Physics-inspired structural representations for molecules and materials

Felix Musil,[1,2] Andrea Grisafi,[1] Albert P. Bartók,[3] Christoph Ortner,[4] Gábor Csányi,[5] and Michele Ceriotti[1,2,*]

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

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

[3] *Department of Physics and Warwick Centre for Predictive Modelling, School of Engineering, University of Warwick, Coventry CV4 7AL, United Kingdom*

[4] *University of British Columbia, University of British Columbia, Vancouver, BC, Canada V6T 1Z2*

[5] *Engineering Laboratory, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, United Kingdom*

(Dated: July 18, 2021)

The first step in the construction of a regression model or a data-driven analysis, aiming to predict or elucidate the relationship between the atomic scale structure of matter and its properties, involves transforming the Cartesian coordinates of the atoms into a suitable *representation*. The development of atomic-scale representations has played, and continues to play, a central role in the success of machine-learning methods for chemistry and materials science. This review summarizes the current understanding of the nature and characteristics of the most commonly used structural and chemical descriptions of atomistic structures, highlighting the deep underlying connections between different frameworks, and the ideas that lead to computationally efficient and universally applicable models. It emphasizes the link between properties, structures, their physical chemistry and their mathematical description, provides 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

The last decade has seen a tremendous increase in the use of data-driven approaches for the modeling of molecules and materials. Atomistic simulation has been a particularly fertile field of use; applications range from the analysis of large databases of materials properties,[1] to the design of molecules with the desired behavior for a given application.[2] Machine learning techniques have been applied to devise coarse-grained descriptions of complex molecular systems,[3–9] to build accurate and comparatively inexpensive interatomic potentials,[10–18] and more generally to predict, or rationalize, the relationship between a specific atomic configuration and the properties that can be computed by electronic-structure calculations[19–26].

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 $\{\mathbf{r}_i, a_i\}$, and possibly by the basis vectors of the periodic repeat unit $\mathbf{h}$ – 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 (fit) a 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 used, usually interchangeably, in chemical and materials informatics to indicate heuristically-determined properties that are easier to compute than the quantities one ultimately wants to predict, but correlate strongly with them, facilitating the construction of transferable and accurate models.[27] 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[10,28], Coulomb matrices[19], and the smooth overlap of atomic positions (SOAP)[29]. It is important to note that representations can be expressed using different mathematical entities. In the most straightforward realisation, the space of features takes 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 describe the relationship between pairs of configurations. Such relationship can be expressed in terms of a kernel function $\mathrm{k}(A, A')$ (e.g. the scalar product between feature vectors), or in terms of a distance between configurations $\mathrm{d}(A, A')$ (e.g. the Euclidean distance between associated features). 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, and so can be seen as equivalent to a representation of individual structures, even in cases in which the distance or the kernel are not explicitly computed from a pair of feature vectors.

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[30–32]. 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. It is part of a special issue that covers the many facets of the application of machine learning to chemical simulations, and the interested reader may find, among others, discussions of machine learning models based on Gaussian process regression, using some of the descriptors we discuss here[33], of the construction of potentials for molecules[34,35] and materials[36], the description of excited states[37], and of unsupervised machine learning schemes[38]. 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
$\mathbf{r}_i$	Position of the i-th atom
$\mathbf{r}_{ji}$	Vector separating the i-th atom and its j-th neighbor, $\mathbf{r}_j - \mathbf{r}_i$
Q	Generic continuous index enumerating the components of an atomic representation
q	Generic discrete index enumerating the components of an atomic representation
$\langle Q\vert A\rangle$	A representation of a structure A indexed by an unspecified label or set of labels Q
$\boldsymbol{\xi}(A_i)$	Feature vector (with elements indexed by q) associated with an atom-centered environment, $\xi_q(A_i) = \langle q\vert A_i\rangle$
$\boldsymbol{\Xi}$	Feature matrix combining the features associated with multiple structures/environments
$\boldsymbol{\chi}_q$	A column in a feature matrix, where $(\boldsymbol{\chi}_q)_i = \xi_q(A_i)$
$y(A_i)$	An atom-centered property, or its systematic approximation in terms of an atom-centered representation $\vert A_i\rangle$
$\tilde{y}(\boldsymbol{\xi})$	A non-linear model that approximates $y(A_i)$ using the feature vector $\boldsymbol{\xi}(A_i)$
$\mathrm{k}(A, A')$	A (non-)linear kernel computed between two structures or environments, represented by the corresponding feature vectors $\boldsymbol{\xi}(A)$
$\mathrm{d}(A, A')$	A distance computed between two structures or environments
$\vert\rho\rangle$	Structure representation based on a smooth atom density
$\vert\rho_i\rangle$	Representation of an environment centered on atom i, that can be obtained by symmetrizing $\vert\rho\rangle$ over translations
$\vert\overline{\rho_i^{\otimes\nu}}\rangle$	Symmetrized ν-point correlation of the atomic density built on the atom-centered representation $\vert\rho_i\rangle$
$\vert\delta\rangle$	Dirac-δ limit of the smooth atom density $\vert\rho\rangle$. Analogous symmetrized versions are indicated as $\vert\delta_i\rangle$ and $\vert\overline{\delta_i^{\otimes\nu}}\rangle$
$\vert V\rangle$	Atom-density field representation, suitable to describe long-range correlations

III. REPRESENTATIONS FOR MATERIALS AND MOLECULES

Even though this review has no intention of providing an exhaustive historical account 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,[39] 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[40] descriptors of molecular structure – based on cheminformatics fingerprints,[41,42]chemical-intuition driven descriptors[43], molecular graphs,[44] or indicators obtained from quantum chemical calculations[45] – to the behavior of a selected compound, usually focusing on properties of direct applicative interest[46–48] such as solubility, toxicity,[49] or pharmacological activity.[50,51]

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 simulations[52–55] 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[56,57]. 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 de-

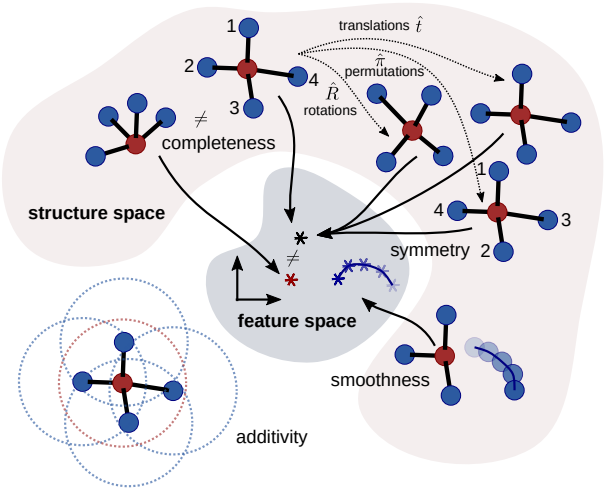


FIG. 1. A schematic overview of the requirements for an effective structural representation. The mapping between structures and feature space should obey fundamental physical symmetries (equivalent structures should be mapped to the same features); should be complete (inequivalent structures should be mapped to distinct features); should be smooth (continuous deformations of a structure should map to a smooth deformation of the associated features). Furthermore, whenever dealing with datasets that are not homogeneous in molecular size, the representation should be additive: a structure should be decomposed in a sum of local environments (usually atom-centered), ensuring transferability and extensivity of predictions.

scribe chemical bonds with Coulomb and $1/r^6$ terms to describe electrostatics and dispersion. Their (few) parameters were determined by matching the values of experimental observables, such as cohesive energies, lattice vectors and elastic constants. The continuous increase in computational power, and the availability of electronic-structure techniques with a better cost-accuracy ratio[58–60] has made it possible to compute extremely accurate energies and properties of specific configurations. This has opened the way to *ab initio* simulations of materials[55], but also provided a viable alternative to empirical functional forms for the construction of interatomic potentials. Starting from the simplest compounds[61], and then gradually increasing in complexity[62], 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[61] – they can predict experiments[63] two decades before measurements become precise enough to verify the theoretical values.[64]

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[56,57,65] aimed at computing averages over the appropriate classical (or quantum[66,67]) distribution of atomic configurations. As a consequence, the representations used as inputs of these surrogate quantum models are usually rather generic, constructed based exclusively on atomic coordinates and chemical species. They aim to establish a precise mapping between a specific structure and the associated atomic-scale quantities, and for this reason have also proven very useful to *analyze* atomistic configurations[68–70], an application we discuss in detail in Section VII. Even though we focus our discussion on this latter class of features, it is worth mentioning the recent, and rather successful, attempts to use descriptors that incorporate information from electronic-structure calculations, that we briefly summarize in Section IX G.

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 en-

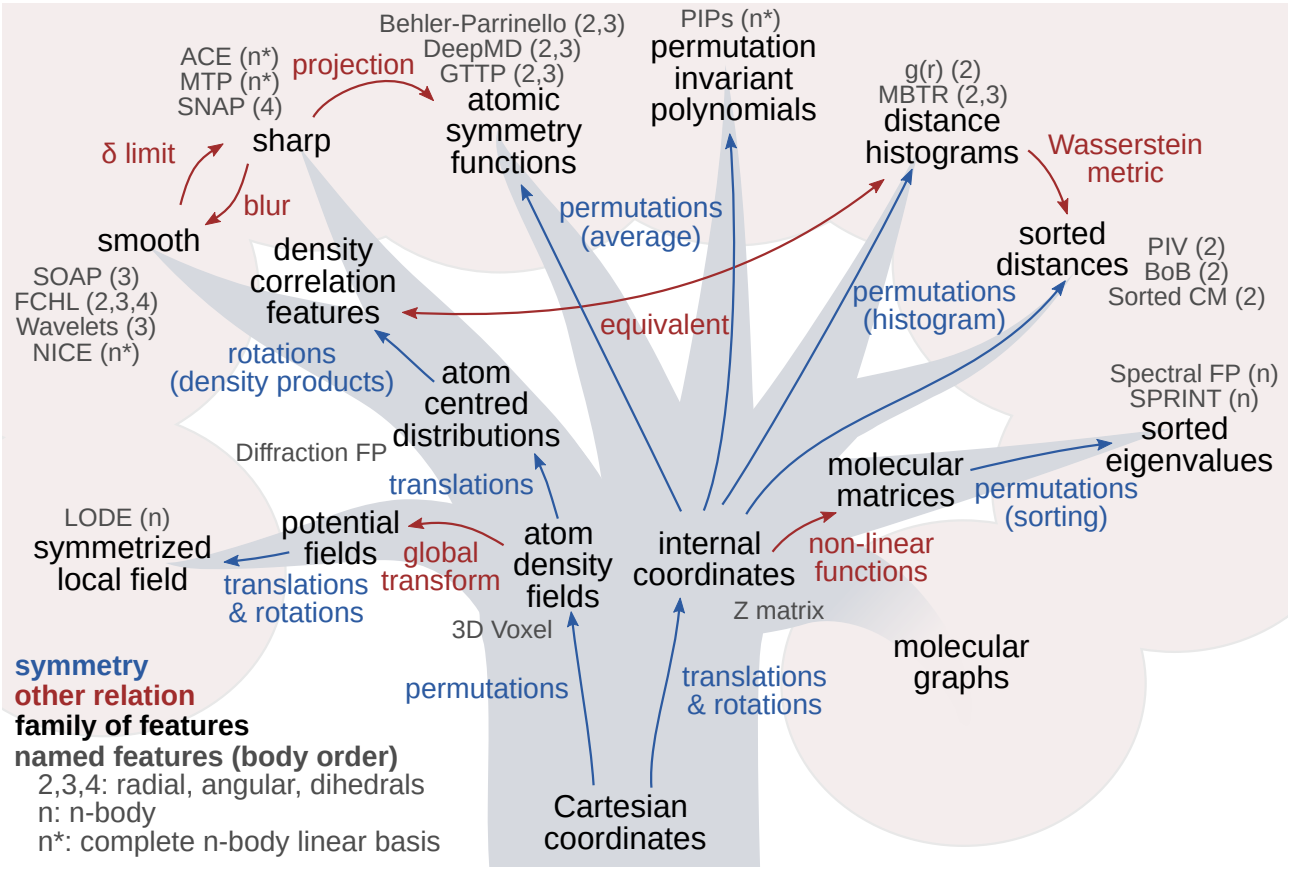


FIG. 2. A phylogenetic tree of structural representations for materials and molecules. Arrows indicate the relationship between different groups of features. Lists of names, in gray, indicate the most common implementations for each class. Classes that appear as "leaves" of the tree are fully symmetric.

vironments are usually associated with higher transferability, reflecting a "divide and conquer" approach to materials modeling[71,72]. 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 VIII.

A. Symmetry

The Cartesian coordinates of the atoms encode all the information that is needed to reconstruct the geometry of a structure. Yet, it is obvious that they cannot be used directly as the input of a regression model. The fact that the Cartesian description of a molecule depends on its absolute position and orientation in space, and the order by which atoms are listed, means that configurations that are completely equivalent can be represented by many different Cartesian values, which makes any regression, classification or clustering scheme inefficient and potentially misleading. Over the years, many different approaches have been proposed by which translations, rotations, inversion and atom permutation symmetries can be enforced, which is reflected in the variety of alternative frameworks to achieve an effective representation to be used of the input of an atomistic machine-learning scheme. In fact, symmetry is such a central principle underpinning these efforts that it can be used to construct a "phylogenetic tree" of representations, organized according to the strategy that is used to incorporate symmetry in their construction, as shown in Figure 2.

