsma and S. E. Wheeler, "Tuning Stacking Interactions between Asp–Arg Salt Bridges and Heterocyclic Drug Fragments", J. Chem. Inf. Model. **59**, 149–158 (2019).

[174]J. S. Murray, T. Brinck, P. Lane, K. Paulsen, and P. Politzer, "Statistically-based interaction indices derived from molecular surface electrostatic potentials: a general interaction properties function (GIPF)", J. Mol. Struct. **307**, 55–64 (1994).

[175]J. S. Murray and P. Politzer, "Statistical analysis of the molecular surface electrostatic potential: an approach to describing noncovalent interactions in condensed phases", J. Mol. Struct. **425**, 107–114 (1998).

[176]A. Volkov, T. Koritsanszky, and P. Coppens, "Combination of the exact potential and multipole methods (EP/MM) for evaluation of intermolecular electrostatic interaction energies with pseudoatom representation of molecular electron densities", Chem. Phys. Lett. **391**, 170–175 (2004).

[177]H. M. Berman, "The Protein Data Bank", Nucleic Acids Res. **28**, 235–242 (2000).

[178]R. H. French, V. A. Parsegian, R. Podgornik, R. F. Rajter, A. Jagota, J. Luo, D. Asthagiri, M. K. Chaudhury, Y.-m. Chiang, S. Granick, S. Kalinin, M. Kardar, R. Kjellander, D. C. Langreth, J. Lewis, S. Lustig, D. Wesolowski, J. S. Wettlaufer, W.-Y. Ching, M. Finnis, F. Houlihan, O. A. von Lilienfeld, C. J. van Oss, and T. Zemb, "Long range interactions in nanoscale science", Rev. Mod. Phys. **82**, 1887–1944 (2010).

[179]P. Debye and E. Hückel, "Zur theorie der elektrolyte. i. gefrierpunktserniedrigung und verwandte erscheinungen", Physikalische Zeitschrift **24**, 185–206 (1923).

[180]R. L. de Carvalho and R. Evans, "The decay of correlations in ionic fluids", Molecular Physics **83**, 619–654 (1994).

[181]R. Kjellander, "Focus article: oscillatory and long-range monotonic exponential decays of electrostatic interactions in ionic liquids and other electrolytes: the significance of dielectric permittivity and renormalized charges", The Journal of Chemical Physics **148**, 193701 (2018).

[182]F. Coupette, A. A. Lee, and A. Härtel, "Screening lengths in ionic fluids", Phys. Rev. Lett. **121**, 075501 (2018).

[183]A. M. Smith, A. A. Lee, and S. Perkin, "The electrostatic screening length in concentrated electrolytes increases with concentration", J. Phys. Chem. Lett. **7**, 2157–2163 (2016).

[184]C. Zhang and G. Galli, "Dipolar correlations in liquid water", J. Chem. Phys. **141**, 084504 (2014).

[185]A. P. Gaiduk and G. Galli, "Local and Global Effects of Dissolved Sodium Chloride on the Structure of Water", J. Phys. Chem. Lett. **8**, 1496–1502 (2017).

[186]D. M. Wilkins, D. E. Manolopoulos, S. Roke, and M. Ceriotti, "Communication: Mean-field theory of water-water correlations in electrolyte solutions", J. Chem. Phys. **146**, 181103 (2017).

[187]L. Belloni, D. Borgis, and M. Levesque, "Screened Coulombic Orientational Correlations in Dilute Aqueous Electrolytes", J. Phys. Chem. Lett. **9**, 1985–1989 (2018).

[188]Y. Chen, H. I. Okur, N. Gomopoulos, C. Macias-Romero, P. S. Cremer, P. B. Petersen, G. Tocci, D. M. Wilkins, C. Liang, M. Ceriotti, and S. Roke, "Electrolytes induce long-range orientational order and free energy changes in the H-bond network of bulk water", Sci. Adv. **2**, e1501891–e1501891 (2016).

[189]S. Yue, M. C. Muniz, M. F. Calegari Andrade, L. Zhang, R. Car, and A. Z. Panagiotopoulos, "When do short-range atomistic machine-learning models fall short?", The Journal of Chemical Physics **154**, 034111 (2021).

