Physics-inspired structural representations for molecules and materials

Felix Musil, 1 Andrea Grisafi, 1 Albert P. Bartók, 2 Christoph Ortner, 3 Gábor Csányi, 4 and Michele Ceriotti 1

1 Laboratory of Computational Science and Modeling, IMX, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland
2 Department of Physics and Warwick Centre for Predictive Modelling, School of Engineering, University of Warwick, Coventry CV4 7AL, United Kingdom
3 University of British Columbia, University of British Columbia, Vancouver, BC, Canada V6T 1Z2
4 Engineering Laboratory, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, United Kingdom
(Dated: January 13, 2021)

The first step in the construction of a regression model or a data-driven analysis framework for matter at the atomic scale involves transforming the Cartesian coordinates that describe the positions of the atoms in the form of a representation that obeys the same symmetries as the properties of interest, and in general reflects the physical nature of the problem. The link between properties, structures, their physical chemistry and their mathematical description is strongest when it comes to applications aimed at determining a precise correspondence between atomic configurations and the quantities that one might compute by a quantum mechanical electronic-structure calculation or measure experimentally. The development of atomic-scale representations have played, and continue to play, a central role in the success of machine-learning methods that rely on this correspondence, such as interatomic potentials, as well as generic property models, structural classifiers and low-dimensional maps of structures and datasets. This review summarizes the current understanding of the nature and characteristics of the most commonly used structural and chemical descriptions of molecules and materials, highlighting the deep underlying connections between different frameworks, and the ideas that lead to computationally efficient and universally applicable models. It gives examples of recent applications to a diverse set of chemical and materials science problems, and outlines the open questions and the most promising research directions in the field.

CONTENTS

I. Introduction 2
II. List of symbols 4
III. Representations for materials and molecules 4
A. Symmetry 4
B. Smoothness 5
C. Local and additivity 6
D. Completeness 7
IV. Symmetrized atomic field representations 8
A. Dirac notation for atomic representations 8
B. Global field representations 10
C. Translational invariance and atom-centered features 11
D. Rotational invariance and body-ordered representations 12
E. Density correlations in an angular momentum basis 12
F. The density trick 14
G. Equivariant representations and tensorial features 14
H. Long-range features 16
V. Representations and models 17
A. Linear models and body-order expansion 17
B. Long-range features and potential tails 21
C. Non-linear models 21
VI. Geometric completeness of representations 23
A. Algebraic completeness: an elementary example 23
B. Geometric completeness of density correlations 24
C. Spectral representations 25
VII. Representations, structures, properties and insights 26
A. Features, distances, kernels 27
B. Measuring structural similarity 28
C. Representations for unsupervised learning 30
D. Analyzing representations and datasets 30
E. Indirect structure-property relationships 31
VIII. Efficiency and effectiveness 32
A. Comparison of features 32
B. Feature selection 34
C. Feature optimization 36
D. Efficient implementation 38
E. Packages to evaluate atom-density representations 41
IX. Applications 43
A. Best match kernels for ligand binding 43
B. Tensorial features and polarizability 44
C. Long-range and non-local responses 45
I. INTRODUCTION

The last decade has seen a tremendous increase in the use of data-driven approaches for the modeling of molecules and materials. Atomistic modeling has been a particularly fertile field of use; applications range from the analysis of large databases of materials properties to the design of molecules with the desired behavior for a given application. Machine learning techniques have been applied to devise coarse-grained descriptions of complex molecular systems, to build accurate and comparatively inexpensive interatomic potentials, and more generally to predict, or rationalize, the relationship between a specific atomic configuration and the properties that can be computed by electronic-structure calculation.

All of these applications to atomic-scale systems share the need to map an atomic configuration $A$ – identified by the positions and chemical identity of its $N$ atoms $\{r_i, a_i\}$, and possibly by the basis vectors of the periodic repeat unit $h$ – into a more suitable representation. This mapping associates $A$ with a point in a feature space, which is then used to construct a machine-learning model to regress the structure-property relation, to cluster (group together) configurations that share similar structural patterns, or to further map the conformational landscape of a data set onto a low-dimensional visualization. The terms descriptor or fingerprint are often used in chemical and materials informatics to indicate heuristically-determined properties that are easier to compute but reflect more directly the relationship between a structure and the quantities one wants to predict. Examples of descriptors include the fractional composition of a compound, the electronegativity of its atoms, a low-level-of-theory determination of the HOMO-LUMO gap of a molecule. In this review we focus on a more systematic class of mappings that use exclusively atomic composition and geometry as inputs, and aim to characterize precisely the instantaneous arrangement of the atoms, for which we use the term representation. We will be especially interested in those representations that apply geometric and algebraic manipulations to the Cartesian coordinates, to transform them in a way that fulfills physically-informed requirements: smoothness and symmetry with respect to isometries. Commonly used representations include atom-centered symmetry functions and the smooth overlap of atomic positions (SOAP). It is important to note that representations can be expressed using different mathematical entities: at the most basic level, the space of features can take the form of a vector space, in which each configuration is associated with a finite-dimensional vector whose entries are explicitly computed by the mapping procedure. Depending on the application, however, it may be simpler or more natural to express the relationship between pairs of configurations in terms of a distance metric $D(A, A')$, or a kernel function $K(A, A')$. As we will see, distance or kernel-based formulations implicitly define a feature space, that in most cases can be expressed (at least approximately) in terms of a vector of features, ad so can be seen as equivalent to a representation of individual structures.

While one can trace the origins of different representations to specific subfields of computational chemistry and materials science, the fact that representations should describe precisely the nature and positions of each atom means that they often are not specialized to a given application, but can be used with little modification for any atomistic system, from gas-phase molecules to bulk solids. This generality, however, does not mean that representations are completely abstract or disconnected from physical and chemical concepts. Over the past few years, it has become clear that representations that reflect more closely some fundamental principles – such as locality, the multi-scale nature of interactions, the similarities in the behavior of elements from the same group in the periodic table – usually yield models that are more robust, transferable and data-efficient. The link between a representation and the physical concepts it incorporates is usually mediated by the strategy one uses to fit the desired structure-property relations: it is often possible to show an explicit relationship between linear regression models built on the representation of a structure and well-known empirical forms of interatomic potentials (such as body-ordered, or multipole expansions), and more complex, non-linear machine-learning schemes built on the same features improve the flexibility in describing structure-property relations, albeit at the price of a less transparent interpretation of their behavior.

Given the central role of structural representations in the application of data-driven methods to atomistic modeling, it is perhaps not surprising that considerable effort is being dedicated to understanding and improving their properties. These efforts follow several directions. First, the efficient, scalable, and parallel implementation of the construction of a given set of features is essential to ensure computational efficiency. Second, reduction in the number of features...
that is used to describe the system reduces the computational effort, and often improves the robustness of the model: feature selection aims at identifying the most expressive, yet concise, description of the system at hand. Third, it is often desirable to fine-tune a representation so that it facilitates training a model on a small number of reference structures, by incorporating more explicitly the available prior knowledge.

This review aims to summarize recent work on the construction of efficient and mathematically sound representations of atomic and molecular structures, with a particular focus on the use for the regression of atomic-scale properties. Rather than focusing on a historical overview, we intend to provide a snapshot of the current insights on what makes a good representation, supporting our considerations with recent publications, and providing a perspective of the most promising research directions in the field.

II. LIST OF SYMBOLS

\[ A \] An atomic structure  
\[ A_i \] An environment centered on the \( i \)-th atom of the structure \( A \)  
\[ r_i \] Position of the \( i \)-th atom  
\[ r_{ij} \] Vector separating the \( i \)-th atom and its \( j \)-th neighbor, \( r_{ij} = r_i - r_j \)  
\[ Q \] Generic continuous index enumerating the components of an atomic representation  
\[ q \] Generic discrete index enumerating the components of an atomic representation  
\[ \langle Q|A \rangle \] A representation of a structure \( A \) indexed by an unspecified label or set of labels \( Q \)  
\[ \Phi(A_i) \] Feature vector (with elements indexed by \( q \)) associated with an atom-centered environment, \( \phi_q(A_i) = \langle q|A_i \rangle \)  
\[ \Phi \] Feature matrix combining the features associated with multiple structures/environments  
\[ \varphi_{q,i} \] A column in a feature matrix, where \( \langle \varphi_{q,i} \rangle \) is an unspecified label or set of labels \( Q \)  
\[ F(A_i) \] An atom-centered property, or its systematic approximation in terms of an atom-centered representation  
\[ \tilde{F}(\Phi) \] A non-linear model that approximates \( F(A_i) \) using the feature vector \( \Phi(A_i) \)  
\[ K(A,A') \] A (non-)linear kernel computed between two structures or environments, represented by the corresponding feature vectors \( \Phi(A) \)  
\[ D(A,A') \] A distance computed between two structures or environments  
\[ |\rho\rangle \] Structure representation based on a smooth atom density  
\[ |\rho_i\rangle \] Representation of an environment centered on atom \( i \) based that results from symmetrizing \( |\rho\rangle \) over translations  
\[ |\rho_{\nu}^{\text{cov}}\rangle \] Symmetrized \( \nu \)-point correlation of the atomic density built on the atom-centered representation \( |\rho_i\rangle \)  

III. REPRESENTATIONS FOR MATERIALS AND MOLECULES

Even though this review has no intention of providing a comprehensive discussion of the development of descriptors for atomic structures, it is worth providing a brief overview. A “data-driven” philosophy emerged early in the field of chemical and molecular science, where the combinatorial extent of the space of possible molecules\(^{31}\) and the possibility of accessing this space with comparatively simple synthetic strategies, encouraged the development of quantitative structure/property relationships (QSPR) techniques, attempting to map\(^{32}\) descriptors of molecular structure – based on chemical-intuition driven descriptors\(^{33}\) or indicators obtained from quantum chemical calculations\(^{34}\) – to the behavior of a selected compound, usually focusing on properties of direct applicative interest\(^{35}\) such as solubility, toxicity\(^{36}\) or pharmacological activity\(^{37}\)
This approach should be contrasted with that of “bottom-up” predictions, that aim to use models of the interactions between the atomic constituents of a material to simulate the behavior of the system on an atomic time and length scale. Starting from the early days of molecular simulation,[44–47] the objective was to predict the energy, the forces, or any other observable of interest, for a specific molecular configuration, and use them to search for (meta-)stable configurations, or to simulate the evolution of the system by molecular dynamics[48,49]. In the absence of reliable reference values for the properties of specific atomic configurations, interatomic potentials (also called empirical force fields) were built using physically-inspired functional forms, combining harmonic terms to describe chemical bonds with Coulomb and 1/r⁶ terms to describe electrostatics and dispersion. Their (few) parameters were determined by matching the values of experimental observables, such as cohesive energies, lattice parameters and elastic constants. The continuous increase in computational power, and the availability of electronic-structure techniques with a better cost-accuracy ratio[50–52] has made it possible to compute extremely accurate energies and properties of specific configurations. This has opened the way to ab initio simulations of material[53], but also provided a viable alternative to empirical functional forms for the construction of interatomic potentials. Starting from the simplest compounds[53], and then gradually increasing in complexity[54], molecular potential energy surfaces fitted by interpolating between a comparatively small number of ab initio reference calculations provided the first practical applications of this idea. The possibility of combining very accurate calculations of the electronic structure of atomic systems with sampling of the statistics and dynamics of the nuclei on the electronic potential energy surface has allowed theoretical predictions that do not only agree with experimental results[53] – they can predict experiments[55] two decades before measurements become precise enough to verify the theoretical values[56].

Even though the ultimate goal of QSPR models and machine-learned potentials is the same – predicting scientifically and/or technologically relevant properties of molecules and materials – the approaches they follow to achieve this goal are quite different, which is reflected in the way an atomic structure is translated into an input for a machine-learning model. Cheminformatics descriptors, or fingerprints, are built ad-hoc incorporating both descriptors of molecular structure and composition, and easy-to-estimate molecular properties. They usually rely on
a considerable amount of prior knowledge, are often system and problem specific, and are meant to label a compound rather than a specific configuration of its atoms. This is a logical consequence of the fact that QSPR aims for an end-to-end description of a thermodynamic property, which is not an attribute of an individual configuration, but of a thermodynamic state of matter. In the case of bottom-up modeling, instead, one aims first at building a very accurate surrogate model that is capable of reproducing precisely and inexpensively the outcome of quantum calculations for a specific configuration of the atoms. The end goal of predicting thermodynamic properties is achieved by coupling these prediction with statistical-sampling methods, aimed at computing averages over the appropriate classical (or quantum) distribution of atomic configurations. As a consequence, the representations used as inputs of these surrogate quantum models are usually rather generic, are constructed using exclusively atomic coordinates and chemical species, and aim to establish a precise mapping between a specific structure and the associated atomic-scale quantities. We focus our discussion on this latter class of features, and also do not attempt to cover recent, and rather successful, attempts to use descriptors that rely on low levels of quantum mechanical theory to predict high-end, accurate molecular properties.

In the rest of this section, we discuss the properties that are desirable for a representation used in atomistic machine learning, which are graphically summarized in Figure 1. The mapping between structures and features should be consistent with basic symmetries – i.e. reflect the fact that the properties associated with a structure do not change when the reference system or the labelling of identical atoms are modified; be smooth, so that models built on the features inherit a regular behavior with changing atomic coordinates; be complete, so that fundamentally distinct configurations are never mapped to the same set of features. Furthermore, many machine-learning tasks benefit greatly from being based on local features, which describe atoms or groups of atoms. Even though this is a less stringent requirement and, as we discuss below, global descriptors have been used very successfully, representations based on local environments are usually associated with higher transferability, reflecting a “divide and conquer” approach to materials modeling. Finally, less fundamental but not less important requirements are the numerical stability and computational efficiency of the structure-representation mapping, which we discuss in Section V.
The Empirical Valence Bond (EVB) method has been used to model bond-breaking events, but the generality of the EVB approach is limited as the possible assignments need be pre-determined. This led to the development of representations that are intrinsically independent on the ordering of the atoms, such as permutation-invariant polynomials (PIPs) which are obtained by summing functions of the internal coordinates over all possible orderings. In their original implementation, the exponentially increasing cost of evaluating these sums limited their applicability to molecules with a small number of degrees of freedom.

The idea that atomic structures had to be properly symmetrized before being passed to a regression algorithm was understood early in the context of constructing molecular potential energy surfaces. It was made central to the construction of interatomic potentials for condensed-phase materials by Behler and Parrinello, who in Ref. introduced the concept of atom-centered symmetry functions (ACSF). Similarly to PIPs, ACSF are translationally and rotationally invariant because they are functions of angles and distances, and permutationally invariant because they are summed over all possible atomic pairs and triplets within an atomic environment. The computational cost of ACSF is kept under control by restricting the range of interactions (which we discuss further in subsection [II C] and the body order of the correlations considered. Despite these restrictions, ACSF models have been shown to achieve comparable accuracy to that reached by PIPs. Indeed, the recently proposed atomic PIPs use the same polynomial basis as global PIPs, but avoid the unfavorable scaling with increasing molecule size by combining locality (via a distance cutoff) and a truncation of the order of the expansion.

Internal coordinates are also the fundamental building block of molecular matrix representations, which are based on functions of the interatomic distances within a structure. Coulomb matrices, which list the formal electrostatic interactions \( q_i q_j / r_{ij} \) between each atomic pair in a structure, have been extensively explored in applications of the machine learning of molecular properties, with the main limitation being connected to the lack of permutation invariance, which has also been tackled by approximate symmetrization, summing over a manageable number of randomized orderings of the atomic indices. We discuss alternative approaches to symmetrizing Coulomb matrices, as well as other representations based on molecular matrices, in Subsection [III B].

The phylogenetic tree in Fig. 2 shows that a large number of existing representation take a different strategy to achieve symmetrization: rather than using internal coordinates that are inherently invariant to rotations and translations, they first – implicitly or explicitly – describe the system as an atom density \( \sum_i g(x - r_i) \), obtained by summing over localized functions centered on the positions \( r_i \) of all atoms in the system. Such a density is naturally invariant to permutations, and only at a later stage one proceeds to symmetrize it over translations and rotations. We discuss in great detail this second approach in Section [IV]. It suffices to say, at this point, that even if the construction of symmetrized density representations is conceptually very different from those based on internal coordinates, there are many direct and indirect links between the two branches, sketched in Figure [2] which we will discuss when reviewing specific classes of representations.

### B. Smoothness

The overwhelming majority of atomic-scale properties are continuous, smooth functions of the atomic coordinates. Function regularity is crucial for creating efficient ML models, and is therefore one of the requirements for a good structural representation. Features constructed from a symmetrized atom density are naturally smooth functions of atomic coordinates, and it is usually not a problem to maintain this regular behavior upon symmetrization over translations and rotations. The level of smoothness can be adjusted by smearing the atomic density, or by expanding it on a smooth basis (effectively a Fourier smoothing), as we discuss more extensively in Section [IV]. Internal coordinates are also usually smooth, but the process of manipulating them to achieve a permutation invariant representation can affect the smoothness of the mapping.

One way to obtain permutation invariance without incurring the exponential scaling of the cost associated with enumerating all possible permutations of atomic indices involves sorting the entries in a distance or Coulomb matrix, an approach that has also been used with permutation invariant vectors (PIV) and “bag of bonds” features (BoB). Computing the eigenvalues of (functions of) interatomic distances, which underlies the SPRINT method, as well the overlap matrix eigenvalue fingerprint also effectively achieves permutation invariance by similar means, since the vector of eigenvalues is taken to be sorted in ascending or descending order. The earliest implementation of the DeepMD scheme also relied on sorting a local distance matrix. However, the sorting operation introduces derivative discontinuities in the mapping between Cartesian coordinates and features, because the order of the distance vector changes as atoms are displaced in the structure.

Figure [3] illustrates the discontinuity of the derivatives of a function that is built from an ordered list of features. Consider a system of 3 atoms that is
FIG. 3. Toy model demonstrating a non-smooth property (solid line) and its discontinuous derivative (dashed line) that are defined as functions of the ordered list of interatomic distances for a three-atom cluster.

uniquely defined by the 3 interatomic distances $r_i$, where the index $i$ denotes the position of the interatomic distance $r_i$ in the ordered list of distances. We define a smooth function of the sorted distances, $f = \sum_i c_i (r_i - r_0)^2$ parameterized by $c$ and $r_0$. The function $f$ is indeed invariant to the permutations of the atom order in the trimer, but at the price of introducing kinks in $f$ and discontinuities in its derivative when the distance ordering changes. Fitting any smooth function of the trimer geometry by optimizing the parameters $c$ and $r_0$ would necessarily lead to poor approximation accuracy.

The lack of regularity has implications for the accuracy and stability of machine-learning models built on such features, as has been shown recently by using a Wasserstein metric to compare Coulomb matrices in a permutation-invariant manner. In this context it is worth noting the remarkable connection linking the Euclidean distance between vectors of sorted distances and the Wasserstein distance between radial distribution functions (Section III.F in Ref. [40]), which builds a formal bridge between conceptually unrelated families of atomic-scale representations.

C. Locality and additivity

The overwhelming majority of empirical interatomic potentials are expressed as an additive combination of local terms, or of long-range pairwise contributions. Early models built to fit molecular potential energy surfaces were built explicitly as a function of the coordinates of all atoms in the system. Besides the issues of computational cost, this approach is problematic, as it hinders the application of the potential to a molecule with a different number of atoms, or chemical composition. The work of Behler and Parrinello did not only have the merit of emphasizing the importance of symmetries in atomistic machine learning, but it also applied to ML interatomic potentials an additive expansion of the molecular energy $E(A)$, writing it as a sum of atom-centered contributions, $E(A) \approx \sum_{i \in A} E(A_i)$. The notion of an additive decomposition of properties, which is also implicit in the functional forms of most interatomic potentials, has far reaching consequences for data-driven models. Combined with the requirement that the atomic contributions only depend on the position of atoms within a finite range of distances, which is needed for the method to be computationally practical and is supported by fundamental physical principles, the additivity assumption breaks down the problem of predicting the properties of a complex structure into simpler, short-range problems. This has non-trivial consequences in terms of the data efficiency of the model, as discussed in Subsection VIII.C. An additive decomposition is also the most straightforward way to ensure extensivity of predictions, i.e. that the prediction of a property for two copies of a molecule at infinite distance from each other is equal to twice the prediction for a single molecule.

It is not by chance that also in the field of molecular machine learning, for which many of the early representations aimed at a global description of a molecule, most of the recent approaches have moved to additive, atom-centered representations that yield more accurate and transferable models for properties that are expected to be extensive. Oftentimes it is possible, and relatively straightforward, to modify a global representation to describe an atom-centered environment, or to combine atom-centered representations to build a global description, e.g. by summing or averaging the values of all the atom-centered features that are present in the structure, as we discuss in Section VII.B. The connection between locality, additivity, and the nature of the structure-property relation that one wants to model is essential to the construction of effective and transferable machine-learning models.

D. Completeness

The requirements of symmetry, smoothness and locality can be seen as geared towards reducing the complexity of the structural representation, eliminating redundant structures, reducing the resolution to the intrinsic length scale over which the target property exhibits substantial variations, and breaking down complicated compounds into simple fragments. This simplification should not, however, come at the expense of the completeness of the representation, meaning that the mapping between Cartesian and feature spaces should keep inequivalent structures distinct. For example, it has been known for some time that
a histogram of interatomic distances (discarding the identity of the connected atoms) is insufficient to fully characterize a structure composed of more than three atoms. More recently, countereamples have emerged showing that atom-centered correlations – at least those of low order – are also insufficient to preserve the injectivity of the structure-feature mapping (see Ref. [104] and Section VI B for a more thorough discussion).

Besides completeness in terms of the geometric structure-feature mapping, one should also consider whether for a chosen regression scheme the feature-property mapping can be converged to arbitrary accuracy. More complex, non-linear models can often provide good results even when using a representation that involves excessive smoothing, or an highly truncated version of a family of features. The interplay between model and features is discussed in more detail in Section V, and the problem of completeness in Section VI.

IV. SYMMETRIZED ATOMIC FIELD REPRESENTATIONS

As discussed in the previous Section, a multitude of representations have been introduced over the past decade, attempting to incorporate basic principles of symmetry and locality at the very core of atomistic machine learning. The differences between them are much less fundamental than it appears at a first glance, and in fact several works have recently pointed at the existence of a unified framework, in which an explicit formal connection can be established between the vast majority of representations. In this Section we summarize the construction of this class of features, that we refer to as “symmetrized atomic field representations”, emphasizing the role played by symmetry and locality, as well as hint to the connection between this class of features and a linear mapping between structure and properties, which is discussed in more detail in Section VI.

A. Dirac notation for atomic representations

We formalize a notation, extending the one introduced in Refs. [90,105] and used in Ref. [108] to compare different kinds of local and global representations, to express the feature vectors associated with a representation in a way that mimics Dirac notation in quantum mechanics. The notation is convenient to perform manipulations of the features that can be expressed as linear transformations, to emphasize the fact that several widely adopted representations are of the same nature, differing primarily in the choice of basis, and allows for a more expanded layout of complicated feature indices compared to subscript notations [109].

Each structure $A$ is associated with a ket $| A; \text{rep} \rangle$ that also contains a description of the nature of the representation. Indices that enumerate the corresponding features are associated with the bra $\langle Q |$. Both the ket and the bra indices can (and will) be used with some looseness, to emphasize the most relevant elements of a representation while keeping the notation slim. For instance, as shown in Fig. 4, one can indicate explicitly multiple bra indices when their meaning in the definition of a representation is important, separating with a semicolon groups of indices that are conceptually related, or condense them in a compound index when the substructure is irrelevant. Moreover, when discussing the construction of a representation, the reference structure is not important, and so one may drop the structure index from the notation and write $| \text{rep} \rangle$ instead of $| A; \text{rep} \rangle$. Vice versa, when the representation of choice is well-established – e.g. when writing expressions that describe the regression scheme after having discussed the choice of representation – one may omit the specifics of the representation and write $| A \rangle$. The core principle we adhere to
is that the ket indicates the nature of the representation (the object being described, the body order, the symmetries that the representation obeys, the way it is constructed or its hyperparameters) while the bra indicates a discretization that might be a matter of computational convenience (e.g. it might depend on the basis on which $|A\rangle$ is expressed), such that – at least in the limit of a complete basis – the discretization is irrelevant, and different choices are connected by a linear transformation. At the most basic level, this notation can be seen as a way to indicate more expressively the nature of the representation used, and to tidily enumerate the components of the associated feature vector. To give a concrete example, the notation used in Refs. [27,101] to indicate the components of a SOAP feature vector associated with the environment $A$, centered around the $i$-th atom of structure $A$, can be written as

$$a_{n_1 n_2}^{a_1 a_2}(A) = \langle a_1 n_1; a_2 n_2; l | A; \rho_i^{(l)}} \rangle,$$

where $\rho_i^{(l)}$ hints at the fact that SOAP features can be derived as a 2-point correlation of the $i$-centered atom density, $a_{i,l}$ indicates the chemical nature of the atoms, $n_{1,2}$ radial basis channels and $l$ an angular momentum channel. The semicolons indicate that the $(a_{i,1})$ and $(a_{i,2})$ indices are conceptually linked in the construction. The $l$ index is associated with both densities, as a consequence of the process that makes the SOAP power spectrum invariant to rotations. When the nature of the indices is unimportant, we use generic indices, namely $\langle Q \rangle$ when substituting continuous indices, and $|q\rangle$ when substituting one or more discrete indices.

Much like Dirac notation in quantum mechanics, however, the real value of this formalism is seen in how it helps avoiding errors when manipulating representations. For example, a change in the basis that is used to practically compute a representation can be written as a linear transformation,

$$\langle T | A \rangle = \int dQ \langle T | Q \rangle \langle Q | A \rangle,$$

where $\langle T | Q \rangle$ indicates the coefficients that enact the change of basis. Note that, much as it is the case in quantum chemistry, non-orthogonal bases introduce some ambiguity in the bra-ket notation, because the coefficients in the expansion of a function in the basis differ from the scalar product between the basis and the function – the two being related by the overlap matrix of the basis. Thus, the notation we use here captures the gist of the various manipulations, but should be treated with some care when translating it into a practical implementation if the basis used is not orthonormal. Furthermore, features could be closely related without being linked to each other by a linear transformation, for instance by the value of a hyperparameter $r_{\text{cut}}$. In these cases, one could indicate in the ket the parameters that distinguish the two representations, e.g. $|r_{\text{cut}} = 3\text{Å}\rangle$, reserving the bra for the indices associated with the expansion of the abstract representation onto a concrete basis.

The scalar product between the features of two structures $A$ and $A'$ (as it could be used to define a linear kernel $K(A, A')$ that expresses the similarity between the configurations) reads

$$K(A, A') \equiv \langle A | A' \rangle = \int dQ \langle Q | A \rangle^* \langle Q | A' \rangle = \int dQ \langle A | Q \rangle \langle Q | A' \rangle,$$

where we use the convention $\langle A | Q \rangle = \langle Q | A \rangle^*$ to obtain an expression that is reminiscent of a completeness relation $\int dQ |Q\rangle \langle Q| = 1$. A similar expression can be written to describe a linear model predicting the value of a target property $E$ for a structure $A$. When using a representation $|A; \text{rep}\rangle$ to describe structures, $E$ can be approximated as

$$E(A) \approx \langle E | A \rangle \approx \int dQ \langle E; \text{rep} | Q \rangle \langle Q | A; \text{rep} \rangle,$$

with a slight abuse of notation because usually it does not make sense to see $\langle Q | E; \text{rep} \rangle$ as an inner product, and so $\langle Q | E; \text{rep} \rangle$ indicates just the regression weights for a model based on $|A; \text{rep} \rangle$. The bra-ket notation we use implicitly assumes linearity in the transformation between different choices of basis, and in the modeling of target properties. All of the expressions discussed here as integrals over a continuous index can be formulated as sums over (finitely or infinitely) countable, discrete indices

$$\int dQ |Q\rangle \langle Q| \sim \sum_q |q\rangle \langle q|.$$

Even though the features can be used as an input of an arbitrarily complex nonlinear regression scheme, we will often investigate their behavior in the context of linear models, because they reveal more transparently how a given representation reflects structure-property relations.

A pattern we use frequently in what follows, and that mimics a construction used in quantum mechanics, is the combination of multiple kets to build a tensor-product space, e.g.

$$| (A; \text{rep}) \otimes (A'; \text{rep}') \rangle = |A; \text{rep}\rangle \otimes |A'; \text{rep}'\rangle.$$

The construction of a tensor-product representation is well-defined even without indicating explicitly the basis used to describe either side of Eq. (6), and it is often possible to use either an explicit Cartesian product of the bases on the right-hand side, or a combined basis

$$\langle Q_1; Q_2 | A \otimes A \rangle \equiv \langle Q_1 | A \rangle \langle Q_2 | A \rangle \rightarrow \langle T | A \otimes A \rangle,$$
using only $|A\rangle$ as a special case of Eq. (3) in which $A \equiv A'$, and rep $\equiv \text{rep}'$ can be omitted.

Finally, we can consider the action of an “operator” on a ket, that is to be interpreted as a linear map that transforms the atomic structure, or the nature of the representation. Taking for instance the operator $i$ associated with inversion symmetry, $i|A\rangle$ indicates the representation associated with structure $A$ after the coordinates of all atoms have been reflected relative to the origin. By summing over the operators associated with a symmetry group, an operation which is also referred to as Haar integration\cite{110}, one can build symmetrized representations that are invariant under the actions of the elements of the group, e.g. for the $C_1$ point group,

$$|\langle A \otimes A \rangle_{C_1}\rangle = |A\rangle \otimes |A\rangle + i|A\rangle \otimes i|A\rangle. \quad (8)$$

When the resulting symmetric representation is used often, and the symmetry group is clear from the context, we indicate the averaging with an overline and omit the explicit indication of the group it has been symmetrized over, e.g., $|\langle A \otimes A \rangle_{C_1}\rangle \rightarrow \overline{|A\rangle \otimes |A\rangle} \rightarrow \overline{|A\rangle}. \overline{|A\rangle}$.