The need to remove the trivial symmetries, namely the dependency of the Cartesian coordinates on the origin and orientation of the reference system, has been recognized very early in the field of chemical and materials modeling. Different sets of internal coordinates[73] (bonds, angles, torsions) have been proposed, based on chemical intuition, as invariant descriptors of molecular geometry, and most of the molecular forcefields that have been so effective in the modeling of biological systems[74–77] rely on internal coordinates to define bonded interactions. A collection of internal coordinates that is sufficient to fully characterize the geometry of a structure, often referred-to as the Z-matrix, is a paradigmatic example of this class of representations. Even though the efficiency of this approach has often been questioned[78,79], particularly because there is no unique way to define the Z-matrix, internal coordinates are still ubiquitous, and

are effective whenever the system being studied has a well-defined, persistent bonding pattern (see Ref. 80 for a recent review). In these cases, internal coordinates can be seen as the initial step in the construction of discretized molecular representations, such as a molecular graph. Even though very widely used in chemical machine learning[2,81], these graph based schemes are not meant to describe the exact arrangement of the atoms, but just their bonding pattern, and so fall outside the scope of this review.

The limitations of an internal-coordinates description become most apparent when one wants to model a chemically-active system, as the bonding patterns can change during the course of a simulation, and therefore the invariance to atom index permutations becomes crucial to achieve a consistent model. The Empirical Valence Bond (EVB) method[82] has been used to simulate 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)[11,83–86] 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. It is worth mentioning that the problem of fitting molecular potential energy surfaces, particularly for applications to gas-phase physical chemistry, has led to approaches that anticipate several of the ideas that have become central to modern machine-learning techniques: the need to symmetrize appropriately atomic structures,[87] the systematic fitting to databases of configurations computed with high levels of quantum chemistry,[61] and even the use of "neural network potentials"[88,89] are just a few examples of the pioneering contributions from this field.

In the condensed phase, a similar pioneering role was played by the construction of systematic expansions of the potential energy of alloys[90], and of bond order potentials based on the moments of the density of states[91–93]. Both anticipate the use of an atom-centered description of the energy, the role of symmetry, and the notion of building a systematic expansion of the target property in terms of a convergent hierarchy of terms of increasing complexity. The first successful attempt of explicitly bringing machine-learning ideas to the construction of interatomic potentials for condensed-phase materials can be attributed to Behler and Parrinello, who in Ref. 10 introduced the concept of atom-centered symmetry functions (ACSF), which rely on a local expansion of the energy and on the construction of a symmetric description of atomic environments. 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 III 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[94]. Indeed, the recently proposed *atomic PIPs*[95] 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 $q_i q_j / r_{ji}$ between each atomic pair in a structure, have been extensively explored in early applications of the machine learning of molecular properties[19], with the main limitation being connected to the lack of permutation invariance[96], which has also been tackled by approximate symmetrization, summing over a manageable number of randomized orderings of the atoms[97,98]. 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 $\sum_i g(\mathbf{x}-\mathbf{r}_i)$, obtained by summing over localized functions centered on the positions $\mathbf{r}_i$ 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[97,99], an approach that has also been used with permutation invariant vectors (PIV)[100], "bag of bonds" features (BoB)[101]. Similar descriptors based on sorted distances have been also used to identify recurring structures in structure optimization algorithms[102,103], and more recently generalized to lexicographically-sorted lists of k-neighbors distances[104]. Computing the eigenvalues of (functions of) interatomic distances, which underlies the SPRINT method[105] as well the overlap matrix eigenvalue fingerprints[68,106], 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[107] 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 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, $f = \sum_i c_i \left(r_i - r_i^0\right)^2$ parameterized by $\mathbf{c}$ and $\mathbf{r}^0$. 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 $\mathbf{c}$ and $\mathbf{r}^0$ 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[108]. 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. 109), which builds a formal bridge between conceptually unrelated families of atomic-scale representations.

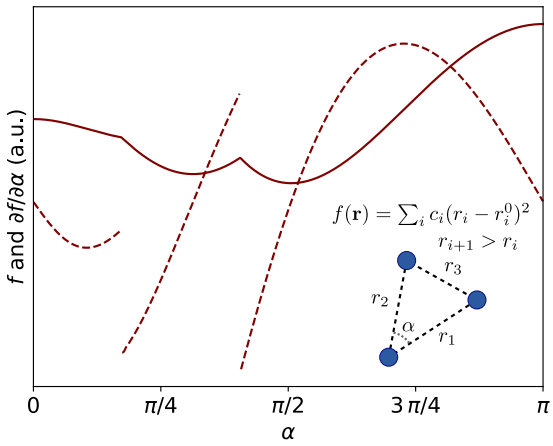


FIG. 3. Toy model demonstrating a non-smooth property (solid line) and its discontinuous derivative (dashed line) that are defined as functions of the ordered list of interatomic distances for a three-atom cluster.

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.[61,110,111] 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[10] 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 implicit in the functional forms of most interatomic potentials, has far reaching consequences in terms of the data efficiency of the model, as discussed in Subsection VIII C. 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[112], the additivity assumption breaks down the problem of predicting the properties of a complex structure into simpler, short-range problems. An additive decomposition is also the most straightforward way to ensure extensivity of predictions[113], 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[19,31,114,115], most of the recent approaches have moved to additive, atom-centered representations[116,117], that yield more accurate and transferable models, at least for extensive properties[118]. Oftentimes it is possible, and relatively straightforward, to modify a global representation to describe an atom-centered environment[68,95,119], or to combine atom-centered representations to build a global description[69], 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. In fact, one could regard the list of atom-centered features for all the atoms in a structure as an *equivariant* global representation of the structure – one in which the entries in the feature vector transform according to the permutation of the atomic indices. This notion underlies for instance the concept of self-attention[120,121], which has been very fruitfully applied in the construction of neural networks and models for cheminformatics. The connection between symmetry, 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[29,122,123]. More recently, counterexamples 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. 124 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 (largely open) 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.[109,125–127] In this Section we summarize the construction of a 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 V.

A. Dirac notation for atomic representations

We formalize a notation, that extends the one introduced in Refs. 109,125 and used in Ref. 128 to compare different kinds of local and global representations, which expresses the feature vectors associated with the representation of a structure in a way that mimics Dirac notation in quantum mechanics. At the most basic level, this notation can be seen as a way to indicate expressively the nature of the representation used, and to tidily enumerate the components of the associated feature vector. Much like in the quantum case, the real value of the formalism is that it emphasizes the basis-set independence of the class of representations we concentrate on, and that it provides visual cues that help recognizing at a glance the linear operations that occur in the construction and manipulation of the feature vectors, and of the models built on them.[129] We will use this notation consistently throughout this review as a neutral medium to express general results that reflect concepts shared by many of the most widespread representation, but occasionally make a link to the different notations that have become established to describe specific frameworks.

Representations in bra-ket notation We use a ket $|A\rangle$ to indicate an abstract feature vector associated

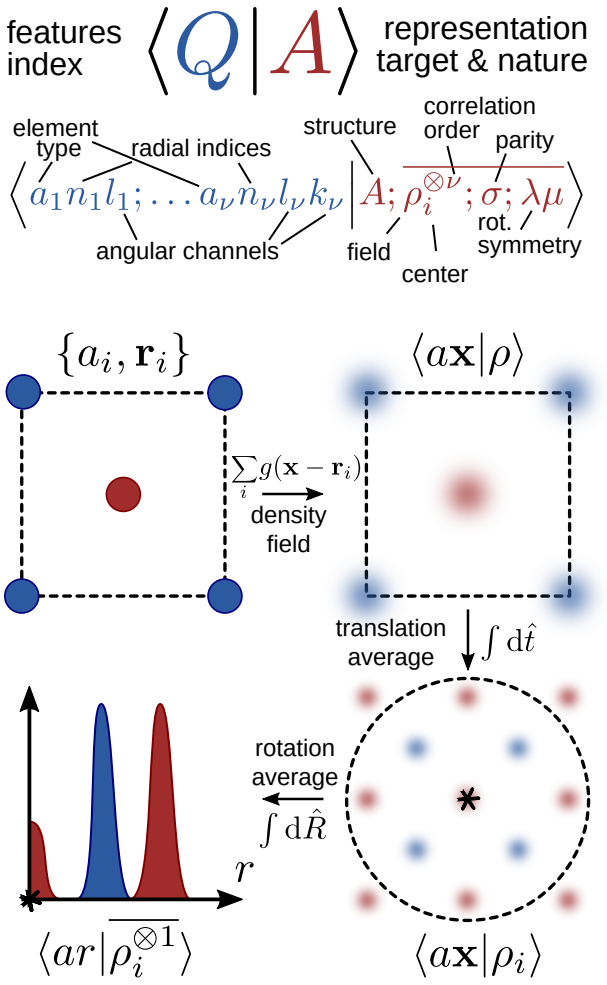


FIG. 4. Top: overview of the notation we use to indicate the features that represent an atomistic structure; bottom: summary of the steps in a symmetrized field construction.

with a structure A, and – when necessary – complement the indication of the structure with one or more symbols and indices (e.g. $|A;\alpha\rangle$) that describe the nature of the representation. These indices might specify the portion of the structure the representation refers to, its symmetry properties, or serve as a reminder of the way the representation was constructed. When we need to explicitly enumerate the elements of the feature vector, we use one or more indices in the bra, leading to expressions of the form $\langle Q|A\rangle$. In this review, we use Q to indicate a generic continuous index, and q to indicate a discrete feature index.

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. Occasionally, e.g. when juxtaposing different choices of basis functions, one may also include qualifiers in the bra, e.g. $\langle n;\mathrm{GTO}|$ to indicate that Gaussian type orbitals are used as a basis. 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 $|\alpha\rangle$ instead of $|A;\alpha\rangle$. Conversely, 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 simply $|A\rangle$.

The indices and the qualifiers that are associated with the structure index (typically in the ket) describe the essential nature of the representation and will be reflected in the architecture of a model built on it. The indices in the bra, instead, simply enumerate features that are of homogeneous nature, are usually manipulated together in the construction of the model, and can be transformed, contracted or sub-selected in a way that does not change the fundamental properties of the representation. In many cases, it is possible to describe the construction of a representation as a combination of kets, without indicating explicitly the use of a particular basis.

This notation can be applied in a way that yields usage patterns that are very similar to those that are common in quantum mechanics, e.g. bra and ket can be interchanged using the convention $\langle A|Q\rangle = \langle Q|A\rangle^\star$. However, much as in the case of the formalism we take inspiration from, a rigorous characterization of the mathematical relations between bras and kets is problematic[130]. It is better to see this notation as a form of symbolic calculus that facilitates memorizing and applying correctly recurring operations and transformations. Let us give a few examples, which also provide a reference of how the notation will be applied in this review.

Change of basis. A change in the basis that is used to practically compute a representation can be written as a linear transformation,

$$\langle T|A\rangle = \int \mathrm{d}Q\ \langle T|Q\rangle\ \langle Q|A\rangle\,, \tag{1}$$

where $\langle T|Q\rangle$ indicates the coefficients that enact the change of basis. This kind of manipulations will be used in Section IV E to convert between a real-space description of the atom-centred density and one based on radial functions and spherical harmonics.[131] All of the expressions discussed here as integrals over a continuous index can be formulated as sums over (finitely or infinitely) countable, discrete indices

$$\int \mathrm{d}Q\ |Q\rangle\ \langle Q| \sim \sum_q |q\rangle\ \langle q|\,. \tag{2}$$

Scalar product and kernels. The scalar product between the features of two structures A and A' can

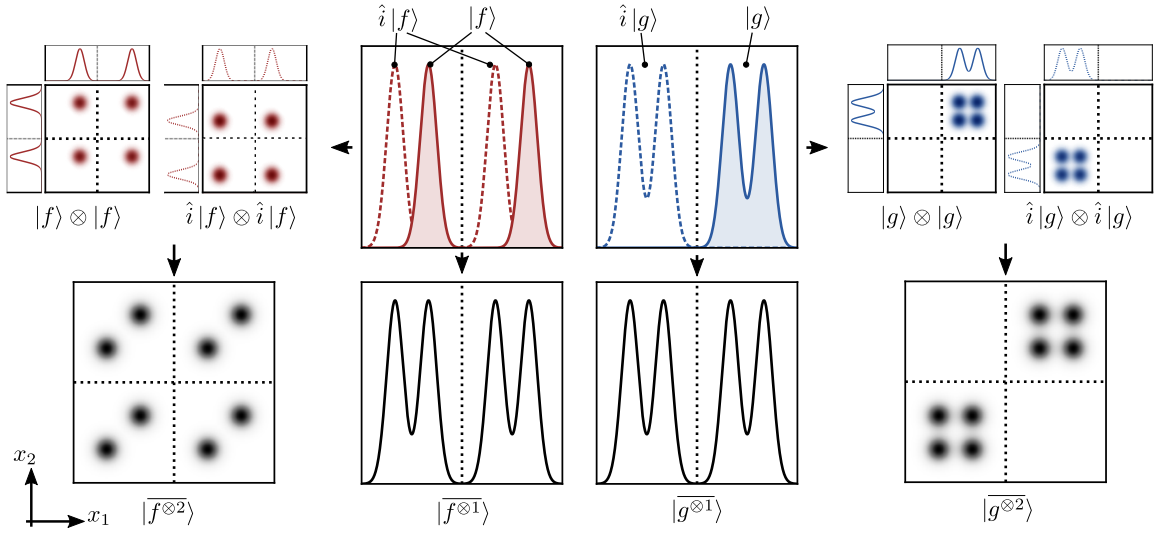


FIG. 5. To obtain features that are invariant to inversion with respect to the vertical dotted line, Haar integration over the symmetry group in this case just corresponds to summing over two symmetry related images. Starting from two distinct functions $|f\rangle$ (left panels, red) and $|g\rangle$ (right panels, blue), the functions (full lines) and their mirror transformation (dotted lines) are summed to obtain invariant features (bottom row). Direct symmetrization is depicted in the central panels, yielding $|\overline{f^{\otimes 1}}\rangle$, while the external panels visualize the construction of tensor-product features, their symmetrization and summation, yielding $|\overline{f^{\otimes 2}}\rangle$.