[190]Z. Guo, F. Ambrosio, W. Chen, P. Gono, and A. Pasquarello, "Alignment of redox levels at semiconductor–water interfaces", Chemistry of Materials **30**, 94–111 (2018).

[191]J. D. Elliott, A. Troisi, and P. Carbone, "A qm/md coupling method to model the ion-induced polarization of graphene", J. Chem. Theory Comput. **16**, 5253–5263 (2020).

[192]J. I. Siepmann and M. Sprik, "Influence of surface topology and electrostatic potential on water/electrode systems", J. Chem. Phys. **102**, 511–524 (1995).

[193]R. Jorn, R. Kumar, D. P. Abraham, and G. A. Voth, "Atomistic modeling of the electrode–electrolyte interface in li-ion energy storage systems: electrolyte structuring", The Journal of Physical Chemistry C **117**, 3747–3761 (2013).

[194]C. Merlet, C. Péan, B. Rotenberg, P. A. Madden, P. Simon, and M. Salanne, "Simulating supercapacitors: can we model electrodes as constant charge surfaces?", J. Phys. Chem. Lett. **4**, 264–268 (2013).

[195]T. Dufils, G. Jeanmairet, B. Rotenberg, M. Sprik, and M. Salanne, "Simulating electrochemical systems by combining the finite field method with a constant potential electrode", Phys. Rev. Lett. **123**, 195501 (2019).

[196]F. London, "The general theory of molecular forces", Trans. Faraday Soc. **33**, 8b–26 (1937).

[197]A. M. Reilly and A. Tkatchenko, "Role of dispersion interactions in the polymorphism and entropic stabilization of the aspirin crystal", Phys. Rev. Lett. **113**, 055701 (2014).

[198]A. Ambrosetti, N. Ferri, R. A. DiStasio, and A. Tkatchenko, "Wavelike charge density fluctuations and van der Waals interactions at the nanoscale", Science **351**, 1171–1176 (2016).

[199]B. Parsaeifard, D. S. De, A. S. Christensen, F. A. Faber, E. Kocer, S. De, J. Behler, A. von Lilienfeld, and S. Goedecker, "An assessment of the structural resolution of various fingerprints commonly used in machine learning", Mach. Learn.: Sci. Technol., 10.1088/2632-2153/abb212 (2020).

[200]A. Marrazzo and R. Resta, "Locality of the anomalous hall conductivity", Phys. Rev. B **95**, 121114 (2017).

[201]A. Marrazzo and R. Resta, "Local theory of the insulating state", Phys. Rev. Lett. **122**, 166602 (2019).

[202]J. Behler, "Atom-centered symmetry functions for constructing high-dimensional neural network potentials", J. Chem. Phys. **134**, 10.1063/1.3553717 (2011).

[203]A. Thompson, L. Swiler, C. Trott, S. Foiles, and G. Tucker, "Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials", Journal of Computational Physics **285**, 316–330 (2015).

[204]A. V. Shapeev, "Moment Tensor Potentials: A Class of Systematically Improvable Interatomic Potentials", Multiscale Model. Simul. **14**, 1153–1173 (2016).

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

[206]M. Welborn, L. Cheng, and T. F. Miller, "Transferability in Machine Learning for Electronic Structure via the Molecular Orbital Basis", J. Chem. Theory Comput. **14**, 4772–4779 (2018).

[207]K. Rossi, V. Jurásková, R. Wischert, L. Garel, C. Corminbœuf, and M. Ceriotti, "Simulating Solvation and Acidity in Complex Mixtures with First-Principles Accuracy: The Case of CH $_3$ SO $_3$ H and H $_2$ O $_2$ in Phenol", J. Chem. Theory Comput. **16**, 5139–5149 (2020).

[208]Z. Deng, C. Chen, X.-G. Li, and S. P. Ong, "An electrostatic spectral neighbor analysis potential for lithium nitride", npj Computational Materials **5**, 75 (2019).

[209]T. Morawietz, A. Singraber, C. Dellago, and J. Behler, "How van der waals interactions determine the unique properties of water", Proc. Natl. Acad. Sci. U. S. A. **113**, 8368–8373 (2016).