Figure 5 illustrates the notation and the Haar-integration in one dimension. Two distinct functions, $f$ and $g$ are plotted using their usual real-space features, $f(x) \equiv \langle x|f\rangle$, and $g(x) \equiv \langle x|g\rangle$. Applying inversion yields $\langle x|i|f\rangle = f(-x)$. An inversion-invariant feature can be created by symmetrizing: $\overline{|f\rangle} = |f\rangle + i|f\rangle$, but our choice of $f$ and $g$ leads to a degenerate feature as $\overline{|f\rangle} = |g\rangle$. A second-order feature may be obtained by generating the tensor product of the functions, e.g. $|f\rangle \otimes |f\rangle$, which in real space results in $\langle x_1;x_2|f \otimes f\rangle \equiv f(x_1)f(x_2)$. Symmetrizing this tensor product yields features $\overline{|f\rangle \otimes |f\rangle}$ and $\overline{|f\rangle \otimes |f\rangle}$ that are also inversion-invariant, but are still able to distinguish between the two functions.

### B. Global field representations

The starting point for the construction of a symmetry-adapted field representation is a field that describes the structure in terms of the distribution of its atoms -- or, more generally, of points that are associated with the building blocks of the material, as one would have in a coarse-grained model. In the simplest possible case, one would take localized functions $g$ centered on each atomic position $r_i$ and define

$$\langle x|A;\rho\rangle \equiv \sum_{i \in A} \langle x - r_i|g\rangle, \quad (9)$$

where $\langle x|g\rangle \equiv g(x)$ is a localized function (e.g. a Gaussian) and the $\rho$ in the ket indicates the kind of field used to describe the structure. As we discuss in more detail in Section 11.2 the atomic density functions can be either finite-width Gaussians, which leads to representations akin to SOAP feature,\cite{111} or Dirac $\delta$ distributions, which recovers representations similar to the moment tensor potential,\cite{111} or the atomic cluster expansion.\cite{112} To indicate the $g \rightarrow \delta$ limit, we use the notation $|\rho\rangle \rightarrow |\delta\rangle$. Atoms (or more generally, “point particles”) can be further characterized.
by internal attributes, that could be discrete (e.g. the chemical nature of an atom, \(a_i\)) or continuous (e.g. a molecular dipole \(u_i\))

\[
\langle u|A; \rho u \rangle \equiv \sum_{i \in A} \delta_{ua_i} \langle u - u_i|g \rangle \langle x_i - r_i|g \rangle .
\] (10)

In this form, Eq. (10) can be seen as an abstraction of the many real-space “voxel” representations of material.\(^{24,25}\) The ket \(|A; \rho\rangle\) defined by expressions like (9) or (10) could be equally well expressed in a different basis, e.g. expanded in plane waves

\[
|k|A; \rho \rangle = \frac{1}{(2\pi)^{3/2}} \int d\mathbf{x} e^{-i\mathbf{k} \cdot \mathbf{x}} \langle x|A; \rho \rangle .
\] (11)

Eqs. (9) and (11) contain the same amount of information. Even though the choice of a basis can be very important to simplify analytical derivations or practical implementation, representations can be regarded as abstract objects that can be defined independently of the basis set, much as it is the case for the wavefunction in quantum mechanics.

C. Translational invariance and atom-centered features

One way to make \(\langle x|\rho \rangle\) translationally invariant is to sum over the continuous translation group, \(\int d\hat{t} \langle x|\hat{t}|\rho \rangle\). As discussed in Ref. 90 summing directly over the atom density eliminates all structural information. Information loss is a usual issue with Haar integration, as exemplified in Figure 5\(^\text{B}\). One can avoid or reduce it by summing over tensor products of the atom density field. Considering the case in which atoms are described only by their position and chemical identity, integrating over translations \(\hat{t}\) yields a two-point density correlation function

\[
\langle a_1 x_1; a_2 x_2 | \rho \otimes \rho \rangle \equiv \langle a_1 x_1; a_2 x_2 | \rho^\otimes 2 \rangle = \int d\hat{t} \langle a_1 x_1 | \hat{t} | \rho \rangle \langle a_2 x_2 | \hat{t} | \rho \rangle = \sum_{ij} \delta_{a_1 a_j} \delta_{a_2 a_i} \int d\mathbf{g} \langle x_i - r_j - t | g \rangle \langle x_2 - r_i - t | g \rangle \sim \sum_{ij} \delta_{a_1 a_j} \delta_{a_2 a_i} \langle (x_i - x_2) - (r_j - r_i) | \tilde{g} \rangle
\] (12)

where \(\tilde{g}\) indicates the cross-correlation of two of the localized density functions. We indicate that the representation has been obtained by averaging over translations the tensor product of two density fields with the superscript notation \(\rho^\otimes 2\), and we separate with a semicolon groups of feature indices that are associated with each field, as discussed in Section III A. Note that the representation in Eq. (12) has a large null space, as it depends only on \(x_1 - x_2\). One could then re-define it by labelling features using a single position vector, or transform it in a plane wave basis:

\[
\langle a_1; a_2 | k | \rho^\otimes 2 \rangle = \int d\mathbf{x} \langle a_1 0; a_2 x | \rho^\otimes 2 \rangle (k|x) = \langle a_1 k | \rho \rangle^* \langle a_2 k | \rho \rangle
\] (13)

where \(\langle k|\rho \rangle \equiv e^{-i\mathbf{k} \cdot \mathbf{x}}\), and where the second equality is a consequence of the convolution theorem. One sees that the translationally-symmetrized density is equivalent to the diffraction pattern of the atomic structure, that has been used as a descriptor to classify crystalline configurations.\(^{23}\)

This construction can be taken as an inspiration to introduce an atom-centered representation

\[
\langle a x | A; \rho_i \rangle = \sum_{j \in A} \delta_{u a_j} \langle x - r_j | \tilde{g} \rangle,
\] (14)

where \(r_{ji} = r_j - r_i\). The fact it is atom centered (and hence translationally invariant) is hinted at by the subscript notation \(\rho_i\), and so in what follows we only use this subscript to distinguish it from its non-symmetrized counterpart (9) and simultaneously to indicate the central atom index. When expressing a representation centered around atom \(i\) without emphasis on its precise nature, we will use the notation \(|A_i\rangle\).

Writing the symmetrized two-point density correlation in terms of Eq. (13) clarifies how an atom-centered representation is a natural consequence of the translational symmetrization:

\[
(a_1 x_1; a_2 x_2 | A; \rho^\otimes 2 \rangle = \sum_{i \in A} \delta_{a_2 a_i} \langle a_1 (x_1 - x_2) | A; \rho \rangle .
\] (15)

When building a linear model, this expression implies an additive decomposition of the target property, as well as the use of separate models depending on the nature of the central atomic species:

\[
\langle E | A \rangle \approx \sum_{i \in A} \langle E; a_i | A_i \rangle = \sum_{i \in A} \sum_{a} \int d\mathbf{x} \langle E; a_i; \rho_i | a \mathbf{x} \rangle \langle a \mathbf{x} | A; \rho_i \rangle .
\] (16)

As discussed in Section V A, this expression can be taken as the prototype of all pair potentials, and higher-order of many-body interaction can be incorporated by taking higher tensor powers before symmetrization, or in the subsequent step of rotational averaging. Localization can be enforced by introducing a cutoff function in the definition (14). This is far from being an inconsequential operation, as it introduces an error (atomic energies and properties cannot depend on neighbors farther than this limit) but often results in more robust models, which perform better in the data-poor regime. We discuss this in more detail in Section VIII C.
D. Rotational invariance and body-ordered representations

The atom-centered representation \[[14]\] is translationally invariant, but does depend on the orientation of the structure. One should then proceed to perform Haar integration over the rotation group and (possibly) over inversion.

We can define the \((\nu + 1)\)-body order symmetrized field representation as

\[
|\rho_i^{\text{sym}}\rangle \equiv |(\rho_i \otimes \cdots \otimes \rho_i)_{SO(3)}\rangle = \sum_{k=0,1} \int_{SO(3)} d\hat{R} \hat{i}^k \hat{R} |\rho_i\rangle \otimes \cdots \otimes \hat{i}^k \hat{R} |\rho_i\rangle. \tag{17}
\]

This can be expanded on an explicit position basis (Fig. 6)

\[
\langle a_1 x_1; \ldots; a_v x_v |\rho_i^{\text{sym}}\rangle = \sum_{k=0,1} \int_{SO(3)} d\hat{R} \langle a_1 x_1 |\hat{i}^k \hat{R} |\rho_i\rangle \cdots \langle a_v x_v |\hat{i}^k \hat{R} |\rho_i\rangle, \tag{18}
\]

making it clear how \(|\rho_i^{\text{sym}}\rangle\) corresponds to a symmetrized, \(\nu\)-point correlation of the atom density centered on the \(i\)-th atom — a \((\nu + 1)\)-point correlation function, in the language used in statistical mechanics to describe the structure of liquids. \[114\] Similar to the case of Eq. (12), this object has a large null space (e.g. it only depends on \(x_1 = |x_1|\)). As discussed in Ref. \[60\], one can choose a more concise enumeration of the real-space correlations in terms of distances and angles, that reduces in the limit \(g \to \delta\) of \(\delta\)-like atom density to a sum over distances and angles between atoms. For instance, for the \(\nu = 2\) case one can write

\[
\langle a_1 r_1; a_2 r_2; \omega |\delta^{\text{sym}}\rangle \propto \sum_{jj'} \delta_{a_1 a_2} \delta_{a_2 a_j} \delta(r_1 - r_{jj'}) \delta(r_2 - r_{jj'}) \delta(\omega - \hat{r}_{jj'} \cdot \hat{r}_{jj'}), \tag{19}
\]

where we use \(\rho \to \delta\) to indicate that the correlation function is built on the \(g \to \delta\) limit of the atom density field. Expressions of this kind reveal the close connection between symmetrized-field representations and atom-centered symmetry functions \[10,115\] as well as equivalent constructions such as that used in DeepMD \[116\] and the FCHL features \[107,108\]. Features that describe a chemical environment are written as a sum over tuples of neighbors of appropriate functions of their distances and angles, and can be seen as just a different choice of basis set for Eq. (19)

\[
\langle a_1 a_2 k |\delta^{\text{sym}}\rangle = \int dr_1 dr_2 d\omega \times f^{(3)}(r_1, r_2, \omega) \langle a_1 r_1; a_2 r_2; \omega |\delta^{\text{sym}}\rangle. \tag{20}
\]

Note that we choose to symmetrize the atom-centered description \(|A; \rho_i\rangle\) — given that this is the procedure that recovers most of the existing representations — but one could as well proceed by averaging over tensor products of the translationally invariant representation of the full structure \(|A; \rho^{\text{iso}}\rangle\)

\[
|\langle \rho \otimes \rho \rangle_{SO(3)} \rangle \sim \int_{SO(3)} d\hat{R} \sum_i \hat{R} |\rho_i\rangle \otimes \hat{R} |\rho_i\rangle, \tag{21}
\]

as it was done for instance in Ref. \[117\]. Doing so results in the appearance of cross terms involving correlations between densities centered on different atoms, which could be used to systematically incorporate in this framework machine-learning approaches based on convolutional, and message-passing, neural networks that combine information centered on neighboring atoms. \[121,114\]

E. Density correlations in an angular momentum basis

More concise (and easier to evaluate) expressions for the density correlation representations can be obtained with a change of basis. Using orthonormal radial functions \(R_n(x) \equiv \langle x|n\rangle\) and spherical harmonics \(Y_{\ell m}(\hat{x}) \equiv \langle \hat{x}|\ell m\rangle\) yields a discrete set of coefficients that transform as spherical harmonics

\[
\langle anl m |\rho_i\rangle = \int dx \langle n|x\rangle \langle lm|\hat{x}\rangle \langle \hat{x}|a\rangle |\rho_i\rangle = \sum_{j \in A_i} \delta_{a\alpha} \int dx \langle n|x\rangle \langle lm|\hat{x}\rangle \langle \hat{x} - r_{jj'}|g\rangle,
\]

\[\tag{22}
\]

where \(\langle nlm; r_{jj'}|g\rangle\) are functions of \(r_{jj'}\) enumerated by the indices \((n, l, m)\), that can be evaluated numerically or analytically, depending on the choice of basis (see Section \[VIII.D\] for a few examples).
The use of spherical harmonics $|lm\rangle$ for the angular basis makes it easy to evaluate the rotational integral of Eq. (17) analytically, because the matrix elements $\langle lm| R |l'm'\rangle = \delta_{ll'} D_{mm'}^{(l)}(R)$ correspond to Wigner-D matrices, an irreducible representation of $SO(3)$. Well-known results from the theory of angular momentum \cite{Kleinhans} such as the orthonormality and the product reduction formula for Wigner-D matrices, allow deriving explicit expressions for the symmetrized field representations of order $\nu = 1, 2, 3$

$$\langle a_{1n_1 l_1 m_1} | \rho^{\overline{1}} \rangle = \frac{8\pi^2}{2l_1 + 1} \langle a_{1n_1 l_1 m_1} | \rho \rangle \delta_{l_1,0} \delta_{m_1,0},$$

(23)

$$\langle a_{1n_1 l_1 m_1; a_{2n_2 l_2 m_2} \rho^{\overline{2}}} \rangle = \delta_{l_1,2} \delta_{m_1,2} \frac{8\pi^2}{2l_1 + 1} \sum_s (-1)^{s-1} \langle a_{1n_1 l_1 s} | \rho \rangle \langle a_{2n_2 l_2} | \rho \rangle,$$

(24)

$$\langle a_{1n_1 l_1 m_1; a_{2n_2 l_2 m_2}; a_{3n_3 l_3 m_3} \rho^{\overline{3}}} \rangle = \frac{8\pi^2}{2l_1 + 1} (-1)^{s_1} \langle a_{2n_2 l_2 s_2} | \rho \rangle \langle a_{3n_3 l_3 s_3} | \rho \rangle,$$

(25)

where $\langle l_1 m_1; l_2 m_2 | LM \rangle$ is a Clebsch–Gordan coefficient.

Much as it was the case for the real-space versions of the density correlation representations, there are several redundant indices in these expressions, resulting from the rotational averaging that leaves the $m_i$ as free parameters. We can then re-label the invariant features, in a way that emphasizes the connection to existing representations, by coupling the angular basis and absorbing some of the consequent constant factors. For the case $\nu = 1$ one can define

$$\langle an | \rho^{\overline{1}} \rangle = \langle an | \rho \rangle$$

(26)

which corresponds to a discretized version of a pair correlation function $g_{ab}(r)$:

$$\langle an | \rho^{\overline{1}} \rangle = \int dr \langle n | r \rangle \langle 00 | r \rangle \langle a(r) | \rho \rangle$$

$$\propto \int dr \, r^2 \langle n | r \rangle \int dr \langle a(r) | \rho \rangle$$

$$\equiv \int dr \, r^2 \langle n | r \rangle \langle ar | \rho \rangle \propto \int dr \, r^2 R_{an}(r) g_{ab}(r)$$

(27)

For the $\nu = 2$ case, Eq. (21) can be redefined as

$$\langle a_{1n_1 l_1 m_1; a_{2n_2 l_2 m_2} \rho^{\overline{2}}} \rangle = \frac{(-1)^{l_1}}{\sqrt{2l_1 + 1}} \sum_m (-1)^m \langle a_{1n_1 l_1 m} | \rho \rangle \langle a_{2n_2 l} | (-m) | \rho \rangle$$

(28)

This corresponds – modulo irrelevant constants – to the rotation invariant 3D shape descriptor\cite{Klug} and to the SOAP features, which would be written, in the notation of Refs.\cite{Roth} and\cite{Ginibre} as

$$P^{a_1 a_2}_{n_1 n_2 l} = \frac{1}{\sqrt{2l + 1}} \sum_m c^{a_1}_{n_1 l m} (c^{a_2}_{n_2 l m})^*,$$

(29)

where $c^{a \rho}_{n l m} = \langle anl | \rho \rangle$ indicate the density expansion coefficients following the same notation. The $\nu = 2$ representation can be written on a real-space basis as $\langle a_{1r_1 a_{2r_2} \omega | \rho \rangle}$, indicating explicitly its nature as three-body density correlation function that depends on two distances $r_1$, $r_2$ and the cosine $\omega$ of the angle between the directions along which they are evaluated. The 4-body order invariant representation becomes

$$\langle a_{1n_1 l_1; a_{2n_2 l_2}; a_{3n_3 l_3} \rho^{\overline{3}}} \rangle = \frac{(-1)^{l_1}}{\sqrt{2l_1 + 1}} \sum_{n_1 l_1 n_2 l_2} (-1)^{m_1} \langle l_1 n_1 l_2 n_2 | l_3 l_3 \rangle \langle a_{1n_1 l_1 m_1} | \rho \rangle \langle a_{2n_2 l_2} | \rho \rangle \langle a_{3n_3 l_3} | (-m_3) | \rho \rangle$$

(30)

corresponding to the SOAP bispectrum\cite{Ginibre}

$$b^{a_1 a_2 a_3}_{n_1 n_2 n_3 l_{n_1 n_2 n_3 l_3} l_{n_1 n_2 n_3 l_3}} = \frac{1}{\sqrt{2l + 1}} \sum_{n_1 l_1 n_2 l_2 n_3 l_3} \langle l_1 n_1 l_2 n_2 | l_3 l_3 \rangle \langle a_{1n_1 l_1 m_1} | \rho \rangle \langle a_{2n_2 l_2 m_2} | \rho \rangle \langle a_{3n_3 l_3 m_3} | \rho \rangle,$$

(31)

and closely related to the bispectrum used in the spectral neighbor analysis method\cite{Klug} which is essentially equivalent to a different choice of basis. As discussed in more detail in Ref.\cite{Klug} and in the next sections, the relationship between the redundant expressions (23)\cite{Klug} (24)\cite{Klug} (25)\cite{Klug} that arise from the integral over rotations, and the more concise versions (26)\cite{Klug} (25)\cite{Klug} (30)\cite{Klug} can be seen as a transformation from the uncoupled to the coupled angular momentum basis, and starting from the $\nu = 4$ additional indices $k_\nu$ must be included to account for the different ways the coupling can be realized. Taking the $g \rightarrow \delta$ limit of $\rho$ yields the atomic cluster expansion (ACE), with $\langle nlm | \delta \rangle$ being equivalent to the $A_{nlm}$ coefficients of Ref.\cite{Klug} and the $B^{(\nu)}_{\nu l_{\nu_1} \cdots l_{\nu_{\nu}}}^{m_{\nu_1} \cdots m_{\nu_{\nu}}}$ coefficients corresponding to $\langle n_{\nu} l_{\nu_1} \cdots n_{\nu_{\nu}} l_{\nu_{\nu}} | \delta \rangle$; i.e., each $B^{(\nu)}_{\nu l_{\nu_1} \cdots l_{\nu_{\nu}}}^{m_{\nu_1} \cdots m_{\nu_{\nu}}}$ indicates a group of basis functions indexed by $k_1, \ldots, k_{\nu}$. Through a further linear transformation (change of basis) made explicit in Refs.\cite{Klug} (24) the moment tensor potential (MTP) of Ref.\cite{Klug} can also be related to this construction. However, the philosophy behind the density correlation features is different from that behind MTPs and ACE, in that these methods were at least originally thought of as bases for polynomial regression. While these basis functions can be equally used as symmetry-adapted features there are subtleties to be considered that we discuss in Sec.\cite{Klug} and
Note that even though the contracted basis $\{(n_{1i},k_{1i})_{1=1...\nu}\}$ eliminates some of the redundant indices that are present in the tensor-product basis, the indices do not label a set of linearly independent features. Symmetries and selection rules – some of which, listed in Ref. [123] can be derived from results of angular momentum theory [123] – restrict greatly the number of independent entries that need to be computed. However, the non-trivial interaction between the radial and angular basis component makes this list incomplete. A mixed algebraic/numerical precomputation step can further reduce the required features [124, 125].

Finally, the global SOAP-like descriptors introduced in Ref. [117] corresponding to Eq. (21), can be readily expressed in an angular momentum basis as

$$\langle a_1n_1; a_2n_2; l|A;\rho^{\otimes 2}\rangle = \frac{(-1)^l}{\sqrt{2l+1}} \sum_m (-1)^m \langle a_1n_1lm|\rho^{\otimes 2}\rangle \langle a_2n_2l(-m)|\rho^{\otimes 2}\rangle,$$

where we recall that $\langle anlm|A;\rho^{\otimes 2}\rangle = \sum_{i\in A} (anlm|\rho_i)$.

F. The density trick

A crucial point in comparing different representations is that with an appropriate discretization of the angular basis one can evaluate symmetrized high-order correlations as sum of products of the density coefficients defined in Eq. (22). This ensures that the cost of computing all coefficients of a given order $\nu$, scales only linearly with the number of neighbors included within the cutoff around atom $i$, even though it scales exponentially with $\nu$ in terms of the number of basis functions, at least with a naive choice of basis. This is to be contrasted with atom-centered symmetry functions (ACSF) [118, 119] and permutation invariant polynomials (PIP) [111] in which function are evaluated over all possible tuples composed of $\nu$ neighbors of the central atom (or on all the possible tuples in a structure to yield a global descriptor). In these frameworks, the cost depends linearly on the number of basis functions, but exponentially with $\nu$ in terms of the number of neighbors. This crucial difference makes density-expansion frameworks more convenient when one wants to ramp up the value of $\nu$, and there are many neighbors. A priori sparsification schemes, exemplified in [95], and feature selection schemes, discussed in Section VIII B allow one to keep only the most important basis functions can eliminate the exponential scaling with $\nu$ altogether.

Despite this rather fundamental difference in philosophy and computational cost, the two families of representations compute entities that are essentially equivalent, which we see by writing explicitly Eq. 28.

in the $g \rightarrow \delta$ limit as a sum over neighbors $j$ and $j'$

$$\langle a_1n_1; a_2n_2; l|\rho^{\otimes 2}\rangle \propto \sum_{jj'} \delta_{a_1a_j} \delta_{a_ia_j'} \langle n_1|r_{ji}\rangle \langle n_2|r_{j'i}\rangle$$

$$\times \sum_m (-1)^m \langle lm|r_{j'i}\rangle \langle l(-m)|r_{j'i}\rangle.$$

By using the addition formula of the spherical harmonics we get the equivalent formulation

$$\langle a_1n_1; a_2n_2; l|\rho^{\otimes 2}\rangle \propto \sqrt{2l+1}$$

$$\sum_{jj'} \delta_{a_1a_j} \delta_{a_ia_j'} R_{n_1}(r_{ji}) R_{n_2}(r_{j'i}) P_l(r_{j'i} \cdot r_{j'i}),$$

in which $P_l(\omega)$ is a Legendre polynomial of order $l$. In Eq. (34), the $\nu = 2$ density correlation coefficients are computed as a function of the distances and angles between triplets of atoms including the central atom $i$. By plugging this expression for $\langle a_1n_1; a_2n_2; l|\rho^{\otimes 2}\rangle$ into Eq. (20), that evaluates the value of an arbitrary atom-centered symmetry function, one sees that this result is not specific to the choice of $P_l$ as angular functions: in the limit of a complete basis set, it is equally possible to compute any ACSF using a sum over neighbor tuples or a contraction of density coefficients, drawing an explicit link between the SOAP power spectrum features [212, 107] Behler-Parrinello symmetry functions [101, 120], the DeepMD framework [118] and FCHL features [22]. Similar expressions could be derived for higher-order atom-centered symmetry functions, showing the complete equivalence – but dramatically different computational scaling with the number of neighbors – of the two frameworks.

G. Equivariant representations and tensorial features

The previous construction is suitable to represent any rotationally-invariant atomic property. In many circumstances, however, one is interested in representing vector-valued or general tensorial quantities $F$. In this case, the prescribed transformations that the tensor undergoes under the symmetry operations of the $O(3)$ group (e.g. $F(\hat{R}A) = \hat{R}F(A)$) have to be incorporated into the atomic representation in the form of equivariant, rather than simply invariant, features, so that the representation follows the same transformation as the target property, $\hat{R}A = \hat{R}A$. Equivariance can be enforced by comparing environments and defining the local contribution to the target relative to a pre-defined local reference frame, which has been used to build machine-learning models of tensorial properties in molecular systems [127, 133]. A more general approach for achieving this goal consists in endowing the representation with the symmetries of spherical harmonics $\langle \hat{x}|\lambda\mu\rangle = Y_{\lambda}^{\mu}(\hat{x})$, as well as the
desired parity under the action of the inversion operator \( i \), which we associate with a ket \( |\sigma\rangle \) such that \( i|\sigma\rangle = \sigma|\sigma\rangle \). The eigenvalue \( \sigma \) is 1 for polar tensors, and \(-1\) for pseudotensors. Features that transform as \( |\sigma\rangle \otimes |\lambda\rangle \) can be achieved by including two additional fields within the symmetrized tensor product of Eq. (17), i.e.,

\[
|\rho_{i}^{\otimes \nu} \otimes |\sigma\rangle \otimes |\lambda\rangle \rangle_{O(3)} = \sum_{k=0,1} \int_{SO(3)} d\hat{R}^{k} \langle \sigma \rangle \otimes \hat{R}^{k} |\lambda\rangle \rangle \\
\quad \otimes \hat{R}^{k} |\rho_{i}\rangle \otimes \cdots \otimes \hat{R}^{k} |\rho_{i}\rangle .
\]

(35)

The operation is depicted in Fig. 7, showing how the \( \lambda \mu \) ket corresponds to the evaluation of a set of spherical harmonics that anchors the atom-centered density to a reference frame. The scalar and rotationally-invariant case is recovered by taking \( |\sigma\rangle \otimes |\lambda\rangle = |1; 00\rangle \).

This construction represents a particularly convenient framework to target the prediction of any Cartesian tensor \( F \) in terms of its irreducible spherical components \( F_{\mu}^{\lambda} \), that transform under rotation and inversion as

\[
F_{\mu}^{\lambda}(RA) = \sum_{m} D_{\mu m}(\hat{R}) F_{m}^{\lambda}(A),
\]

\[
F_{\mu}^{\lambda}(iA) = \sigma(-1)^{\lambda} F_{m}^{\lambda}(A).
\]

(36)

Within a linear regression model, they can be written as the combination of equivariant representations of the proper order \( \lambda \) and parity \( \sigma \) with a set of rotationally-invariant weights \( \langle Q|F; \sigma; \lambda \rangle \):

\[
F_{\mu}^{\lambda}(A) = \langle F|A; \sigma; \lambda \rangle \\
\quad \approx \sum_{i} \int dQ \langle F; \sigma; \lambda \rangle \hat{\rho}_{i}^{\otimes \nu} \langle Q|A; \hat{\rho}_{i}^{\otimes \nu}; \sigma; \lambda \mu \rangle .
\]

(37)

Each irreducible spherical component of \( F \) gives rise to a separate equivariant model, and the appropriate transformation rules are ensured by the fact that each equivariant feature \( \langle Q|A; \hat{\rho}_{i}^{\otimes \nu}; \sigma; \lambda \mu \rangle \) separately transforms as the spherical harmonics \( \langle lm | \rangle \). Much like the case of invariant symmetrized fields features, Eq. (35) can be most effectively computed by first expanding the atom-centered field on a basis of spherical harmonics. An equivariant extension of the atomic cluster expansion[23] or the moment tensor potentials can be recovered taking the \( g \to \delta \) limit.

A concrete example of these features is given by the density coefficients themselves: in fact, one can see that the \( \nu = 1 \) equivariant reads simply

\[
\langle n|\rho_{i}^{\otimes 1}; \sigma; \lambda \mu \rangle \equiv \langle n\lambda(-\mu)|\rho_{i}\rangle \rho_{s1}.
\]

(38)

Note how in the bra-ket notation the \( (\lambda, \mu) \) indices on the two sides of this equation carry a different meaning. When used in the bra of the density expansion \( \langle n|\rho_{i}\rangle \), they identify one of many rotationally variant components; there is no explicit link to their symmetry properties, and one could build a rotationally variant models by selecting only some of the \( \mu \) values for a given \( (n, \lambda) \). When used in the ket of an equivariant feature \( \langle n|\rho_{i}^{\otimes 1}; \sigma; \lambda \mu \rangle \), they label groups of features that should be taken together, because they transform in a specific way under the symmetries of the \( O(3) \) group. It is easy to see that by using \( \langle n|\rho_{i}^{\otimes 1}; \sigma; \lambda \mu \rangle \) features in Eq. (37) one obtains a model that fulfills (36), because acting on the ket with \( \hat{R} \) yields a product with the associated Wigner matrix – with the caveat that pseudotensors cannot be described by \( \nu = 1 \) features.