be written using a complete basis indexed by Q as

$$\langle A|A'\rangle = \int \mathrm{d}Q\ \langle A|Q\rangle\ \langle Q|A'\rangle\,, \tag{3}$$

where one recognizes an expression that is reminiscent of a completeness relation $\int \mathrm{d}Q\ |Q\rangle\ \langle Q| = 1$. This definition only holds for a complete, orthogonal basis and might entail an approximation when computed with a finite basis. The notation $\langle A|A'\rangle$ can also be used to refer to a kernel $\mathrm{k}(A, A')$ that expresses the similarity between two configurations; this is obvious when considering a linear kernel, but can also be used for non-linear kernels, keeping in mind that it might not be possible to write explicitly the features that correspond to the Hilbert space that reproduces the kernel.[132]

Linear models. The bra-ket notation implicitly assumes linearity in the transformation between different choices of basis, and in the modeling of target properties. Even though the features can be used as an input of an arbitrarily complex nonlinear regression scheme (see Section V D), 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. When using a representation $|A;\alpha\rangle$ to describe structures, a linear model for a property $y(A)$ can be written as

$$y(A) \equiv \langle y|A\rangle \approx \int \mathrm{d}Q\ \langle y;\alpha|Q\rangle\ \langle Q|A;\alpha\rangle\,, \tag{4}$$

where $\langle y;\alpha|Q\rangle$ indicates the regression weights for a model based on $|A;\alpha\rangle$. Leaving aside (important) issues related to regularization, this expression emphasizes that one can transform simultaneously the weights and the features to a different basis, and the predicted value is unchanged. The expression $\langle y|A\rangle$ can also be seen as a hint of the fact that a collection of properties could be used as descriptors for a structure A, although this is an approach we only discuss briefly in this review.

Tensor product. 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.

$$|(A;\alpha)\otimes(A';\alpha')\rangle = |A;\alpha\rangle \otimes |A';\alpha'\rangle\,. \tag{5}$$

The construction of a tensor-product representation is well-defined even without indicating explicitly the basis used to describe either side of Eq. (5), and it is often possible to use either an explicit Cartesian product of the bases on the right-hand side, or a combined basis

$$\langle Q_1;Q_2|A\otimes A\rangle \equiv \langle Q_1|A\rangle\ \langle Q_2|A\rangle \to \langle T|A\otimes A\rangle\,, \tag{6}$$

using only $|A\rangle$ as a special case of Eq. (5) in which $A \equiv A'$, and $\alpha \equiv \alpha'$ can be omitted.

Operators and symmetry averages. 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. Taking for instance the operator $\hat{i}$ associated with inversion symmetry, $\hat{i}\,|A\rangle$ indicates the representation associated with structure A after the coordinates of all atoms have been reflected relative to the origin. Much as in quantum mechanics, the operator can also be applied to the bra, where it corresponds to a transformation of the basis. In terms of

symmetry operations, this corresponds to the active or passive transformations, acting on the structure or on the reference frame. By summing over the operators associated with a symmetry group, an operation which is also referred to as Haar integration[133], one can build symmetrized representations that are covariant under the actions of the elements of the group, e.g. for the C_i point group,

$$|\langle A\otimes A\rangle_{C_i};\sigma\rangle = |A\rangle\otimes|A\rangle + \sigma(\hat{i}\,|A\rangle\otimes\hat{i}\,|A\rangle). \quad (7)$$

The index σ takes the value -1 for representations that change sign under inversion, and $+1$ for invariant features; in the invariant case, σ may be omitted. 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., $|\langle A\otimes A\rangle_{C_i}\rangle \to |\overline{A\otimes A}\rangle \to |\overline{A^{\otimes 2}}\rangle$. 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 $\langle x|\,\hat{i}\,|f\rangle = f(-x)$. An inversion-invariant feature can be created by symmetrizing: $|\overline{f^{\otimes 1}}\rangle = |f\rangle + \hat{i}\,|f\rangle$, but our choice of f and g leads to a degenerate description, as $|\overline{f^{\otimes 1}}\rangle = |\overline{g^{\otimes 1}}\rangle$. 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 $|\overline{g^{\otimes 2}}\rangle$ and $|\overline{f^{\otimes 2}}\rangle$ that are also inversion-invariant, but are still able to distinguish between the two functions.

An example: SOAP in bra-ket notation To give a concrete example of the use of this formalism, let us compare the functional notation used in Refs. 29,69 to indicate the components of a SOAP feature vector with the corresponding bra-ket notation. The reader who is unfamiliar with the SOAP construction will find the remainder of this Section, and in particular Section IV E, to give a very detailed account of this family of features, and might better skip this brief overview, that assumes knowledge of the derivation from Ref. 29. The SOAP power spectrum describes the two-point correlations between the atom density centered around the i-th atom of structure A, expanded in terms of atomic species (labeled by the indices $a_{1,2}$), radial basis functions (labeled by $n_{1,2}$) and angular momentum channels (labeled by l). The density expansion coefficients can be written as

$$\begin{aligned}\langle anlm|A;\rho_i\rangle &= \int \mathrm{d}\mathbf{x}\ \langle n|x\rangle\ \langle lm|\hat{\mathbf{x}}\rangle\ \langle a\mathbf{x}|A;\rho_i\rangle \\ &\Vert \\ c^{i,a}_{nlm} &= \int \mathrm{d}\mathbf{x}\, R_n(x)^\star Y_l^m(\hat{\mathbf{x}})^\star \rho_i^a(\mathbf{x}).\end{aligned} \quad (8)$$

In this expression, $\langle a\mathbf{x}|A;\rho_i\rangle \equiv \rho^{i,a}(\mathbf{x})$ indicates the atom-centred density, $\langle x|n\rangle \equiv R_n(x)$ an orthonormal set of radial functions, and $\langle\hat{\mathbf{x}}|lm\rangle \equiv Y_l^m(\hat{\mathbf{x}})$ the spherical harmonics.

The SOAP features for the environment A_i can be written as

$$\begin{aligned}\langle a_1n_1;a_2n_2;l|A;\overline{\rho_i^{\otimes 2}}\rangle &\propto \\ &\sum_m \langle A;\rho_i|a_2n_2lm\rangle\ \langle a_1n_1lm|A;\rho_i\rangle \\ &\Vert \\ p^{i,a_1a_2}_{n_1n_2l} &\propto \sum_m c^{i,a_1}_{n_1lm}(c^{i,a_2}_{n_2lm})^\star.\end{aligned} \quad (9)$$

In the functional notation, one relies on the convention that c corresponds to the density expansion coefficients and p to the power spectrum, while the Dirac notation uses the more expressive symbols ρ_i to indicate the i-centered atom density, and $\overline{\rho_i^{\otimes 2}}$ as a reminder that SOAP features can be derived as a symmetry-averaged 2-point correlation of $|\rho_i\rangle$. This expanded notation is indicative of the place of the SOAP powerspectrum in the hierarchy of density-correlation features, and is useful to distinguish between different kinds of features (radial correlations, power spectrum, bispectrum ...). When it is clear that one is only using one type of representation, the compact (and generic) form $|A_i\rangle$ can be used instead. When it comes to the indices labelling different features, the functional notation mixes the indices (a) associated with the chemical species of the neighbors and the index i of the central atom, separating them from those associated with the radial channel (n). This reflects how SOAP was originally introduced to describe single-element systems. In the Dirac notation, on the other hand, the (a_1n_1) and (a_2n_2) indices are grouped together to indicate that they are conceptually linked in the construction as a tensor product of two densities, and the index indicating the identity of the central atom is associated with the ket.

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 $\mathbf{r}_i$ and define

$$\langle\mathbf{x}|A;\rho\rangle \equiv \sum_{i\in A}\langle\mathbf{x}|\mathbf{r}_i;g\rangle\,, \quad (10)$$

where $\langle\mathbf{x}|\mathbf{r}_i;g\rangle \equiv g(\mathbf{x}-\mathbf{r}_i)$ is a localized function (e.g. a Gaussian) centered on the i-th atom, and the ρ in the ket indicates the kind of field used to describe the structure. As we discuss in more detail in Section IV E, the atomic density functions can be either

finite-width Gaussians, which leads to representations akin to SOAP features[29], or Dirac δ distributions, which recovers representations similar to the current implementation of moment tensor potentials[134] or the atomic cluster expansion[126]. To indicate the $g \to \delta$ limit, we use the notation $|\rho\rangle \to |\delta\rangle$. Atoms, or more generally, "point particles" such as those one could associate to a coarse grained description of a molecular system, can be further characterized by internal attributes, that could be discrete (e.g. the chemical nature of an atom, or a molecule, which we indicate as a_i) or continuous (e.g. an atomic or molecular dipole $\mathbf{u}_i$)

$$\langle a\mathbf{u}\mathbf{x}|A;\rho u\rangle \equiv \sum_{i\in A} \delta_{aa_i}\,\langle \mathbf{u}|\mathbf{u}_i;g\rangle\,\langle \mathbf{x}|\mathbf{r}_i;g\rangle\,. \tag{11}$$

In this form, Eq. (11) can be seen as an abstraction of the many real-space "voxel" representations of materials,[135,136] that are used often in the context of generative models and reinforcement learning[137]. The ket $|A;\rho\rangle$ defined by expressions like (10) or (11) could be equally well expressed in a different basis, e.g. expanded in plane waves

$$\langle \mathbf{k}|A;\rho\rangle = \frac{1}{(2\pi)^{3/2}}\int \mathrm{d}\mathbf{x}\, e^{-\mathrm{i}\mathbf{k}\cdot\mathbf{x}}\,\langle \mathbf{x}|A;\rho\rangle = \sum_{i\in A}\langle \mathbf{k}|\mathbf{r}_i;g\rangle\,, \tag{12}$$

which also shows how the change of basis can be applied directly to the atom-centred density contributions. Eqs. (10) and (12) contain the same amount of information, and can be seen as special cases of a formal definition of the representation for the structure A as a sum of atomic representations,

$$|A;\rho\rangle = \sum_{i\in A}|\mathbf{r}_i;g\rangle\,. \tag{13}$$

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 $\langle \mathbf{x}|\rho\rangle$ translationally invariant is to sum over the continuous translation group, $\int \mathrm{d}\hat{t}\,\langle \mathbf{x}|\,\hat{t}\,|\rho\rangle$. Summing directly over the atom density eliminates all structural information, because $\int \mathrm{d}\hat{t}\,\langle \mathbf{x}|\,\hat{t}\,|\mathbf{r}_i;g\rangle = \int \mathrm{d}\mathbf{t}\, g(\mathbf{t}-\mathbf{r}_i) = 1$. Information loss is a usual issue with Haar integration, as exemplified in Figure 5. 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

$$\begin{aligned}\langle a_1\mathbf{x}_1; a_2\mathbf{x}_2|\,\langle\rho\otimes\rho\rangle_{\mathbb{R}^3}\rangle &\equiv \langle a_1\mathbf{x}_1; a_2\mathbf{x}_2|\overline{\rho^{\otimes 2}}\rangle \\ &= \int \mathrm{d}\hat{t}\,\langle a_1\mathbf{x}_1|\,\hat{t}\,|\rho\rangle\,\langle a_2\mathbf{x}_2|\,\hat{t}\,|\rho\rangle \\ &= \sum_{ij}\delta_{a_1a_j}\delta_{a_2a_i}\int \mathrm{d}\hat{t}\,\langle \mathbf{x}_1-\mathbf{t}|\mathbf{r}_j;g\rangle\,\langle \mathbf{x}_2-\mathbf{t}|\mathbf{r}_i;g\rangle \\ &\propto \sum_{ij}\delta_{a_1a_j}\delta_{a_2a_i}\,\langle(\mathbf{x}_1-\mathbf{x}_2)|(\mathbf{r}_j-\mathbf{r}_i);\tilde{g}\rangle\end{aligned} \tag{14}$$

where $\tilde{g}$ indicates the cross-correlation of two of the localized density functions. In the case of a Gaussian density, $\tilde{g}$ is simply a Gaussian with twice the variance, and outside this section we will use just g to indicate the atomic density both in $|\rho\rangle$ and $|\rho_i\rangle$. As a remindet that the representation has been obtained by averaging over translations the tensor product of two density fields, we use the superscript notation $\overline{\rho^{\otimes 2}}$, and we separate with a semicolon groups of feature indices that are associated with each factor in the tensor product, as discussed in Section IV A. Note that the representation in Eq. (14) has a large null space, as it depends only on $\mathbf{x}_1-\mathbf{x}_2$. One could then re-define it by labelling features using a single position vector, or transform it in a plane wave basis:

$$\begin{aligned}\langle a_1; a_2; \mathbf{k}|\overline{\rho^{\otimes 2}}\rangle &= \int \mathrm{d}\mathbf{x}\, e^{-\mathrm{i}\mathbf{k}\cdot\mathbf{x}}\,\langle a_1\mathbf{0}; a_2\mathbf{x}|\overline{\rho^{\otimes 2}}\rangle \\ &= \langle a_1\mathbf{k}|\rho\rangle^{\star}\,\langle a_2\mathbf{k}|\rho\rangle\end{aligned} \tag{15}$$

where the second equality is a consequence of the convolution theorem. One sees that the translationally-symmetrized density is essentially equivalent to the diffraction pattern of the atomic structure $I(\mathbf{k})$, that has been already used as a descriptor to classify crystalline configurations.[138]