[210]N. Artrith, T. Morawietz, and J. Behler, "High-dimensional neural-network potentials for multicomponent systems: Applications to zinc oxide", Phys. Rev. B **83**, 153101 (2011).

[211]P. Bleiziffer, K. Schaller, and S. Riniker, "Machine learning of partial charges derived from high-quality quantum-mechanical calculations", Journal of Chemical Information and Modeling **58**, 579–590 (2018).

[212]B. Nebgen, N. Lubbers, J. S. Smith, A. E. Sifain, A. Lokhov, O. Isayev, A. E. Roitberg, K. Barros, and S. Tretiak, "Transferable dynamic molecular charge assignment using deep neural networks", J. Chem. Theory Comput. **14**, 4687–4698 (2018).

[213]K. Yao, J. E. Herr, D. W. Toth, R. Mckintyre, and J. Parkhill, "The tensormol-0.1 model chemistry: a neural network augmented with long-range physics", Chem. Sci. **9**, 2261–2269 (2018).

[214]S. A. Ghasemi, A. Hofstetter, S. Saha, and S. Goedecker, "Interatomic potentials for ionic systems with density functional accuracy based on charge densities obtained by a neural network", Phys. Rev. B **92**, 045131 (2015).

[215]S. Faraji, S. A. Ghasemi, S. Rostami, R. Rasoulkhani, B. Schaefer, S. Goedecker, and M. Amsler, "High accuracy and transferability of a neural network potential through charge equilibration for calcium fluoride", Phys. Rev. B **95**, 104105 (2017).

[216]T. W. Ko, J. A. Finkler, S. Goedecker, and J. Behler, "A fourth-generation high-dimensional neural network potential with accurate electrostatics including non-local charge transfer", Nature Communications **12**, 2041–1723 (2021).

[217]C. Böttcher, O. van Belle, P. Bordewijk, and A. Rip, *Theory of electric polarization* (Elsevier Scientific Pub. Co., 1978).

[218]N. Marzari and D. Vanderbilt, "Maximally localized generalized Wannier functions for composite energy bands", Phys. Rev. B **56**, 12847–12865 (1997).

[219]L. Zhang, M. Chen, X. Wu, H. Wang, W. E, and R. Car, "Deep neural network for the dielectric response of insulators", Phys. Rev. B **102**, 041121 (2020).

[220]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).

[221]H. Huo and M. Rupp, "Unified representation for machine learning of molecules and crystals", ArXiv Prepr. ArXiv170406439 **13754** (2017).

[222]M. Eickenberg, G. Exarchakis, M. Hirn, and S. Mallat, "Solid harmonic wavelet scattering: Predicting quantum molecular energy from invariant descriptors of 3D electronic densities", Adv. Neural Inf. Process. Syst. **2017-Decem**, 6541–6550 (2017).

[223]R. Dreizler and E. Gross, *Density functional theory: an approach to the quantum many-body problem* (Springer Berlin Heidelberg, 2012).

[224]P. P. Ewald, "Die berechnung optischer und elektrostatischer gitterpotentiale", Annalen der Physik **369**, 253–287 (1921).

[225]S. Plimpton, "Fast Parallel Algorithms for Short-Range Molecular Dynamics", J. Comput. Phys. **117**, 1–19 (1995).

[226]R. Jinnouchi, F. Karsai, C. Verdi, R. Asahi, and G. Kresse, "Descriptors representing two- and three-body atomic distributions and their effects on the accuracy of machine-learned inter-atomic potentials", J. Chem. Phys. **152**, 234102 (2020).

[227]D. J. Griffiths, *Introduction to electrodynamics; 4th ed.* Re-published by Cambridge University Press in 2017 (Pearson, Boston, MA, 2013).

[228]A. Tkatchenko and M. Scheffler, "Accurate Molecular Van Der Waals Interactions from Ground-State Electron Density and Free-Atom Reference Data", Phys. Rev. Lett. **102**, 073005 (2009).

[229]B. Huang and O. A. Von Lilienfeld, "Communication: Understanding molecular representations in machine learning: The role of uniqueness and target similarity", J. Chem. Phys. **145**, 10.1063/1.4964627 (2016).