Scalar products of these equivariant features generate kernels that are suitable for symmetry-adapted Gaussian process regression[133, 134], e.g. \( \lambda \)-SOAP kernels of the form

\[
k_{\mu \nu}^{\sigma \lambda}(A, A') = \sum_{i \in A, i' \in A'} \int dQ \langle A; \hat{\rho}_{i}^{\otimes \nu}; \sigma; \lambda \mu | \langle Q|A'; \hat{\rho}_{i'}^{\otimes \nu}; \sigma; \lambda \mu \rangle ,
\]

(39)

which generalize the covariant kernels used in Ref. [135] to learn interatomic forces. The fact that equivariant features of the form (35) follow \( O(3) \) transformation rules means that they can be combined using established relationships in the quantum theory of angular momentum. In particular, the coupled-basis representation used in the definition of Eqs. (26–30) can be formulated for an arbitrary value of \( \nu \), and in this form it is possible to express succinctly[123] a recursive formula to evaluate \( \langle \rho_{i}^{\otimes \nu}; \sigma; \lambda \mu \rangle \) based on lower order terms:

\[
\langle \ldots n_{k} l_{k} k_{i} n_{k} l_{k} k_{i} | n_{l} l_{l} l_{l} k_{l}; \lambda \mu | \langle Q|A; \hat{\rho}_{i}^{\otimes \nu}; \sigma; \lambda \mu \rangle \approx \delta_{\sigma \sigma} \delta_{\lambda \lambda} \times \sum_{m} \langle lm; k|\lambda \mu \rangle \langle n|\rho_{i}^{\otimes \nu}; \sigma; k \rangle \langle \ldots n_{k} l_{k} k_{i} n_{l} l_{l} l_{l} k_{l}; k \rangle .
\]

(40)
For $\nu = 2$, the recursion yields the original expression for $\lambda$-SOAP equivariants:\textsuperscript{133}

$$
\langle n_1l_1; n_2l_2 |^\nu \rho_{02} ; \sigma ; \lambda \mu \rangle = \frac{\delta_{\mu\mu}}{\sqrt{2\lambda + 1}} \sum_m \langle l_1m; l_2(m-m) |\lambda \mu \rangle \times \langle n_1l_1; n_2l_2 |m-m \rangle |\rho_1 \rangle \langle n_2l_2; n_2l_2 |m-m \rangle |\rho_2 \rangle .
$$

(41)

Similar recursive expressions have been independently proposed to efficiently compute invariant features\textsuperscript{147,111} that can be obtained by taking $|\sigma; \lambda \mu \rangle = |1; 00\rangle$ in Eq. (40). The possibility of combining equivariant features using angular momentum rules is also exploited in the construction of covariant neural networks.\textsuperscript{118,119}

It is also worth mentioning that it is possible to build models that are imbued with the appropriate transformation properties in an indirect fashion, by learning atom-centered scalars and combining them with the atomic positions to evaluate formal (or actual) molecular multipoles. This is easily seen for the case of the dipole moment of a neutral molecule, that can be computed as

$$
\mu(A) = \sum_{i \in A} q_i (A_i) \mathbf{r}_i ,
$$

(42)

Models of this form have been used since the early days of the construction of molecular potential and dipole moment surfaces,\textsuperscript{128,138} combined with neural-network potentials to compute IR spectra in the condensed phases,\textsuperscript{139} and more recently combined with tensorial models, to describe the interplay of atomic charges and polarization contributing to the total dipole moment.\textsuperscript{130} Assigning constant formal charges $q_i$ to atoms has also been used to derive covariant kernels, in the so-called operator machine learning framework\textsuperscript{141}, which is also similar in spirit to the tensorial embedded atom neural network\textsuperscript{142}. The gist of the idea (although expressed in a feature rather than kernel (or NN) language) is that one can define a translationally-invariant representation that depends formally on an applied electric field, e.g.

$$
\langle x | A_i ; E \rangle = \sum_{j \in A} g_j (x - r_{ji}) (r_{ji} , E) q_j
$$

(43)

Deriving with respect to one of the components of $E$ brings a dependency on the corresponding component of $r_{ji}$, that upon rotational averaging (keeping in mind that $\hat{R}$ acts on atomic coordinates and not on the external field) plays the same role as $|\lambda \mu \rangle$ in Eq. (35), providing a basis of features that can be used to learn vectors covariantly. The use of local interatomic vectors to build a covariant reference system is similar to the approach adopted in Ref.\textsuperscript{143} to define a general atomic neighborhood fingerprint, and in Ref.\textsuperscript{144} to machine-learn the position of electronic Wannier centers in condensed phases. Despite the superficial similarity with the environment-dependent point-charge model of Eq. (42), this scheme is more closely related to a framework based on atomic dipoles, since its predictions can be decomposed as a sum of atom-centered equivariant terms.

H. Long-range features

Introducing a cutoff in the definition of the local density is not only necessary to reduce the cost of evaluating the expansion coefficients, or the number of terms that have to be included to obtain a converged expansion of the density correlations. Increasing the range of the environment makes the model more complex, which often results in slower learning when limited training data is available.\textsuperscript{23} One pragmatic solution is to build models that explicitly incorporate a physically-motivated functional form as a baseline, which could take the form of an existing model\textsuperscript{147,148} or atomic multipoles\textsuperscript{142}.

A representation that aims to describe, in a data-driven manner, properties that depend on multiple length scales, should be similarly multi-scale in nature. This idea has been implemented by combining local representations with different cutoffs,\textsuperscript{23} scaling atomic contributions according to distance,\textsuperscript{23,150} (as we discuss in more detail in Section VIII C), treating separately intra- and inter-molecular correlations,\textsuperscript{142,151} as well as by building global structural representations based on an intrinsically multi-scale wavelet scattering transform.\textsuperscript{152} An alternative approach, that provides a good example of how the symmetrized atomic field construction can be extended beyond the use of an atomic density as the starting point, defines a Coulomb-like potential field based on the smoothed atomic density (Figure 5).

$$
\langle a x | A ; V \rangle = \int d x' \frac{\langle a x | A ; \rho \rangle}{|x - x'|} .
$$

(44)

Symmetrizing this field in the same way as for $|\rho \rangle$ leads to an atom-centered potential,

$$
\langle a x | A ; V_i \rangle = \int d x' \frac{\langle a x | A ; \rho_i \rangle}{|x - x'|} + \int d x' \frac{\langle a x | A ; \rho_i^c \rangle}{|x - x'|} 
\equiv \langle a x | A ; V_i^c \rangle + \langle a x | A ; V_i^c \rangle ,
$$

(45)

where we introduce the short-range density $|\rho_i^c \rangle$, restricted to the region within the cutoff, and the far-field density $|\rho_i^c \rangle$, restricted outside the cutoff, and the corresponding local and non-local fields $|V_i^c \rangle$ and $|V_i^c \rangle$. Crucially, these features incorporate information on atoms outside the cutoff, yet their complexity can be kept under control by restricting the range of the spherical environment over which they are computed. Just as for $|\rho_i \rangle$, the ket can be discretized by expanding it on an orthogonal basis of radial functions and spherical harmonics to obtain $\langle anlm | V_i \rangle$. 


FIG. 8. Relationship between Cartesian coordinates, local and long-range fields, for a hypothetical “doped graphene” system. (Left) Reference structure; middle: atom-density field, divided in three elemental channels, color-coded; right: atom-density potential, color-coded. Adapted from Ref. 137

FIG. 9. A multi-scale equivariant representation combining atom-centered density fields $|\rho_i\rangle$, long-range fields $|V_i\rangle$ and a set of spherical harmonics. One can then build features that are fully equivariant by averaging $|V_i\rangle$ over the symmetry operation of the $O(3)$ group, leading to $\nu$-point correlations analogous to those discussed above. Furthermore, one can combine local and long-range fields, as in Figure 9, constructing a family of multi-scale long-distance equivariants (LODE) features of the form

$$\langle \rho \otimes V \otimes \sigma \otimes \lambda | R \rangle = \sum_{k=0}^{\nu} \int dR \hat{R}^{k} | \sigma \rangle \otimes \hat{R}^{k} | \rho \rangle \otimes \hat{R}^{k} | V \rangle \otimes \hat{R}^{k} | \sigma \rangle \otimes \hat{R}^{k} | \rho \rangle \otimes \hat{R}^{k} | V \rangle$$

(46)

The simplest multi-scale representation $|\rho \otimes V\rangle$ can be linked to physics-based models using an atom-centered multipole expansion of electrostatic interactions, as we discuss further in Section V B but are effective to learn a multitude of long-ranged interactions, from permanent electrostatics to polarization and dispersion. Figure 10 shows that, when trying to represent long-range interactions between molecular fragments, a model based on local $|\rho_{\text{local}}\rangle$ features produces a completely unphysical behavior, with the interaction reaching a plateau when the molecules are separated by more than the cutoff distance. Multi-scale LODE features, instead, can describe the asymptotic tail even when using a 3Å cutoff in the definition of the atom-centered environments, and are capable of representing interactions of very different chemical nature. Using a non-local field as the starting point of the symmetrization procedure provides interesting opportunities to incorporate long-range, many-body interactions in atomistic machine learning.

V. REPRESENTATIONS AND MODELS

Even though this review focuses on the problem of representing atomic structures in terms of a vector of features, one cannot ignore the intimate connection between the choice of features and how they are used to construct representations of symmetric properties, such as site energies, which are then used in the context of regression schemes. The purpose of this section is therefore to discuss the interplay between representations and models. Given a set of symmetric features $\langle q | \lambda | A_i \rangle$ of an atomic environment $A_i$, we explore how to use it to represent a symmetric property $F(A_i)$. We discuss linear approximations,

$$F(A_i) \approx \sum_q \langle F | q \rangle \langle q | A_i \rangle$$

(47)

and show that the family of features we introduced in Section IV lead to natural generalisations of well-established models of interactions between atoms and molecules in terms of a body-ordered expansion. These relatively simple models put stringent requirements on the quality of the feature sets. We then go on to review how highly non-linear models may provide more flexibility in describing the relationship between a structure and its properties, and yield satisfactory results even with a rather simple, imperfect choice of features.
Here, and in the following, we always understand implicitly that equality in these approximations can only be attained in the limit of an infinite cutoff radius and suitably converged parameterisation.

A. Linear models and body-order expansion

An advantage of linear models is that they can often be connected to classical physics-inspired frameworks, and bring to light physical-chemical insights. An example of this connection involves the construction of interatomic potentials in terms of a body-ordered hierarchy of atom-centered energy terms

\[ E(A) = \sum_{i \in A} E(A_i) = \sum_{\nu} \sum_{i \in A} E^{(\nu+1)}(A_i), \quad (48) \]

in which each term can be written as a sum over \( \nu \) neighbors of the central atom

\[ E^{(\nu+1)}(A_i) = \sum_{j_1 < \cdots < j_\nu} f^{(\nu+1)}(r_{j_1}, \ldots, r_{j_\nu}). \quad (49) \]

This kind of expansion underlies the vast majority of empirical force fields, that are customarily written as a combination of pair potentials, and short-range 2, 3, and 4-body bonded terms.

Most potentials truncate this expansion at body-order three, i.e. \( \nu = 2 \) – a notable exception being the dihedral angle potentials used in force fields, that are four-body but involve selected groups of atoms rather than a sum over all possible triplets. This is because the cost of a naive evaluation of the sum \( \sum_{j_1 < \cdots < j_\nu} \) scales exponentially with the body order \( \nu \), i.e. as \( \mathcal{O}(N^\nu) \) for an environment containing \( N \) atoms. More sophisticated ways of symmetrizing the body-ordered terms, such as those discussed in Refs. [138] and [78], alleviate this behavior. In the following paragraphs we demonstrate, in particular, how this exponential scaling can be overcome by using the density correlation representations discussed in Section [V].

The three-body case. It is illuminating to first discuss in full detail the representation of a 3-body site potential, written traditionally in internal coordinates, in the form

\[ E(A_i) = \sum_j f^{(2)}(r_{ji}) + \sum_{j_1 < j_2} f^{(3)}(r_{ji}, r_{j_1i}, \omega_{j_1j_2i}), \quad (50) \]

where \( \omega_{j_1j_2i} := \mathbf{r}_{ji} \cdot \mathbf{r}_{j_1i} \). In order to connect to the atomic density correlations we first rewrite this as

\[ E(A_i) = \sum_j \left( f^{(2)}(r_{ji}) - \frac{1}{2} f^{(3)}(r_{ji}, r_{ji}, 0) \right) + \frac{1}{2} \sum_{j_1} f^{(3)}(r_{ji}, r_{j_1i}, \omega_{jj_1i}) + \sum_{j_1 < j_2} f^{(3)}(r_{ji}, r_{j_1i}, \omega_{j_1j_2i}), \quad (51) \]

adding and subtracting a self-interaction from the 3-body term.

Approximating \( u^{(2)}(r) \) in terms of a radial basis...
\[ R_n(r) = \langle r | n \rangle \text{ yields} \]

\[ E^{(2)}(A_i) = \sum_j u^{(2)}(r_{ji}) \approx \sum_j \sum_n \langle u^{(2)}| n \rangle \langle n|r_{ji} \rangle \]

\[ = \sum_n \langle u^{(2)}| n \rangle \int d\mathbf{r} \langle n|r \rangle \sum_j \delta(r - r_{ji}) = \sum_n \langle u^{(2)}| n \rangle \langle n|\delta^{(2)} \rangle \]

(52)

where \( |\delta_i\rangle \) is the \( g \rightarrow \delta \) limit of the atom-centered density \( |\rho_i\rangle \).

For the three-body term we revisit (20): first, we approximate \( u^{(3)} \) in terms of the radial basis \( R_n(r) = \langle r|n \rangle \) and the Legendre polynomials \( P_l(\omega) = \langle \omega|l \rangle \),

\[ u^{(3)}(r_{ji}, r_{j'i}, \omega_{ijj'}) \approx \sum_{nn'l} \langle u^{(3)}|nn'l \rangle \langle n|r_{ji} \rangle \langle n'|r_{j'i} \rangle \langle l|\omega_{ijj'} \rangle. \]  

(53)

Applying Legendre’s addition theorem to expand the \( P_l \) in terms of spherical harmonics \( Y^m_l(\hat{r}) = \langle \hat{r}|lm \rangle \),

\[ \langle l|\omega_{ijj'} \rangle = \frac{4\pi}{2l + 1} \sum_{m = -l}^{l} (-1)^m \langle lm|r_{ji} \rangle \langle l(-m)|r_{j'i} \rangle, \]

absorbing the \( \frac{4\pi}{2l + 1} \) into the weights \( \langle u^{(3)}|nn'l \rangle \) and reordering the summation yields

\[ u^{(3)}(r_{ji}, r_{j'i}, \omega_{ijj'}) = \sum_{nn'l} \langle u^{(3)}|nn'l \rangle \langle n|r_{ji} \rangle \langle n'|r_{j'i} \rangle \times \sum_{m = -l}^{l} (-1)^m \langle lm|r_{ji} \rangle \langle l(-m)|r_{j'i} \rangle \]

\[ = \sum_{nn'l} \langle u^{(3)}|nn'l \rangle \sum_{m} (-1)^m \langle nlm|r_{ji} \rangle \langle n'l(-m)|r_{j'i} \rangle. \]  

(54)

Finally, we sum over all \((j, j')\) and reorder the summation to arrive at

\[ \sum_{jj'} \langle u^{(3)}|n|n'n'\rangle = \sum_{nn'l} \langle u^{(3)}|nn'l \rangle \]

\[ \sum_{m} (-1)^m \sum_{jj'} \langle nlm|r_{ji}\delta \rangle \sum_{j''} \langle n'l(-m)|r_{j'i}\delta \rangle \]

\[ = \sum_{nn'l} \langle u^{(3)}|nn'l \rangle \sum_{m} (-1)^m \langle nlm|\delta_{ij} \rangle \langle n'l(-m)|\delta_{ij} \rangle \]

\[ = \sum_{nn'l} \langle u^{(3)}|nn'l \rangle \langle n'l'|\delta^{(2)} \rangle. \]  

(55)

In summary, we have written an arbitrary 3-body site potential in terms of 1- and 2-correlations of the atomic density,

\[ E(A_i) = \sum_n \langle u^{(2)}|n \rangle \langle n|\delta^{(2)} \rangle + \sum_{nn'l} \langle u^{(3)}|nn'l \rangle \langle n'l'|\delta^{(2)} \rangle \]  

(56)

Aside from connecting classical body-ordered interatomic potentials and \( \nu \)-correlations of the atomic density this formulation has significant advantages in terms of computational complexity which we discuss below after generalising the argument to arbitrary body-order.

**General \((\nu + 1)\)-body order potentials.** The systematic expansion to arbitrary body orders has been applied to the description of alloys in terms of a cluster expansion, a procedure that was very early shown to provide a complete description of the problem\(^3\), to the rationalization of fragment-based electronic structure methods\(^4\) and to the construction of last-generation potentials for water and aqueous systems\(^5\).

We adopt the generalisation of (51) that includes self-interaction,

\[ E^{(\nu + 1)}(A_i) = \sum_{j_1, \ldots, j_{\nu}} u^{(\nu + 1)}(\mathbf{r}_{j_1}, \ldots, \mathbf{r}_{j_{\nu}}), \]  

(57)

which can be obtained from the more natural formulation\(^3\) by correcting the self-interaction terms into the \( \nu \)-body-order energy similarly to Eq. (51).

To connect (57) to the density correlations we represent the rotationally invariant \((\nu + 1)\)-body function \( u^{(\nu + 1)} \) as

\[ u^{(\nu + 1)}(\mathbf{r}_{j_1}, \ldots, \mathbf{r}_{j_{\nu}}) = \int_{O(3)} d\hat{R} \int dQ \langle u^{(\nu + 1)}|Q \rangle \langle Q | \hat{R} | \mathbf{r}_{j_1}, \ldots, \mathbf{r}_{j_{\nu}} \rangle, \]  

(58)

where we use \( Q \) as a shorthand for \((\mathbf{x}_1; \ldots; \mathbf{x}_\nu)\), so that \( \langle Q | \mathbf{r}_{j_1}, \ldots, \mathbf{r}_{j_{\nu}} \rangle = \prod_{k=1}^{\nu} \delta(\mathbf{x}_k - \mathbf{r}_{j_k}) \). The rotation can be made to act on the atomic positions or on the basis, depending on convenience. The \((\nu + 1)\)-order site energy is obtained by summing over clusters of neighbors

\[ E^{(\nu + 1)}(A_i) \approx \sum_{j_1, \ldots, j_{\nu}} \int_{O(3)} d\hat{R} \int dQ \langle u^{(\nu + 1)}|Q \rangle \langle Q | \hat{R} | \mathbf{r}_{j_1}, \ldots, \mathbf{r}_{j_{\nu}} \rangle \]

\[ = \int dQ \langle u^{(\nu + 1)}|Q \rangle \int_{O(3)} d\hat{R} \sum_{j_1, \ldots, j_{\nu}} \langle Q | \hat{R} | \mathbf{r}_{j_1}, \ldots, \mathbf{r}_{j_{\nu}} \rangle. \]  

(59)

The symmetrized sum can be reordered to show that it corresponds to the \( \nu \)-point density correlation

\[ \int_{O(3)} d\hat{R} \sum_{j_1, \ldots, j_{\nu}} \langle \mathbf{x}_1; \ldots; \mathbf{x}_\nu | \hat{R} | \mathbf{r}_{j_1}, \ldots, \mathbf{r}_{j_{\nu}} \rangle \]

\[ = \int_{O(3)} d\hat{R} \sum_{j_1, \ldots, j_{\nu}} \prod_k \delta(\hat{R}\mathbf{x}_k - \mathbf{r}_{j_k}) \]

\[ = \int_{O(3)} d\hat{R} \prod_k \sum_{j_k} \delta(\hat{R}\mathbf{x}_k - \mathbf{r}_{j_k}) \]

\[ = \int_{O(3)} d\hat{R} \prod_k \langle \hat{R}\mathbf{x}_k | \delta \rangle = \langle \mathbf{x}_1; \ldots; \mathbf{x}_\nu | \delta^{(\nu)} \rangle, \]  

(60)
which is precisely Eq. [18] written in the $g \rightarrow \delta$ limit. Thus we have explicitly represented $E^{(\nu+1)}$ in terms of the symmetry-adapted density correlations. We emphasize again that this calculation required the inclusion of the self-interactions as the starting point.

For a practical implementation we can choose a finite, discrete basis, approximating $E^{(\nu+1)}$ as

$$E^{(\nu+1)}(A_i) \approx \sum_q \langle u^{(\nu+1)}|q \rangle \langle q|\delta^n_{i}\rangle . \quad (61)$$

Any complete implementation of $\nu$-order density correlation features provides a basis to expand $u^{(\nu+1)}$ and approximate a $(\nu + 1)$-order term, that contributes to the body-ordered expansion of $E(A)$. The foregoing discussion shows that these bases are complete: An (infinite) collection of symmetrized features $\{\langle q|A_i\rangle\}_{q \in \mathbf{q}_{\text{total}}}$ is a complete linear basis if there exists a sequence of finite subsets $\mathbf{q} \subset \mathbf{q}_{\text{total}}$ such that

$$F(A_i) \approx F_\mathbf{q}(A_i) := \sum_{q \in \mathbf{q}} \langle F|q \rangle \langle q|A_i\rangle , \quad (62)$$

i.e., $F_\mathbf{q}$ approximates $F$ to within arbitrary accuracy in the limit as the number of features tends to infinity. We stress here that the weights $\langle F|q \rangle$ depend on the entire choice of feature set $\mathbf{q}$ and not just the single index $q$. Therefore the density correlation features provide a universal, complete linear basis to approximate body-ordered potentials and, more generally, body-ordered expansions of properties that can be meaningfully written as a sum of atom-centered contributions.

For the specific choice

$$\langle q|A_i\rangle = \otimes_{\nu=1}^\nu \langle n_i \alpha \lambda \mu \alpha \delta_i\rangle , \quad (63)$$

Eq. (61) is the ACE model. Note that the symmetrized correlations $\langle q|\delta^n_{i}\rangle$ can be efficiently and conveniently evaluated as already hinted at in Section IV. Since MTPs provide an alternative basis set for the same space, they are complete as well, and in the same sense. We also emphasize that a rigorous proof of completeness of MTPs was already given by Shapeev. The “density trick,” i.e., expanding in terms of the density correlations, ensures linear scaling in terms of the number of neighbors $N_i$ rather than the $(N_i \nu)^{\nu+1}$ scaling of the naive representation, which enables modeling very high body-orders. A recursive evaluation of the $\nu$-correlations implemented by the MTP and ACE bases, or by the NICE formalism, avoids an unfavorable scaling of the evaluation of the high-order terms (see Section VIIID for a summary of these techniques).

Density smearing. The real-space view of the density correlation features may be more intuitive when considering finite smearing of the atomic contributions to $|\rho_{i}\rangle$, that gives rise to a smooth function that can be seen as a proxy for the electronic density, and is reminiscent of the atoms-in-molecules description of the electronic structure of a molecule or a condensed-phase system as a collection of atom-centered densities.

Since we derived the link between density correlations and body-ordered potentials, and in particular the proof of the completeness of the linear expansion, only in the limit of a sharp density we now discuss whether a similar formal guarantee holds for a general $|\rho_{i}\rangle$, admitting in particular smearing of the atomic contributions. With tensor-product bases, all statements derived for higher correlation orders can eventually be reduced to a one-dimensional description, that is sufficient to reveal the essential features of the problem.

Given a function $f(x)$ which can be expanded into polynomials $p_n(x) = x^n :$

$$f(x) \approx f_{n_{\text{max}}}(x) = \sum_{n=0}^{n_{\text{max}}} c_n x^n \rightarrow_{n_{\text{max}} \rightarrow \infty} f(x), \quad (64)$$

we want to check whether we can also represent $f$ in terms of smeared polynomials,

$$p_n^\sigma(x) = g \ast x^n = \int (t - x)^n e^{-t^2/2\sigma^2}/\sqrt{2\sigma\pi} \, dt \quad (65)$$

For this particular choice of Gaussian smearing we can evaluate this expression explicitly and obtain

$$p_n^\sigma(x) = x^n + \text{lower order terms}, \quad (66)$$

i.e., $p_n^\sigma$ is in fact still a polynomial with leading-order term $x^n$ and this means it forms a basis. In particular we can now again represent $f_{n_{\text{max}}}(x)$ exactly as

$$f_{n_{\text{max}}}(x) = \sum_{n=0}^{n_{\text{max}}} c_n p_n^\sigma(x) \quad (67)$$

And in the limit $n_{\text{max}} \rightarrow \infty$ we recover $f$.

In the more general case, suppose that we have an arbitrary complete basis $\{\phi_j\}$. Then we can approximate $x^n \approx \sum_j b_{nj} \phi_j(x)$. The smearing operator $g \ast \cdot$ is bounded, which allows us to write

$$p_n^\sigma = g \ast x^n \approx \sum_j b_{nj} g \ast \phi_j = \sum_j b_{nj} \phi_j^\sigma. \quad (68)$$

Given that $p_n^\sigma$ are dense, it follows that also the smeared basis functions $\phi_j^\sigma = g \ast \phi_j$ are dense. From these arguments it is reasonable to conjecture that the smeared density correlations also form a complete linear basis. There is however an important caveat: The inverse of the smearing operator is unbounded, which implies that the coefficients of the expansion of $f$ in terms of the smoothed polynomial basis necessarily blow up when the order is increased, even if $f$ has a stable expansion in a polynomial basis. Therefore, in practice, the smoothening of the density, the truncation of the basis, and the regularisation of the regression, must be carefully coordinated and adapted to the natural scale of the variations of the target function $f$, i.e., to its “natural” smoothness.
B. Long-range features and potential tails

A similar formal correspondence with well-established functional forms of physical interactions can be derived when using (scalar) multiscale LODE features \(^{46}\) within an additive, linear learning model, using as target the electrostatic energy \(U(A)\),

\[
U(A) = \sum_{i \in A} U(A_i) = \sum_{i \in A} \int dQ \langle U(Q) \rangle \langle Q; A; \rho_i \otimes V_i \rangle.
\]

The fact that the representation is linear both in the density and in the potential fields allows one to derive rigorous asymptotic relationships for the interaction between two distant portions of the system, that resemble the electrostatic interactions between the multipoles of a localized charge density distribution and any other charge that is located arbitrarily far away.\(^{127}\)

Focusing only on the long-range contribution \(U^\gamma\) to \(U(A_i)\), that is associated with the part of \([A; V_i]\) generated by the far-field density, \([A; V_i^\gamma]\), one can write

\[
U^\gamma(A_i) = \sum_{a_1a_2} \sum_{l=0}^{l_{\text{max}}} \sum_{m=-l}^{l} \int_{r_{\text{cut}}} dr \left( \langle \rho_i^a \rangle a_2 \langle l m \rangle \frac{1}{r^{l+1}} \right) \times \langle a_1 a_2 ; l m | M_i^{\gamma}(U) \rangle.
\]

In this expression, in which the reader can recognize the similarity with the multipole expansion of the electrostatic potential \(^{128}\) \(|\rho_i^a\rangle\) indicates the atom density outside the cutoff, which is not computed explicitly but is encoded in the expansion of the local atomic potential \(^{44}\). The coefficients \(\langle a_1 a_2 ; l m | M_i^{\gamma}(U) \rangle\) can be written as a combination of the regression weights \(\langle a_1 n_1; a_2 n_2; l | U \rangle\) and the local density coefficients \(\langle a_1 l m | \rho_i^a \rangle\), and can be interpreted as adaptive multipole coefficients that depend in a general manner on the atomic distribution within the environment.

Given that the atomic densities and potentials are not the physical charge density and electrostatic potential of the system, it is the role of the regression procedure to modulate the multipoles so as to reproduce the reference data for the electrostatic energy. In Fig. 11 we report an example where this is demonstrated by extrapolating the long-range interaction between a pair of rigid \(\text{H}_2\text{O}\) and \(\text{CO}_2\) molecules, upon training the multiscale LODE model on the long-range, yet not asymptotic, interaction profiles associated with 33 different reciprocal orientations of the two molecules. The figure compares the asymptotic extrapolation performance upon centering the representation on different atomic centers, as well as by truncating the angular expansion at different \(l_{\text{max}}\). It is apparent that the angular cutoff chosen reflects the number of multipoles introduced in the expansion of Eq. (70) and thus determines sharp crossovers of the prediction accuracy across critical \(l_{\text{max}}\) values. For instance, a model that uses only features centered on the oxygen atom of \(\text{H}_2\text{O}\) improves dramatically its performance when \(l_{\text{max}}\) is increased from zero to one. A model using the carbon atom of \(\text{CO}_2\) as the only environment shows a similar, sharp improvement in accuracy when going from \(l_{\text{max}} = 1\) to \(l_{\text{max}} = 2\). This is consistent with the primarily dipolar nature of the electrostatic field generated by a water molecule, and with the quadrupolar nature of the center-symmetric carbon dioxide. Even though this example showcases the link between a linear model based on \(|\rho_i \otimes V_i\rangle\) and multipole electrostatics, the representation is sufficiently flexible to describe also other kinds of interactions, as demonstrated in Figure 10.

C. Non-linear models

Historically, linear representations used basis sets in internal coordinates (typically interatomic distances or simple transformations of them) that exploded in size with body order, see e.g. Refs. \(^{70}\)\(^{135}\) and with exponential scaling in their computational cost of prediction due to the need to sum over all \(\nu\)-clusters in a configuration or atomic environment. Moreover, it is clear that high body orders would...
be needed to obtain the desired accuracy, especially for models of materials. About a decade ago, non-linear fits using low body order ($\nu = 2$) descriptors appeared, with the surprising result that a few hundred degrees of freedom were enough to get good potential$^{111,112}$. Contrary to linear modeling where the symmetry-adapted features $\langle q | A_i \rangle$ are used as a basis, in the context of non-linear regression, they are best thought of as a coordinate transformation $\phi(A_i)$. In a linear setting the choice of a basis, and the details of the implementation, are a matter of computational performance but can be converged to a well-defined, basis-set independent limit. When used as the input of a non-linear model, instead, the entries of the feature vector $\phi_q(A_i) = \langle q | A_i \rangle$ must always be precisely defined, because there is no complete basis set limit in which the models become equivalent. If $F(A_i)$ is a symmetric property such as a site energy, we aim to construct approximations of the general form

$$F(A_i) \approx \tilde{F}(\phi(A_i)).$$

(71)

The two most commonly used models for $\tilde{F}$ are artificial neural network (ANN) and kernel ridge regression (KRR) models. In KRR models, one builds a kernel matrix $K$ with elements

$$K_{ij} = K(\phi(A_i), \phi(A_j)),$$

(72)

which provides a similarity measure between the environments $A_i$ and $A_j$, measured in terms of the similarity between the corresponding feature vectors $\phi(A_i)$ and $\phi(A_j)$. Useful kernel functions, $K$, are nonlinear, e.g. polynomials, Gaussians, etc.$^{113}$ The kernel inherits the symmetry of the feature vectors, and therefore a model for a symmetric property $F(A_i)$ can be obtained as

$$F(A_i) = \sum_{j \in M} c_j K(\phi(A_i), \phi(M_j)),$$

(73)

where, in the simplest setting, the $M_j$ are scattered interpolation points, but more generally are simply a collection of “centers” which induce a basis $\{K(\cdot, \phi(M_j))\}$ in the symmetrized feature space. The weights $c_j$ are then obtained by a linear regression. Kernel models have two main advantages over “naive” linear regression using the same features. (1) They introduce implicitly a non-linear mapping between the inputs and a “reproducing kernel Hilbert space” $|A_i\rangle \rightarrow |A_i; K\rangle$, which has a larger (often infinite) dimensionality, allowing for a more flexible approximation of $F(A_i)$. (2) Given that the basis is centered on the training points, it is adapted to the geometry of the data set in feature space. For example, if the centers $|A_i; K\rangle$ in feature space fall on (or close to) a low-dimensional manifold then the KRR model naturally exploits this. For a comprehensive discussion of the use of kernel methods in atomistic modeling, see Ref.$^{119}$. In the context of body ordered features discussed above, the non-linearity in the kernel effectively increases the body order of the features used in the regression model, but in rather special way: only those high body order terms are present that can be obtained as functions of low body order features. See Section VI on completeness for a more detailed discussion.