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

$$\langle a\mathbf{x}|A;\rho_i\rangle = \sum_{j\in A}\delta_{aa_j}\,\langle \mathbf{x}|\mathbf{r}_{ji};\tilde{g}\rangle\,, \tag{16}$$

where $\mathbf{r}_{ji} = \mathbf{r}_j - \mathbf{r}_i$. The fact that $|A;\rho_i\rangle$ is atom centered (and hence translationally invariant) is hinted at by the subscript notation ρ_i, and so in what follows we only use this subscript to distinguish it from its non-symmetrized counterpart (10) 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. (16) clarifies how an atom-centered representation is a natural consequence of the translational symmetrization:

$$\langle a_1\mathbf{x}_1; a_2\mathbf{x}_2|A;\overline{\rho^{\otimes 2}}\rangle = \sum_{i\in A}\delta_{a_2a_i}\,\langle a_1(\mathbf{x}_1-\mathbf{x}_2)|A;\rho_i\rangle\,. \tag{17}$$

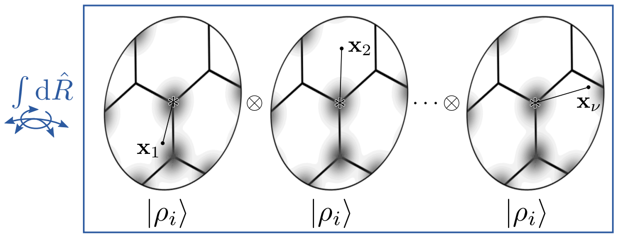


FIG. 6. Graphical scheme of the construction of a $SO(3)$-symmetrized tensor product representation. Copies of the atom-centered density are evaluated at ν separate points, and the tensor product is averaged by simultaneously rotating all densities.

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:

$$\begin{aligned}\langle y|A\rangle &\approx \sum_{i\in A} \langle y; a_i|A_i\rangle \\ &= \sum_{i\in A}\sum_{a} \int \mathrm{d}\mathbf{x}\ \langle y; a_i|a\mathbf{x}\rangle\ \langle a\mathbf{x}|A;\rho_i\rangle\,.\end{aligned} \tag{18}$$

Note that in this case we assume that only the regression weights depend on the nature of the central atom, but one might as well fine-tune the atom-centred features depending on the central atom. 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 (16). 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, as one can measure in terms of the locality of the response of forces to atomic displacements of neighbors[15]. However, introducing a relatively short-range cutoff 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 (16) 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

$$\begin{aligned}|\overline{\rho_i^{\otimes\nu}}\rangle &\equiv |\,\langle\underbrace{\rho_i\otimes\cdots\otimes\rho_i}_{\nu\ \text{times}}\rangle_{O(3)}\rangle \\ &= \sum_{k=0,1}\int_{SO(3)} \mathrm{d}\hat{R}\ \hat{i}^k\hat{R}\,|\rho_i\rangle\otimes\ldots\otimes\hat{i}^k\hat{R}\,|\rho_i\rangle\,.\end{aligned} \tag{19}$$

This can be expanded on an explicit position basis

$$\begin{aligned}&\langle a_1\mathbf{x}_1;\ldots a_\nu\mathbf{x}_\nu|\overline{\rho_i^{\otimes\nu}}\rangle \\ &= \sum_{k=0,1}\int_{SO(3)} \mathrm{d}\hat{R}\ \langle a_1\mathbf{x}_1|\,\hat{i}^k\hat{R}\,|\rho_i\rangle\ldots\langle a_\nu\mathbf{x}_\nu|\,\hat{i}^k\hat{R}\,|\rho_i\rangle\,,\end{aligned} \tag{20}$$

emphasizing that $|\overline{\rho_i^{\otimes\nu}}\rangle$ corresponds to a symmetrized, ν-point correlation of the atom density centered on the i-th atom (Fig. 6) – a $(\nu+1)$-point correlation function, in the language used in statistical mechanics to describe the structure of liquids[139,140]. Similar to the case of Eq. (14), this object has a large null space (e.g. in the $\nu=1$ case it only depends on $x_1=|\mathbf{x}_1|$). As discussed in Ref. 109, one can choose a more concise enumeration of the real-space correlations in terms of distances and angles, that reduces in the limit $g\to\delta$ to a sum over distances and angles between atoms. For instance, for the $\nu=2$ case one can write

$$\begin{aligned}&\langle a_1 r_1; a_2 r_2;\omega|\overline{\delta_i^{\otimes 2}}\rangle \\ &\propto \sum_{jj'}\delta_{a_1 a_j}\delta_{a_2 a_{j'}}\delta(r_1-r_{ji})\delta(r_2-r_{j'i})\delta(\omega-\hat{\mathbf{r}}_{ji}\cdot\hat{\mathbf{r}}_{j'i}),\end{aligned} \tag{21}$$

where we use $\rho\to\delta$ to indicate that the correlation function is built on the Dirac-δ limit of the atom density field. Expressions of this kind reveal the close connection between symmetrized-field representations and atom-centered symmetry functions[10,20,141], as well as equivalent constructions such as those used in the ANI[20] and DeepMD[142] frameworks, and the FCHL features[116,117]. 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. (21)

$$\begin{aligned}\langle a_1 a_2 k|\overline{\delta_i^{\otimes 2}}\rangle &= \int \mathrm{d}r_1\ \mathrm{d}r_2\ \mathrm{d}\omega \\ &\times\ \langle k; G^3|r_1 r_2\omega\rangle\ \langle a_1 r_1; a_2 r_2;\omega|\overline{\delta_i^{\otimes 2}}\rangle \\ &\equiv \\ &\sum_{jj'}\delta_{a_1 a_j}\delta_{a_2 a_{j'}}G^3_k(r_{ji},r_{j'i},\hat{\mathbf{r}}_{ji}\cdot\hat{\mathbf{r}}_{j'i}),\end{aligned} \tag{22}$$

that demonstrates the connection between density correlations and atom-centered symmetry functions

computed as a sum over groups of neighbors following the notation used in Ref. 141.

Note that we choose to symmetrize the atom-centered description $|\rho_i\rangle$ – 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 $|\overline{\rho^{\otimes 2}}\rangle$

$$|\,\langle\,\langle\rho\otimes\rho\rangle_{\mathbb{R}^3}\otimes\langle\rho\otimes\rho\rangle_{\mathbb{R}^3}\rangle_{SO(3)}\rangle\sim \int_{SO(3)}\mathrm{d}\hat{R}\,\sum_{ii'}\hat{R}\,|\rho_i\rangle\otimes\hat{R}\,|\rho_{i'}\rangle\,, \quad (23)$$

as it was done for instance in Ref. 143. 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.[21,144]

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 $Y_l^m(\hat{\mathbf{x}})\equiv\langle\hat{\mathbf{x}}|lm\rangle$ yields a discrete set of coefficients that transform as spherical harmonics

$$\begin{aligned}\langle anlm|A;\rho_i\rangle &= \int \mathrm{d}\mathbf{x}\,\langle n|x\rangle\,\langle lm|\hat{\mathbf{x}}\rangle\,\langle a\mathbf{x}|A;\rho_i\rangle\\ &= \sum_{j\in A_i}\delta_{aa_j}\int \mathrm{d}\mathbf{x}\,\langle n|x\rangle\,\langle lm|\hat{\mathbf{x}}\rangle\,\langle x\hat{\mathbf{x}}|\mathbf{r}_{ji};g\rangle\\ &= \sum_{j\in A_i}\delta_{aa_j}\,\langle nlm|\mathbf{r}_{ji};g\rangle\,, \quad (24)\end{aligned}$$

where $\langle nlm|\mathbf{r}_{ji};g\rangle$ corresponds to the expansion in radial functions and spherical harmonics of a Gaussian centered on the interatomic vector $\mathbf{r}_{ji}$. These expansion coefficients can be seen as functions of $\mathbf{r}_{ji}$, 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 is natural, and makes it easy to evaluate the rotational integral of Eq. (19) analytically, because the matrix elements $\langle lm|\hat{R}|l'm'\rangle=\delta_{ll'}D^l_{m'm}(\hat{R})$ correspond to Wigner-D matrices, an irreducible representation of $SO(3)$. Well-known results from the theory of angular momentum,[145] such as the orthonormality and the product reduction formula for Wigner-D matrices, allow deriving explicit expressions for the symmetrized field representations of order $\nu=1,2,3$

$$\langle a_1n_1l_1m_1|\overline{\rho_i^{\otimes 1}}\rangle=\frac{8\pi^2}{2l_1+1}\,\langle a_1n_1l_1m_1|\rho_i\rangle\,\delta_{l_10}\delta_{m_10} \quad (25)$$

$$\langle a_1n_1l_1m_1;a_2n_2l_2m_2|\overline{\rho_i^{\otimes 2}}\rangle=\delta_{l_1l_2}\delta_{m_1m_2}\frac{8\pi^2}{2l_1+1}\sum_s(-1)^{s-m1}\,\langle a_1n_1l_1s|\rho_i\rangle\,\langle a_2n_2l_2(-s)|\rho_i\rangle\,, \quad (26)$$

$$\begin{aligned}\langle a_1n_1l_1m_1;a_2n_2l_2m_2;a_3n_3l_3m_3|\overline{\rho_i^{\otimes 3}}\rangle = &\\ \frac{8\pi^2}{2l_1+1}(-1)^{-m_1}\,\langle l_2m_2;l_3m_3|l_1(-m_1)\rangle &\\ \sum_{s_1s_2s_3}(-1)^{-s_1}\,\langle l_2s_2;l_3s_3|l_1(-s_1)\rangle\,\langle a_1n_1l_1s_1|\rho_1\rangle &\\ \langle a_2n_2l_2s_2|\rho_2\rangle\,\langle a_3n_3l_3s_3|\rho_3\rangle\,, \quad (27)\end{aligned}$$

where $\langle l_1m_1;l_2m_2|LM\rangle$ 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 some of 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 inconsequential constant factors. For the case $\nu=1$ one can define

$$\langle an|\overline{\rho_i^{\otimes 1}}\rangle=\langle an00|\rho_i\rangle \quad (28)$$

which corresponds to a discretized version of a pair correlation function

$$\begin{aligned}\langle an|\overline{\rho_i^{\otimes 1}}\rangle &= \int \mathrm{d}\mathbf{x}\,\langle n|x\rangle\,\langle 00|\hat{\mathbf{x}}\rangle\,\langle a(x\hat{\mathbf{x}})|\rho_i\rangle\\ &\propto \int \mathrm{d}x\,x^2\,\langle n|x\rangle\int\mathrm{d}\hat{\mathbf{x}}\,\langle a(x\hat{\mathbf{x}})|\rho_i\rangle\\ &\sim \int \mathrm{d}r\,r^2R_n(r)^\star g_a(r), \quad (29)\end{aligned}$$

in which we use the usual notation $g_a(r)$ to indicate the distribution of a atoms (although in this case it is restricted to an i-centered environment rather than averaged over an equilibrium distribution). For the $\nu=2$ case, Eq. (26) can be redefined as

$$\langle a_1n_1;a_2n_2;l|\overline{\rho_i^{\otimes 2}}\rangle=\frac{(-1)^l}{\sqrt{2l+1}}\sum_m(-1)^m\,\langle a_1n_1lm|\rho_i\rangle\,\langle a_2n_2l(-m)|\rho_i\rangle \quad (30)$$