[230]R. A. Bell, M. C. Payne, and A. A. Mostofi, "Does water dope carbon nanotubes?", The Journal of Chemical Physics **141**, 164703 (2014).

[231]Y. Litman, D. Donadio, M. Ceriotti, and M. Rossi, "Decisive role of nuclear quantum effects on surface mediated water dissociation at finite temperature", J. Chem. Phys. **148**, 102320 (2018).

[232]D. Maksimov, C. Baldauf, and M. Rossi, "The conformational space of a flexible amino acid at metallic surfaces", Int J Quantum Chem, 10.1002/qua.26369 (2020).

[233]M. W. Finnis, R. Kaschner, C. Kruse, J. Furthmuller, and M. Scheffler, "The interaction of a point charge with a metal surface: theory and calculations for (111), (100) and (110) aluminium surfaces", J. Phys.: Condens. Matter **7**, 2001–2019 (1995).

[234]J. Neugebauer and M. Scheffler, "Adsorbate-substrate and adsorbate-adsorbate interactions of na and k adlayers on al(111)", Phys. Rev. B **46**, 16067–16080 (1992).

[235]L. Mørch Folkmann Garner, "Machine-learning the polarizabilities and permanent dipole moments of amino acids", MA thesis (Department of Chemistry, University of Copenhagen, 2020).

[236]M. J. Willatt, F. Musil, and M. Ceriotti, "Feature optimization for atomistic machine learning yields a data-driven construction of the periodic table of the elements", Phys. Chem. Chem. Phys. **20**, 29661–29668 (2018).

[237]F. M. Paruzzo, A. Hofstetter, F. Musil, S. De, M. Ceriotti, and L. Emsley, "Chemical shifts in molecular solids by machine learning", Nat. Commun. **9**, 4501 (2018).

[238]V. Kapil, D. M. Wilkins, J. Lan, and M. Ceriotti, "Inexpensive modeling of quantum dynamics using path integral generalized Langevin equation thermostats", J. Chem. Phys. **152**, 124104 (2020).

[239]Y. Zhang, H. B. de Aguiar, J. T. Hynes, and D. Laage, "Water structure, dynamics, and sum-frequency generation spectra at electrified graphene interfaces", The Journal of Physical Chemistry Letters **11**, 624–631 (2020).

[240]S. J. A. van Gisbergen, J. G. Snijders, and E. J. Baerends, "A density functional theory study of frequency-dependent polarizabilities and van der waals dispersion coefficients for polyatomic molecules", The Journal of Chemical Physics **103**, 9347–9354 (1995).

[241]J. C. Slater and G. F. Koster, "Simplified LCAO Method for the Periodic Potential Problem", Phys. Rev. **94**, 1498–1524 (1954).

[242]K. T. Schütt, M. Gastegger, A. Tkatchenko, K.-R. Müller, and R. J. Maurer, "Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions", Nat Commun **10**, 5024 (2019).

[243]M. Kobayashi and H. Nakai, "Divide-and-conquer approaches to quantum chemistry: theory and implementation", in *Linear-scaling techniques in computational chemistry and physics: methods and applications*, edited by R. Zalesny, M. G. Papadopoulos, P. G. Mezey, and J. Leszczynski (Springer Netherlands, Dordrecht, 2011), pp. 97–127.

[244]"Learning (from) the electron density: transferability, conformational and chemical diversity", CHIMIA International Journal for Chemistry **74**, 232–236 (2020).

[245]A. Fabrizio, K. R. Briling, D. D. Girardier, and C. Corminboeuf, "Learning on-top: regressing the on-top pair density for real-space visualization of electron correlation", The Journal of Chemical Physics **153**, 204111 (2020).

[246]C. Lin, E. Durant, M. Persson, M. Rossi, and T. Kumagai, "Real-space observation of quantum tunneling by a carbon atom: flipping reaction of formaldehyde on cu(110)", The Journal of Physical Chemistry Letters **10**, 645–649 (2019).

[247]Y. Shao, L. Knijff, F. M. Dietrich, K. Hermansson, and C. Zhang, "Modelling bulk electrolytes and electrolyte interfaces with atomistic machine learning", Batteries & Supercaps **n/a**, https://doi.org/10.1002/batt.202000262 (2020).