While nonlinear models are by their very nature more flexible in representing complex high-dimensional features, linear models come with different advantages. As we have shown in Section VA and Section VB they tend to be more easily “interpretable”, e.g. in terms of a body-ordered expansion of the target properties, or in terms of physically-motivated asymptotic forms of the interactions. But are nonlinear models necessary to achieve high accuracy? This notion is challenged by the SOAP$^{27}$, the SNAP$^{121,122}$, the MTP$^{113}$, the ACh$^{129}$ and the NICE$^{123}$ representations: the “density trick” and its generalisations to higher body-orders, replacing polynomials with correlations of the atom-centered density, circumvents both the explicit symmetrization as well as the summation of all $\nu$-clusters of traditional body ordered expansions. Particularly when using density correlations above $\nu = 2$, it is critical to fully exploit the computational cost gains offered by permutation symmetric properties. Even if one were to initially specify a model in terms of the “natural” body-order expansion$^{47}$, one should convert it for computationally efficient evaluation to a representation in terms of $\nu$-correlations, e.g. within the ACh$^{126,127}$, MTP$^{113}$ or NICE$^{123}$ frameworks. By employing the recursive evaluations introduced in Refs.$^{107,111,123}$ this transformation makes it possible to truncate at
very high body-orders without significant penalty in computational cost, as discussed in more detail in Section V IID. As an illustration of how a linear fit based on high-quality density-correlation representations can compete with non-linear models we show in Fig. 12 the learning curves resulting from the regression of the atomization energy for a very large and geometrically diverse database of CH$_4$ configurations (generated by randomly displacing the H atoms around the central carbon, in a sphere with a radius of 3.5Å). The plot reflects a tradeoff between model complexity and the availability of training data. Saturation of the learning curves indicates that the model does not have sufficient flexibility to describe fully the underlying structure-property relations.

Thus, linear models based on NICE features incorporating higher and higher body order are capable of describing the structure-property relations to a higher degree of accuracy, which is apparent in the delayed saturation of the learning curve. One sees that a ν = 4 model starts saturating around $n_{\text{train}} = 10^5$, even though the system is composed of 5 atoms, and so the body-ordered expansion should be fully converged.

This is because a linear model requires a complete basis, while here we select only a few 1000s invariants at each body order. A NN model can be designed to be more flexible and beat this saturation, at the expense, however, of performance in the small data set limit - which, in a more chemically diverse regression exercise, usually translates to poorer transferability.

VI. GEOMETRIC COMPLETENESS OF REPRESENTATIONS

In Section IV we discussed two classes of models built from invariant features: linear models, for which the representation plays the role of a basis to expand the target property, and nonlinear models, where the representation $\Phi(A_i) = \{(q(A_i))_q\}$ plays the role of a coordinate transformation generating a finite-dimensional feature vector used as the argument of a non-linear function. In order to guarantee systematic convergence of these models in suitable limits we require that the employed set of features is complete. We already hinted in Section IV C that these two scenarios lead to different requirements on the notion of completeness. In this section we provide a more in-depth discussion of the completeness issue in the nonlinear setting, and point out open problems.

Suppose we are given a finite collection of symmetry adapted features $\Phi(A)$ which we wish to use as a descriptor for atomic structures or environments, for example symmetrized correlations of the density as described in foregoing sections. We say that this descriptor is geometrically complete if the mapping $A \mapsto \Phi(A)$ is injective. Colloquially, this means that any two atomic configurations that are not related by symmetry are mapped to different descriptors. The consequence of this definition is that, if a feature vector $\phi(A)$ is geometrically complete then a symmetric property $F(A)$ can be written as

$$F(A) = \tilde{F}(\phi(A)), \quad (74)$$

where $\tilde{F}$ is now a generic non-linear function in feature space which can be constructed by any arbitrary regression scheme. We then say that the feature vector $\phi(A)$ is algebraically complete, which is a less stringent requirement than the linear completeness discussed in Section IV A.

A. Algebraic completeness: an elementary example

The ideal goal would be to have complete finite feature sets, that allow to approximate any symmetric function of the coordinates to arbitrary accuracy. As an elementary introduction to how such a construction might be achieved in principle, we consider a collection of $N$ particles in 1D, $\{x_i\}_{i=1}^N$. As a concrete example, one can take two particles with positions $(x_1, x_2)$. In the absence of an angular component, we only need to consider the projection of the density $\rho(x) = \sum_i \delta(x - x_i)$ onto the monomial basis $x^n$:

$$\langle n| \rho \rangle = \sum_{i=1}^N x_i^n, \quad n \in \mathbb{N} \quad (75)$$

For example, if $N = 2$, $\langle 1| \rho \rangle = x_1 + x_2$, $\langle 2| \rho \rangle = x_1^2 + x_2^2$, etc. In this simple setting, one sees easily how the $\nu$-point density correlations form a basis of symmetric polynomials

$$\langle n_1 \cdots n_\nu | \rho^{\otimes \nu} \rangle = \sum_{i_1 \cdots i_\nu} x_{i_1}^{n_1} \cdots x_{i_\nu}^{n_\nu} = \prod_{k=1}^{\nu} \langle n_k| \rho \rangle \quad (76)$$

which is linearly complete because it contains all possible symmetrized monomials. In analogy to what we did in Section IV A we use the “self-interaction” formulation in which the sum extends over all the tuples of particle indices. For the case of two particles, linear combinations of $\langle n_1 n_2 | \rho^{\otimes 2} \rangle = x_1^n x_2^n + x_1^n x_2^n + x_1^n x_2^n$ are sufficient to write any symmetric polynomial of the particle positions.

Thus, if we allow for algebraic operations on the $\langle n| \rho \rangle$, it is clear that the $\nu = 1$ coefficients provide a sufficient basis, because the elements of the linear basis (76) can be obtained as a product, e.g. $\langle n_1 n_2 | \rho^{\otimes 2} \rangle = \langle n_1| \rho \rangle \langle n_2| \rho \rangle$. In fact, well-established results from the theory of symmetric polynomials allow making an even stronger statement. The first $N$ power sum polynomials $(\langle n| \rho \rangle)_{n=1}^N$ provide an algebraically-complete basis to write any symmetric polynomial function of the coordinates of $N$ particles.
For instance, for $N = 2$ we can express the $n = 3$ term as a polynomial of $\langle 1|\rho \rangle$ and $\langle 2|\rho \rangle$

$$\langle 3|\rho \rangle = x_1^3 + x_2^3 = \frac{3}{2}(x_1 + x_2)(x_1^2 + x_2^2) - \frac{1}{2}(x_1 + x_2)^3$$

$$= \frac{3}{2} \langle 1|\rho \rangle \langle 2|\rho \rangle - \frac{1}{2} \langle 1|\rho \rangle^3 \quad (77)$$

This result implies, in general, that the mapping

$$\{x_i\}_{i=1}^N \mapsto \Phi = \{\langle n|\rho \rangle\}_{i=1}^N \quad (78)$$

is injective: knowledge of the first $N$ features $\langle n|\rho \rangle$ allows us to uniquely reconstruct the configuration (but not the index of the atoms). That is, this minimal feature set $\Phi(A)$ is indeed geometrically complete. It is not too difficult to construct similar complete and finite feature sets for finitely many particles in two and three dimensions as long as only permutational symmetry is considered. However, incorporating also rotational symmetry into the equivalence of particle configurations makes this much more challenging as we discuss next.

**B. Geometric completeness of density correlations**

In general, for three-dimensional atom configurations it is clear that taking all $\nu$-correlations provides a complete set of features (after all, they are even complete in the linear sense), however, this is not useful in practice. As we explain next, it remains an open problem how to construct a minimal complete feature set in this general setting.

It is clear just based on dimensionality arguments that a descriptor that has fewer than $3N - 6$ components (the number of elements in the Cartesian position vectors, subtracting the degrees of freedom associated to translations and rotations) cannot be complete for $N$ particles. On the other hand, the descriptors based on $\nu$-point correlations have a number of components that scales with $N^\nu$. But having more than the necessary minimum number of components does not ensure that a descriptor is complete.

Although it was appreciated for a long time that two-correlations for entire structures are not complete, i.e. knowing the set of distances between points is not enough to reconstruct the point set $\mathcal{S}$. That is, this minimal feature set $\Phi(A)$ is indeed geometrically complete. It is not too difficult to construct similar complete and finite feature sets for finitely many particles in two and three dimensions as long as only permutational symmetry is considered. However, incorporating also rotational symmetry into the equivalence of particle configurations makes this much more challenging as we discuss next.

![FIG. 13. A pair of environments that are not distinguished by two-correlations (sets of distances from the origin and central angles), formed from the blue atoms (1-3) and either one of 4 or 4'. The angle $\alpha$ is arbitrary. The tables on the right show the angles (or equivalently, distances) between the numbered particles in each configuration. The two environments are not related by symmetry, but the sets of distances are identical, only a pair are swapped, that are highlighted with the gray background.](image-url)
but the relationship between these two cases is subtle. Given an entire structure, it can be considered to be the “environment” of the point at the origin, and the same symmetries apply. However, specific representations appear differently in the two views. For example, the two-correlation for an entire structure, the set of interparticle distances, when considered as an environment of the origin is the two-correlation of neighbors, and we naturally think of it as a three-body correlation, i.e. a set of distances from the origin and central angles. Note that the problem of completeness for entire structures is exactly the same as the problem of reconstructing point sets.

One way to break the degeneracy between the representations of two entire structures involves combining information on different environments. For instance, one can describe the entire structure using an additive combination of atom-centered features analogous to Eq. 115. Following the above reasoning, a pair of environments that degenerate in terms of the list of distances and angles are also structures that are degenerate in terms of the list of distances. However, these structures are not necessarily degenerate in terms of the combined list of distance and angle histograms of each local environment. Thus, taking non-linear transformations of atom-centered features cannot resolve the environment-level degeneracies, but can provide a way to differentiate entire structures The construction of injective yet concise representations for environments and structures is still an open problem, whose solution may help to improve the accuracy and computational efficiency of machine-learning models.

Note that in this discussion we are implicitly taking atomic structures related by symmetry as identical, and we focus on whether the injectivity holds for the domain of the descriptor map being the original atomic structures. The case of whether the same consideration holds for general scalar fields (e.g. those arising in the LODE construction) is a separate problem. For the case of translation symmetry (torus geometry) it is well-known that no finite correlation order suffices to reconstruct all signals, however most signals can be reconstructed already from the bi-spectrum ($\nu = 3$). To the best of our knowledge it is an open problem whether analogous results hold for the case of rotational symmetry of 3D spherical geometry.

C. Spectral representations

As we explained above the set of all $(N-1)$-correlations is complete for $N$ particles, because it is equivalent to the completeness of polynomial basis sets such as MTP, PIP, aPPI and NICE (see also Section V A). Any of these bases can be expressed in terms of the $\nu$-correlations via a linear transformation, and vice-versa. Even for fixed maximum polynomial degree, these are enormous representations. Depending on how $\nu$-correlation features are chosen their number might scale as rapidly as $(\eta_{\text{max}}+\nu)^N$, where $\eta_{\text{max}}$ is the number of one-particle features.

There is a class of much lower-dimensional descriptor maps based on the eigenspectra of overlap matrices that lifts the degeneracy for the known examples, although their actual completeness is unknown. The construction of these “spectral representations” proceeds as follows: First, one constructs an artificial overlap matrix based on the positions of atoms within the $i$-centered environment $A_i$:

$$T_{jj'} = f_{\text{cut}}(r_{jj}) f_{\text{cut}}(r_{jj'}) t(r_{jj'})$$

where $t: \mathbb{R} \to \mathbb{R}$. Then, one computes the ordered spectrum $(\tau_k)_{k=1}^N$ of $T$. If $T$ is invariant (or covariant) $(\tau_k)_{k=1}^N$ is an invariant descriptor of $A_i$. Due to eigenvalue crossings, the mapping $A_i \mapsto (\tau_k)_{k=1}^N$ is non-smooth, hence one should project it on a smooth basis, e.g. polynomial,

$$\phi_k := \sum_i (\tau_i)^k.$$  

(80)

The features (or, fingerprints as they are also called) correspond to the moments of the histogram of eigenvalues, and contain precisely the same information.

An alternative way to write $\phi_k$ is

$$\phi_k = \text{Tr} T^k,$$  

(81)

which is not computationally more efficient, but highlights the close connection between $(\phi_k)_{k=1}^N$ and the body-ordered features we discussed in previous sections. From (81) we observe that

$$\phi_1 = Nt(0),$$

$$\phi_2 = \sum_{j_1,j_2} t(r_{j_1j_2})^2 \cdot f_{\text{cut}}(r_{j_1}) f_{\text{cut}}(r_{j_2})$$

$$\phi_3 = \sum_{j_1,j_2,j_3} t(r_{j_1j_2}) t(r_{j_2j_3}) t(r_{j_3j_1}) \cdot \prod_{\alpha=1}^3 f_{\text{cut}}(r_{j_\alpha}),$$

(82)

and so forth. That is, the feature (fingerprint) $\phi_k$ contains the projection of the histogram of $k$-simplices onto a single basis function. In other words, the cutoff function $f_{\text{cut}}$ and the overlap function $t$ play the role of $R_n$ and $P_l$ in (34). Thus, the $\phi_k$ provide invariant high body-order features at relatively low computational cost.

If one takes $t$ to be scalar (as we have done here) then there are at most $N$ invariant features for $N$ neighbors, but $3N-6$ independent coordinates – so that the spectral features must be grossly undercomplete. This source of incompleteness is easily lifted by simply taking multiple overlap matrices with different $t$ functions, or taking $t$ to be matrix-valued, as
done in Ref. \[87\]. However, even with that modification in mind, it is not at all understood whether these features are complete or can be made complete with limited modifications. For example it can be shown that most high-body order features are actually polynomials of low body-order features, which means that they do not contain genuine high correlation information. This can be observed very easily with a seemingly trivial modification to the spectral representation construction. Consider \(N\) particles on the unit circle at positions \(r_{ij}\), as in Fig. 13. In particular we then have only \(N - 1\) independent variables, which means that a scalar \(t\) is in principle sufficient to identify the configuration. However, choosing \(T_{j j'} := \cos \theta_{ij j'}\)

it is straightforward to see that the two overlap matrices \(T\) for the two configurations of Figure 13 have eigenvalues \(\{0, 0.9, 1, 3\}\). That is, this particular choice of spectral descriptor is unable to distinguish them nor any two configurations for different \(\alpha\).

Even for a general atomic environment, \(T_{j j'} = r_{ji} t_{j' i} \cos \theta_{ij j'}\) is the Gram matrix of the environment, which has at most three non-zero eigenvalues and hence the collection \((\phi_k)_{k=1}^N\) contains at most three independent features even though formally, \(\phi_k = \text{Tr} T^k\) has body-order \(k\). For a configuration in which the neighbors lie on a sphere, this case can be written as an overlap matrix by choosing an appropriate, monotonically decreasing \(t(r_{ij'})\), and for the general case with an appropriate (albeit contrived) choice of \(f_{\text{cut}}\) and \(t\). The purpose of these examples is to highlight that, although spectral descriptors offer some attractive features such as their computationally cheap high body-order nature, understanding under which conditions they are complete is subtle and requires a much deeper investigation.

**VII. REPRESENTATIONS, STRUCTURES, PROPERTIES AND INSIGHTS**

A mathematical representation of the structure of an atomic configuration is not only useful as the starting point of supervised-learning algorithms, aimed at predicting its energy and properties. It can also be used, in combination with unsupervised learning schemes, to compare structures in search for repeating atomic patterns.\[119\] to obtain low-dimensional projections that help visualize complex datasets\[111,119\], and more generally to describe the lie of the land in (free)energy landscapes and interpret structure-property relationships in complex systems.\[126\] There is a long-standing tradition of developing domain-specific descriptors to use in the automatic analysis of structural data. For instance, simulations of polypeptides have been interpreted in terms of backbone dihedral angles.\[200\] Discrete secondary-structure categories as well as sophisticated continuous fingerprints of secondary structure and backbone chirality\[201-204\]. Simulations of clusters and condensed-phase systems have often used more general indicators, such as Steinhardt order parameter,\[205\] cubic harmonics,\[206\] radial distribution functions (either directly\[202,210\] or in the form of entropy-inspired fingerprint,\[208\], histograms of coordination numbers,\[211\] that can be seen as precursors of the atom-density correlation representations that we discuss in Section 1V.D. More broadly, general-purpose descriptors that can be understood, more or less transparently, as a special case of the density-correlation features \(\rho^{\phi(\mathbf{r})}\) have been developed and used in unsupervised-learning contexts as much as in the context of regression models. A few examples include the diffraction-based fingerprints of Ziletti et al.\[210\], the local order metric of Martelli et al.\[211\] the spectral representations of Sadeghi et al.\[209\], the Minkowski structure metric of Mickel et al.\[212\] (that closely resembles and anticipates the construction of the moment tensor potentials).

Understanding the way a representation converts the Cartesian coordinates of atoms into features is necessary to make sense of any subsequent analysis, because any explicit or implicit assumption made in the structure-feature map will be reflected in the unsupervised analyses based on those features.\[213\] An example of this is given in Figure 14 that shows the effect of using rotationally variant or invariant features (respectively, \(\langle nlm|\rho_i\rangle\) and \(\langle n_1 n_2|\rho^{\phi(\mathbf{r})}\rangle\)) to analyze a simulation of undercooled iron.\[215\] Atoms are colored according to a two-dimensional projection describing the associated environments, in this case obtained using a kernel principal component analysis\[214\] built on the feature vectors \(\Phi(A_i)\). Using orientation-dependent features makes it possible to distinguish more clearly the presence of multiple grains, and would be useful, for instance, to investigate the texture of the nanocrystalline sample, much like one would do with an electron backscattering diffraction analysis. Using invariant features highlights that all nanocrystals have the same structure, and makes it possible to recognize the disordered environments at the grain boundaries. Conversely, investigating the effect of different representation choices on the unsupervised analysis of a well-understood system can be useful to better appreciate the relation between structures and features.

In this Section we summarize recent developments, and identify clear insights, related to the use of representations to determine the similarity between structures, to perform clustering and dimensionality reductions analyses, and to build models that go beyond the injective structure-property map that we have used this far.
A. Features, distances, kernels

Before delving into the use of structural representations to visualize and classify atomic configurations, let us recall the link between feature vectors \( \Phi(A_i) \), that are associated to structures or environments, and distances or kernels, that express the relationship between two of these entities. For example, given a feature vector \( \Phi \), it is possible to define a distance using e.g. a Euclidean metric, \( D(A_i, A'_i) = \| \Phi(A_i) - \Phi(A'_i) \|^2 \), and use as a kernel the scalar product \( K(A_i, A'_i) = \Phi(A_i) \cdot \Phi(A'_i) \), or a non-linear
function, e.g., an exponential of a squared distance $K(A, A') = \exp(-\gamma D(A, A')^2)$.

The opposite is also true: for a given set of configurations $M$, and any (negative definite) distance or (positive definite) kernel $K$, it is possible to construct a set of features that generate the kernel by taking their scalar product—a practical implementation of the concept of reproducing kernel Hilbert space that underlies kernel methods. One only needs to construct the kernel matrix $K_{ij} = K(M_i, M_j)$, and find its eigenvalues and eigenvectors $Ku_i^{(j)} = \lambda_j u_i^{(j)}$. It is easy to see that the scalar product between the reproducing features

$$\phi^K_j(A) = \sum_{i \in M} K(A, M_i)u_i^{(j)}/\sqrt{\lambda_j} \quad (83)$$

computed for two members of the reference dataset yields exactly the value of the kernel function between the two configurations.

### B. Measuring structural similarity

Most unsupervised learning algorithms rely on the definition of a metric to tell apart structures depending on their similarity. At the most basic level, a metric that is capable of identifying identical structures is extremely useful in all the applications that aim at automating the search of materials or molecules with desirable properties. This is not an entirely trivial task: in molecular searches, a mismatch in the simple ordering of atomic indices can lead to the failure of metrics based on the alignment of conformers, such as the root mean square distance (RMSD), and the exact calculation of a permutation invariant version would involve combinatorially increasing computational effort.

In the case of condensed phases, one needs to deal with the problem that the same periodic structure can be described by different choices of unit cell size and orientation. The requirements for a metric to compare atomic structures are similar to those discussed in Section III and have been discussed in great detail in Ref. 86: a good metric needs to be invariant to rotations, translations, and permutations, and still be capable of telling distinct structures apart. The comparison between the resolving power of different metrics has been often determined using distance-distance correlation maps, such as those shown in Figure 15, that compare the distance between pairs of structures in a reference dataset, as computed by two metrics. In the most extreme case, one observes pairs of structures that are identical based on a metric, and distinct based on another—indicating the presence of a manifold of degenerate structures that are distinct, but cannot be told apart by one of the distances.

An important aspect when defining a metric for structural comparison is the fact one is often interested in measuring the dissimilarity between entire structures, $D(A, A')$. Most of the representations we discussed this far are designed to compare atom-centered environments, and therefore yield $D(A_i, A'_i)$. As a practical example, we define $D$ as the Euclidean distance between the representations,

$$D^2(A_i, A'_i) \equiv (|A'_i| - |A_i|)^2 \quad (84)$$

where we use the abstract notation $|A|$ rather than $\phi(A_i)$ to highlight the connection with the definition of the global representation $|A; \rho^{(2)}|$ as the sum of environmental $|A; \rho_i|$ introduced in Section IV C. Different ways of combining atom-centered representa-
tions to obtain a structure-level comparison are discussed and benchmarked in Ref. [101], using a construction based on the definition of global kernels. Here we present the same strategies, but express them directly in terms of distances. The two formulations are equivalent when using the kernel-induced distance.

The simplest global distance can be defined as a mean over all environment pairs,

$$D^2(A, A') = \frac{1}{N_A N_{A'}} \sum_{i \in A, i' \in A'} D^2(A_i, A'_i) =$$

$$\frac{1}{N_A N_{A'}} \sum_{i \in A, i' \in A'} (|A'_i| - |A_i|)^2 =$$

$$\left[ \sum_{i \in A} \frac{|A_i|}{N_A} - \sum_{i' \in A'} \frac{|A'_i|}{N_{A'}} \right]^2 \equiv (|A| - |A'|)^2 \quad (85)$$

where the last line shows that the average environment distance $D^2(A, A')$ can be computed by taking the Euclidean distance between the mean of the environment’s features in the two structures. This construction is very natural, and consistent with an additive decomposition of properties in a regression model, but potentially lacks resolving power: two structures with very different environments could end up having a similar value of the average feature vector.

An alternative way to determine a global metric involves finding the best match between the environments of the two structures, defining

$$\hat{D}^2(A, A') = \arg\min_{P \in \mathbb{U}^{N_A \times N_{A'}}} \sum_{i \in A, i' \in A'} D^2(A_i, A'_i)P_{i'i'} \quad (86)$$

where $\mathbb{U}^{N_A \times N_{A'}}$ is the set of $N_A \times N_{A'}$ doubly-stochastic matrices, i.e. matrices with positive entries such that sums of rows and columns all equal 1/$N_A$ and 1/$N_{A'}$ respectively. When $N_A = N_{A'}$, the optimal $P$ contains only zeros and 1/$N_A$, and the problem can be construed as a linear assignment problem, and solved in $O(N_A^3)$ time using the Hungarian algorithm [226]. Much like the case of the use of sorted interatomic distances as a structural representation (Section IIIB), the process of matching entries in the environment distance matrix introduces discontinuities in the derivatives of the distance metric. One can solve this problem, obtaining at the same time a scheme with a cost that scales as $O(N_A^3)$ and that can be applied to the comparison of structures of different sizes, by introducing an entropy regularization in Eq. $(87)$

$$\hat{D}^2(A, A')^2 = \arg\min_{P \in \mathbb{U}^{N_A \times N_{A'}}} \sum_{i \in A, i' \in A'} P_{i'i'}(D^2(A_i, A'_i)+\gamma \ln P_{i'i'}), \quad (87)$$

controlled by the magnitude of the parameter $\gamma$. This approach was introduced in Ref. [226] for the general problem of solving optimal transport problems and of evaluating the Wasserstein distance between probability distributions, and was first applied in Ref. [101] to atomistic problems in terms of regularized entropy match (REMatch) kernels. By introducing a non-additive combination of the environments, REMatch kernels offer an increased resolving power compared to the plain average distance $(85)$, as demonstrated in Figure 16. The figure also shows that Eq. $(87)$ interpolates between the average and the best-match metrics, to which it tends respectively for $\gamma \to \infty$ and $\gamma \to 0$.

![Figure 17](image-url)
C. Representations for unsupervised learning

As stressed in the introduction of this Section, in performing cluster analysis or dimensionality reduction, the choice of featurization is not a neutral one, but introduces bias that will be visible in the end result of the analysis\cite{comment1}. While sometimes this bias is desirable, such as in Fig. 14 in which a judicious choice of features makes it possible to emphasize, or ignore, the orientation of grains in a polycrystalline sample, one should resist the temptation to fine-tune parameters that do not have an obvious meaning to obtain a result that reflects a preconceived interpretation of the data. The top row of Fig. 17 shows how different choices of the hyperparameters of the SOAP powerspectrum (cutoff radius $r_{\text{cut}}$, density smearing $\sigma_p$, and the types of atoms that are used as environment centers) change unpredictably the distribution of the points on the 2D map obtained by principal components analysis of the features. In the first panel, in particular, one can recognize a degree of correlation between the position of the points, and intuitive structural and energetic properties, such as the number of H-bonds, and the lattice energy. The correlation is however far from perfect, and with other reasonable choices of hyperparameters it disappears almost completely.

One possible approach to make unsupervised models less dependent on the details of the underlying featurization is to combine them with an element of supervised learning. This includes, for instance, combining or contrasting density-based clustering with (kernel) support vector machines classification\cite{comment2}. Even more explicitly, one can combine a variance-maximization scheme analogous to PCA with the regression of a target property, as in principal covariates regression (PCovR)\cite{comment3}. In PCovR one minimizes a loss built as a mixture of a PCA and a linear regression of a target property, as in principal covariates regression of a target property, one minimizes

$$\ell = \sum_i \alpha \| \Phi - \Phi P_{\Phi T} P_{T \Phi} \|^2 + (1 - \alpha) \| Y - \Phi P_{\Phi T} P_{T Y} \|^2.$$  \hspace{1cm} (88)

The matrix $P_{\Phi T}$ projects from the feature space to a low-dimensional latent space, $P_{T \Phi}$ reconstructs an approximation of the full-dimensional feature vector based on its latent-space embedding, and $P_{T Y}$ regresses the property matrix $Y$ using the latent-space coordinates as inputs. By explicitly looking for a latent-space projection that allows to regress linearly a target property, one forces the dimensionality reduction to identify a subspace of the chosen features that correlates well with one or more quantities of interest. The lower row of Fig. 17 is obtained using a recent kernel extension of this method (KPCovR)\cite{comment4}, attempting simultaneously to maximise the spread of data and the kernel regression of the lattice energy, giving equal weight to the two components ($\alpha = 0.5$). Not only points on the resulting map correlate very well with the target: one observes that also structural parameters such as the H-bond counts are now clearly separated between different regions, and the appearance of further groups of well-clustered structures that correspond to similar isomers of azaphenacene\cite{comment5}. What is perhaps more important, introducing an explicit supervised learning target leads to maps that are more consistent across different choices of hyperparameters. Thus, (K)PCovR reduces the arbitrariness of the description, and mitigates the risk of implicitly introducing an unknown bias.

![FIG. 18.](image)

**FIG. 18.** (a) A sketch-map\cite{comment6} representation of the QM7 molecular dataset\cite{comment7} based on a SOAP kernel distance. Each point corresponds to one molecule. Left: points are colored according to the atomization energy; right: points are colored according to composition. Adapted from Ref.\cite{comment8}  (b) Principal component analysis (PCA) on the multiple layers of a deep NN learning simultaneously the 14 properties of the QM7 molecular dataset, using Coulomb matrix features as representation. Each point (molecule) is colored according to the rule: $E$ and HOMO (highest occupied molecular orbital energy) large $\rightarrow$ red; $E$ large and HOMO small $\rightarrow$ blue; $E$ small and HOMO large $\rightarrow$ green; $E$ and HOMO small $\rightarrow$ black. The NN extracts, layer after layer, a representation of the chemical space that better captures the multiple properties of the molecule. Adapted from Ref.\cite{comment9}.

D. Analyzing representations and datasets

The unsupervised or semi-supervised analysis of a dataset helps building an intuitive understanding of complicated structure-property relations for a material or a class of materials. Given the “black box” nature of many machine-learning models (and the fact that even the rigorously-defined density correlation features we focus on in this review have a high-dimensional nature and non-trivial relationship to the actual atomic structure) low-dimensional projections of the feature space can also be useful to gain a bet-
FIG. 19. 2D maps obtained applying the t-SNE di-

nymensionality reduction algorithm [229] to three different 
molecular datasets – the systematic enumeration of 9-
non-H-atoms molecules in QM9 [230], the conformers of aminoacids in the Berlin aminoacid dataset AA [231] and the large molecules extracted from the Cambridge struc-
tural dataset of the OE dataset [232]. Panel (a) uses a 
Coulomb matrix representation, panel (b) uses the MBTR 
features [29]. Reproduced from Ref. 96.

ter understanding of the structure of feature space. For example, Fig. 18a tells us less about the QM7 
dataset [222] (that contains small organic molecules con-
taining C, H, N, O, S, Cl) than about the SOAP fea-
tures that underlie the representation: the unsupervised analysis shows that the chemical composition is the most clear-cut differentiating characteristic when looking at this dataset with SOAP lenses. Fig. 18b visualizes the same QM7 data using a different repre-
sentation, based on the Coulomb matrix, and shows 
how successive layers of a neural network transform 
these features into non-linear combinations that cor-
relate very well with the target properties. Thus, this 
visualization helps understand how a highly-nonlinear 
function transforms a description of the system into 
combinations that can be more easily used for regres-
sion, and diagnose the inner workings of the deep neu-
network.

A final “introspective” application of this kind of 
analysis involves examining the structure of a dataset – not as a way to learn about the atomistic configu-
rations it contains, but about its makeup, or the relation-
ship with other datasets. An example is given in 
Fig. 19 showing the comparison between the chemical 
space covered by three databases of organic molecules, 
with QM9 and AA being mostly disjoint, and the more 
diverse OE molecules encompassing both the other 
sets. Other examples of this kind of analysis are dis-
cussed in Section IX.