This corresponds – modulo irrelevant constants – to the rotation invariant 3D shape descriptor[146] and to the SOAP features, which would be written, in the notation of Refs. 29,69 as

$$p^{i,a_1a_2}_{n_1n_2l}=\frac{1}{\sqrt{2l+1}}\sum_m c^{i,a_1}_{n_1lm}(c^{i,a_2}_{n_2lm})^\star, \quad (31)$$

where $c^{i,a}_{nlm} = \langle anlm|\rho_i\rangle$ indicate the density expansion coefficients following the same notation. The $\nu = 2$ representation can also be written on a real-space basis as $\langle a_1 r_1; a_2 r_2; \omega|\overline{\rho_i^{\otimes 2}}\rangle$, emphasizing its nature as three-body density correlation function that depends on two distances r_1, r_2 and the cosine ω of the angle between the directions along which they are evaluated. The 4-body order invariant representation becomes

$$\langle a_1 n_1 l_1; a_2 n_2 l_2; a_3 n_3 l_3|\overline{\rho_i^{\otimes 3}}\rangle = \frac{(-1)^{l_3}}{\sqrt{2l_3+1}} \sum_{m_1 m_2 m_3} (-1)^{m_3} \langle l_1 m_1; l_2 m_2|l_3 m_3\rangle \langle a_1 n_1 l_1 m_1|\rho_i\rangle \langle a_2 n_2 l_2 m_2|\rho_i\rangle \langle a_3 n_3 l_3(-m_3)|\rho_i\rangle \quad (32)$$

corresponding to the SOAP bispectrum[29]

$$b^{i,a_1 a_2 a_3}_{n_1 l_1 n_2 l_2 n_3 l_3} = \frac{1}{\sqrt{2l+1}} \sum_{m_1 m_2 m_3} \langle l_1 m_1; l_2 m_2|l_3 m_3\rangle c^{i,a_1}_{n_1 l_1 m_1} c^{i,a_2}_{n_2 l_2 m_2} (c^{i,a_3}_{n_3 l_3 m_3})^\star, \quad (33)$$

and closely related to the bispectrum used in the spectral neighbor analysis method[147,148], which is essentially equivalent to a different choice of basis. As discussed in more detail in Ref. 149 and in the next sections, the relationship between the redundant expressions (25, 26, 27) that arise from the integral over rotations, and the more concise versions (28, 30, 32) can be seen as a transformation from the uncoupled to the coupled angular momentum basis, and starting from the $\nu = 4$ additional indices k_ν must be included to account for the different ways the coupling can be realized. A practical implementation of these higher body order features is given by the the atomic cluster expansion (ACE), which is usually computed based on the $g \to \delta$ limit of ρ. The coefficients of the atom density are indicated as $\langle nlm|\delta_i\rangle \equiv A_{inlm}$ following the notation of Ref. 126, and

$$\langle n_1 l_1 k_1 \cdots n_\nu l_\nu k_\nu|\overline{\delta_i^{\otimes\nu}}\rangle \equiv B^{(\nu)}_{\substack{in_1\cdots n_\nu \\ l_1\cdots l_\nu}} \quad (34)$$

correspond to the features associated with ν-order neighbor clusters. Note that each $B^{(\nu)}_{\substack{in_1\cdots n_\nu \\ l_1\cdots l_\nu}}$ indicates a group of basis functions indexed by $k_1, \ldots, k_\nu$. An equivalent construction, that emphasizes the connection with angular momentum theory, is provided by the N-body iterative contraction of equivariants[149], that is discussed in Section IV G. Through a further linear transformation (change of basis) made explicit in Refs. 127,150 the moment tensor potential (MTP) of Ref. 134 can also be related to this construction. 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. VI and in Sec. VIII B. Note that even though the contracted basis $\langle (a_i n_i l_i k_i)_{i=1\ldots\nu}|$ 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. 149, can be derived from results of angular momentum theory[151] – 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[127].

Finally, the global SOAP-like descriptors introduced in Ref. 143, corresponding to Eq. (23), can be readily expressed in an angular momentum basis as

$$\langle a_1 n_1; a_2 n_2; l|A; \overline{\rho^{\otimes 2}} \otimes \overline{\rho^{\otimes 2}}\rangle = \frac{(-1)^l}{\sqrt{2l+1}} \sum_m (-1)^m \langle a_1 n_1 lm|\overline{\rho^{\otimes 2}}\rangle \langle a_2 n_2 l(-m)|\overline{\rho^{\otimes 2}}\rangle, \quad (35)$$

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

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. (24). This ensures that the cost of computing *all* coefficients of a given order ν, scales only linearly with the number of neighbors included within the cutoff around atom i, even though it scales exponentially with ν 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),[20,141,142] and permutation invariant polynomials (PIP),[11] in which function are evaluated over all possible tuples composed of ν 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 ν 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 ν, and there are many neighbors. A-priori sparsification schemes, exemplified in (106), and feature selection schemes, discussed in Section VIII B, allow one to keep only the most important basis functions, and eliminate the exponential scaling with ν 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. (30) in the $g \to \delta$ limit as a sum over neighbors j and j'

$$\langle a_1 n_1; a_2 n_2; l|\overline{\delta_i^{\otimes 2}}\rangle \propto \sum_{jj'} \frac{\delta_{a_1 a_j}\delta_{a_1 a_{j'}}}{\sqrt{2l+1}} \langle n_1|r_{ji}\rangle \langle n_2|r_{j'i}\rangle \times \sum_m (-1)^m \langle lm|\hat{\mathbf{r}}_{ji}\rangle \langle l(-m)|\hat{\mathbf{r}}_{j'i}\rangle . \quad (36)$$

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

$$\langle a_1 n_1; a_2 n_2; l|\overline{\delta_i^{\otimes 2}}\rangle \propto \sqrt{2l+1} \sum_{jj'} \delta_{a_1 a_j}\delta_{a_1 a_{j'}} \langle n_1|r_{ji}\rangle \langle n_2|r_{j'i}\rangle \langle l|\hat{\mathbf{r}}_{ji}\cdot\hat{\mathbf{r}}_{j'i}\rangle , \quad (37)$$

in which $\langle\omega|l\rangle \equiv P_l(\omega)$ is a Legendre polynomial of order l. In Eq. (37), 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 $\langle a_1 n_1; a_2 n_2; l|\overline{\delta_i^{\otimes 2}}\rangle$ into Eq. (22), 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,[29,69] Behler-Parrinello symmetry functions,[20,152] the DeepMD framework,[142] and FCHL features[116]. 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 $\mathbf{y}$. In this case, the prescribed transformations that the tensor undergoes under the symmetry operations of the $O(3)$ group (e.g. $\mathbf{y}(\hat{R}A) = \hat{R}\mathbf{y}(A)$) have to be incorporated into the atomic representation in the form of covariant, rather than simply invariant, features, so that the representation follows the same transformation as the target property, $|\hat{R}A\rangle = \hat{R}|A\rangle$. Equivariance (the general concept that indicates symmetry-adapted behavior, encompassing both invariance and covariance) 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[153–156]. A more general approach for achieving this goal consists in endowing the

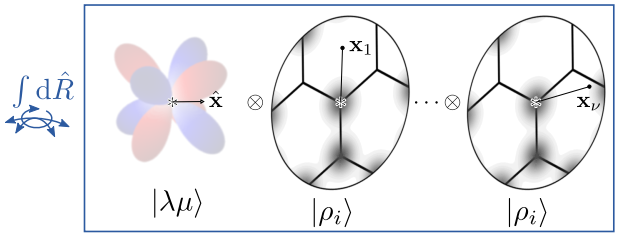


FIG. 7. Graphical scheme of the construction of a $SO(3)$ equivariant tensor product representation. Copies of the atom-centered density are evaluated at ν separate points, together with a set of spherical harmonics that provide a basis to expand the components of a tensorial property. The tensor product is averaged by simultaneously rotating all densities and the $|\lambda\mu\rangle$ term.

representation with the symmetries of spherical harmonics $\langle\hat{\mathbf{x}}|\lambda\mu\rangle = Y_\lambda^\mu(\hat{\mathbf{x}})$, as well as the desired parity under the action of the inversion operator $\hat{i}$, which we associate with a ket $|\sigma\rangle$ such that $\hat{i}|\sigma\rangle = \sigma|\sigma\rangle$. The eigenvalue σ is 1 for polar tensors, and -1 for pseudotensors. Features that transform as $|\sigma\rangle \otimes |\lambda\mu\rangle$ can be achieved by including two additional fields[109,157] within the symmetrized tensor product of Eq. (19), i.e.,

$$|\langle \rho_i^{\otimes\nu} \otimes \sigma \otimes \lambda\mu\rangle_{O(3)}\rangle \equiv |\overline{\rho_i^{\otimes\nu}; \sigma; \lambda\mu}\rangle = \sum_{k=0,1} \int_{SO(3)} \mathrm{d}\hat{R}\, \hat{i}^k |\sigma\rangle \otimes \hat{i}^k\hat{R}|\lambda\mu\rangle \otimes \hat{i}^k\hat{R}|\rho_i\rangle \otimes \cdots \otimes \hat{i}^k\hat{R}|\rho_i\rangle . \quad (38)$$

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 $|\sigma;\lambda\mu\rangle = |1;00\rangle$.

This construction represents a particularly convenient framework to target the prediction of any Cartesian tensor $\mathbf{y}$ in terms of its irreducible spherical components,[158] namely $y_\mu^{\sigma\lambda}$, that transform under rotation and inversion as

$$\begin{aligned} y_\mu^{\sigma\lambda}(\hat{R}A) &= \sum_m D^\lambda_{\mu m}(\hat{R}) y_m^{\sigma\lambda}(A), \\ y_\mu^{\sigma\lambda}(\hat{i}A) &= \sigma(-1)^\lambda y_m^{\sigma\lambda}(A). \end{aligned} \quad (39)$$

Within a linear regression model, they can be written as the combination of equivariant representations of the proper order λ and parity σ with a set of rotationally-invariant weights $\langle Q|\mathbf{y};\sigma;\lambda\rangle$:

$$y_\mu^{\sigma\lambda}(A) = \langle \mathbf{y}|A;\sigma;\lambda\mu\rangle \approx \sum_i \int \mathrm{d}Q\, \langle \mathbf{y};\sigma\lambda;|Q\rangle \langle Q|A;\overline{\rho_i^{\otimes\nu};\sigma;\lambda\mu}\rangle . \quad (40)$$

Each irreducible spherical component of $\mathbf{y}$ gives rise to a separate equivariant model, and the appropriate transformation rules are ensured by the fact that

each equivariant feature $\langle Q|A;\overline{\rho_i^{\otimes\nu};\sigma;\lambda\mu};\rangle$ separately transforms as the spherical harmonics $|lm\rangle$ and the parity function $|\sigma\rangle$. Much like the case of invariant symmetrized fields features, Eq. (38) can be most effectively computed by first expanding the atom-centered field on a basis of spherical harmonics, and is equivalent to an equivariant extension of the atomic cluster expansion[150] or the moment tensor potentials, that are usually evaluated in 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

$$\langle n|\overline{\rho_i^{\otimes 1};\sigma;\lambda\mu}\rangle \equiv \langle n\lambda(-\mu)|\rho_i\rangle\,\delta_{\sigma 1}. \tag{41}$$

Note how in the bra-ket notation the (λ,μ) indices on the two sides of this equation carry a different meaning. When used in the bra of the local density expansion $\langle n\lambda\mu|\rho_i\rangle$, they identify one of many components that are translationally invariant, but are not required to be rotationally equivariant; there is no explicit link to their behavior under rotation, and one could build a model by selecting only some of the μ values for a given (n,λ). When used in the ket of an equivariant feature $\langle n|\overline{\rho_i^{\otimes 1};\sigma;\lambda\mu}\rangle$, 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. By using $\langle n|\overline{\rho_i^{\otimes 1};\sigma;\lambda\mu}\rangle$ features in Eq. (40) one obtains a model that fulfills (39) (with the caveat that pseudotensors cannot be described by $\nu = 1$ features) because acting on the spherical harmonics with $\hat{R}$ yields a product with the associated Wigner matrix

$$\begin{aligned}\sum_n \langle \mathbf{y};\lambda|n\rangle\ \langle n|\hat{R}A;\overline{\rho_i^{\otimes 1};\lambda\mu}\rangle\ &= \\ &= \sum_n \langle \mathbf{y};\lambda|n\rangle\ \langle nl(-\mu)|\hat{R}A;\rho_i\rangle \\ &= \sum_n \langle \mathbf{y};\lambda|n\rangle \sum_m D^\lambda_{\mu m}(\hat{R})\,\langle nl(-m)|A;\rho_i\rangle \\ &= \sum_m D^\lambda_{\mu m}(\hat{R}) \sum_n \langle \mathbf{y};\lambda|n\rangle\ \langle n|A;\overline{\rho_i^{\otimes 1};\lambda m}\rangle\ .\end{aligned} \tag{42}$$

The same covariant property applies to all density-correlation features,

$$|\hat{R}A;\overline{\rho_i^{\otimes\nu};\sigma;\lambda\mu}\rangle = \sum_m D^\lambda_{\mu m}(\hat{R})\,|A;\overline{\rho_i^{\otimes\nu};\sigma;\lambda m}\rangle\,. \tag{43}$$

Scalar products of these equivariant features generate matrix-valued kernels, that are suitable for symmetry-adapted Gaussian process regression – for example λ-SOAP kernels[159,160]. Each entry in the kernel describes the coupling between the μ channels associated with the two environments,

$$\begin{aligned}\mathrm{k}^{\sigma\lambda}_{\mu\mu'}(A_i,A'_{i'}) = \int \mathrm{d}Q \\ \times\ \langle A;\overline{\rho_i^{\otimes\nu};\sigma;\lambda\mu}|Q\rangle\ \langle Q|A';\overline{\rho_{i'}^{\otimes\nu};\sigma;\lambda\mu}\rangle\,.\end{aligned} \tag{44}$$

The symmetry properties of the features translate into the a kernel that transforms under rotations of the environments as

$$\begin{aligned}\mathrm{k}^{\sigma\lambda}_{\mu\mu'}(\hat{R}A_i,\hat{R}'A'_{i'}) = \\ \sum_{mm'} D^\lambda_{\mu m}(\hat{R})\,\mathrm{k}^{\sigma\lambda}_{mm'}(A_i,A'_{i'})D^\lambda_{\mu' m'}(\hat{R}')^\star,\end{aligned} \tag{45}$$

which generalizes the covariant property for kernels introduced by Glielmo et al. for the case of Cartesian vectors.[161]