[248]N. G. Hörmann, O. Andreussi, and N. Marzari, "Grand canonical simulations of electrochemical interfaces in implicit solvation models", The Journal of Chemical Physics **150**, 041730 (2019).

[249]K. Kaufmann and W. Baumeister, "Single-centre expansion of Gaussian basis functions and the angular decomposition of their overlap integrals", J. Phys. B At. Mol. Opt. Phys. **22**, 1–12 (1989).

[250]U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, "A smooth particle mesh ewald method", The Journal of Chemical Physics **103**, 8577–8593 (1995).

Andrea Grisafi

Curriculum Vitae

✉ *andrea.grisafi@epfl.ch*

Education

2016–Now **PhD**, Doctoral assistant in the Laboratory of Computational Science and Modeling (COSMO), EPFL. Supervisor: Prof. Michele Ceriotti.

2014–2016 **Master**, Degree in Physical Chemistry at Scuola Normale Superiore of Pisa and the University of Pisa with full marks. Thesis: *Solute-solvent interactions and solvation free-energy: a density functional investigation*. Supervisor: Prof. Vincenzo Barone.

2010–2014 **Bachelor**, Degree in Chemistry at the University of Pisa with full marks. Thesis: *Potential energy surfaces via Quantum Monte Carlo methods*. Supervisor: Prof. Claudio Amovilli.

Publications

1. **Article**, Amovilli C., Floris F., Grisafi A., Localized polycentric orbital basis set for Quantum Monte Carlo calculations derived from the decomposition of Kohn-Sham optimized orbitals, *Computation*, 4, 10 (**2016**).
2. **Master Thesis**, Grisafi A., Solute-solvent interaction and solvation free-energy: a density functional investigation, (**2016**).
3. **Article**, Liang C., Tocci G., Wilkins D. M., Grisafi A., Roke S., Ceriotti M., Solvent fluctuations and nuclear quantum effects modulate the molecular hyperpolarizability of water, *Physical Review B*, 96, 041407 (**2017**).
4. **Article**, Grisafi A., Wilkins D. M., Csànyi G., Ceriotti M., Symmetry-adapted machine learning for tensorial properties of atomistic systems, *Physical Review Letters*, 120, 036002 (**2018**).
5. **Article**, Grisafi A., Fabrizio A., Meyer B., Wilkins D. M., Corminboeuf C., Ceriotti M., Transferable machine learning model of the electron density, *ACS Central Science*, 5, 57-64 (**2019**).
6. **Article**, Wilkins D. M., Grisafi A., Yang Y., Lao K. U., DiStasio R. A., Ceriotti M., Accurate molecular polarizabilities with coupled-cluster theory and machine learning, *Proceedings of the National Academy of Sciences*, 116, 3401-3406 (**2019**).
7. **Book Chapter**, Grisafi A., Wilkins D. M., Willatt, M. J., Ceriotti M., Atomic-scale representation and statistical learning of tensorial properties, *Machine Learning in Chemistry: Data-Driven Algorithms, Learning Systems, and Predictions*, Chap.1, pp.1-21 (**2019**).
8. **Article**, Raimbault N., Grisafi A., Ceriotti M., Rossi M., Using Gaussian process regression to simulate the vibrational Raman spectra of molecular crystals, *New Journal of Physics*, 21, 105001 (**2019**).
9. **Article**, Fabrizio A., Grisafi A., Meyer B., Ceriotti M., Corminboeuf C., Electron density learning of non-covalent systems, *Chemical Science*, 10, 9424-9432 (**2019**).
10. **Article**, Grisafi A., Ceriotti M., Incorporating long-range physics in atomic-scale machine learning, *The Journal of Chemical Physics* 151, 204105 (**2019**).

11. **Article**, Fabrizio A., Briling K., Grisafi A., Corminboeuf C., Learning (from) the electron density: transferability, conformational and chemical diversity, *CHIMIA International Journal of Chemistry* 74 (4), 232-236 (**2020**).
12. **Article**, Grisafi A., Nigam J., Ceriotti M., Multi-scale approach for the prediction of atomic scale properties, *Chemical Science*, accepted, - (**2020**).
13. **Preprint**, Musil F., Grisafi A., Bartók A., Ortner C., Csányi G., Ceriotti M., Physics-inspired structural representations for molecules and materials, *arXiv:2101.04673*, (**2021**).