E. Indirect structure-property relationships

The one-to-one mapping between an atomic struc-
ture and its representation is one of the key require-
ments to achieve accurate “surrogate quantum mod-
elts” of atomic-scale properties. However, it can also be 
a limitation whenever one wants to describe properties 
that are not strictly associated with the specific config-
uration at hand. For example, consider the databases 
of molecular properties (e.g. the QM9 dataset [230]) that 
have been extensively used as a benchmark, and have 
been a powerful driving force behind the development 
of the representations we describe here. The typical 
benchmark involves taking a structure whose geome-
try has been optimized at the DFT level and use it to 
predict the DFT energy – and exercise that is mani-
festly of little practical utility. What one would like 
to do, instead, is not having to specify the optimized 
geometry, and use a non-optimized structure to pre-
dict the properties of the nearest local configurational 
optimum. As shown in Fig. 20a, this is conceptually 
problematic, because we are now trying to achieve a 
many-to-one mapping. A possible solution is to map 
each distorted geometry to an idealized one, or to use 
a lower level of theory to determine an unique struc-
ture $A_0$. Thus, the many-to-one mapping is realized 
by the local optimization procedure, and the corre-
sponding representation $|A_0\rangle$ can be used to uniquely 
identify the entire basin of attraction of the local min-
imum. Only for the training structures, this geometry 
is optimized further at a higher level of theory, obtaining 
the structure $A_0$ for which properties are meant 
to be computed. When the model is fitted, the rela-
tionship between $A_0$ and its high-quality counterpart
is learned implicitly. This kind of “indirect” model has been used, for instance, in Ref. [28] where structures optimized at the semiempirical PM7 level were used to predict CCSD energetics computed for a DFT-optimized version of the same compound. While the error was almost twice as large as a model using directly $|A_0\rangle$ as input, chemical accuracy could be reached when discarding from the training set structures for which the DFT-optimized structure was too different from the PM7-optimized geometry.

A similar conceptual problem arises when one wants to build models for properties that are associated with a thermodynamic state rather than a precise structure, such as a melting point or solubility of a material. Typically, and particularly if the ensemble consists in relatively small fluctuations around equilibrium, one might take a representative structure (e.g. the minimum energy configuration) and use the corresponding representation $|A_0\rangle$ as a proxy of the thermodynamic state (Fig. 20b). Alternatively, consider the case in which the target property can be estimated as an ensemble average, e.g. over a Boltzmann distribution at inverse temperature $\beta$, $P(A) = e^{-\beta E(A)}/Z$,

$$\langle F\rangle_\beta \equiv \frac{1}{Z} \int dA e^{-\beta E(A)} F(A)$$ (89)

where $Z = \int dA e^{-\beta E(A)}$ is the canonical partition function. One can exploit the linear nature of the representation and define an “ensemble ket”

$$|A;\beta\rangle \equiv \frac{\int dA e^{-\beta E(A)} |A\rangle}{\int dA e^{-\beta E(A)}}.$$ (90)

With this definition, one could use a linear model for $F(A)$ with weights $\langle q|F\rangle$ and see that

$$\langle F\rangle_\beta \approx \sum_q \langle F|q\rangle \langle q|A;\beta\rangle,$$ (91)

which is convenient because it allows using properties of configurations and of ensembles on the same footing – and possibly combining them in a single training exercise. Note also that the same approach could be applied in a kernel setting, computing the ensemble average of the reproducing kernel Hilbert space vector associated with the structures.

**VIII. EFFICIENCY AND EFFECTIVENESS**

We have discussed in Section III how most of the existing choices of representations share profound similarities, and shown, in Section IV that many alternative schemes can be formally related to each other by means of a linear transformation, smoothening or a limit operation. However, this is not to say that in practical applications they are entirely equivalent. The computational cost of evaluating them, and their performance in classifying structures, and in regressing their properties, is determined by the choice of basis functions. Even for formally equivalent representations, the condition number of the linear transformation between them and their corresponding bases have significant impact on the numerical behavior of the computed coefficients and the quantities derived from these coefficients.

**A. Comparison of features**

A preliminary question when comparing alternative choices of features for the description of atomic structures and/or environments is that of establishing an objective way of assessing their relative merits. The performance when used in the regression of useful atomic-scale properties is an obvious criterion, but such a comparison is intimately intertwined with the target property and the regression algorithm.

Very recent efforts have attempted to characterize different representations in terms of their information content – for instance through the eigenvalue spectrum of the covariance or kernel matrix associated with a dataset, the decrease in accuracy when reducing the number of features or the sensitivity of the features to atomic displacements.

This latter approach can be realized by directly comparing the separation in feature space against finite displacements of the atoms or through an analysis of the Jacobian $J_{jk} = \partial \langle k|A_i\rangle / \partial x_j$. The sensitivity of the features to small changes of the atomic positions indicates their usability and performance in regression of classification tasks. Onat et al. [172] analysed the effect of random perturbations in crystalline environments, finding that, for features based on atomic density correlations, displacements of atoms in the environment usually cause a linear response. One notable deviation from this trend are perturbations along some high-symmetry directions in atomic environments carved from perfect crystals, where the response to displacements is second-order, implying that the representations cannot capture these types of deformations (Fig. 21). However, as discussed in reference [172] the types of symmetric deformations applied in the study correspond to reflection operations. Due to the body-correlation order considered, features are invariant to mirror symmetry, and so the observed loss of sensitivity is not unexpected. Analyzing the response of the features to perturbations in terms of the Jacobian, as in Ref. [176] has the advantage of characterizing fully the sensitivity at a given point. The Jacobian should have six zero principal values, corresponding to rigid rotations and translations of the environment. Additional zeros could be associated with the presence of a continuous manifold of degenerate structures. However, as demonstrated by the finite-displacement deformation...
FIG. 21. (a) A $4 \times 4 \times 4$ Si cell, that is taken as the reference structure for radial and tangential perturbation of the neighbors of the central atom. (b,c) Norm of difference of atomic descriptors on atom $i$ as a neighboring atom $j$ is perturbed from its reference position. (b) A radial perturbation yields a linear change in the features; (c) a tangential perturbations in a high symmetry direction for the first shell yields a quadratic change in the features. Adapted from Ref. 172

in Fig. 21), high-symmetry configurations can result in directions with zero gradient that have no adverse effect on the accuracy of a model built on the density correlation features.

Another comparison between different bases is to analyse the landscape defined by the similarity or distance between environments, $D(A_i, A'_i)$ where the environment $A_i$ is kept fixed. The distance between the atom-centered environments $A_i$ and $A'_i$ can be defined as the Euclidean distance between feature vectors, Eq. (54). It is clear that the field defined by the Cartesian coordinates of $A'_i$ will have a global minimum manifold where the field is exactly zero, corresponding to equivalent environments $A_i$ and $A'_i$ that are related by symmetry operations. Whether there are other manifolds at exactly $D(A_i, A'_i) = 0$, corresponding to the same features resulting from symmetrically nonequivalent environments is related to the question of completeness (Section VI B). In practical applications, the shape of the global minimum manifold has also implications for the numerical evaluation. In particular, one could examine how different $A_i$ and $A'_i$ may be for $D(A_i, A'_i) < \epsilon$ where $\epsilon$ is a small number. Using a random search approach, the numerical sensitivity of the feature landscape has been analysed in Ref. 27. Reference structures $A_i$ were perturbed and then reconstructed by minimising the distance $D(A_i, A'_i)$, and the optimised structures compared to the reference ones. For small numbers of neighbors in the reference environment, all the examined representations performed similarly well, but only SOAP was capable of accurately reconstructing the reference environments of more than 12 neighbors. As we have seen in earlier sections, this differences can be attributed to the choice of basis functions other representations use, although it should be noted that SOAP distances and similarities converge in the limit of a complete basis, therefore the actual form of the basis might affect the convergence, and the computational cost of the representation, but does not impact its resolving power.

A more explicit comparison between pairs of representations can be obtained by evaluating the error

FIG. 22. The table shows the error (GFRE, Eq. (92)) and the distortion (GFRD, see Ref. 236) computed between density correlation features of different order, as well as the NICE features$^{123}$ up to $\nu = 4$, for a data set of random CH₄ configurations$^{237}$ and hypothetical carbon allotropes predicted by AIRSS$^{218,238}$. Adapted from Ref. 236

33
one incurs when using a set of features, arranged in a feature matrix $\Phi$ in which each row corresponds to a sample in a reference dataset, to linearly reconstruct a second featurization of the same structures or environments $\Phi'$, defining a global feature space reconstruction error

$$\text{GFRE}(\Phi, \Phi') = \min_{\textbf{P}} \sqrt{\|\Phi'_{\text{test}} - \Phi_{\text{test}} \textbf{P}\|^2 / n_{\text{test}}}.$$  

(92)

$\textbf{P}$ is a linear regression weight matrix obtained on a training subset of the rows of $\Phi$ and $\Phi'$, and both sets of features are assumed to be standardised.\textsuperscript{22}

The GFRE can be extended to also incorporate non-linearity in the mapping, either by a locally-linear approach, or by using a kernelized version, and is not symmetric. $\text{GFRE}(\Phi, \Phi') \ll \text{GFRE}(\Phi', \Phi)$ indicates that the featurization underlying $\Phi$ is more informative than that used to build $\Phi'$, and vice versa. $\text{GFRE}(\Phi, \Phi') \approx \text{GFRE}(\Phi', \Phi) \approx 0$ implies that the two featurizations contain similar information, but are not necessarily equivalent. One could emphasize more some structural correlations than others: imagine for instance multiplying by a large constant the entries of one column. This kind of distortions, which can have a substantial impact on the performance of models built on $\Phi$ or $\Phi'$, can be measured by defining a global feature space distortion (GFRD)

$$\text{GFRD}(\Phi, \Phi') = \min_{\textbf{Q} \in \mathbb{U}} \sqrt{\|\Phi_{\text{test}} \textbf{P} - \Phi_{\text{test}} \textbf{Q}\|^2 / n_{\text{test}}}.$$  

(93)

$\textbf{P}$ is the same projection matrix that enters the definition of the GFRE (so that $\Phi \textbf{P} \approx \Phi'$), and $\textbf{Q}$ is the unitary transformation that best aligns $\Phi$ and the best linear approximation of $\Phi'$.

If both GFRE and GFRD are zero, then the linearly independent components of $\Phi$ and $\Phi'$ are related by a unitary transformation, which implies that distances and scalar products between feature vectors are equal between the two representations being compared. Figure 22 demonstrates the use of these measures to compare $|\rho^{\nu}_{\nu'}|$ features of different body order. The asymmetry is very clear, with higher-order features containing more information than their lower-order counterparts. Note that – in view of the linear nature of the mapping – this is not entirely obvious: formally, $\nu = 1$ features are not linearly dependent on higher-$\nu$ features, and so these observations reflect the specific nature of the atom-density field whose correlations are being represented, and the nature of the structures in the benchmark datasets. The figure also includes invariants built with the N-body iterative contraction of equivariants (NICE) framework, that are designed to capture most of the information up to high body orders. The truncation of the expansion, that is necessary to keep the evaluation of $\nu = 4$ order features affordable, leads to a small residual GFRE when reconstructing $\nu = 3$ features. The GFRD is rather large between all featurizations, reflecting the fact that – even though higher-order features contain sufficient information to describe lower-order correlations – they weight the information differently, which is why it is often beneficial to treat different orders of correlation separately in the construction of inter-atomic potentials.\textsuperscript{[hypertext]}

B. Feature selection

Numerical feature vectors $\Phi(A_i)$ are the result of a basis set expansion of atomic environments, which are, for practical purposes, truncated. A concrete discretization of the symmetrized expansion of atomic environments, which are, for practical purposes, truncated. A concrete discretization of the symmetrized $\nu$-correlations is obtained by choosing a finite subset from the set of all possible features,

$$\mathbf{q} \subset \mathbf{q}_{\text{total}} := \{(n_\alpha l_\alpha)_{\alpha=1}^\nu, \nu \in \mathbb{N}\}.  \quad (94)$$

(A choice of $(n_\alpha l_\alpha)_{\alpha}$ naturally induces a choice of $m_\alpha$ and symmetrized features.) The role of the discretization $\mathbf{q}$ is very different for linear and nonlinear models and therefore warrants a brief comment: For nonlinear models we typically only require geometric completeness (see Section VI), which means that the feature set can be chosen minimal but in a way so that all possible configurations, or at least all configurations of interest (e.g. from a training set) can be distinguished in a stable and smooth way. While it is an open problem to characterize precisely what this entails, we generally expect that relatively small feature sets on the order of hundreds for single-species scenarios could be sufficient.

On the other hand, converging a linear model requires eventually letting the discretisation $\mathbf{q}$ converge to the full feature set $\mathbf{q}_{\text{total}}$, which in practice leads to a much larger set $\mathbf{q}$ and in particular higher correlation-orders $\nu$ to achieve a desired accuracy, e.g. on the order $\mathcal{O}(10^000)$ features for single-species models. The additional cost in training and evaluating the features is of course off-set by the fact there is no additional cost in evaluating the nonlinear models. Due to the large feature sets the selection of effective subset of $\mathbf{q}$ may be even more important in the linear setting. In particular it will be crucial to a priori choose sparse subsets of $\mathbf{q}_{\text{total}}$ rather than tensor-product sets due to the combinatorial explosion of the number of features with high $\nu$ (curse of dimensionality). For example, a total-degree $D$ discretisation,

$$\mathbf{q}(\nu_{\text{max}}, D) = \{(n_\alpha l_\alpha)_{\alpha=1}^\nu : \nu \leq \nu_{\text{max}}, \quad \sum_\alpha n_\alpha + l_\alpha \leq D\} \quad (95)$$

was used by Bachmayr et al.,\textsuperscript{102} while closely related a priori sparsifications were used by Braams and Bowman,\textsuperscript{103} Shapeev\textsuperscript{113} in all cases demonstrating accuracy/performance competitive with or outperforming nonlinear models.
### Data-driven selections

When using high-body order features, some of the components can be related by non-trivial linear dependencies, that can be enumerated numerically. The construction does not ensure that there is no other linear dependence that is specific to a given dataset, meaning that feature vectors could potentially be compressed even further without noticeable deterioration in the quality of the representation. The benefits of the compression are clear: if only a few components need to be evaluated, significant efficiency gains may be realised both in computational effort and storage requirements.

Thus, the objective of feature selection or truncation is to find a subset of features that retain the information content of the original, untruncated representation. This is to be contrasted with dimensionality reduction techniques, that apply a linear transformation on the full feature vector to generate a lower dimensional representation. These only reduce the computational cost of operations that are applied on the reduced feature vectors: the whole feature vector must be evaluated first, before being able to determine its projections.

An example of a feature selection strategy is the farthest point sampling (FPS) technique. One chooses an initial column \( \varphi_{c_0} \) (indexed by \( c_0 \)) of the feature matrix, and then iterate selecting the columns that maximize the Hausdorff distance to the previously selected columns

\[
c_{m+1} = \arg\max_j \left\{ \min_{1 \leq c_m} \| \varphi_j - \varphi_c \| \right\},
\]

instead, generates a low-rank approximation of the feature matrix \( \Phi \), in the form

\[
\Phi \approx \text{CUR}.
\]

Unlike singular value decomposition, CUR uses the actual columns \( (C) \) and rows \( (R) \) of \( \Phi \). To make the selection, a score is associated with each feature \( c \)

\[
\pi_c = \frac{1}{k} \sum_{i=1}^{k} \langle v_i \rangle^2.
\]

based on the right singular vectors \( v_i \) of the singular value decomposition of \( \Phi \). \( k \) is the approximate rank of \( \Phi \). Features may be selected in a probabilistic procedure or simply based on their score. Imbalzano et al. argued that the scores associated with feature vector components which are linearly dependent are close, therefore the selection can easily result in a redundant set. Instead, in Ref. a greedy algorithm based on the CUR decomposition was suggested, where features were selected iteratively. The feature with the highest score is selected, and the columns of \( \Phi \) are orthogonalised relative to the column corresponding to the selected feature. The scores are updated in each step, so the linear dependence of already selected features are removed. This iterative scheme often performs better when using a very small value of \( k \) in constructing the leverage scores.

Fig. 23 shows that a data-driven selection of the most relevant/diverse features makes it possible to achieve models with an accuracy that approaches that of the full model while reducing the number of components by a factor of about 3 (for linear regression) or 10 (for KRR). Particularly for intermediate sizes of the selection, the improvement in accuracy with respect to a random selection can be dramatic. Both FPS and CUR methods can be improved further by incorporating information on the properties associated

---

**Table 1:**

| \( n_{\text{features}} \) | LR on Feature Selections | KRR on Feature Selections |
|--------------------------|--------------------------|---------------------------|
| 10,000                   | LR                      | KRR                       |
| 20,000                   | LR                      | KRR                       |
| 50,000                   | LR                      | KRR                       |
| 100,000                  | LR                      | KRR                       |

**Fig. 23:** Test-set error in the prediction of a linear regression (left) and kernel ridge regression (right) models of the nuclear chemical shieldings of atoms in a set of molecular materials, as a function of the number of features used in the model. Features are selected using the FPS and CUR methods (full lines) from a set of 2520 SOAP features. The shaded areas indicate the range of values obtained varying the mixing parameter \( \alpha \) in a principal covariate-augmented version of the methods. Adapted from Ref. 230.
with the structures as in Eq. (88). Including a supervised component by setting \( \alpha < 1 \) in the feature selection usually leads to more performing models, as shown in Fig. 23. Feature selection methods can be applied to any flavor of density correlation features. Imbalzano et al. used a reference data set on liquid water and a large set of systematically generated ACSFs. Evaluating the RMSE of the predicted energies and forces revealed that automatic selections performed by a CUR or FPS approach may achieve similar performance to features selected based on chemical intuition and heuristics, while keeping approximately the selection size. A dramatic reduction in numbers of features is also possible for the SOAP power spectrum, and a data-driven selection of the most important components has quietly become commonplace to accelerate SOAP-based ML models. A more systematic investigation of the effectiveness of feature selection for many commonly used atomic descriptors has been recently reported by Onat et al., who analysed how accurately the original feature vector can be reconstructed from the reduced set, as well as the performance on a practical regression task.

C. Feature optimization

As discussed in Section V C, non-linear models optimize the description of their inputs by generating new features that are best correlated with the target property, or that are adapted to the structure of the dataset. For instance, taking products of 2-body features results in an effective representation that incorporates some, but not all, features of body order 3, 4... In some cases it is possible to find an expression for the effective representation associated with a kernel model while other cases (most notably deep neural network models) put less focus on the interpretability of the intermediate features, and act largely as data-driven 'black boxes'. Alternatively, feature optimization can be performed explicitly on the representations presented in Section IV E. Such optimization could take the form of the choice of basis functions. In the Behler-Parrinello framework, it is customary to select a small number of atom-centered symmetry functions based on experience and heuristic. An optimization of the hyperparameters by gradient descent has also been proposed to obtain more accurate models based on atom-centered symmetry functions.

When considering systematically-convergent implementations of density-correlation features, the optimization of the basis set is less crucial, although one may want to reduce the size of the basis for the sake of computational efficiency, as discussed in Section VIIB. That is not to say that the details of the practical implementation of the features does not change the behavior of a model built upon them. Optimizing hyperparameters such as cutoff radius, density smearing, basis set cutoff affect how naturally the features correlate with the target property, which is one of the factors determining how quickly a regression model becomes capable of performing accurate prediction. For example, the smearing of the atom density, or the truncation of the basis set, should reflect the natural scale over which the target properties vary. Similarly, the size of the local environment determined by \( r_{\text{cut}} \) relates to the typical decay length of interactions, as mentioned in Section VII C, but it also changes the effective dimensionality of feature space, which affects the accuracy of the model in a non-trivial way. Consider the learning curves shown in Fig. 24 that report on the prediction accuracy, as a function of the train set size, for a kernel ridge regression model of molecular atomization energies, based on SOAP features that differ by the value of \( r_{\text{cut}} \). A very large cutoff \( r_{\text{cut}} = 5\text{Å} \) does not yield the best performance, despite providing information on a wider range of dis-
In order to compare results to Fig. 24, they ordinates must be divided by 18. Adapted from Ref. 97.

Optimization of a radial scaling function has become commonplace, and most recent applications based on the SOAP power spectrum rely on it to achieve consistently optimal performance in both the data-poor and data-rich regime.

Rather than optimizing the correlations between geometric features and the target properties, one can attempt to build features that incorporate a notion of chemical similarity between different elements. The idea was introduced in terms of an alchemical similarity kernel in Ref. 101 that is also a core component of the FCHL framework but has been implemented in different forms in the context of atom-centered symmetry function and of generic atom-density...
is beneficial in the data-poor regime, as it allows the
dial correlations. A low-dimensional alchemical space
similar data/complexity interplay as observed for ra-
rent degrees of compression, one sees that there is a
structures. By comparing learning curves with differ-

\[ \langle \alpha | = \sum_a \langle \alpha | a \rangle \langle a | , \]

where the coefficients \( \langle \alpha | a \rangle \) enact the projection be-
 tween the elemental and the “alchemical” basis. The
 reduction in the dimensionality of the feature space
can be substantial: for powerspectrum \( (\nu = 2) \) fea-
tures, the number of components scales quadratically
with the number of species, and so even just halv-
ing the dimension of the chemical space reduces the
number of powerspectrum features by 75%:

\[ \langle \alpha_1 n_1; \alpha_2 n_2; l | \rho^{\nu}_{l,2} \rangle = \sum_{a_1, a_2} \langle \alpha_1 | a_1 \rangle \langle \alpha_2 | a_2 \rangle \langle a_1 n_1; a_2 n_2; l | \rho^{\nu}_{l,2} \rangle . \]

Figure 27 demonstrates how reducing the dimen-
sionality of chemical space helps achieving a transfer-
able, accurate model with a small number of training
structures. By comparing learning curves with differ-
ent degrees of compression, one sees that there is a
similar data/complexity interplay as observed for ra-
dial correlations. A low-dimensional alchemical space
is beneficial in the data-poor regime, as it allows the

D. Efficient implementation

Despite encoding similar information content, the
differences in formulation of competing structural rep-
resentations may lead to large variations in imple-
mentation and performance. A first fundamental di-
vide is between evaluation of features by summing
over clusters of \( \nu \) neighbors and computing \( \nu \) ten-
sor product of atomic densities (see Section IV.F).
Consider the case of evaluating “SOAP-like” ACSF
by cluster sum (cost \( n_{\text{max}}^2 l_{\text{max}}^2 n_{\text{neigh}}^2 \)) and by den-
sity expansion (cost \( n_{\text{neigh}} n_{\text{max}}^2 l_{\text{max}}^2 \)) for the density,
and \( n_{\text{max}}^2 l_{\text{max}}^2 \) for the SOAP evaluation). Despite the adverse scaling of ACSF computed as a sum over clusters of neighbors, these representations can be implemented efficiently \(^{255,256,257}\) by relying on a careful selection of the features (discussed in Section VII D), reuse of parts of the computations, parallelism and GPU acceleration \(^{255,256}\). In fact, when computing a linear model which is explicitly equivalent to a \((n+1)\)-body order potential, the low-order terms can be more efficiently evaluated as a sum over neighbors \(^{107,230}\).

In line with the general focus of this review, we concentrate in particular on the efficient implementation of atom-density representations. As we shall see, roughly the same considerations apply to both those representations that are built on a smooth atom density \(^{208,105,123}\) that generalize the construction of the SOAP powerspectrum and bispectrum \(^{29}\) and those that are built on a \(\delta\)-like density, such as ACI \(^{110,255,256}\) and MTP \(^{111,259}\). Indeed, both families of representations rely on three steps: (i) expansion of the local atom density on a suitable basis, e.g. Eq. (22), (ii) computation of \(\nu\) tensor products of the expansion, and then (iii) contraction over the correlations to obtain equivariant features. While these three steps have been implemented in different ways, their efficient implementation relies on similar considerations.

**Atomic density expansion** Equation (18) provides the blueprints for a broad class of \((n+1)\)-body atom-density representation. Practical implementations differ by the type of localized function used to construct the local atom density (see Eq. (14)), and by the radial and angular basis used for its expansion. As discussed in Section IV E, spherical harmonics are a natural angular basis, but other choices are possible. For instance, the MTP representation projects the atomic density onto a tensor product of direction vectors leading to the covariant moment tensor \(^{111,212}\)

\[
\mathbf{M}_n^{\otimes \nu}(\rho_i) = \sum_{j \in \mathbb{A}} P_{a,\nu}(r_{ji}) \mathbf{\hat{r}}_{ji} \otimes \mathbf{\hat{r}}_{ji} \ldots \otimes \mathbf{\hat{r}}_{ji}
\]

where \( P_{a,\nu}(r_{ji}) \) is a radial function. Invariant components can be obtained by combining and contracting products of the elements of these tensors. The tensor product basis is directly related to spherical harmonics, as shown in appendix B.2 of Ref. 107. The performance of the MTP representation \(^{223}\), which relies on an efficient recursive evaluation of the basis function \(^{113}\), is a testament to the effectiveness of this basis choice.

As shown in Section I V E, the choice of an angular basis of spherical harmonics simplifies greatly the evaluation of Eq. (18), that can be written in terms of contractions of density coefficients \(\langle nlm|\rho_i \rangle\) (cf. Eq. (22)). If the environment-centered density is written in terms of a sum of density functions \(g(x-r_{ji}) = (x-r_{ji}|g)\), peaked at the neighbors positions, the expansion coefficients can be written as the accumulation

\[
\langle nlm|\rho_i \rangle = \sum_{j \in \mathbb{A}} f_{\text{cut}}(r_{ji}) \langle nlm; r_{ji}|g \rangle \tag{102}
\]

of terms that correspond to an expansion over a basis of radial functions \(\langle x|nl \rangle\) and spherical harmonics \(\langle \hat{x}|lm \rangle\) of contributions coming from Gaussians centered on each neighbor

\[
\langle nlm; r_{ji}|g \rangle = \int d\mathbf{x} \langle nl|x \rangle \langle lm|\hat{x} \rangle (\mathbf{x} - r_{ji}|g) \tag{103}
\]

In the \(g \rightarrow \delta\) limit, the contribution from the \(j\)-th neighbor amounts simply to a product of the radial and angular functions evaluated at \(r_{ji}\),

\[
\langle nlm; r_{ji}|\delta \rangle = \langle nl|r_{ji} \rangle \langle lm|\hat{r}_{ji} \rangle \tag{104}
\]

Each of these terms can be computed very efficiently, exploiting in particular the fact that all orders of the spherical harmonics and their derivatives can be computed using recursion relations \(^{106,107,206}\).

It might appear that using a smooth atom density \(g\) complicates substantially the evaluation of Eq. (103). However when \(g\) is a spherical Gaussian with standard deviation \(\sigma\), the integral over \(d\mathbf{x}\) can be computed analytically \(^{221}\),

\[
\int d\mathbf{x} \langle lm|\hat{x} \rangle (\mathbf{x} - r_{ji}|g) = \langle l; x;r_{ji}|g \rangle \langle lm|\hat{r}_{ji} \rangle \tag{105}
\]

where

\[
\langle l; x;r_{ji}|g \rangle = 4\pi e^{-r^2/2\sigma^2} e^{-x^2/2\sigma^4} \mu_i((x-x_{\sigma})^2) \tag{106}
\]

so one gets

\[
\langle nlm; r_{ji}|g \rangle = \langle lm|\hat{r}_{ji} \rangle \int dx \langle nl|x \rangle \langle l; x;r_{ji}|g \rangle \tag{107}
\]

The radial part of the integral

\[
\int dx \langle nl|x \rangle \langle l; x;r_{ji}|g \rangle = \langle nl;r_{ji}|g \rangle \tag{108}
\]

can be computed numerically for any form of the radial basis resulting in \(n_{\text{max}}^2 l_{\text{max}}^2 \mu_{\text{grid}}\) evaluations of special functions. For instance, the original implementation of the SOAP representation uses a numerically orthogonalized, equispaced Gaussian basis \(^{22}\). Alternatively, this integral might also be performed analytically by using Gaussian type orbitals (GTO) as the radial basis \(^{436,437,438}\). This choice makes it possible to compute the coefficients of the smeared density as easily as for the \(g \rightarrow \delta\) case

\[
\langle nlm; r_{ji}|g;\text{GTO} \rangle = \langle lm|\hat{r}_{ji} \rangle \langle nl;r_{ji}|g;\text{GTO} \rangle \tag{109}
\]

where the only overhead comes from having to compute \(O(n_{\text{max}}l_{\text{max}})\) terms for the radial part and its orthonormalization. This is asymptotically cheaper.
than combining radial and angular terms – which requires $O(n_{\text{max}}l_{\text{max}}^2)$ multiplications per neighbor – but can be substantial in practical cases, because the analytical integrals in Eqs. (105) and (108) yield non-standard special functions.

To reduce this overhead, one can choose a form of the atomic density that is symmetric about $r_i$ instead of $r_j$:

$$\langle x - r_{ji} | \tilde{g} \rangle = \exp \left[ -\frac{(x - r_{ji})^2}{2\sigma_i^2} - \frac{r_{ji}^2}{\sigma_\perp} (1 - \hat{r}_{ji} \cdot \hat{x}) \right].$$

Together with a choice of radial functions that do not depend explicitly on $l$, this allows factorizing the radial integrals (108) as

$$\int dx \langle n|x \rangle \langle x - r_{ji}; l|\tilde{g} \rangle = \langle n; r_{ji}|\tilde{g} \rangle \langle l; r_{ji}|\tilde{g} \rangle.$$ (111)

Coupled with the polynomial basis proposed in Ref. [27] these expansion coefficients can be computed efficiently using recurrence relations in the radial and angular coefficients. More in general, the cost of evaluating the radial integrals $\langle n;l|\rho\rangle$ can be made negligible by using splines to approximate the value of the special functions resulting from the integrals, or the numerical integration of basis functions for which there is no analytical expression. Another aspect that does not affect the asymptotic scaling of the expansion, but can significantly influence the prefactor, involves the evaluation of spherical harmonics.

Several well-established techniques can be used to speed up the calculation of $Y^m_l$, including the use of real-valued spherical harmonics, the use of recurrence relations, and the use of formulations that are entirely written in terms of the Cartesian components of $\hat{r}_{ji}$.

**Symmetrized n-body correlations** The density coefficients $\langle anlm|\rho \rangle$ are then combined to compute invariant (or covariant) features. Formally, the evaluation of the symmetry-adapted features – both those built using only local $|\rho_i\rangle$ features, and the multi-scale features that combine $|\rho_i\rangle$ and $|V_i\rangle$ – involves a tensor product of $\nu$ sets of density coefficients to yield density correlations in the uncoupled basis $\langle (a_n l_m k)_{\sigma \lambda \ell} \rangle$, and then a contraction along the $m_i$ indices, that generates the equivariant features expressed in the coupled basis $\langle (a_n l m)_{\sigma \lambda \ell} \rangle$. A technical difficulty one has to keep in mind when implementing the calculation of equivariant features is that the angular $(l, m)$ indices have an irregular memory layout, with $-l \leq m \leq l$. Depending on the hardware architecture, it might be beneficial to store the coefficients in a regular $(l_{\text{max}} + 1) \times (2l_{\text{max}} + 1)$ array, padded with zeros.