The fact that equivariant features of the form (38) 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. (28–32) can be formulated for an arbitrary value of ν, and in this form it is possible to express succintly[149] a recursive formula to evaluate $|\overline{\rho_i^{\otimes\nu};\sigma;\lambda\mu}\rangle$ based on lower order terms:

$$\begin{aligned}\langle \dots n_\nu l_\nu k_\nu; nlk|\overline{\rho_i^{\otimes(\nu+1)};\sigma;\lambda\mu}\rangle \propto \delta_{s\sigma((-1)^{l+k+\lambda})}\times \\ \sum_m \langle lm;k(\mu-m)|\lambda\mu\rangle\ \langle n|\overline{\rho_i^{\otimes 1};lm}\rangle \\ \langle \dots;n_\nu l_\nu k_\nu|\overline{\rho_i^{\otimes\nu};s;k(\mu-m)}\rangle\,.\end{aligned} \tag{46}$$

For $\nu = 2$, the recursion yields the original expression for λ-SOAP equivariants[159]

$$\begin{aligned}\langle n_1 l_1; n_2 l_2|\overline{\rho_i^{\otimes 2};\sigma;\lambda\mu}\rangle = \\ \frac{\delta_{\sigma(-1)^{l_1+l_2+\lambda}}}{\sqrt{2\lambda+1}} \sum_m \langle l_1 m; l_2(\mu-m)|\lambda\mu\rangle \\ \times\ \langle n_1 l_1(-m)|\rho_i\rangle\ \langle n_2 l_2(m-\mu)|\rho_i\rangle\,.\end{aligned} \tag{47}$$

Similar recursive expressions have been independently proposed to efficiently compute *invariant* features[127,134], that can be obtained by taking $|\sigma;\lambda\mu\rangle = |1;00\rangle$ in Eq. (46). The possibility of combining equivariant features using angular momentum rules is also exploited in the construction of covariant neural networks[144,162]

One can also 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

$$\boldsymbol{\mu}(A) = \sum_{i\in A} q(A_i)\mathbf{r}_i. \tag{48}$$

Models of this form have been used since the early days of the construction of molecular potential and dipole moment surfaces[62,164], combined with neural-network potentials to compute IR spectra in the condensed phases[165], and more recently combined with

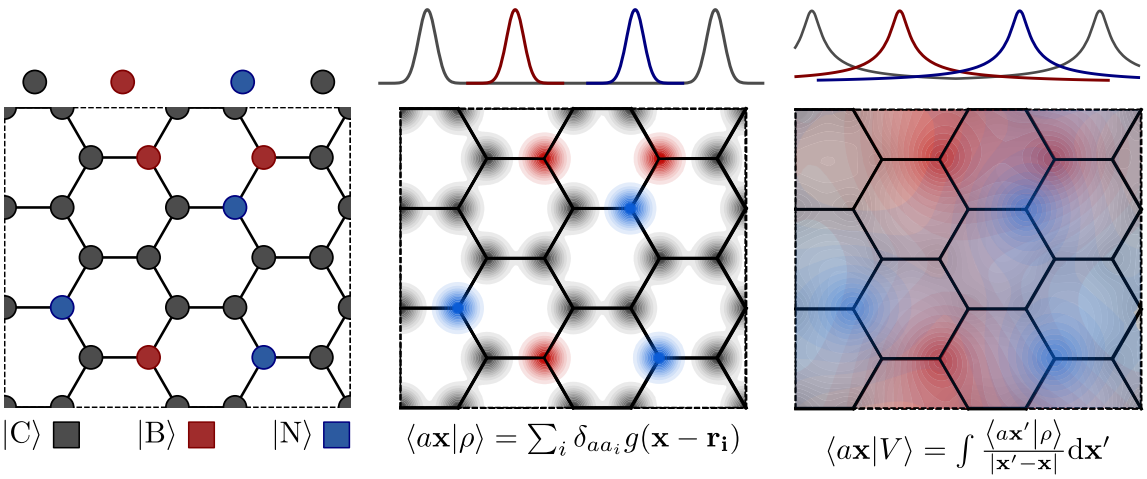


FIG. 8. Relationship between Cartesian coordinates, local and long-range fields. The top row shows a 1D cartoon, and the second row a more realistic, hypothetical "doped graphene" system in 2D. Left: Reference structure; middle: atom-density field, divided in three elemental channels, color-coded; right: atom-density potential, color-coded. Adapted with permission from Ref. 163. Copyright 2020 Royal Society of Chemistry.

tensorial models, to describe the interplay of atomic charges and polarization contributing to the total dipole moment[166]. Assigning constant formal charges q_i to atoms has also been used to derive covariant kernels, in the so-called operator machine learning framework[167], which is also similar in spirit to the tensorial embedded atom neural network[168]. 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.

$$\langle \mathbf{x}|A_i;\mathbf{E}\rangle = \sum_{j\in A} \langle \mathbf{x}|\mathbf{r}_{ji};g\rangle\,(\mathbf{r}_{ji}\cdot\mathbf{E})q_j \qquad (49)$$

Deriving with respect to one of the components of $\mathbf{E}$ brings a dependency on the corresponding component of $\mathbf{r}_{ji}$, 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. (38), 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. 169 to define a general atomic neighborhood fingerprint, and in Ref. 170 to learn the position of electronic Wannier centers. Despite the superficial similarity with the environment-dependent point-charge model of Eq. (48), this scheme more closely resembles 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.[30] The problem is particularly evident when studying systems with a prominent electrostatic component[159,173], but long-range physics is ubiquitous[174], and ultimately limits the accuracy and transferability of machine-learning models[24,166,173]. 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[175,176], an electrostatic scheme based on machine-learned partial charges[165,177,178] or atomic multipoles[154,179]. Alternatively, one may attempt to construct representations that are multi-scale in nature, and are therefore suitable to describe, in a data-driven manner, properties that depend on multiple length scales. This idea has been implemented by combining local representations with different cutoffs[30], scaling atomic contributions according to distance[116,125] (see also Section VIII C), treating separately intra- and intermolecular correlations[180,181], as well as by building global structural representations based on an intrinsically multi-scale wavelet scattering transform[182].

A recently-proposed, more radical take to the problem, extends the symmetrized atomic field construction beyond the use of the atomic density as the starting point. In order to describe more naturally the long-range behavior that is typical of electrostatic interactions, it defines a Coulomb-like potential field based on the smoothed atomic density (Figure 8)

$$\langle a\mathbf{x}|A;V\rangle = \int \mathrm{d}\mathbf{x}'\,\frac{\langle a\mathbf{x}|A;\rho\rangle}{|\mathbf{x}-\mathbf{x}'|}. \qquad (50)$$

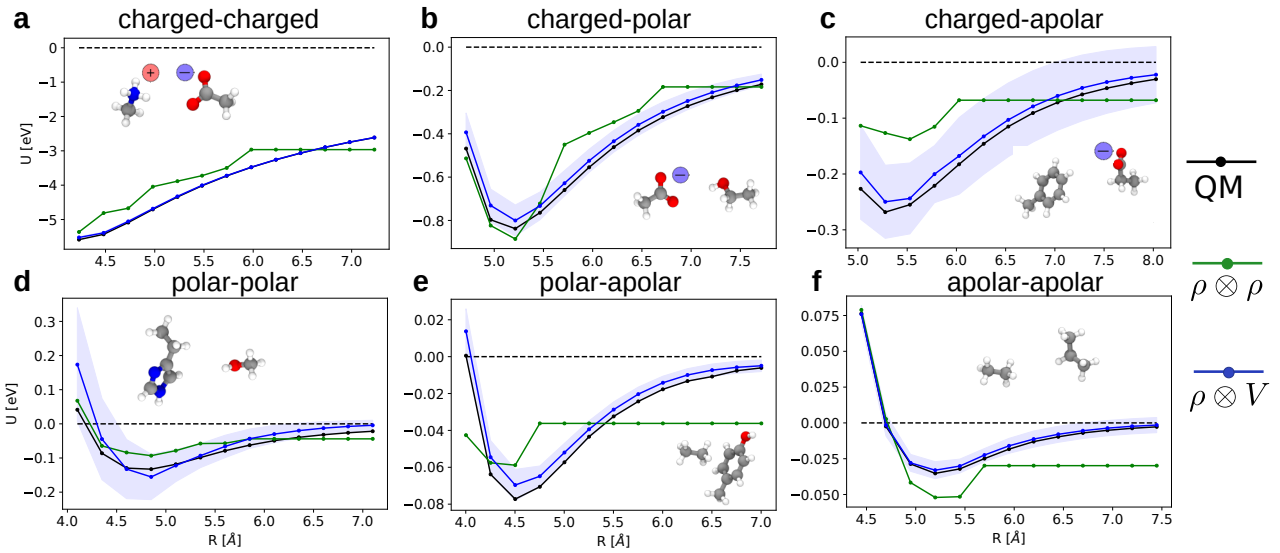


FIG. 9. Median-error binding curves for six different classes of intermolecular interactions, involving charged, polar and apolar molecules extracted from the BioFragment Database[171]. (black lines) reference quantum-mechanical calculations. (green lines) predictions of a local ($|\overline{\rho_i^{\otimes 2}}\rangle$-based) model. (blue lines) predictions of a multi-scale ($|\overline{\rho_i \otimes V_i}\rangle$-based) model. The shaded area indicates the confidence interval for the prediction estimated from a committee model[172]. Reproduced from Ref. 163.

This is a *global* operation, which can however be performed efficiently by transforming the density in plane waves, using one of the many different schemes that are routinely used to model electrostatics. Symmetrizing $|V\rangle$ in the same way as for $|\rho\rangle$ leads to an atom-centered potential,

$$\langle a\mathbf{x}|A;V_i\rangle = \int \mathrm{d}\mathbf{x}' \frac{\langle a\mathbf{x}|A;\rho_i^<\rangle}{|\mathbf{x}-\mathbf{x}'|} + \int \mathrm{d}\mathbf{x}' \frac{\langle a\mathbf{x}|A;\rho_i^>\rangle}{|\mathbf{x}-\mathbf{x}'|} \equiv \langle a\mathbf{x}|A;V_i^<\rangle + \langle a\mathbf{x}|A;V_i^>\rangle \ , \quad (51)$$

where we introduce the short-range density $|\rho_i^<\rangle$, restricted to the region within the cutoff, and the far-field density $|\rho_i^>\rangle$, restricted outside the cutoff, and the corresponding local and non-local fields $|V_i^<\rangle$ and $|V_i^>\rangle$. 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 $\langle anlm|V_i\rangle$.

One can then build features that are fully equivariant by averaging $|V_i\rangle$ over the symmetry operations of the $O(3)$ group, leading to ν-point correlations analogous to those discussed above. Furthermore, one can combine local and long-range fields, as in Figure 10, constructing a family of multi-scale long-distance equivariants (LODE) features[163], that in the most general form can be written as $|\overline{\rho_i^{\otimes\nu} \otimes V_i^{\otimes\nu'};\sigma;\lambda\mu}\rangle$:

$$|\,\langle \rho_i^{\otimes\nu} \otimes V_i^{\otimes\nu'} \otimes \sigma \otimes \lambda\mu\rangle_{O(3)}\rangle = \sum_{k=0,1} \int_{SO(3)} \mathrm{d}\hat{R}\ \hat{\imath}^k\,|\sigma\rangle \otimes \hat{\imath}^k\hat{R}\,|\lambda\mu\rangle \otimes \underbrace{\hat{\imath}^k\hat{R}\,|\rho_i\rangle \cdots \otimes \hat{\imath}^k\hat{R}\,|\rho_i\rangle}_{\nu \text{ times}} \otimes \underbrace{\hat{\imath}^k\hat{R}\,|V_i\rangle \cdots \otimes \hat{\imath}^k\hat{R}\,|V_i\rangle}_{\nu' \text{ times}} \ . \quad (52)$$

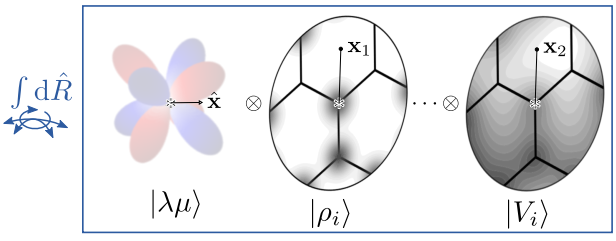


FIG. 10. A multi-scale equivariant representation combining atom-centered density fields $|\rho_i\rangle$, long-range fields $|V_i\rangle$ and a set of spherical harmonics.

The simplest multi-scale representation $|\overline{\rho_i \otimes V_i}\rangle$ can be linked to physics-based models using an atom-centered multipole expansion of electrostatic interactions, as we discuss further in Section V C, but are effective to learn a multitude of long-ranged interactions, from permanent electrostatics, to polarization and dispersion. When trying to represent long-range interactions between molecular fragments, a model based on local $|\overline{\rho_i^{\otimes 2}}\rangle$ features produces a completely unphysical behavior, with the interaction reaching a plateau when the molecules are separated by more than the cutoff distance (Figure 9). 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 repre-

senting 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 models of symmetric properties, such as site energies, which are then used in the context of regression schemes.[10,21,81,95,126,134,144,159,163,183–185] The purpose of this section is therefore to discuss the interplay between representations and models. Given a set of symmetric features $\langle q|A_i\rangle$ of an atomic environment A_i, we explore how to use it to represent a symmetric property $y(A_i)$. We discuss linear approximations,

$$y(A_i) \approx \sum_q \langle y|q\rangle \, \langle q|A_i\rangle \tag{53}$$

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 on the nature of the underlying representations An example of this connection involves the construction of interatomic potentials in terms of a body-ordered hierarchy of atom-centered energy terms

$$E(A) = \sum_{i\in A} E(A_i) = \sum_\nu \sum_{i\in A} E^{(\nu+1)}(A_i), \tag{54}$$

in which each term can be written as a sum over ν neighbors of the central atom

$$E^{(\nu+1)}(A_i) = \sum_{j_1<\cdots<j_\nu} v^{(\nu+1)}(\mathbf{r}_{ji_1},\dots \mathbf{r}_{ji_\nu}). \tag{55}$$

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 ν, i.e. as $\mathcal{O}(N_i^\nu)$ for an environment containing N_i atoms. More sophisticated ways of symmetrizing the body-ordered terms, such as those discussed in Refs. 186 and 95, 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 IV.