Workshops and Conferences

June 2020 **Online Workshop on Modern Approaches to Coupling Scales In Materials Simulations, Bavaria (Germany)**, Invited talk: "A multi-scale atomistic representation with consistent long-range behaviour".

November 2019 **Research Day of Materials Department, EPFL (Switzerland)**, Invited talk: "On the importance of symmetry and locality in atomic-scale machine learning".

October 2019 **Science Day, Starling Hotel, Lausanne (Switzerland)**, Contributed talk: "Learning tensorial properties in molecules and condensed-phases: from dielectric responses to electron densities".

August 2019 **Hands-on Workshop on Density Functional Theory and Beyond, University of Barcelona (Spain)**, Poster: "A transferable machine learning model for all-electron charge densities".

June 2019 **Marvel Junior Seminar, École Polytechnique Fédérale de Lausanne (Switzerland)**, Contributed talk: "Symmetry-adapted learning of tensorial properties in molecules and condensed-phases".

June 2019 **Skoltech International Workshop, Skolkovo Institute of Science and Technology, Moscow (Russia)**, Contributed talk: "Learning tensorial properties in molecules and condensed-phases: from dielectric responses to full-electron charge-densities".

February 2019 **SACC Spring Meeting, University of Geneva (Switzerland)**, Poster: "A transferable machine learning model of the electron density".

January 2019 **IMS Symposium on Water at Interfaces, Okazaki Conference Center (Japan)**, Contributed talk: "Statistical learning of dielectric response functions in molecules and condensed-phases".

September 2018 **International Workshop on Computational Design and Discovery of Novel Materials, SwissTech Convention Center, Lausanne (Switzerland)**, Poster: "A transferable machine learning model of the electron density".

March 2018 **CECAM Workshop on Electrostatics in Concentrated Electrolytes, Lausanne (Switzerland)**, Poster: "Machine-learning of electrical response tensors".

March 2018 **DPG Meeting, Berlin (Germany)**, Contributed talk: "Symmetry-adapted machine learning for tensorial properties of atomistic systems".

July 2016 **Workshop on Water and Water Systems, Ettore Majorana Foundation and Centre for Scientific Culture, Erice (Italy)**, Poster: "Systematic prediction of the screening lengths beyond the Debye-Hückel limit: a classical-DFT approach".

Supervision of students

Summer 2018 - Fall 2018 **Semester project**, Supervision of a student on a semester project regarding the tensorial machine learning of the polarizability of paracetamol crystals.

Fall 2019 - Spring 2020 **Master thesis**, Co-supervision of a student on a Master thesis project regarding long-range interactions methods in atomistic machine-learning.

Teaching activities

Spring 2017-2019 **Course assistant**, Assistant to the course of *Statistical Mechanics* held by Prof. Michele Ceriotti at EPFL, including the preparation of weekly exercises and computational laboratories.

Fall 2017-2019 **Course assistant**, Assistant to the course of *Surfaces and Interfaces* held by Prof. Michele Ceriotti at EPFL, including the preparation of weekly exercises, as well as of the midterm and final exams.

Scientific review activities

Fall 2019 **Review of article**, Review activity for *The Journal of Chemical Theory and Computation*.

Spring 2020 **Review of article**, Review activity for *The Journal of Chemical Theory and Computation*.

Software contributions

TENSOAP, Program for doing symmetry-adapted GPR of any tensorial property expanded on a spherical harmonics basis: `https://github.com/dilkins/TENSOAP`.

SALTED, Program for doing symmetry-adapted learning of three-dimensional electron densities:: `https://github.com/andreagrisafi/SALTED`.

Awards

2014 - 2016 **Scholarship**, Merit scholarship in Chemistry at Scuola Normale Superiore of Pisa (`https://www.sns.it/en`).

Computational skills

Programming, Working knowledge of Linux scripting (bash), Fortran, C, Python, Jupyter Notebook and Mathematica. Use of GitHub in collaborative programming projects.

Scientific communication, LaTeX typesetting for scientific papers and presentations.

Languages

Italian, mother tongue.

English, fluent.

French, intermediate.