A more substantial challenge associated with the increase of the body order is that both the number of linearly independent features and the cost of evaluating each of them based on a naive contraction of the tensor products of density coefficients (e.g. based on the expressions in Ref.[108]) increase exponentially with $\nu$. Even though the exponential scaling is related to the expansion parameters $l_{\text{max}}$ and $n_{\text{max}}$, and not on the number of neighbors, as it would be the case for the calculation of the features as a sum over clusters of $\nu$ atoms (see Sec. [IVF] and [V]), it makes the enumeration of a complete linear basis prohibitively expensive. The recurrence relations of Eq. (110) or the equivalent ones for the invariant features proposed in Ref.[123] make it possible to evaluate individual equivariant features with a cost that scales only linearly with $\nu$. To beat completely the exponential scaling, these recursive expressions should be combined with feature selection schemes such as those discussed in Section [VIII]. For example, the $n$-body iterative contraction of equivariant (NICE) features incorporates a selection/contraction step at each level of the iteration. For each equivariant component $\langle q'|\tilde{\rho}^{\nu \lambda}_{\sigma \mu}\rangle$, one determines (e.g. by principal component analysis, or just by dropping some components) a set of coefficients $U^{\nu \sigma \lambda}_{q q'}$ that can be used to reduce the dimensionality of the features

$$\langle q'|\tilde{\rho}^{\nu \lambda}_{\sigma \mu}\rangle = \sum_{q'} U^{\nu \sigma \lambda}_{q q'} \langle q'|\tilde{\rho}^{\nu \lambda}_{\sigma \mu}\rangle.$$ (112)

Given that this operation only mixes features with the same equivariant behavior, it is then possible to perform an iteration equivalent to Eq. (10) to increase
the body order further
\[ \langle q^\sigma(\nu+1); \sigma; \lambda \mu \rangle \equiv \langle q^\sigma(\nu+1); n\lambda k; \sigma; \lambda \mu \rangle = \\
\delta_{\sigma(\nu+1)+1+\gamma} \sum_{\ell k} \langle lm; (k-m) | \lambda \mu \rangle \\
\times \langle n | \rho^\sigma(\nu+1); lm \rangle \langle q^\sigma; \lambda \mu \rangle \langle q; (k-m) \rangle. \] (113)

Note that in the first line we use the loose definition of the indices in the bra-ket notation (Section [IV A]). The \( \nu + 1 \) term can be indexed explicitly, with a notation that recalls the lower-order terms that are combined to obtain it; once it is computed, the granularity of the indexing becomes irrelevant, and a flat index can be used to streamline the notation. With this combination of expansion and contraction only the components that contribute significantly to the description of the structural diversity of the dataset, or to the prediction of the target properties, are retained to evaluate higher-order correlations.

An alternative perspective for developing efficient implementations is to represent invariant or equivariant properties \( F(A_i) \) in terms of the unsymmetrized correlations,
\[ F(A_i) \approx \sum_{n_1 l_1 m_1 \ldots n_l l_m m} \langle F | n_1 l_1 m_1 \ldots n_l l_m m \rangle \times \langle n_1 l_1 m_1 \ldots n_l l_m m | \rho_i^{\sigma(\nu+1)} \rangle, \]
with the desired symmetries imposed through constraints on the coefficients \( \langle F | n_1 l_1 m_1 \ldots n_l l_m m \rangle \). While this perspective imposes additional complexity on regression schemes it is convenient for fast evaluation of a fitted model (with coefficients now ensuring the correct symmetries) since the coupling coefficients need not be stored or evaluated anymore. An efficient evaluation now requires a recursion for the unsymmetrized correlations
\[ \langle n_1 l_1 m_1 \ldots n_l l_m m | \rho_i^{\sigma(\nu+1)} \rangle = \prod_{\alpha=1}^{\nu} \langle n_\alpha l_\alpha m_\alpha | \rho_i \rangle, \]
which is relatively straightforward to construct, the key challenge being to retain only the \( (n_\alpha, l_\alpha, m_\alpha) \) features that give rise to non-zero coefficients.

### E. Packages to evaluate atom-density representations

To provide a practical example of the use of software to compute representations, we compare three packages, namely quippy\(^{262}\) Descriptor\(^{262}\) and librascal\(^{262}\) that are open source, and can be easily used in a Python code. We do not discuss the internals of the implementations, but show code snippets that can be readily used to evaluate descriptors of atomic structures, primarily focusing on the SOAP powerspectrum. All examples use the Atomic Simulation Environment\(^{262}\), and atomic structures are assumed to be stored in the variable \( \text{structures} \), an instance of ASE’s \texttt{Atoms} object. In all of these implementations, the descriptor vectors are returned as \texttt{numpy.array} objects, from which kernel values may be obtained by computing the dot products between descriptor vectors.

The quippy python package is based on the QUIP suite with the GAP extension, which provides the descriptors module. QUIP must be downloaded and built using a Fortran compiler before quippy, which uses f90wrap to access the compiled functions in QUIP via python interfaces. In the GAP implementation, Gaussian radial basis functions are used, placed at equal intervals, and orthogonalized.

```python
from quippy.descriptors import Descriptor
soap = Descriptor("soap,cutoff=3.5,
    cutoff_transition_width=0.0,atom_sigma=0.3,
    n_max=8,l_max=6")
# returns a dictionary containing the features
# and connectivity information
features = soap.calc(structures)
# features["data"] is a numpy array with the
# shape (n_environments,n_features)
```

The \texttt{Descriptor} object is initialised using a string containing the kernel parameters in a \texttt{key=value} format, with some keys being mandatory.

The \texttt{dscribe} package\(^{262}\) implements multiple descriptors, including SOAP, MBTR\(^{262}\) and ACSF. A python interface is used to interact with calculator functions written in C/C++, ensuring efficient evaluation. The main difference between the SOAP implementation of \texttt{quippy} and \texttt{dscribe} are the choice of radial basis functions, which are spherical primitive Gaussian Type Orbitals (GTOs), orthogonalised using the method suggested by Löwdin\(^{262}\). Alternatively, cubic or higher order polynomials may also be chosen. In analogy with the definition of GTOs used in quantum chemistry, the radial basis has an explicit dependence on \( l \).

```python
from dscribe.descriptors import SOAP
species = np.unique(structures.
    get_chemical_symbols())
species = np.unique(structures.
    get_chemical_symbols())
soap = SOAP(
    rcut=3.5,
    nmax=8,
    lmax=6,
    sigma=0.3,
    species=["H","C"]
)
# returns a (n_environments,n_features) numpy array
X = soap.create(structures)
```

The \texttt{get} python object providing the descriptor is constructed from the class \texttt{SOAP} and specifying the parameters in the initialisation arguments.

The package librascal also provides a variety of descriptors, but chiefly focuses on the calculation of
density-based representations, including SOAP and the $\nu = 1$ and $\nu = 3$ correlations. The back-end, written in C++, can be accessed from python interfaces. Exploiting the spirit of the general construction of $[\rho_{\text{GTO}}^{\text{avr}}]$ features, librascal implements two kinds of radial functions, namely a family of GTO-like radial functions as well as a discrete variable representation (DVR) basis, corresponding to a real-space evaluation of the symmetrized density using a Gauss-Legendre quadrature rule.

from rascal.representations import SphericalInvariants

 hypers = {
    'soap_type': 'PowerSpectrum',
    'interaction_cutoff': 3.5,
    'radial_basis': 'GTO', # or 'DVR'
    'max_radial': 8,
    'max_angular': 6,
    'gaussian_sigma_constant': 0.3,
    'gaussian_sigma_type': 'Constant',
    'cutoff_smooth_width': 0.0,
    'normalize': False
}

soap = SphericalInvariants(**hypers)
# returns a (n_environments,n_features) numpy array
X = soap.transform(structures).get_features(soap)

The SphericalInvariants object uses the transform method to compute SOAP features, that are stored internally in a sparse format, in which each dense block corresponds to a $(a,a')$ pair of elemental densities. These features can be used to compute scalar-product kernels between two environments, or cast to a dense array through the get_features method.

These three packages all compute “SOAP” features, but differ in the choice of basis functions. Much as with electronic structure codes, that often yield results that differ significantly despite performing nominally the same type of calculations, one cannot expect to be able to combine the features computed by one package with the regression weights computed by another. It is however important to assess whether the features are equivalent in a less stringent sense, e.g. whether they contain analogous information, and whether they converge to the same limit when the expansion parameters $(\tau_{\text{max}},l_{\text{max}})$ are increased. Figure 29 demonstrates the convergence of the GFRE and GFRD (see Section VIII A and Ref. 236) between small-$(\tau_{\text{max}},l_{\text{max}})$ features and a highly converged $\Phi_{\text{full}}$ featurization. In all cases we consider GFRE($\Phi_{\text{full}}$, $\Phi$) is at least one order of magnitude smaller than GFRE($\Phi$, $\Phi_{\text{full}}$). One sees that, reassuringly, in all cases the feature reconstruction errors converge towards zero. For $\tau_{\text{max}} = 16$ all choices of radial bases are essentially converged, and the residual error is due to the convergence of the angular channels. Since all implementations use equivalent spherical harmonics expansions, the convergence with the angular cutoff $l_{\text{max}}$ is nearly identical. The convergence rate of the radial bases, however, is not the same. The GTO bases in librascal and dscribe have similar amounts of information, and converge faster than the bases used in quippy and the librascal DVR implementation. The GFRD also converges to zero for most implementations – meaning that in the complete basis set limit the corresponding features become equivalent. The implementation in dscribe is an exception, with a GFRD saturating at approximately 0.1, suggesting that implementation details lead to persistent differences in the weighting of different kinds of correlations even when $(\tau_{\text{max}},l_{\text{max}})$ increase beyond the values that are typically used in practice.
IX. APPLICATIONS

In this Section we want to report some representative applications that highlight different aspects of the representations discussed in this Review – demonstrating how an understanding of the nature and properties of the structure/features mapping can be used to construct efficient and insightful machine-learning models.

![ROC curve](image)

FIG. 30. ROCs of binary classifiers based on a SOAP kernel, applied to the prediction of the binding behavior of ligands and decoys taken from the DUD-E database. Each ROC corresponds to one specific protein receptor, and plots the fraction of true positives $p(\cdot\mid\cdot)$ against the fraction of false negatives $p(\cdot\mid\cdot)$. The red curve is the average over the individual ROCs. The dashed line corresponds to receptor FGFR1, which contains inconsistent data in the version of the DUD-E at the time of the original publication. Inset: AUC performance measure as a function of the number of ligands used in the training, for the “best match”-SOAP kernel (MATCH) and average molecular SOAP kernel (AVG). Adapted from Ref. 28.

A. Best match kernels for ligand binding

Contrary to the problem of predicting interatomic potentials, or other extensive proteins, the affinity between a protein and a small drug-like molecule does not fit well into the mold of an additive property model. The structure of the ligand must allow for the active portion of the molecule to fit in the binding pocket of the target protein, and the nature of the chemical groups in this “warhead” portion are more important to determine the strength of the interaction than peripheral portions of the molecule. Figure 30 shows the accuracy of a classifier based on SOAP features, that aims to distinguish active components from decoys for a given target protein. The targets and the ligands, as well as their “ground truth” binding behavior are taken from the database of useful decoys, enhanced (DUD-E). The performance of the classifier is represented in terms of the receiver operating characteristic (ROC) curves (the ROC curve of a perfect classifier would run along the left and top margins of the plot, while a classifier that is as good as random would run along the diagonal), and their area under the curve (AUC) (the AUC is the integral of the ROC, and roughly corresponds to the fraction of molecules that are classified correctly). The AUC plot, in the inset of Fig. 30 shows that a model based on an average metric – that describes each molecule as the average of its environments, Eq. (85) – performs rather poorly, which is unsurprising given the highly non-additive nature of the binding affinity. Using a “best-match” kernel (equivalent to the distance in Eq. 86, and implemented in practice as the small-$\gamma$ limit of the REMatch kernel) improves dramatically the accuracy of the classifier, bringing the AUC to well above 0.95. A judicious choice of the training structures, based on farthest point sampling, accelerates even further the convergence of the classifier with train set size. This application provides an example of how local representations can be combined in a non-additive way, resulting in a dramatic improvement of the machine-learning performance for a problem in which non-additive behavior is to be expected.

![Atomic contributions](image)

FIG. 31. Predicted atomic contributions to the total CCSD polarizability tensor for a selection of the showcase dataset as reported in Refs. 23,268. The ellipsoids are aligned along the principal axes of $\alpha_{ij}$, 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). Figure reproduced from Ref. 23.
B. Tensorial features and polarizability

Some of the early examples of machine-learning models leveraging covariant features focused on the prediction of dielectric response functions, such as the dipole moment $\mu$ (and the equivalent bulk quantity, polarization), polarizability $\alpha$ (and the closely-related electronic dielectric constant) as well as higher-order terms, such as the first hyperpolarizability $\beta$. We discuss the case of the static dipole polarizability $\alpha$ as a representative case that highlights many of the current ideas and applications. In its Cartesian form, $\alpha$ is a symmetric tensor, fully determined by six components ($\alpha_{xx}, \alpha_{yy}, \alpha_{zz}, \alpha_{xy}, \alpha_{xz}, \alpha_{yz}$). In order to build a machine-learning model based on equivariant density correlation features, it is more convenient to apply a unitary transformation that casts it into its irreducible spherical components (ISCs). The spherically symmetric term, $\alpha^{(0)}$, corresponds to the trace of the tensor, while the 5 anisotropic components, $\alpha^{(2)}(\lambda=-2,-1,0,1,2)$ transform collectively as $\lambda=2$ spherical harmonics, and can be computed using recursive relationships that are explicitly reported in Ref. 132.

A clear advantage of this construction is that, unlike the components of the Cartesian tensor, the two ISCs of $\alpha$ can be independently represented by the equivariant density-based features corresponding to $\lambda=0$ and $\lambda=2$, relying on a linear prediction model similar to the one reported in Eq. (37). The inherent locality of the model means that the tensor prediction can be broken down in the sum of individual atomic contributions $\alpha = \sum_i \alpha_i$. These local components can be combined to make predictions on larger, and more complex molecules than those included in the training set. This transferability was exploited in the AlphaML model222,269 to fit against coupled-clusters (CCSD) reference values, computed on small organic molecules from the QM7b database222,269, and predict on 52 larger “showcase” molecules that are at the limit of what is computable with state-of-the-art quantum chemistry methods. On these molecules, the error of AlphaML against the CCSD reference (0.24 a.u./atom) was less than half the discrepancy between CCSD and DFT (0.57 a.u./atom).

An additive model also provides predictions for the local contributions to $\alpha$, which are represented in Fig. 31 in terms of ellipsoids aligned along the principal axes of $\alpha(A_i)$. Even though these components do not have to be physically meaningful – given that the only training target is given by total polarizabilities – the local $\alpha(A_i)$ reflect some chemical insights, e.g. the model predicts large components when centering the representation on the highly-polarizable sulfur atoms, as well as along the directions where the molecules are highly polarizable. Highly conjugated molecules are also interesting because they exhibit a non-additive behavior of the polarizability, due to the vanishing HOMO-LUMO gap.

The prediction of $\alpha$ using equivariant features can also be extended to the condensed phase and provides a crucial ingredient to compute Raman spectra. An example of this is reported in Ref. 133, where the polarizability of crystal polymorphs of paracetamol are predicted along a full molecular dynamics trajectory, thus allowing for the calculation of the Raman intensity in terms of the polarizability correlation spectrum. As shown in Fig. 33 given the local

![Figure 32](https://example.com/figure32.png)

**FIG. 32.** Polarizability per carbon atom ($\alpha/nc$) vs. number of carbons (nc) for the series of $s$-trans alkenes (from C$_6$H$_{14}$ to C$_{22}$H$_{44}$, full line) and acenes (from benzene to pentacene, dotted line), as well as fullerene (C$_{60}$). The green squares (and error bars) indicate the experimental measurements for C$_{60}$. Results are provided from DFT (blue) and CCSD (red) calculations, as well as the corresponding AlphaML models. Figure reproduced from Ref. 23.

![Figure 33](https://example.com/figure33.png)

**FIG. 33.** (black line) Raman spectrum prediction of paracetamol form-I averaged over 16 different training models. Each training model is obtained by a random subselection of 2000 configurations over a total of 2500. (shaded area) Standard deviation of the predicted spectra over the 16 models, calibrated with a likelihood maximization procedure described in Ref. 136 (blue line) Reference ab initio Raman spectrum. Figure adapted from Ref. 133.
C. Long-range and non-local responses

The clear breakdown of a ML model based on local features that is apparent in Fig. 32 is representative of a general limitation of density-based features. There are essentially two approaches one can take to tackle the issue of the non-locality of the structure-property relations, both of which are illustrated in Figure 34. One approach is to learn a proxy of the target property, which has a more localized nature, and which can then be easily manipulated to obtain the end result. The top panel of Fig. 34, adapted from Ref. 133 is an example of this approach. The electronic dielectric response $\epsilon_\infty$ of bulk water is affected by a collective, macroscopic electrostatic effect that is captured, in the continuum limit, by well-known expressions such as the Clausius-Mossotti relation, $\alpha = \frac{V}{\epsilon - 1}/(\epsilon + 2)$, that links $\epsilon_\infty$ to an effective molecular polarizability. This effective $\alpha$ is more readily learnable by a local model, leading to better accuracy and transferability in predicting $\epsilon_\infty$.

A different approach is needed when there is no obvious transformation of the target property to a more local version, as is the case for the polarizability of conjugated hydrocarbons. In these cases, one needs a model that is able to describe arbitrary non-local correlations. Long-range representations such as multiscale LODE features (Eq. (46)) are particularly attractive, in that they combine a long-range character (coming from the potential field) with an additive decomposition that provides the transferability needed to extend the prediction to systems of increasing size. This is demonstrated in Fig. 34, where the multiscale LODE model is tested for predicting the isotropic component of the polarizability of a series of polypeptides of increasing length. While the prediction at small peptide lengths share a similar accuracy as that obtained using a pure density-based representation, the inclusion of the potential field greatly decreases the prediction error when considering longer molecular chains. A large, overall improvement of the prediction accuracy is observed when adopting an optimized, weighted combination between local and LODE features. These result suggests that the inclusion of long-range features within the regression model provides a better description of the intermediate-range interactions, and that by adjusting the relative importance of local and delocalized terms the model can be trained only on small molecules, and extrapolate reliably across systems of increasing size. Very similar findings were reported on the transferability of models of molecular dipoles 140 – where however the splitting between local and long-range physics was achieved by combining different regression models rather than by different choices of features.

**FIG. 34.** Top: Learning curves of the components of the $\lambda = 0$ and $\lambda = 2$ components of the dielectric response tensor $\epsilon_\infty$ of water, through direct learning (red and green lines, respectively) and indirect learning going through the Clausius-Mossotti relation (blue and gray). 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 hexagonal ice structures using the ML model trained on liquid water. Bottom: 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. The curves correspond to a LODE model (blue) a squared-kernel SOAP model (green) and a hybrid model mixing the two kernels (red). Adapted from Refs. 133 and 137.

nature of the polarizability response in this kind of systems, accurate Raman intensities and lineshapes can be predicted for the entire range of frequencies. The low cost associated with computing dielectric response functions by ML models using symmetry-adapted features makes it possible to routinely evaluate condensed-phases infrared and Raman spectra including also a description quantum mechanical nature of the nuclei 274 – a task that until very recently required enormous computational effort 272.
predicted using equivariant features of the corresponding order – for instance with a linear model

\[
    c_{n\lambda\mu}(A_i) \approx \sum_q \langle c_{n\lambda}|Q\rangle \langle Q|A_i; \rho^{\sigma\nu}; \sigma; \lambda\mu \rangle, \tag{115}
\]
even though current implementations used a kernel regression scheme. The non-orthogonality of the basis used to represent \( n_e(r) \), implies that the learning phase has now to be performed considering all the different density components at the same time. While this may sound as a computational drawback of the model, it also improves the locality of the coefficients, which underlies its remarkable transferability across vast conformational and chemical spaces, since the electron density can be effectively learned as a collection of local contributions. This is well exemplified in Figs. 35, where the electron density prediction of \( C(8) \) hydrocarbon\(^{22,276}\) is tested upon having trained the model on much smaller compounds, with only 4 carbon atoms. This approach has since been applied to more complex systems, such as oligopeptides\(^{276}\) and to the prediction of other scalar fields such as the on-top density.

\[
\text{FIG. 35. Extrapolation results for the valence electron density of one octane (left) and one octatetraene (right) conformer, using a model trained on butadiene and butane. (top) DFT/PBE density isosurface at 0.25, 0.1, 0.01 Bohr}^{-3}, (middle) machine-learning prediction isosurface at 0.25, 0.1, 0.01 Bohr}^{-3}, (bottom) machine-learning error, red and blue isosurfaces refer to \pm 0.005 Bohr}^{-3} respectively. Figure adapted from Ref. 22 (Copyright 2018 American Chemical Society).
\]

**D. Electronic charge densities**

Another relevant scenario where the data-driven prediction of a quantum property benefits from a representation that relies on the use of local and equivariant features is the electron density \( n_e(r) \) of an atomic structure. The density is a scalar field, and has been modeled with some success by predicting its value at a specific point by an invariant representation centered on that point\(^{22,276}\). Given that (particularly in the case of an all-electron calculation) atomic nuclei are a natural vantage point to decompose the overall electron density, one may want instead to model \( n_e(r) \) as a sum of atom-centered contributions, \( n_e(r) = \sum_i n_e(A_i; r) \). These atom-centered terms can then be conveniently decomposed in terms of local functions, at the price of adopting a multi-centered non-orthogonal basis for the expansion i.e.,

\[
    n_e(A_i; r) = \sum_{n\lambda\mu} c_{n\lambda\mu}(A_i) R_n(|r - r_i|) Y^\lambda \left( \frac{r - r_i}{r - r_i} \right), \tag{114}
\]

where \( R_n \) represent some suitably optimized radial functions (for instance those used in resolution of the identity methods in quantum chemistry\(^{276}\)) and \( c_{n\lambda\mu}(A_i) \) correspond to the non-orthogonal expansion coefficients, that depend on the arrangement of atoms in the environment \( A_i \). These coefficients must transform in a covariant fashion with a rotation of the environment, and each \( \lambda \)-component can be independently

\[
\text{FIG. 36. Scatter plot and histograms based on Steinhardt order parameter,} q_\text{ref} \text{, } q_\text{ref} \text{, computed for simulations of liquid water and different phases of ice. Reproduced from Ref. 278.}
\]

**E. Structural classification and structural landscapes**

As discussed in Section [VII](#) the choice of a representation to describe atomic structures determines the “lens” through which they are interpreted, which in turn has a strong impact on the way unsupervised learning schemes, such as clustering and dimensionality reduction, bring to light recurring patterns, and structure-property relations. The potential of general-purpose, atom-density correlation features for these tasks has been recognized rather early. Figure 36 adapted from Ref. 278 shows a classification of snap-
FIG. 37. Sketch map of the structural similarity of 15,869 distinct PBE-DFT geometry-optimised ice structures, as computed by the Euclidean distance in SOAP space. The sketch-map coordinates are obtained by minimizing the error in reproducing this similarity in terms of distances between points on a 2D projection. The density and static lattice energy of each structure are encoded by the size and colour of the respective point, and correlate strongly with the position on the map. Known ice phases are labelled in blue. 34 new candidates are labelled in black and numbered in order of increasing dressed energy relative to a generalized convex hull construction. Reproduced from Ref. 280.

FIG. 38. Sketch-map describing the makeup of the structures included in the train set of an accurate and transferable potential for carbon. Selected structures are identified for graphite, diamond, hexagonal diamond (Lonsdaleite), amorphous carbon and fullerenes. Points are coloured according to their energy, while contours indicate the density of the database population in a particular region. Adapted from Ref. 281.

An even more comprehensive mapping of the phase diagram of water – in which crystalline and amorphous phases from across the phase diagram, as well as transition pathways between them were considered – was produced in Ref. 187 using permutation-invariant vectors as global descriptors for the different configurations. Abstract structural descriptors are particularly useful when applied to datasets that contain hypothetical structures, generated by a high-throughput procedure. In combination with a dimensionality-reduction scheme, and with a generalized convex hull construction that attempts to estimate the synthesizability of materials by considering jointly their predicted stability, and the structural similarity to other potential candidates, a SOAP representation has been able to rediscover all known (meta)stable ice phases, as well as to propose another 34 structures which might be also stabilizable by pressure, doping, or co-crystallization (see Fig. 37).

An incomplete list of applications that use general-purpose features for structural analysis and classification includes: the construction of structure-property maps for small organic molecules, molecular materials, and corrosion inhibitors; the identification and characterization of defects in solids and self-assembled polymers; the classification of secondary-structure patterns in polypeptides and the building blocks of zeolites; the classification of different phases in multi-phase materials; the search of stable phases of materials; the determination of the convergence of microsolvation studies of the hydration free energy.
There are also several examples, besides those given in Section VII where low-dimensional maps have been used as a tool to understand the structure of a data set or the nature of a representation. In Ref. [51] a PCA map was used to understand the effect of randomizing the atom ordering on the feature space associated with a Coulomb matrix description of molecules, emphasizing the information loss associated with sorting of the elements – an alternative route to achieve permutation invariance. In Ref. [101] maps based on different kinds of SOAP kernels provided an understanding of the effect of different approaches to combining environment-level kernels, and of different definitions of an alchemical kernel between chemical elements, on the similarity between molecules as measured by the representation. In Ref. [77] maps of a dataset of water oligomers were used to compare the performance of different ML scheme to build 2 and 3-body models of the energy of water clusters.

The use of low-dimensional representations to visualize the structure of a dataset, showing the relationship between different kinds of training structures, identifying regions that are poorly sampled, and determining how new configurations relate to the data the ML model has been fitted, is also gaining traction. An example of such map is given in Fig. 38 showing the diversity of the structures used to train a transferable machine-learning potential for carbon. Adopting the same type of representations used for the regression model as the basis of this kind of analysis ensures that the maps describe the same feature space that underlies the fit.

F. 3D representations for QSPR

Even though the focus of this review is on descriptors of the 3D structure of materials applied to the construction of surrogate models of quantum mechanical properties, there is also growing interest in their application to QSPR tasks. As we briefly discuss in Section [11] the descriptors that have been traditionally used in cheminformatics are based on a collection of molecular properties, or on molecular graph descriptors that do not depend on the particular conformation. From a conceptual point of view, their coarseness is an advantage, because it is compatible with the definition of thermodynamic properties that are not associated with a single specific configuration, such as solvation and ligand binding free energies. Nevertheless, there is growing evidence that the use of descriptors incorporating information on the 3D geometries can improve the accuracy of QSPR models, especially for difficult cases that involve very flexible molecules, as well as for data analytics approaches for materials informatics. One of the core challenges in these efforts is the determination of the conformer geometries that should be used to evaluate the 3D descriptors, an operation for which several strategies have been explored to enhance the accuracy of QSPR models. As we briefly discuss in Section VII one of the most promising research directions involves combining the high fidelity of density based representations with a well-principled construction of ensembles of features. This is still a very active subject of research, with very encouraging results having been recently demonstrated for the prediction of the solubility of small molecules or the computational screening for antiviral drugs.

CONCLUSIONS AND OUTLOOK

The description of atomic structures in terms of mathematically sound, computationally efficient, and physically-inspired representations has largely driven the extraordinarily successful application of machine-learning schemes to atomic-scale modeling. Independently-developed representations have undergone a process of convergent evolution to fulfill a concurrent set of requirements, such as symmetry with respect to translations and rotations, smoothness and injectivity – a clear indication of the importance of these properties to obtain efficient machine-learning models. Over the past few years, a more systematic study of the problem of representing atomic structures has clarified the connections between most of the successful representations, and between these and well-established concepts in statistical physics of liquids and alloys (the cluster expansion), as well as with the construction of potential energy surfaces for molecules and the condensed phase. A better understanding of the mathematical properties of representations is likely to lead, in the near future, to more robust and better performing implementations.

A formal treatment of symmetries also enabled the development of equivariant features that are suitable to build models that automatically obey the same transformation rules as vectors and tensors, making it possible to learn efficiently properties such as dipole moments, polarizability and density fields. This equivariant formulation can also be used to iteratively increase the body order of a structural representation – an important open question is how to best treat these high-body-order terms, whether by linear models that explicitly include dedicated high-order features, or by non-linear models that generate (some of) them algebraically. The answer rests both on practical considerations and on the very fundamental, highly non-trivial issue of whether a representation of limited body order provides a complete (injective) description of an atomic structure.

Another open challenge is how to deal with non-additivity, and with properties that depend on long-
range interactions between far-away atoms. Particularly promising is a long-distance equivariant framework, that can be formulated as as a rather straightforward extension of the same density-correlations scheme that underlies local features, and can be related to a multipole expansion of interactions. It is yet to be seen whether it can describe more subtle physical phenomena such as quantum delocalization, polarization and charge transfer, and how it compares with more explicitly physically-motivated “hybrid” models. A better control of the multi-scale nature of the interactions, including the use of “multi-resolution” features, is likely to be one of the focal point of feature-engineering efforts, which may lead to an incremental – but nevertheless important – increase of the accuracy of ML models of matter. The optimization of features for a specific problem may however impact their general applicability, which is one of the critical advantages of the class of abstract, generic representations we focus on in this review, that can be seen as the point of convergence of molecular potential energy surfaces and condensed-phase potentials.

One of the most recent research directions aims at extending even further the reach of this framework, by resolving the divide between three dimensional continuous representations and discrete fingerprints, for applications to quantitative structure-property relations. The challenge here is to reconcile the superior resolving power of 3D, atom-density-correlation representations with the fact that traditional cheminformatics tasks aim to predict “macroscopic” properties, such as solubility or toxicity, that are not associated with an individual configuration, but rather with the ensemble of conformers that are associated with a specific thermodynamic state point. Another traditional application of cheminformatics is the inverse design of molecules with prescribed (or optimized) properties, and the construction for generative models. While one could envisage to use 3D representations for this task, a substantial hurdle would be the fact that the map between structure and density-correlation features is not bijective: there are feature vectors that do not correspond to any structure, and even feature vectors that cannot be obtained as a symmetrized correlation of an arbitrary scalar field. Thus, the unconstrained search for the “optimal feature vector” might result in a set of features that do not correspond to an actual structure. Until this issue is better understood, efforts to use atom-density representations for inverse design should rely on approaches that do not require an inverse feature map.