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

$$E(A_i) = \sum_j v^{(2)}(r_{ji}) + \sum_{j<j'} v^{(3)}(r_{ji}, r_{j'i}, \omega_{ijj'}), \tag{56}$$

where $\omega_{ijj'} := \hat{\mathbf{r}}_{ji}\cdot\hat{\mathbf{r}}_{j'i}$. In order to connect to the atomic density correlations we first rewrite this as

$$\begin{aligned} E(A_i) &= \sum_j \Big(v^{(2)}(r_{ji}) - \frac{1}{2}v^{(3)}(r_{ji}, r_{ji}, 0)\Big) \\ &\qquad + \frac{1}{2}\sum_{jj'} v^{(3)}(r_{ji}, r_{j'i}, \omega_{ijj'}) \\ &=: \sum_j u^{(2)}(r_{ji}) + \sum_{jj'} u^{(3)}(r_{ji}, r_{j'i}, \omega_{ijj'}), \end{aligned} \tag{57}$$

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

Approximating $u^{(2)}(r)$ in terms of a radial basis $\langle r|n\rangle \equiv R_n(r)$ yields

$$\begin{aligned} E^{(2)}(A_i) &= \sum_j u^{(2)}(r_{ji}) \\ &\equiv \sum_j \langle u^{(2)}|r_{ji}\rangle \approx \sum_j\sum_n \langle u^{(2)}|n\rangle \, \langle n|r_{ji}\rangle \\ &= \sum_n \langle u^{(2)}|n\rangle \int \mathrm{d}r \, \langle n|r\rangle \sum_j \delta(r-r_{ji}) \\ &= \sum_n \langle u^{(2)}|n\rangle \, \langle n|\overline{\delta_i^{\otimes 1}}\rangle \end{aligned} \tag{58}$$

where $|\delta_i\rangle$ is the $g\to\delta$ limit of the atom-centered density $|\rho_i\rangle$. As in Eq. (4), the use of the Dirac notation to express the pair potential highlights the fact that (atom-centered) properties can be seen as a type of representation, and that in this sense a linear model is nothing but an expansion in a discrete basis of $\langle u^{(2)}|r\rangle \equiv u^{(2)}(r)$.

For the three-body term we revisit (22): first, we approximate $u^{(3)}$ in terms of the radial basis $\langle r|n\rangle \equiv R_n(r)$ and the Legendre polynomials $\langle\omega|l\rangle \equiv P_l(\omega)$,

$$u^{(3)}(r_{ji}, r_{j'i}, \omega_{ijj'}) \approx \sum_{nn'l} \langle u^{(3)}|nn'l\rangle \, \langle n|r_{ji}\rangle \, \langle n'|r_{j'i}\rangle \, \langle l|\omega_{ijj'}\rangle \,. \quad (59)$$

Applying Legendre's addition theorem to expand the P_l in terms of spherical harmonics $\langle\hat{\mathbf{r}}|lm\rangle \equiv Y_l^m(\hat{\mathbf{r}})$,

$$\langle l|\omega_{ijj'}\rangle = \frac{4\pi}{2l+1} \sum_{m=-l}^{l} (-1)^m \, \langle lm|\hat{\mathbf{r}}_{ji}\rangle \, \langle l(-m)|\hat{\mathbf{r}}_{ji}\rangle \,,$$

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

$$\begin{aligned} u^{(3)}(r_{ji}, r_{j'i}, \omega_{ijj'}) &= \sum_{nn'l} \langle u^{(3)}|nn'l\rangle \, \langle n|r_{ji}\rangle \, \langle n'|r_{j'i}\rangle \\ &\quad \times \sum_{m=-l}^{l} (-1)^m \, \langle lm|\hat{\mathbf{r}}_{ji}\rangle \, \langle l(-m)|\hat{\mathbf{r}}_{j'i}\rangle \\ &= \sum_{nn'l} \langle u^{(3)}|nn'l\rangle \sum_{m} (-1)^m \, \langle nlm|\mathbf{r}_{ji}\rangle \, \langle n'l(-m)|\mathbf{r}_{j'i}\rangle \,. \end{aligned} \quad (60)$$

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

$$\begin{aligned} \sum_{jj'} u^{(3)}(r_{ji}, r_{j'i}, \omega_{ijj'}) &= \sum_{nn'l} \langle u^{(3)}|nn'l\rangle \\ \sum_{m} (-1)^m \sum_{j} \langle nlm|\mathbf{r}_{ji};\delta\rangle \sum_{j'} \langle n'l(-m)|\mathbf{r}_{j'i};\delta\rangle \\ = \sum_{nn'l} \langle u^{(3)}|nn'l\rangle \sum_{m} (-1)^m \, \langle nlm|\delta_i\rangle \, \langle nl(-m)|\delta_i\rangle \\ = \sum_{nn'l} \langle u^{(3)}|nn'l\rangle \, \langle nn'l|\overline{\delta_i^{\otimes 2}}\rangle \,. \end{aligned} \quad (61)$$

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

$$\begin{aligned} E(A_i) = \sum_{n} \langle u^{(2)}|n\rangle \, \langle n|\overline{\delta_i^{\otimes 1}}\rangle \\ + \sum_{nn'l} \langle u^{(3)}|nn'l\rangle \, \langle nn'l|\overline{\delta_i^{\otimes 2}}\rangle \end{aligned} \quad (62)$$

Aside from connecting classical body-ordered interatomic potentials and ν-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[90], to the rationalization of fragment-based electronic structure methods[187], and to the construction of last-generation potentials for water and aqueous systems[175].

We adopt the generalisation of (57) that includes self-interaction,

$$E^{(\nu+1)}(A_i) = \sum_{j_1,\ldots,j_\nu} u^{(\nu+1)}(\mathbf{r}_{j_1 i} \ldots \mathbf{r}_{j_\nu i}), \quad (63)$$

which can be obtained from the more natural formulation (55) by incorporating the self-interaction terms into the ν-body-order energy similarly to Eq. (57).

To connect (63) to the density correlations we represent the rotationally invariant $(\nu+1)$-body function $u^{(\nu+1)}$ as

$$\begin{aligned} &u^{(\nu+1)}(\mathbf{r}_{j_1 i}, \ldots, \mathbf{r}_{j_\nu i}) \\ &= \int_{O(3)} \mathrm{d}\hat{R} \int \mathrm{d}Q \, \langle u^{(\nu+1)}|Q\rangle \, \langle Q|\, \hat{R}\, |\mathbf{r}_{j_1 i}, \ldots, \mathbf{r}_{j_\nu i}\rangle \,, \end{aligned} \quad (64)$$

where we use Q as a shorthand for $(\mathbf{x}_1; \ldots \mathbf{x}_\nu)$, so that $\langle Q|\mathbf{r}_{j_1 i}, \ldots, \mathbf{r}_{j_\nu i}\rangle \equiv \prod_{k=1}^{\nu} \delta(\mathbf{x}_k - \mathbf{r}_{j_k i})$. 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

$$\begin{aligned} &E^{(\nu+1)}(A_i) \\ &\approx \sum_{j_1,\ldots,j_\nu} \int_{O(3)} \mathrm{d}\hat{R} \int \mathrm{d}Q \, \langle u^{(\nu+1)}|Q\rangle \, \langle Q|\, \hat{R}\, |\mathbf{r}_{j_1 i} \cdots \mathbf{r}_{j_\nu i}\rangle \\ &= \int \mathrm{d}Q \, \langle u^{(\nu+1)}|Q\rangle \int_{O(3)} \mathrm{d}\hat{R} \sum_{j_1,\ldots,j_\nu} \langle Q|\, \hat{R}\, |\mathbf{r}_{j_1 i} \cdots \mathbf{r}_{j_\nu i}\rangle \,. \end{aligned} \quad (65)$$

The symmetrized sum can be reordered to show that it corresponds to the ν-point density correlation

$$\begin{aligned} &\int_{O(3)} \mathrm{d}\hat{R} \sum_{j_1,\ldots,j_\nu} \langle \mathbf{x}_1; \ldots \mathbf{x}_\nu|\, \hat{R}\, |\mathbf{r}_{j_1 i} \cdots \mathbf{r}_{j_\nu i}\rangle \\ &\quad = \int_{O(3)} \mathrm{d}\hat{R} \sum_{j_1 \ldots j_\nu} \prod_{k} \delta(\hat{R}\mathbf{x}_k - \mathbf{r}_{j_k i}) \\ &\quad = \int_{O(3)} \mathrm{d}\hat{R} \prod_{k} \sum_{j_k} \delta(\hat{R}\mathbf{x}_k - \mathbf{r}_{j_k i}) \\ &= \int_{O(3)} \mathrm{d}\hat{R} \prod_{k} \langle \hat{R}\mathbf{x}_k|\delta_i\rangle = \langle \mathbf{x}_1; \ldots \mathbf{x}_\nu|\overline{\delta_i^{\otimes \nu}}\rangle \,, \end{aligned} \quad (66)$$

which is precisely Eq. (20) written in the $g \to \delta$ 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 (63) – even though, if one wishes so, they can be removed from the final result[188].

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

$$E^{(\nu+1)}(A_i) \approx \sum_q \langle u^{(\nu+1)}|q\rangle \, \langle q|\overline{\delta_i^{\otimes\nu}}\rangle \,. \qquad (67)$$

Any complete implementation of ν-order density correlation features[126,127,134,149] provides a basis to expand $u^{(\nu+1)}$ and approximate the $(\nu+1)$-order term, that contributes to the body-ordered expansion of $E(A)$. The foregoing discussion shows that these bases are complete in the following sense. An (infinite) collection of symmetrized features $\{\langle q|A_i\rangle\}_{q\in\mathbf{q}_\mathrm{total}}$ is a *complete linear basis* if there exists a sequence of finite subsets $\mathbf{q} \subset \mathbf{q}_\mathrm{total}$ such that

$$y(A_i) \approx y_\mathbf{q}(A_i) := \sum_{q\in\mathbf{q}} \langle y|q\rangle \, \langle q|A_i\rangle \,, \qquad (68)$$

i.e. $y_\mathbf{q}$ approximates y to within arbitrary accuracy in the limit as the number of features tends to infinity. We stress here that the weights $\langle y|q\rangle$ 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

$$\langle q| = \otimes_{\alpha=1}^{\nu} \langle n_\alpha l_\alpha m_\alpha| \qquad (69)$$

Eq. (67) is the ACE model[126,127]. Note that the symmetrized correlations $\langle q|\overline{\delta_i^{\otimes\nu}}\rangle$ can be efficiently and conveniently evaluated as already hinted at in Section IV E. 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 [134], and the essence of the idea can be traced back to the cluster expansion theory of alloys[90]. 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 $\binom{N_i}{\nu}$ scaling of the naive representation (55), which enables modeling very high body-orders. A recursive evaluation of the ν-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 VIII D for a summary of these techniques).

B. 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[189] description of the electronic structure of a molecule or a condensed-phase system as a collection of atom-centered densities. In the literature using SOAP features, the width of the atom-centrered Gaussians has been often indicated as a hyperparameter with an important influence on the robustness[190] and accuracy[191,192] of the resulting machine-learning models. 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. Note that the following discussion provides only *theoretical guarantees*; we explain below that excessive smearing creates severe numerical ill-conditioning which must be carefully considered in practical implementations.

We begin by noting that the expansion of a smeared density in a basis $\langle x|n\rangle$ is identical to the expansion of a δ-like density in the corresponding smeared (a.k.a. *mollified*) basis $\langle x|n;g\rangle \equiv \int \mathrm{d}x' \, \langle n|x'\rangle \, g(x-x')$:

$$\begin{aligned}\langle n|\rho\rangle &= \int \mathrm{d}x \, \langle n|x\rangle \sum_i g(x-x_i) \\ &= \int \mathrm{d}x \sum_i \delta(x-x_i) \int \mathrm{d}x' \, \langle n|x'\rangle \, g(x-x') \\ &= \int \mathrm{d}x \sum_i \delta(x-x_i) \, \langle n;g|x\rangle = \langle n;g|\delta\rangle \,. \end{aligned} \qquad (70)$$

With this observation in hand showing that $\langle x|n;g\rangle$ inherits completeness from $\langle x|n\rangle$ is sufficient to ensure that all our results apply also to smeared densities.