In the quest for more accurate and efficient machine-learning models of the structure and properties of atomistic systems, physically-motivated concepts have been incorporated into the mathematical representation of atomic configurations, resulting in striking similarities with traditional modeling frameworks. When treading the fine line between data-driven and physics-based approaches, the core question is how to achieve a natural description of well-understood phenomena without giving up the flexibility to model unexpected, complex effects – and how to build features that can be optimized for a specific application, while still being universally applicable.

ACKNOWLEDGMENTS

The authors would like to thank Yasushi Shibuta for providing the structures used in Fig. 14 and Stefan Goeckeler for providing Fig. 15 and the many colleagues and friends who discussed with us about this review, and the ideas it summarizes. FM, MC and AG acknowledge support by the National Center of Competence in Research MARVEL, funded by the Swiss National Science Foundation.

AUTHOR BIOGRAPHIES

Félix Musil studied physics at the EPFL and received his MSc in applied physics in 2015, with a thesis on the modeling of plasma in a fusion reactor. For his PhD he joined in 2016 the group of Prof. Michele Ceriotti at the EPFL to develop and apply methods to investigate structure–property relationships in materials using atomistic modeling and machine learning techniques.

Andrea Grisafi studied chemistry at the University of Pisa and Scuola Normale Superiore of Pisa. In 2016, he received his MSc in physical chemistry with a thesis on the statistical mechanics of simple ionic liquids. Since then, he is a PhD student in the group of Prof. Michele Ceriotti at EPFL, where he works on the development of atomic-scale representations that are suitable to incorporate physical symmetries and long-range effects within machine-learning models of molecular and materials properties.

Albert P. Bartók is an Assistant Professor at the University of Warwick. He earned his PhD degree in physics from the University of Cambridge in 2010, his research having been on developing interatomic potentials based on ab initio data using machine learning. He was a Junior Research Fellow at Magdalene College, Cambridge and later a Leverhulme Early Career Fellow. Before taking up his current position, he was a Research Scientist at the Science and Technology Facilities Council. His research focuses on developing theoretical and computational tools to understand atomistic processes.

Christoph Ortner is Professor of Mathematics at the University of British Columbia (Canada). After obtaining his doctorate in numerical analysis in 2007 at the University of Oxford (UK) and remaining there as an RCUK fellow, he moved to the University of Warwick in 2011 and to UBC in 2020. His main
interests revolve around mathematical and computational aspects of atomistic and multi-scale modeling.

Gábor Csányi is Professor of Molecular Modeling at the University of Cambridge (UK). He obtained his doctorate in computational physics (2001) from the Massachusetts Institute of Technology (USA), having worked on electronic structure problems in materials science and chemistry. He was in the group of Mike Payne in the Cavendish Laboratory before joining the faculty of the Engineering Laboratory at Cambridge. He is developing algorithms and data-driven numerical methods for atomic scale problems in materials science and chemistry.

Michele Ceriotti is Associate Professor at the Institute of Materials at the École Polytechnique Fédérale de Lausanne. He received his Ph.D. in Physics from ETH Zürich in 2010, under the supervision of Professor Michele Parrinello. He spent three years in Oxford as a Junior Research Fellow at Merton College, and joined EPFL in 2013, where he leads the laboratory for Computational Science and Modeling. His research interests focus on the development of methods for molecular dynamics and the simulation of complex systems at the atomistic level, as well as their application to problems in chemistry and materials science—using machine learning both as an engine to drive more accurate and predictive simulations, and as a conceptual tool to investigate the interplay between data-driven and physics-inspired modeling.

---

1. Olexandr Isayev, Denis Fourches, Eugene N. Muratov, Corey Oses, Kevin Rasch, Alexander Tropsha, and Stefano Curtarolo, “Materials cartography: Representing and mining materials space using structural and electronic fingerprints,” Chem. Mater. 27, 735–743 (2015).
2. Benjamin Sanchez-Lengeling and Alán Aspuru-Guzik, “Inverse molecular design using machine learning: Generative models for matter engineering,” Science 361, 360–365 (2019).
3. Payel Das, Mark Moll, Hernán Stamati, Lydia E. Kavraki, and Cecilia Clementi, “Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction,” Proc. Natl. Acad. Sci. U. S. A. 103, 9885–9890 (2006).
4. Michele Ceriotti, Gareth A. Tribello, and Michele Parrinello, “Simplifying the representation of complex free-energy landscapes using sketch-map,” Proc. Natl. Acad. Sci. U. S. A. 108, 13023–13028 (2011).
5. Vojtˇ ech Spiwok and Blanka Kr´ alov´ a, “Metadynamics in the conformational space nonlinearly dimensionally reduced by Isomap,” J. Chem. Phys. 135, 224504 (2011).
6. Mary A Rohrdanz, Wenwei Zheng, and Cecilia Clementi, “Discovering mountain passes via torchlight: Methods for the definition of reaction coordinates and pathways in complex macromolecular reactions,” Annu. Rev. Phys. Chem. 64, 295–316 (2013).
7. Kiran H. Kanékal and Tristan Bereau, “Resolution limit of data-driven coarse-grained models spanning chemical space,” Journal of Chemical Physics 151, 10.1063/1.5119101 (2019). arXiv:1907.04982
8. Nicholas E. Jackson, Alec S. Bowen, Lucas W. Antony, Michael A. Webb, Venkatram Vishwanath, and Juan J. de Pablo, “Electronic structure at coarse-grained resolutions from supervised machine learning,” Science Advances 5, eaav1190 (2019).
9. Jianh Wang, Simon Olsson, Christoph Wehmeyer, Adrià Pérez, Nicholas E. Charron, Gianni De Fabritiis, Frank Noé, and Cecilia Clementi, “Machine Learning of Coarse-Grained Molecular Dynamics Force Fields,” ACS Central Science 5, 755–767 (2019).
10. Jörgh Behler and Michele Parrinello, “Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces,” Phys. Rev. Lett. 98, 146401 (2007).
11. Bastiaan J. Braams and Joel M. Bowman, “Permutationally invariant potential energy surfaces in high dimensionality,” Int. Rev. Phys. Chem. 28, 577–606 (2009).
12. Albert P. Bartók, Mike C. Payne, Risi Kondor, and Gábor Csányi, “Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons.” Phys. Rev. Lett. 104, 136403 (2010).
13. Gabriele C. Sosso, Giacomo Menci, Sebastiano Caravati, Jörgh Behler, and Marco Bernasconi, “Neural network interatomic potential for the phase change material GeTe,” Phys. Rev. B 85, 174103 (2012).
14. Jörgh Behler, “Perspective: Machine learning potentials for atomistic simulations,” J. Chem. Phys. 145, 170901 (2016).
15. Volker L. Deringer and Gábor Csányi, “Machine learning based interatomic potential for amorphous carbon,” Phys. Rev. B 95, 094203 (2017).
16. Daniele Dragoni, Thomas D. Daff, Gábor Csányi, and Nicola Marzari, “Achieving DFT accuracy with a machine-learning interatomic potential: Thermomechanics and defects in bcc ferromagnetic iron,” Phys. Rev. Materials 2, 013808 (2018).
17. Bingqing Cheng, Guglielmo Mazzola, Chris J. Pickard, and Michele Ceriotti, “Evidence for supercritical behaviour of high-pressure liquid hydrogen,” Nature 585, 217–220 (2020).
18. Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O. Anatole von Lilienfeld, “Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning,” Phys. Rev. Lett. 108, 095301 (2012).
19. J S Smith, O Isayev, and A E Roitberg, “ANI-1: An extensible neural network potential with DFT accuracy at force field computational cost,” Chem. Sci. 8, 3192–3203 (2017).
20. 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).
25 Federico M. Paruzzo, Albert Hofstetter, Félix Musil, Sandip De, Michele Ceriotti, and Lyndon Ensley, “Chemical shifts in molecular solids by machine learning,” Nat. Commun. 9, 4501 (2018)
26 Andrea Grisafi, Alberto Fabrizio, Benjamin Meyer, David M. Wilkins, Clemence Corminboeuf, and Michele Ceriotti, “Transferable Machine-Learning Model of the Electron Density,” ACS Cent. Sci. 5, 57–64 (2019)
27 David M. Wilkins, Andrea Grisafi, Yang Yang, Ka Un Liao, Robert A. DiStasio, and Michele Ceriotti, “Accurate molecular polarizabilities with coupled cluster theory and machine learning,” Proc. Natl. Acad. Sci. U. S. A. 116, 3401–3406 (2019)
28 K. T. Schütz, M. Gasteiger, 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)
29 Rumin OuYang, Stefano Curtarolo, Emre Ahmeti, Matthias Scheffler, and Luca M. Ghiringhelli, “SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates,” Phys. Rev. Mater. 2, 083802 (2018)
30 Jörg Behler, Sönke Lorenz, and Karsten Reuter, “Representing molecule-surface interactions with symmetry-adapted neural networks,” The Journal of Chemical Physics 127, 014705 (2007)
31 Albert P. Bartók, Risi Kondor, and Gábor Csányi, “On representing chemical environments,” Phys. Rev. B 87, 184115 (2013)
32 Albert P. Bartók, Sandip De, Carl Poelking, Noam Bernstein, James R. Kermode, Gábor Csányi, and Michele Ceriotti, “Machine learning unifies the modeling of materials and molecules,” Sci. Adv. 3, e1701816 (2017)
33 Haoyan Huo and Matthias Rupp, “Unified representation for machine learning of molecules and crystals,” ArXiv Prepr. ArXiv170406439 13754 (2017).
34 Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyne Ping Ong, “Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals,” Chem. Mater. 31, 3564–3572 (2019)
35 Lorenz C. Blum and Jean-Louis Reymond, “970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13,” J. Am. Chem. Soc. 131, 8732–8733 (2009)
36 Mati Karelson, Molecular Descriptors in QSAR/QSPR (Wiley-Interscience, New York, 2000).
37 David Weininger, “SMILES, a chemical language and information system. I. Introduction to methodology and encoding rules,” J. Chem. Inf. Model. 28, 31–36 (1988)
38 Roberto Todeschini and Viviana Consonni, Molecular Descriptors for Chemoinformatics, Methods and Principles in Medicinal Chemistry, Vol. 2 (Wiley, 2010) pp. 1–252.
39 Todd J. Wills, Dmitrii A. Polshakov, Matthew C. Robinson, and Alpha A. Lee, “Impact of Chemist-In-The-Loop Molecular Representations on Machine Learning Outcomes,” J. Chem. Inf. Model. 60, 4449–4456 (2020)
30 Gisbert Schneider and Ulrich Fechner, “Computer-based de novo design of drug-like molecules,” Nat Rev Drug Discov 4, 649–663 (2005)
31 Mati Karelson, Victor S. Lobanov, and Alan R. Katritzky, “Quantum-Chemical Descriptors in QSR/QSPR Studies,” Chem. Rev. 96, 1027–1044 (1996)
32 A. Gautlon, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, and J. P. Overington, “ChEMBL: a large-scale bioactivity database for drug discovery,” Nucleic Acids Research 40, D1100–D1107 (2012)
33 Teague Sterling and John J. Irwin, “ZINC 15 – Ligand Discovery for Everyone,” Journal of Chemical Information and Modeling 55, 2324–2337 (2015)
34 Sungwhan Kim, Paul A. Thiessen, Evan E. Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A. Shoemaker, Jiayao Wang, Bo Yu, Jian Zhang, and Stephen H. Bryant, “PubChem substance and compound databases,” Nucleic Acids Research 44, D1202–D1213 (2016)
35 Zhengqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande, “MoleculeNet: a benchmark for molecular machine learning,” Chemical Science 9, 513–530 (2018)
36 Olga Obrezanova, Gábor Csányi, Joelle M. R. Gola, and Matthew D. Segall, “Gaussian Processes: A Method for Automatic QSAR Modeling of ADME Properties,” J. Chem. Inf. Model. 47, 1847–1857 (2007)
37 Michael M. Mysinger, Michael Carchia, John J. Irwin, and Brian K. Shoichet, “Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking,” Journal of Medicinal Chemistry 55, 6582–6594 (2012)
38 John Edward Jones, “On the determination of molecular fields—II. From the equation of state of a gas,” Proc. R. Soc. Lond. A 106, 463–477 (1924)
39 B. J. Alder, D. M. Gass, and T. E. Wainwright, “Studies in molecular dynamics. VIII. The transport coefficients for a hard-sphere fluid,” The Journal of Chemical Physics 53, 3813–3826 (1970)
40 Frank H Stillinger and Anesur Rahman, “Improved simulation of liquid water by molecular dynamics Comparison of simple potential functions for simulating liquid water: Improved simulation of liquid water by molecular dynamics*,” J. Chem. Phys. J. Chem. Phys. J. Phys. Gen. Method J. Chem. Phys. J. Chem. Phys. J. Chem. Phys. 601, 954–926 (1974)
41 R Car and M Parrinello, “Unified Approach for Molecular Dynamics and Density-Functional theory. R. Car and M. Parrinello.pdf,” Phys. Rev. Lett. 55, 2471–2474 (1985).
42 M P Allen and D J Tildesley, Computer Simulation of Liquids (Oxford University Press, USA, 1990).
43 Daan Frenkel and Berend Smits, Understanding Molecular Simulation, 2nd ed. (Academic Press, London, 2002).
44 Robert G. Parr and Weitao Yang, Density-Functional Theory of Atoms and Molecules, 1st ed., International
Series of Monographs on Chemistry No. 16 (Oxford Univ. Press [n.a.], New York, NY, 1994).

Kieron Burke, “Perspective on density functional theory.” J. Chem. Phys. 136, 150901 (2012).

George H. Booth, Andreas Grüneis, Georg Kresse, Kieron Burke, “Perspective on density functional theory.” J. Comput. Chem. 12, 606–610 (1991).

Jon Baker and Fora Chan, “The location of transition states: A comparison of Cartesian, Z-matrix, and natural internal coordinates.” J. Comput. Chem. 17, 888–904 (1996).

Judith A. Harrison, J. David Schall, Sabina Maskey, Paul T. Mukulski, M. Todd Knippenberg, and Brian H. Morrow, “Review of force fields and intermolecular potentials used in atomistic computational materials research.” Applied Physics Reviews 5, 031104 (2018).

David K Duvendau, Dougal Maclaurin, Jorge Iparaguire, Rafael Bombarel, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams, “Convolutional networks on graphs for learning molecular fingerprints,” in Advances in Neural Information Processing Systems (2015) pp. 2224–2232.

Shina C. L. Kamerlin and Arieh Warshel, “The empirical valence bond model: Theory and applications,” WIREs Comput Mol Sci 1, 30–45 (2011).

J. M. Bowman, B. J. Braams, S. Carter, C. Chen, G. Czakó, B. Fu, X. Huang, E. Kamarchik, A. R. Sharma, B. C. Shepler, Y. Wang, and Z. Xie, “Ab Initio-Based Potential Energy Surfaces for Complex Molecules and Molecular Complexes,” J. Phys. Chem. Lett. 1, 1866–1874 (2010).

Zhen Xie and Joel M. Bowman, “Permutationally Invariant Polynomial Basis for Molecular Energy Surface Fitting via Monomial Symmetrization,” J. Chem. Theory Comput. 6, 26–34 (2010).

Bin Jiang and Hua Guo, “Permutation invariant polynomial neural network approach to fitting potential energy surfaces,” The Journal of Chemical Physics 139, 054112 (2013).

Michael A. Collins and Drew F. Parsons, “Implications of rotation–inversion–permutation invariance for analytic molecular potential energy surfaces,” The Journal of Chemical Physics 99, 6756–6772 (1993).

Helmut Gassner, Michael Probst, Albert Lauenstein, and Kersti Hermansson, “Representation of Intermolecular Potential Functions by Neural Networks,” J. Phys. Chem. A 102, 4596–4605 (1998).

Thuong T. Nguyen, Eszter Székely, Giulio Imbalzano, Jörg Behler, Gábor Csányi, Michele Ceriotti, Andreas W. Götz, and Francesco Paesani, “Comparison of permutationally invariant polynomials, neural networks, and Gaussian approximation potentials in representing water interactions through many-body expansions,” J. Chem. Phys. 148, 241725 (2018).

Cas van der Oord, Genevieve Dousse, Gábor Csányi, and Christoph Ortner, “Regularised atomic body-ordered permutation-invariant polynomials for the construction of interatomic potentials,” Mach. Learn. Sci. Technol. 1, 015004 (2020).

Jonathan E. Moussa, “Comment on “Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning”,” Phys. Rev. Lett. 109, 059801 (2012).

Grégoire Montavon, Katja Hansen, Siamac Fazli, Matthias Rupp, Franziska Biegler, Andreas Ziehe,
Alexandre Tkatchenko, Anatole V. Lilienfeld, and Klaus-Robert 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.

Katja Hansen, Grégoire Montavon, Franziska Biegler, Siamac Fazli, Matthias Rupp, Matthias Scheffler, O. Anatole von Lilienfeld, Alexandre Tkatchenko, and Klaus-Robert Müller, “Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies,” J. Chem. Theory Comput. 9, 3404–3419 (2013)

Matthias Rupp, Raghunathan Ramakrishnan, and O. Anatole von Lilienfeld, “Machine Learning for Quantum Mechanical Properties of Atoms in Molecules,” J. Phys. Chem. Lett. 6, 3309–3313 (2015)

S. Pipolo, M. Salanne, G. Ferlat, S. Klotz, A. \hspace0 167em M Salit, and F. Pietrucci, “Navigating at Will on the Water Phase Diagram,” Phys. Rev. Lett. 119 (2017), 10.1103/physrevlett.119.245701

Katja Hansen, Franziska Biegler, Raghunathan Ramakrishnan, Wiktor Pronobis, O. Anatole von Lilienfeld, Klaus-Robert Müller, and Alexandre Tkatchenko, “Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space," J. Phys. Chem. Lett. 6, 2326–2331 (2015)

Fabio Pietrucci and Wanda Andreoni, “Graph Theory Meets Ab Initio Molecular Dynamics: Atomic Structures and Transformations at the Nanoscale,” Phys. Rev. Lett. 107, 085504 (2011)

Ali Sadeghi, S. Alireza Ghasemi, Bastian Schaefer, Stephan Mohr, Markus A. Lill, and Stefan Goedecker, “Metrics for measuring distances in configuration spaces,” J. Chem. Phys. 139, 184118 (2013)

Li Zhu, Maximilian Amsher, Tobias Fuhrer, Bastian Schaefer, Somayeh Faraji, Samare Rostami, S Alireza Ghasemi, Ali Sadeghi, Migle Grauzinyte, Chris Wolverton, and Stefan Goedecker, “A fingerprint based metric for measuring similarities of crystalline structures,” J. Chem. Phys. 144, 034203 (2016)

Han Wang, Linfeng Zhang, Jiequn Han, and Weinan E, “DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics,” Computer Physics Communications 228, 178–181 (2018)

Onur Çaylak, Anatole von Lilienfeld, and Bjørn Baumeier, “Wasserstein metric for improved quantum machine learning with adjacency matrix representations,” Mach. Learn.: Sci. Technol. (2020), 10.1088/2632-2153/aba048

Michael J. Willalt, Félix Musil, and Michele Ceriotti, “Atom-density representations for machine learning,” J. Chem. Phys. 150, 154110 (2019)

Josef Ischtwan and Michael A. Collins, “Molecular potential energy surfaces by interpolation,” The Journal of Chemical Physics 100, 8080–8088 (1994)

Tak-San Ho and Herschel Rabitz, “A general method for constructing multidimensional molecular potential energy surfaces from ab initio calculations,” The Journal of Chemical Physics 104, 2584–2597 (1996)

E Prodan and W Kohn, “Nearsightedness of electronic matter,” Proc. Natl. Acad. Sci. 102, 11635–11638 (2005)

Hyunwook Jung, Sina Stocker, Christian Kunkel, Harald Oberhofer, Byungchan Han, Karsten Reuter, and Johannes T. Margraf, “Size-Extensive Molecular Machine Learning with Global Representations,” Chem. SystemsChem 2 (2020), 10.1002/sys.201900052

Felix Faber, Alexander Lindmaa, O. Anatole von Lilienfeld, and Rickard Armiento, “Crystal structure representations for machine learning models of formation energies,” Int. J. Quantum Chem. , n/a-n/a (2015)

Annika Stuke, Milica Todorović, Matthias Rupp, Christian Kunkel, Kunal Ghosh, Lauri Himanen, and Patrick Rinke, “Chemical diversity in molecular orbital energy predictions with kernel ridge regression,” J. Chem. Phys. 150, 204121 (2019)

Felix A. Faber, Anders S. Christensen, Bing Huang, and O. Anatole Von Lilienfeld, “Alchemical and structural distribution based representation for universal quantum machine learning,” J. Chem. Phys. 148, 241717 (2018)

Anders S. Christensen, Lars A. Bratholm, Felix A. Faber, and O. Anatole von Lilienfeld, “FCCHL revisited: Faster and more accurate quantum machine learning," J. Chem. Phys. 152, 044107 (2020)

Raghunathan Ramakrishnan and O Anatole von Lilienfeld, “Many Molecular Properties from One Kernel in Chemical Space,” Chim. Int. J. Chem. 69, 182–186 (2015)

Kipton Barros and Yasuyuki Kato, “Efficient Langevin simulation of coupled classical fields and fermions,” Phys. Rev. B 88, 235101 (2013)

Sandip De, Albert P. Bartók, Gábor Csányi, and Michele Ceriotti, “Comparing molecules and solids across structural and alchemical space,” Phys. Chem. Chem. Phys. 18, 13754–13769 (2016)

Mireille Boutin and Gregor Kemper, “On reconstructing n-point configurations from the distribution of distances or areas,” Advances in Applied Mathematics 32, 709–735 (2004)

Bing Huang and O. Anatole von Lilienfeld, “Communication: Understanding molecular representations in machine learning: The role of uniqueness and target similarity,” The Journal of Chemical Physics 145, 161102 (2016)

Sergey N Pozdnyakov, Michael J Willatt, Albert P Bartók, Christoph Ortner, Gábor Csányi, and Michele Ceriotti, “Incompleteless of Atomic Structure Representations,“ Phys. Rev. Lett. 125, 166601 (2020)

Michael J. Willatt, Félix Musil, and Michele 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)

Ralf Drautz, “Atomic cluster expansion for accurate and transferable interatomic potentials,” Phys. Rev. B 99, 014104 (2019)

Markus Bachmayr, Gabor Csanyi, Ralf Drautz, Genevieve Dusson, Simon Etter, Cas van der Oord,
and Christoph Ortner, “Approximation of Atomic Interactions with Spherical Harmonics,” (2019) arXiv:1911.03550.

108 Marcel F. Langer, Alex Göeßmann, and Matthias Rupp, “Representations of molecules and materials for interpolation of quantum-mechanical simulations via machine learning,” arxiv:2003.12081 (2020).

109 To facilitate the use of this notation in LATEX documents, we provide a set of macros at https://github.com/cosmo-epfl/cosmo-tools/tree/master/tex/dirac-rep.

110 L. Nachbin, The Haar integral (R. E. Krieger Pub. Co., 1976).

111 Alexander V. Shapeev, “Moment Tensor Potentials: A Class of Systematically Improvable Interatomic Potentials,” Multiscale Model. Simul. 14, 1153–1173 (2016).

112 Seiji Kajita, Nobuko Oiba, Ryosuke Jinnouchi, and Ryoji Asahi, “A Universal 3D Voxel Descriptor for Solid-State Material Informatics with Deep Convolutional Neural Networks,” Sci. Rep. 7, 1–9 (2017).

113 Juhwan Noh, Jaehoon Kim, Helge S. Stein, and Benjamin Sanchez-Lengeling, John M. Gregoire, Alan Aspuru-Guzik, and Yoonsung Jung, “Inverse Design of Solid-State Materials via a Continuous Representation,” Matter 1, 1370–1384 (2019).

114 David Chandler, Introduction to Modern Statistical Mechanics (Oxford University Press, New York, 1987).

115 Jörg Behler, “Atom-centered symmetry functions for constructing high-dimensional neural network potentials,” The Journal of Chemical Physics 134, 074106 (2011).

116 Linfeng Zhang, Jiequn Han, Han Wang, Roberto Car, and Weinan E, “Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics,” Phys. Rev. Lett. 120, 143001 (2018).

117 Juraj Mavračí, Felix C. Mocanu, Volker L. Deringer, Gábor Csányi, and Stephen R. Elliott, “Similarity Between Amorphous and Crystalline Phases: The Case of TiO2,” J. Phys. Chem. Lett. 9, 2985–2990 (2018).

118 Brandon Anderson, Truong Son Hy, and Risi Kondor, “Cormorant: Covariant Molecular Neural Networks,” in NeurIPS (2019) p. 10.

119 William J. (William Jackson) Thompson, Angular momentum: an illustrated guide to rotational symmetries for physical systems (Wiley-VCH, 2004) p. 461.

120 Michael Kazhdan, Thomas Funkhouser, and Szymon Rusinkiewicz, “Rotation invariant spherical harmonic representation of 3d shape descriptors,” in Proceedings of the 2003 Eurographics/ACM SIGGRAPH Symposium on Geometry Processing, SGP ’03 (Eurographics Association, Goslar, DEU, 2003) p. 156–164.

121 A.P. Thompson, L.P. Swiler, C.R. Trott, S.M. Foiles, and G.J. Tucker, “Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials,” Journal of Computational Physics 285, 316–330 (2015).

122 Mitchell A. Wood and Aidan P. Thompson, “Extending the accuracy of the SNAP interatomic potential form,” The Journal of Chemical Physics 148, 241721 (2018).

123 Jigyasa Nigam, Sergey Pozdnyakov, and Michele Ceriotti, “Recursive evaluation and iterative contraction of N-body equivariant features,” J. Chem. Phys. 153, 121101 (2020).

124 Ralf Drautz, “Atomic cluster expansion of scalar, vectorial, and tensorial properties including magnetism and charge transfer,” Phys. Rev. B 102, 024104 (2020).

125 L. C. Biedenharn and J. D. Louck, The Racah-Wigner Algebra in Quantum Theory 1st ed. (Cambridge University Press, 1984).

126 Jörg Behler, “Neural network potential-energy surfaces in chemistry: A tool for large-scale simulations.” Phys. Chem. Chem. Phys. PCCP 13, 17930–55 (2011).

127 Chris M. Handley and Paul L. A. Popelier, “Dynamically Polarizable Water Potential Based on Multipole Moments Trained by Machine Learning,” J. Chem. Theory Comput. 5, 1474–1489 (2009).

128 Tristan Bereau, Denis Andrienko, and O. Anatole Von Lilienfeld, “Transferable Atomic Multipole Machine Learning Models for Small Organic Molecules,” J. Chem. Theory Comput. 11, 3225–3233 (2015).

129 Chungwen Liang, Gabriele Toacci, David M. Wilkins, Andrea Grisafi, Sylvie Roke, and Michele Ceriotti, “Solvent fluctuations and nuclear quantum effects modulate the molecular hyperpolarizability of water,” Phys. Rev. B 96, 041407 (2017).

130 Christoph Scherer, René Scheid, Denis Andrienko, and Tristan Bereau, “Kernel-Based Machine Learning for Efficient Simulations of Molecular Liquids,” J. Chem. Theory Comput. 16, 3194–3204 (2020).

131 Andrea Grisafi, David M. Wilkins, Michael J. Willatt, and Michele Ceriotti, “Atomic-Scale Representation and Statistical Learning of Tensorial Properties,” in Machine Learning in Chemistry, Vol. 1326, edited by Edward O. Pyzer-Knapp and Teodoro Laino (American Chemical Society, Washington, DC, 2019) pp. 1–21.

132 A J Stone, “Transformation between cartesian and spherical tensors,” Mol. Phys. 29, 1461–1471 (1975).

133 Andrea Grisafi, David M. Wilkins, Gábor Csányi, and Michele Ceriotti, “Symmetry-Adapted Machine Learning for Tensorial Properties of Atoms,” Phys. Rev. Lett. 120, 036002 (2018).

134 Nathaniel Raimbault, Andrea Grisafi, Michele Ceriotti, and Mariana Rossi, “Using Gaussian process regression to simulate the vibrational Raman spectra of molecular crystals,” New J. Phys. 21, 105001 (2019).

135 Aldo Glielmo, Peter Sollich, and Alessandro De Vita, “Accurate interatomic force fields via machine learning with covariant kernels,” Phys. Rev. B 95, 214302 (2017).

136 Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lu-Samm Yang, Li Li, Kai Kohlhoff, and Patrick Riley, “Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds,” arXiv preprint arXiv:1802.08219 (2018).

137 Andrea Grisafi, Jigyasa Nigam, and Michele Ceriotti, “Multi-scale approach for the prediction of atomic scale properties,” Chem. Sci. , 10.1039.D0SC04934D (2021).

138 Qi Yu and Joel M. Bowman, “Classical, Thermostated Ring Polymer, and Quantum VSCF/VCI Calculations of IR Spectra of H 7 O 3 + and H 9 O 4 + (Eigen) and Comparison with Experiment,” J. Phys. Chem. A 123, 1399–1409 (2019).
Michael Gastegger, Jörg Behler, and Philipp Marquetand, “Machine learning molecular dynamics for the simulation of infrared spectra,” Chem. Sci. 8, 6924–6935 (2017).

Max Veit, David M. Wilkins, Yang Yang, Robert A. DiStasio, and Michele Ceriotti, “Predicting molecular dipole moments by combining atomic partial charges and atomic dipoles,” J. Chem. Phys. 153, 024113 (2020).

Anders S. Christensen, Felix A. Faber, and O. Anatole von Lilienfeld, “Operators in quantum machine learning: Response properties in chemical space,” J. Chem. Phys. 150, 064105 (2019).

Yaolong Zhang, Sheng Ye, Jinxiao Zhang, Ce Hu, Jun Jiang, and Bin Jiang, “Efficient and Accurate Simulations of Vibrational and Electronic Spectra with Symmetry-Preserving Neural Network Models for Tensorial Properties,” J. Phys. Chem. B 124, 7284–7290 (2020).

Rohit Batra, Huan Doan Tran, Chiho Kim, James Richard F. W. Bader, Richard F. W. Bader, Tobias Morawietz, and Jörg Gregorius R. Medders, Volodymyr Babin, and Gábor Csányi, “Equation of State of Fluid Methane from First Principles with Machine Learning Potentials,” J. Chem. Theory Comput. 15, 2574–2586 (2019).

Derek P. Metcalf, Alexios Koutsoukas, Steven A. S. Alireza Ghasemi, Albert Hofstetter, Santanu Saha, Tristan Bereau, Robert A. DiStasio, Alexandre Tkatchenko, and Michele Ceriotti, “Fast and Accurate Uncertainty Approximation potentials: A brief tutorial introduction,” Int. J. Quantum Chem. 115, 1051–1057 (2015).

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl, “Neural message passing for quantum chemistry,” (PMLR, International Convention Centre, Sydney, Australia, 2017) pp. 1263–1272.

Stefan Chmiela, Alexander Tkatchenko, Huziel E. Sauceda, Igor Poltavsky, Kristof T. Schütt, and Klaus Robert Müller, “Machine learning of accurate energy-conserving molecular force fields,” Science Advances 3, e1603015 (2017), arXiv:1611.04678.

Gregory R. Bowman, Kyle A. Beauchamp, George Boxer, and Vijay S. Pande, “Progress and challenges in the automated construction of Markov state models for full protein systems,” J. Chem. Phys. 131, 124101 (2009).

J.M. Sanchez, F. Ducastelle, and D. Gràtias, “Generalized cluster description of multicomponent systems,” Physica A: Statistical Mechanics and its Applications 128, 334–350 (1984).

Ryan M. Richard and John M. Herbert, “A generalized many-body expansion and a unified view of fragment-based methods in electronic structure theory,” The Journal of Chemical Physics 137, 064113 (2012).

Richard F. W. Bader, Atoms in Molecules: A Quantum Theory, The International Series of Monographs on Chemistry No. 22 (Clarendon Press ; Oxford University Press, Oxford [England] : New York, 1994).

Kyle Mills, Kevin Ryczko, Iryna Luchak, Adam Domrady, Chris Beeler, and Isaac Tamblyn, “Extensive deep neural networks for transferring small scale learning to large scale systems,” Chemical Science 10, 4129–4140 (2019), arXiv:1708.06686.

Risi Kondor, “N-body Networks: A Covariant Hierarchical Neural Network Architecture for Learning Atomic Potentials,” (2018) arXiv:1803.01588.

Miguel A. Caro, “Optimizing many-body atomic descriptors for enhanced computational performance of machine learning based interatomic potentials,” Physical Review B 100 (2019), 10.1103/physrevb.100.024112 arXiv:1905.02142.

Aldo Glielmo, Claudio Zeni, and Alessandro De Vita, “Efficient nonparametric n-body force fields from ma-
V. Botu and R. Ramprasad, “Learning scheme to predict atomic forces and accelerate materials simulations,” Physical Review B - Condensed Matter and Materials Physics, 92, 094306 (2015).

C. Saunders, A. Gammerman, and V. Vovk, “Ridge Regression Learning Algorithm in Dual Variables,” Proceedings of the 15th International Conference on Machine Learning, 515–521 (1998).

Carl Edward Rasmussen and Christopher K Williams, Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning) (The MIT Press, 2005).

Volker L. Deriinger, Albert P. Bartók, Noam Bernstein, David M. Wilkins, Michele Ceriotti, Gábor Csányi, “Gaussian process regression for materials and molecules,” Submitted (2021).

I. G. Macdonald, Symmetric Functions and Hall Polynomials, reprinted in paperback ed., Oxford Classic Texts in the Physical Sciences (Clarendon Press, Oxford, 2015).

Berk Onat, Christoph Ortner, and James R. Kermode, “Sensitivity and dimensionality of atomic environment representations used for machine learning interatomic potentials,” J. Chem. Phys. 153, 144106 (2020).

Behnam Parsaeifard, Deb Sankar De, Anders Steen Christensen, Felix Andreas Faber, Emir Kocer, Sandip De, Jörg Behler, Anatole von Lilienfeld, and Stefan Goedecker, “An assessment of the structural resolution of various fingerprints commonly used in machine learning,” Mach. Learn.: Sci. Technol. (2020).

John I Yellott and Geoffrey J Iverson, “Uniqueness properties of higher-order autocorrelation functions,” J. Opt. Soc. Am. A, JOSAA 9, 388–404 (1992).

Ramakrishna Kakarla, “The Bispectrum as a Source of Phase-Sensitive Invariants for Fourier Descriptors: A Group-Theoretic Approach,” J. Math. Imaging Vis. 44, 341–353 (2012).

Ramakrishna Kakarla, “The Bispectrum as a Source of Phase-Sensitive Invariants for Fourier Descriptors: A Group-Theoretic Approach,” J Math Imaging Vis. 44, 341–353 (2012).

Atsuto Seko, Atsushi Togo, and Isao Tanaka, “Group-theoretical high-order rotational invariants for structural representations: Application to linearized machine learning interatomic potential,” Phys. Rev. B 99, 214108 (2019).

Yasushi Shibuta, Shintaro Tanaka, Tomohiro Takaki, and Munekazu Ohno, “Submicrometer-scale molecular dynamics simulation of nucleation and solidification from undercooled melt: Linkage between empirical interpretation and atomistic nature,” Acta Materialia 105, 328–337 (2016).

J MacQueen, “Some methods for classification and analysis of multivariate observations,” in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics (University of California Press, Berkeley, Calif., 1967) pp. 281–297.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in KDD 1996 (1996).

A Callisch, “Network and graph analyses of folding free energy surfaces,” Curr. Opin. Struct. Biol. 16, 71–78 (2006).

Jim Ruppert, Will Welch, and Ajay N. Jain, “Automatic identification and representation of protein binding sites for molecular docking,” Protein Sci. 6, 524–533 (2008).

Piero Gasparotto and Michele Ceriotti, “Recognizing molecular patterns by machine learning: An agnostic structural definition of the hydrogen bond,” J. Chem. Phys. 141, 174110 (2014).

Alex Rodríguez and Alessandro Laio, “Clustering by fast search and find of density peaks,” Science 344, 1492–1496 (2014).

Fionn Murtagh and Pedro Contreras, “Algorithms for hierarchical clustering: An overview. II,” Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 7, e1219 (2017).

Piero Gasparotto, Robert Horst Meillner, and Michele Ceriotti, “Recognizing Local and Global Structural Motifs at the Atomic Scale,” J. Chem. Theory Comput. 14, 486–498 (2018).

Fabio Pietrucci and Roman Martoňák, “Systematic comparison of crystalline and amorphous phases: Charting the landscape of water structures and transformations,” J. Chem. Phys. 142, 104704 (2015).

Pablo M. Piaggi and Michele Parrinello, “Predicting polymorphism in molecular crystals using orientational entropy,” Proc. Natl. Acad. Sci. 115, 10251–10256 (2018).

Leonid Kahle, Aris Marcolongo, and Nicola Marzari, “Modeling lithium-ion solid-state electrolytes with a pinball model,” Phys. Rev. Mater. 2, 065405 (2018).

Bingqing Cheng, Ryan-Rhys Griffiths, Simon Wengert, Christian Kunkel, Tamas Stenczel, Bo-nan Zhu, Volker L. Deriinger, Noam Bernstein, Johannes T. Margraf, Karsten Reuter, and Gabor Csányi, “Mapping Materials and Molecules,” Acc. Chem. Res. 53, 1981–1991 (2020).

R. R. Coifman, Stan T. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, “Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps,” Proc. Natl. Acad. Sci. U. S. A. 102, 7426–7431 (2005).

Andrew L Ferguson, Athanassios Z Panagiotopoulos, Pablo G Debernadetti, and Ioannis G Kevrekidis, “Systematic determination of order parameters for chain dynamics using diffusion maps,” Proc. Natl. Acad. Sci. U. S. A. 107, 13597–602 (2010).

Mary A. Rohrdanz, Wenwei Zheng, Mauro Maggioni, and Cecilia Clementi, “Determination of reaction coordinates via locally scaled diffusion map,” J. Chem. Phys. 134, 124116 (2011).

Rampi Ramprasad, Rohit Batra, Ghanshyam Pilania, Arun Mannodi-Kanakkithodi, and Chihoo Kim, “Machine learning in materials informatics: Recent ap-
plications and prospects,” Npj Comput. Mater. 3, 54 (2017).

Tobias Lenke and Christine Peter, “EncoderMap: Dimensionality Reduction and Generation of Molecule Conformations,” J. Chem. Theory Comput. 15, 1209–1215 (2019).

David Wales, Energy Landscapes: Applications to Clusters, Biomolecules and Glasses (Cambridge University Press, 2003).

Mary E. Karpen, Douglas J. Tobias, and Charles L. Brooks, “Statistical clustering techniques for the analysis of long molecular dynamics trajectories: Analysis of 2.2-ns trajectories of YPGDVs,” Biochemistry 32, 412–420 (1993).

Andrew E. Torda and Wilfred F. van Gunsteren, “Algorithms for clustering molecular dynamics configurations,” J. Comput. Chem. 15, 1331–1340 (1994).

Benjamin A. Helfrecht, Rocio Semino, Giovanni Pireddu, Scott M. Auerbach, and Michele Ceriotti, “A new kind of atlas of zeolite building blocks,” J. Chem. Phys. 151, 154112 (2019).

G.N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan, “Stereochemistry of polypeptide chain configurations,” J. of Molecular Biology 7, 95–99 (1963).

Dmitrij Frishman and Patrick Argos, “Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence,” Protein Eng. Des Sel 9, 133–142 (1996).

Wolfgang Kabsch and Christian Sander, “Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features,” Biopolymers 22, 2577–2637 (1983).

Fabio Pietrucci and Alessandro Laio, “A Collective Variable for the Efficient Exploration of Protein Beta-Sheet Structures: Application to SH3 and GB1,” J. Chem. Theory Comput. 5, 2197–2201 (2009).

Adriana Pietropaolo, Davide Branduardi, Massimiliano Bonomi, and Michele Parrinello, “A chirality-based metrics for free-energy calculations in biomolecular systems,” J. Comput. Chem. 32, 2627–2637 (2011).

Paul J Steinhardt, David R Nelson, and Marco Ronchetti, “Bond-orientational order in liquids and glasses,” Phys. Rev. B 28, 784–805 (1983).

Stefano Angioletti-Uberti, Michele Ceriotti, Peter D. Lee, and Mike W. Finnis, “Solid-liquid interface free energy through metadynamics simulations,” Phys. Rev. B - Condens. Matter Mater. Phys. 81, 125416 (2010).

Artem R. Oganov and Mario Valle, “How to quantify energy landscapes of solids,” The Journal of Chemical Physics 130, 104504 (2009).

Mario Valle and Artem R. Oganov, “Crystal fingerprint space – a novel paradigm for studying crystal-structure sets,” Acta Crystallogr A Found Crystalogr 66, 507–517 (2010).

Pablo M. Piaggi and Michele Parrinello, “Entropy based fingerprint for local crystalline order,” The Journal of Chemical Physics 147, 114112 (2017).

Angelo Ziletti, Devinder Kumar, Matthias Scheffler, and Luca M. Ghiringhelli, “Insightful classification of crystal structures using deep learning,” Nat. Commun. 9, 2775 (2018).

Fausto Martelli, Hsin-Yu Ko, Erdal C. Oğuz, and Roberto Car, “Local-order metric for condensed-phase environments,” Phys. Rev. B 97, 064105 (2018).

Walter Mickel, Sebastian C. Kapfer, Gerd E. Schröder-Turk, and Klaus Mecke, “Shortcomings of the bond orientational order parameters for the analysis of disordered particulate matter,” The Journal of Chemical Physics 138, 044501 (2013).

Michele Ceriotti, “Unsupervised machine learning in atomistic simulations, between predictions and understanding,” J. Chem. Phys. 150, 150901 (2019).

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller, “Nonlinear Component Analysis as a Kernel Eigenvalue Problem,” Neural Comput. 10, 1299–1319 (1998).

Marco Cuturi, “Positive Definite Kernels in Machine Learning,” ArXiv Prepr. ArXiv09115367 (2009).

Artem R. Oganov and Colin W. Glass, “Crystal structure prediction using ab initio evolutionary techniques: Principles and applications,” J. Chem. Phys. 124, 244704 (2006).

Maximilian Amsler and Stefan Goedecker, “Crystal structure prediction using the minima hopping method,” J. Chem. Phys. 133, 224104 (2010).

Chris J Pickard and Richard J Needs, “Ab initio random structure searching,” J. Phys. Condens. Matter 23, 053201 (2011).

Farren Curtis, Xiayue Li, Timothy Rose, Álvaro Vázquez-Mayagoitia, Saswata Bhattacharya, Luca M. Ghiringhelli, and Noa Marom, “GAtor: A First-Principles Genetic Algorithm for Molecular Crystal Structure Prediction,” J. Chem. Theory Comput. 14, 2246–2264 (2018).

Artem R. Oganov, Chris J Pickard, Qiang Zhu, and Richard J. Needs, “Structure prediction drives materials discovery,” Nat Rev Mater 4, 331–348 (2019).

Grégoire Ferré, Jean-Bernard Maillet, and Gabriel Stoltz, “Permutation-invariant distance between atomic configurations,” The Journal of Chemical Physics 143, 104114 (2015).

Grégoire Montavon, Matthias Rupp, Vivekanand Gobre, Álvaro Vázquez-Mayagoitia, Katja Hansen, Alexandre Tkatchenko, Klaus Robert Müller, and O. Anatole Von Lilienfeld, “Machine learning of molecular electronic properties in chemical compound space,” New J. Phys. 15, 095003 (2013).

Jack Yang, Sandip De, Josh E. Campbell, Sean Li, Michele Ceriotti, and Graeme M. Day, “Large-Scale Computational Screening of Molecular Organic Semiconductors Using Crystal Structure Prediction,” Chem. Mater. 30, 4361–4371 (2018).

Benjamin A Helfrecht, Rose K Cersonsky, Guillaume Gobre, Alvaro Vazquez-Mayagoitia, Saswata Bhattacharya, Luca M. Ghiringhelli, and Noa Marom, “GAtor: A First-Principles Genetic Algorithm for Molecular Crystal Structure Prediction,” J. Chem. Theory Comput. 14, 2246–2264 (2018).

Artem R. Oganov, Chris J Pickard, Qiang Zhu, and Richard J. Needs, “Structure prediction drives materials discovery,” Nat Rev Mater 4, 331–348 (2019).

Grégoire Ferré, Jean-Bernard Maillet, and Gabriel Stoltz, “Permutation-invariant distance between atomic configurations,” The Journal of Chemical Physics 143, 104114 (2015).

Grégoire Montavon, Matthias Rupp, Vivekanand Gobre, Álvaro Vázquez-Mayagoitia, Katja Hansen, Alexandre Tkatchenko, Klaus Robert Müller, and O. Anatole Von Lilienfeld, “Machine learning of molecular electronic properties in chemical compound space,” New J. Phys. 15, 095003 (2013).

Jack Yang, Sandip De, Josh E. Campbell, Sean Li, Michele Ceriotti, and Graeme M. Day, “Large-Scale Computational Screening of Molecular Organic Semiconductors Using Crystal Structure Prediction,” Chem. Mater. 30, 4361–4371 (2018).

Benjamin A Helfrecht, Rose K Cersonsky, Guillaume Gobre, and Michele Ceriotti, “Structure-property maps with Kernel principal covariates regression,” Mach. Learn.: Sci. Technol. 1, 045021 (2020).

H. W. Kuhn, “The Hungarian method for the assignment problem,” Nav. Res. Logist. Q. 2, 83–97 (1955).

Marco Cuturi, “Sinkhorn Distances: Lightspeed Computation of Optimal Transport,” in Advances in Neural Information Processing Systems 32, edited by C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger (Curran Associates, Inc., 2019) pp.
227 Benjamin A. Helfrecht, Piero Gasparotto, Federico Giberti, and Michele Ceriotti, “Atomic motif recognition in (bio)polymers: Benchmarks from the protein data bank,” Front. Mol. Biosci. 6, 1–14 (2019).
228 Sijmen de Jong and Henk A.L. Kiers, “Principal covariates regression,” Chemometrics and Intelligent Laboratory Systems 14, 155–164 (1992).
229 Laurens van der Maaten and Geoffrey Hinton, “Visualizing data using t-SNE,” J. Mach. Learn. Res. 9, 2579–2605 (2008).
230 Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole Von Lilienfeld, “Quantum chemistry structures and properties of 134 kilo molecules,” Sci. Data 1, 1–7 (2014).
231 Matti Ropo, Markus Schneider, Carsten Baldauf, and Volker Blum, “First-principles data set of 45,892 isolated and cation-coordinated conformers of 20 proteinogenic amino acids,” Sci. Data 3, 160009 (2016).
232 Christoph Schober, Karsten Reuter, and Harald Oberhofer, “Virtual Screening for High Carrier Mobility in Organic Semiconductors,” J. Phys. Chem. Lett. 7, 3973–3977 (2016).
233 James P. Stewart, “Optimization of parameters for semiempirical methods VI: More modifications to the NDDO approximations and re-optimization of parameters,” J. Mol Model 19, 1–32 (2013).
234 Felix A. Faber, Luke Hutchison, Bing Huang, Justin Gilmer, Samuel S. Schoenholz, George E. Dahl, Oriol Vinyals, Steven Kearnes, Patrick F. Riley, and O. Anatole von Lilienfeld, “Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error,” J. Chem. Theory Comput. 13, 5255–5264 (2017).
235 Yuxing Zuo, Chi Chen, Xiangguo Li, Zhi Deng, Yiming Chen, Gábor Csányi, Alexander V. Shapeev, Aidan P. Thompson, Mitchell A. Wood, and Shyne Ping Ong, “Performance and Cost Assessment of Machine Learning Interatomic Potentials,” J. Phys. Chem. A., 9808723 (2020).
236 Alexander Goscinski, Guillaume Fraux, and Michele Ceriotti, “The role of feature space in atomistic learning,” arxiv:2009.02741 (2020).
237 Sergey Pozdnyakov, Michael Willatt, and Michele Ceriotti, “Dataset: Randomly-displaced methane configurations,” (2020).
238 Chris J. Pickard, “AIRSS data for carbon at 10GPa and the C+N+H+O system at 1GPa,” (2020).
239 Rose K. Cersosky, Benjamin A. Helfrecht, Edgar A. Engel, and Michele Ceriotti, “Improving sample and feature selection with principal covariates regression,” arxiv:2012.12253 (2020).
240 Y Eldar, M. Lindembaum, M Porat, and Y Y Zeevi, “The farthest point strategy for progressive image sampling,” IEEE Trans. Image Process. Publ. IEEE Signal Process. Soc. 6, 1305–15 (1997).
241 Michele Ceriotti, Gareth A. Tribello, and Michele Parrinello, “Demonstrating the transferability and the descriptive power of sketch-map,” J. Chem. Theory Comput. 9, 1521–1532 (2013).
242 Michael W. Mahoney and Petros Drineas, “CUR matrix decompositions for improved data analysis,” Proc. Natl. Acad. Sci. U. S. A. 106, 697–702 (2009).
243 Giulio Imbalzano, Andrea Anelli, Daniele Giófré, Sinja Kleess, Jörg Behler, and Michele Ceriotti, “Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials,” J. Chem. Phys. 148, 241730 (2018).
244 Miguel A. Morales, Carlo Pierleoni, Eric Schwegler, and D. M. Ceperley, “Evidence for a first-order liquid-liquid transition in high-pressure hydrogen from ab initio simulations,” Proc. Natl. Acad. Sci. U. S. A. 107, 12799–12803 (2010).
245 Edgar A. Engel, Andrea Anelli, Albert Hofstetter, Federico Paruzzo, Lyndon Emsley, and Michele Ceriotti, “A Bayesian approach to NMR crystal structure determination,” Phys. Chem. Chem. Phys. 21, 23385–23400 (2019).
246 Chiheb Ben Mahmoud, Andrea Anelli, Gábor Csányi, and Michele Ceriotti, “Learning the electronic density of states in condensed matter,” Phys. Rev. B 102, 235130 (2020).
247 Hao Gao, Junjie Wang, and Jian Sun, “Improve the performance of machine-learning potentials by optimizing descriptors,” J. Chem. Phys. 150, 244110 (2019).
248 Michael J. Willatt, Michele Ceriotti, and Stuart C. Althorpe, “Approximating Matsubara dynamics using the planetary model: Tests on liquid water and ice,” J. Chem. Phys. 148, 102336 (2015).
249 Felix A. Faber, Alexander Lindmaa, O. Anatole Von Lilienfeld, and Rickard Armiento, “Machine Learning Energies of 2 Million Elpasolite (ABC2D6) Crystals,” Phys. Rev. Lett. 117, 135502 (2016).
250 Nongnuch Artrith, Alexander Urban, and Gerbrand Ceder, “Efficient and accurate machine-learning interpolation of atomic energies in compositions with many species,” Phys. Rev. B 96, 014112 (2017).
251 M. Gastegger, L. Schwiedrzik, M. Bittermann, F. Berzsenyi, and P. Marquetand, “wACSF—Weighted atom-centered symmetry functions as descriptors in machine learning potentials,” J. Chem. Phys. 148, 241709 (2018).
252 Samare Rostami, Maximilian Amsler, and S. Alireza Ghasemi, “Optimized symmetry functions for machine-learning interatomic potentials of multicomponent systems,” The Journal of Chemical Physics 149, 124106 (2018).
253 A knockout mice is a genetically-modified mouse in which one or more genes have been inactivated. The effect on the development of the animal can be used to understand the role played by the affected gene(s). Aditya Kamath, Rodrigo A. Vargas-Hernández, Roman V. Krems, Tucker Carrington, and Sergei Manzhos, “Neural networks vs Gaussian process regression for representing potential energy surfaces: A comparative study of fit quality and vibrational spectrum accuracy,” Journal of Chemical Physics 148, 241702 (2018).
254 Andreas Singraber, Tobias Morawietz, Jörg Behler, and Christoph Dellago, “Parallel Multistream Training of High-Dimensional Neural Network Potentials,” J. Chem. Theory Comput. 15, 3075–3092 (2019).
255 Xiang Gao, Farhad Ramezanhorbani, Oleandr Isayev, Justin S. Smith, and Adrian E. Roitberg, “TorchANI: A Free and Open Source PyTorch-Based Deep
K. Kaufmann and W. Baumeister, “Single-centre ex-

260

259

258

Denghui Lu, Han Wang, Mohan Chen, Jiduan Liu, Lin
Lin, Roberto Car, Weinan E, Weile Jia, and Linfeng
Zhang, “86 PFLOPS Deep Potential Molecular Dy-
namics simulation of 100 million atoms with ab initio
accuracy,” (2020) arXiv:2004.11658.

257

256

255

254

Sergey Pozdnyakov, Artem R. Oganov, Arslan Maz-
itov, Ivan Kruglov, and Efim Mazhnik, “Fast
general two- and three-body interatomic potential,”
arxiv:1910.07513 (2019).

253

252

251

250

249

248

247

246

245

244

243

242

241

240

239

238

237

236

235

234

233

232

231

230

229

228

227

226

225

224

223

222

221

220

219

218

217

216

215

214

213

212

211

210

209

208

207

206

205

204

203

202

201

200

199

198

197

196

195

194

193

192

191

190

189

188

187

186

185

184

183

182

181

180

179

178

177

176

175

174

173

172

171

170

169

168

167

166

165

164

163

162

161

160

159

158

157

156

155

154

153

152

151

150

149

148

147

146

145

144

143

142

141

140

139

138

137

136

135

134

133

132

131

130

129

128

127

126

125

124

123

122

121

120

119

118

117

116

115

114

113

112

111

110

109

108

107

106

105

104

103

102

101

100

99

98

97

96

95

94

93

92

91

90

89

88

87

86

85

84

83

82

81

80

79

78

77

76

75

74

73

72

71

70

69

68

67

66

65

64

63

62

61

60

59

58

57

56

55

54

53

52

51

50

49

48

47

46

45

44

43

42

41

40

39

38

37

36

35

34

33

32

31

30

29

28

27

26

25

24

23

22

21

20

19

18

17

16

15

14

13

12

11

10

9

8

7

6

5

4

3

2

1

0
Patrick Rowe, Volker L. Deringer, Piero Gasparotto, Gábor Csányi, and Angelos Michaelides, “An accurate and transferable machine learning potential for carbon,” J. Chem. Phys. 153, 034702 (2020)

GA A Gallet and Fabio Pietrucci, “Structural cluster analysis of chemical reactions in solution,” J. Chem. Phys. 139, 74101 (2013)

Sandip De, Félix Musil, Teresa Ingram, Carsten Baldauf, and Michele Ceriotti, “Mapping and classifying molecules from a high-throughput structural database,” J. Cheminformatics 9, 1–14 (2017)

Dmitrii Maksimov, Carsten Baldauf, and Mariana Rossi, “The conformational space of a flexible amino acid at metallic surfaces,” Int J Quantum Chem (2020), 10.1002/qqua.26369

Félix Musil, Sandip De, Jack Yang, Joshua E. Campbell, Graeme M. Day, and Michele Ceriotti, “Machine learning for the structure-energy-property landscapes of molecular crystals,” Chem. Sci. 9, 1289–1300 (2018)

Ganna Gryno’va, Kun-Han Lin, and Clémentine Corminboeuf, “Read between the Molecules: Computational Insights into Organic Semiconductors,” J. Am. Chem. Soc. 140, 16370–16386 (2018).

Jack Yang, “Mapping temperature-dependent energy–structure–property relationships for solid solutions of inorganic halide perovskites,” J. Mater. Chem. C, 10.1039/D0TC04515B (2020)

Tim Würger, Christian Feiler, Félix Musil, Gregor B. V. Feldbauer, Daniel Höche, Svatlana V. Lamaka, Mikhail L. Zheludevich, and Robert H. Meißner, “Data Science Based Mg Corrosion Engineering,” Front. Mater. 6, 53 (2019)

Tristan A. Sharp, Spencer L. Thomas, Ekin D. Cubuk, Samuel S. Schoenholz, David J. Srolovitz, and Andrea J. Liu, “Machine learning determination of atomic dynamics at grain boundaries,” Proc Natl Acad Sci USA 115, 10943–10947 (2018)

Jonathan L. Priedeman, Conrad W. Rosenbrook, Oliver K. Johnson, and Eric R. Homer, “Quantifying and connecting atomic and crystallographic grain boundary structure using local environment representation and dimensionality reduction techniques,” Acta Materialia 161, 431–443 (2018)

Eric R. Homer, Derek M. Hensley, Conrad W. Rosenbrook, Andrew H. Nguyen, and Gus L. W. Hart, “Machine-Learning Informed Representations for Grain Boundary Structures,” Front. Mater. 6, 168 (2019)

Piero Gasparotto, Davide Bochicchio, Michele Ceriotti, and Giovanni M. Pavan, “Identifying and Tracking Defects in Dynamic Supramolecular Polymers,” J. Phys. Chem. B 124, 589–599 (2020)

C. Dietz, T. Kretz, and M. H. Thoma, “Machine-learning approach for local classification of crystalline structures in multphase systems,” Phys. Rev. E 96, 011301 (2017)

Maxwell Fulford, Matteo Salvaglio, and Carla Molteni, “Deep Ice: A Deep Neural Network Approach To Identify Ice and Water Molecules,” J. Chem. Inf. Model. 59, 2141–2149 (2019)

Aleks Reinhardt, Chris J. Pickard, and Bingqing Cheng, “Predicting the phase diagram of titanium dioxide with random search and pattern recognition,” Phys. Chem. Chem. Phys. 22, 12697–12705 (2020)

Yasemin Basdogan, Mitchell C. Groenenboom, Ethan Henderson, Sandip De, Susan B. Rempe, and John A. Keith, “Machine Learning-Guided Approach for Studying Solvation Environments,” J. Chem. Theory Comput. 16, 633–642 (2020)

Bartomeu Monserrat, Jan Gerit Brandenburg, Edgar A. Engel, and Bingqing Cheng, “Liquid water contains the building blocks of diverse ice phases,” Nat Commun 11, 5757 (2020)

Volker L. Deringer, Miguel A. Caro, and Gábor Csányi, “A general-purpose machine-learning force field for bulk and nanostructured phosphorus,” Nat Commun 11, 5461 (2020)

Chao Shen, Junjie Ding, Zhe Wang, Dongsheng Cao, Xiaoxin Ding, and Tingjun Hou, “From machine learning to deep learning: Advances in scoring functions for protein–ligand docking,” WIREs Computational Molecular Science 10 (2020), 10.1002/wcms.1429.

Siavosh Dustmalchi, Maryam Hamzeh-Mivehroud, and Karim Asadpour-Zeynali, “Comparison of different 2D and 3D-QSAR methods on activity prediction of histamine H3 receptor antagonists,” Iranian Journal of Pharmaceutical Research 11, 97–108 (2012)

Karolina Jagiello, Monika Grzonkowska, Marta Swirog, Lucky Ahmed, Bakhtiyor Rasulev, Agge los Avramopoulos, Manthos G. Papadopoulos, Jerzy Leszczynski, and Tomasz Puzyn, “Advantages and limitations of classic and 3D QSAR approaches in nano-QSAR studies based on biological activity of fullerene derivatives,” Journal of Nanoparticle Research 18, 256 (2016)

Kamal Choudhary, Brian DeCost, and Francesca Tavazza, “Machine learning with force-field-inspired descriptors for materials: Fast screening and mapping energy landscape,” Phys. Rev. Materials 2, 083801 (2018)

Víctor E. Kuz’mín, Anatoly G. Artemenko, Pavel G. Polischuk, Eugene N. Muratov, Alexander I. Hromov, Anatoly V. Liahovskiy, Sergey A. Andronati, and Svetlana Yu. Makan, “Hierarchical system of QSAR models (1D-4D) on the base of simplex representation of molecular structure,” Journal of Molecular Modeling 11, 457–467 (2005)

Dmitry Zankov, V., Maria Matveieva, Aleksandra Nikonenko, Ramil Nugmanov, Alexandre Varnek, Pavel Polischuk, and Timur Madzhidov, “QSAR Modeling Based on Conformation Ensembles Using a Multi-Instance Learning Approach,” (2020), 10.26434/CHEMRXIV.13456277.V1

Jan Weinreich, Nicholas J. Browning, and O. Anatole von Lilienfeld, “Machine Learning of Free Energies in Chemical Compound Space Using Ensemble Representations: Reaching Experimental Uncertainty for Solvation,” (2020), arXiv:2012.09722.

Simon Axelrod and Rafael Gomez-Bombarelli, “Molecular machine learning with conformer ensembles,” (2020) arXiv:2012.08452