We first consider the case of standard monomials. Any continuous function $f(x)$ can be expanded to within arbitrary accuracy into polynomials x^n :

$$f(x) \approx f_{n_\mathrm{max}}(x) = \sum_{n=0}^{n_\mathrm{max}} c_n x^n \underset{n_\mathrm{max}\to\infty}{\to} f(x). \qquad (71)$$

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

$$p_n^g(x) = g * x^n = \int (t-x)^n e^{-t^2/2\sigma^2}/\sqrt{2\sigma^2\pi} \, \mathrm{d}t \,. \qquad (72)$$

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

$$p_n^g(x) = x^n + \text{lower order terms}, \qquad (73)$$

i.e., p_n^g 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 $f_{n_{\max}}(x)$ *exactly* as

$$f_{n_{\max}}(x) = \sum_{n=0}^{n_{\max}} c'_n p^g_n(x) \tag{74}$$

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

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

$$\begin{aligned} p^g_n(x) = g * x^n &\approx \sum_j b_{nj} \int \mathrm{d}x'\, g(x-x') \langle x'|j\rangle \\ &= \sum_j b_{nj} \langle x|j;g\rangle \,. \end{aligned} \tag{75}$$

Given that p^g_n are dense, it follows that also the smeared basis functions $\langle x|j;g\rangle \equiv g * \langle x|j\rangle$ are dense. From these arguments it is reasonable to conclude that the smeared density correlations also form a complete linear basis.

As already mentioned above, this is a purely theoretical statement, and there is 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 size of the basis is increased, even if f has a stable expansion in a polynomial basis. Therefore, in practice, the smoothing 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. Failure to do so may result in a representation that has insufficient resolution to describe the response of the target property to structural deformations, or vice versa to one that contains redundant information and is prone to overfitting.

C. 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 (52) within an additive, linear learning model, using as target the electrostatic energy $U(A)$,

$$U(A) = \sum_{i\in A} U(A_i) = \sum_{i\in A} \int \mathrm{d}Q\, \langle U|Q\rangle\, \langle Q|A;\overline{\rho_i \otimes V_i}\rangle \,. \tag{76}$$

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

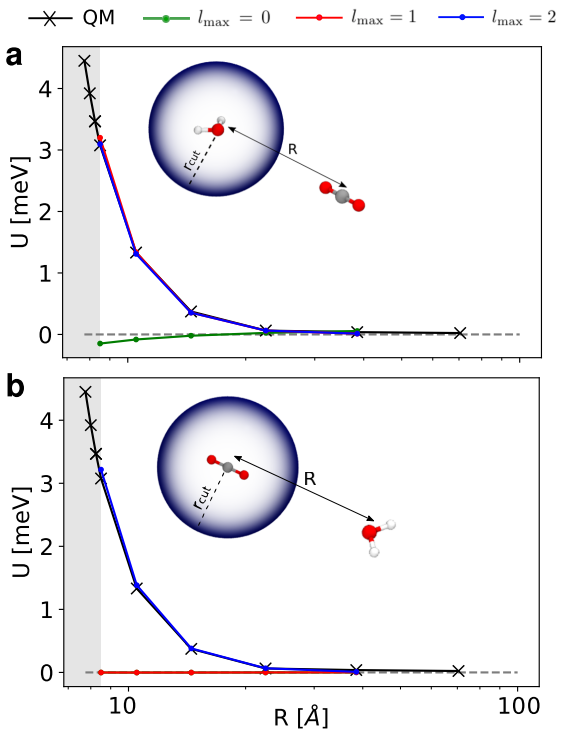


FIG. 11. Extrapolated asymptotic interaction profiles for a given configuration of H_2O and CO_2 at different angular cutoff values $l_{\max}$. Top and bottom panels show the results of the asymptotic extrapolation when centring the representation (a) on the oxygen atom of H_2O and (b) on the carbon atom of CO_2. Adapted with permission from Ref. 163. Copyright 2020 Royal Society of Chemistry.

far away.[163] Focusing only on the long-range contribution $U^>$ to $U(A_i)$, that is associated with the part of $|A;V_i\rangle$ generated by the far-field density, $|A;V_i^>\rangle$, one can write

$$\begin{aligned} U^>(A_i) &= \sum_{l=0}^{l_{\max}} \int \mathrm{d}r_1\, \mathrm{d}r_2\, \langle U|r_1 r_2 l\rangle\, \langle r_1 r_2 l|\overline{\rho_i^< \otimes V_i^>}\rangle \\ &= \sum_{l=0}^{l_{\max}} \sum_{m=-l}^{+l} \int_{r_{\mathrm{cut}}}^{\infty} \mathrm{d}r\, \frac{1}{r^{l+1}} \langle lm|M_i^<(U)\rangle\, \langle \rho_i^>|rlm\rangle \,. \end{aligned} \tag{77}$$

In this expression, in which the reader can recognize the similarity with the multipole expansion of the electrostatic potential,[158] $|\rho_i^>\rangle$ indicates the atom density outside the cutoff, which is not computed explicitly but is encoded in the expansion of the local atomic potential (50). The coefficients $\langle lm|M_i^<(U)\rangle$ can be written as a combination of the regression weights $\langle r_1 r_2 l|U\rangle$ and the local density coefficients $\langle rlm|\rho_i^<\rangle$, 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 H_2O and 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 atoms, as well as by truncating the angular expansion at different $l_{\max}$. It is apparent that the angular cutoff chosen reflects the number of multipoles introduced in the expansion of Eq. (77) and thus determines sharp crossovers of the prediction accuracy across critical $l_{\max}$ values. For instance, a model that uses only features centered on the oxygen atom of H_2O improves dramatically its performance when $l_{\max}$ is increased from zero to one. A model using the carbon atom of CO_2 as the only environment shows a similar, sharp improvement in accuracy when going from $l_{\max} = 1$ to $l_{\max} = 2$. 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 $|\overline{\rho_i \otimes V_i}\rangle$ and multipole electrostatics, the representation is sufficiently flexible to describe also other kinds of interactions, as demonstrated in Figure 9.

D. 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. 87,186, and with exponential scaling in their computational cost of prediction due to the need to sum over all ν-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 potentials[10,12]. 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. 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 taken as the input of a non-linear model, instead, the entries of the feature vector must always be precisely defined, because there is no complete basis set limit in which the models become equivalent. To emphasize that many of the formal manipulations that are possible in a linear context take on a different meaning when features are used for a non-linear model, we abandon the Dirac notation and indicate as $\boldsymbol{\xi}(A_i)$ the feature vector that describes the atom-centred environment A_i, whose components are $\xi_q(A_i) = \langle q|A_i\rangle$. If $y(A_i)$ is a symmetric property such as a site energy, we aim to construct approximations of the general form

$$y(A_i) \approx \tilde{y}\big(\boldsymbol{\xi}(A_i)\big). \tag{78}$$

The two most commonly used models for $\tilde{y}$ are artificial neural networks[20,21,35,36,107,152,165,184,193,194] (ANN) and kernel ridge regression[22,23,30,34,116,179,185,191,195,196] (KRR) models. In KRR models,[197] one builds a kernel matrix $\mathbf{K}$ with elements

$$K_{ij} = \mathrm{k}(\boldsymbol{\xi}(A_i), \boldsymbol{\xi}(A_j)), \tag{79}$$

which provides a similarity measure between the environments A_i and A_j, measured in terms of the similarity between the corresponding feature vectors $\boldsymbol{\xi}(A_i)$ and $\boldsymbol{\xi}(A_j)$. Useful kernel functions, k, are nonlinear, e.g. polynomials, Gaussians, etc.[198]. The kernel inherits the symmetry of the feature vectors, and therefore a model for a symmetry-invariant property $y(A_i)$ can be obtained as

$$\tilde{y}(A_i) = \sum_{j\in M} b_j\, \mathrm{k}(\boldsymbol{\xi}(A_i), \boldsymbol{\xi}(M_j)), \tag{80}$$

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 $\{\mathrm{k}(\cdot, \boldsymbol{\xi}(M_j))\}_j$ in the symmetrized feature space. The weights b_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" $|A_i\rangle \to |A_i;\mathrm{k}\rangle$, which has a larger (often infinite) dimensionality, allowing for a more flexible approximation of $y(A_i)$. (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 $|A_i;\mathrm{k}\rangle$ 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. 33. 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 a 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.

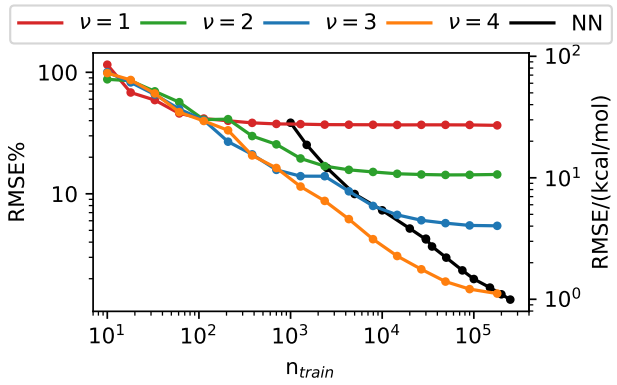


FIG. 12. Learning curves for the formation energy of CH_4 structures using linear models based on NICE features truncated to increasing body order ν ($r_{\rm cut} = 6$Å, $n_{\rm max} = 10$, and $l_{\rm max} = 10$, up to 3200 invariants retained at each body order) and an ANN model using NICE features up to $\nu = 4$. Errors are expressed both in absolute terms and as a percentage of the standard deviation of the dataset. The models are trained using features centered on both C and H. Reproduced with permission from Ref. 149. Copyright 2020 American Institute of Physics.

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 V A and Section V C, 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 SNAP,[147,148], the MTP,[134], the ACE[126] and the NICE[149] 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 ν-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 (55), one should convert it for computationally efficient evaluation to one of the many representation built in terms of ν-correlations. By employing the recursive evaluations introduced in Refs. 127,134,149, 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 VIII D. 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.[30,123] 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 $\nu = 4$ model starts saturating around $n_{\rm train} = 10^5$, 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 and structurally diverse regression exercise, usually translates to poorer transferability.

VI. ALTERNATIVE NOTIONS OF COMPLETENESS

Suppose we are given a finite collection of symmetry adapted features $\boldsymbol{\xi}(A) = \{\langle q|A\rangle\}_q$ which we wish to use as a *descriptor* for atomic structures or environments, for example symmetrized correlations of the density as described in the foregoing sections. In Section V we discussed two classes of models built from such equivariant features: linear models,

$$A \mapsto \sum_q \langle y|q\rangle \, \langle q|A\rangle \,, \tag{81}$$

for which the representation $\boldsymbol{\xi}(A)$ plays the role of a basis to expand the target property; and nonlinear models,

$$A \mapsto \tilde{y}(\boldsymbol{\xi}(A)), \tag{82}$$

where the representation plays the role of a coordinate transformation generating a finite-dimensional feature vector used as the argument of a non-linear function $\tilde{y}$. In order to guarantee systematic convergence of these models to an arbitrary target, in suitable limits, we require that the employed set of features is *complete*. We already hinted in Section V D 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.

Recall from Section V D that for linear models the correct notion of completeness is the well-known and well-understood concept of a complete (linear) basis from linear algebra. In the context of a nonlinear

model $\tilde{y}(\boldsymbol{\xi}(A))$ it is instructive to think of $\tilde{y}$ as a universal approximator in feature space (e.g., an ANN, GP, etc). We then ask the question whether (in a suitable limit) the model can represent an arbitrary symmetric property $y(A)$, i.e., whether

$$y(A) = \tilde{y}(\boldsymbol{\xi}(A)), \tag{83}$$

is achievable. This is the case if and only if the mapping $A \mapsto \boldsymbol{\xi}(A)$ is *injective*: this means that any two atomic configurations that are *not* related by symmetry are mapped to different descriptors. In particular knowledge of $\boldsymbol{\xi}$ would then enable us in principle to reconstruct the configuration A. When this is the case, we say that the descriptor $\boldsymbol{\xi}$ is *geometrically complete*.

A. A pedagogical 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:

$$\langle n|\rho\rangle = \sum_{i=1}^{N} x_i^n, \quad n \in \mathbb{N} \tag{84}$$

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 ν-point density correlations form a basis of symmetric polynomials

$$\langle n_1 \cdots n_\nu|\rho^{\otimes\nu}\rangle = \sum_{i_1 \ldots i_\nu} x_{i_1}^{n_1} \cdots x_{i_\nu}^{n_\nu} = \prod_{k=1}^{\nu} \langle n_k|\rho\rangle \tag{85}$$

which is complete (in the sense of a linear basis) because it contains all possible symmetrized monomials. In analogy to what we did in Section V 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 $\langle n_1 n_2|\rho^{\otimes 2}\rangle = x_1^{n_1+n_2} + x_1^{n_1}x_2^{n_2} + x_2^{n_1}x_1^{n_2} + x_2^{n_1+n_2}$ 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 (85) 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[199] 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 $\langle 2|\rho\rangle$

$$\begin{aligned}\langle 3|\rho\rangle = x_1^3 + x_2^3 &= \frac{3}{2}(x_1 + x_2)(x_1^2 + x_2^2) - \frac{1}{2}(x_1 + x_2)^3 \\ &= \frac{3}{2}\,\langle 1|\rho\rangle\ \langle 2|\rho\rangle - \frac{1}{2}\,\langle 1|\rho\rangle^3\end{aligned} \tag{86}$$

This result implies, in general, that the mapping

$$\{x_i\}_{i=1}^N \mapsto \boldsymbol{\xi} = \{\,\langle n|\rho\rangle\}_{n=1}^N \tag{87}$$

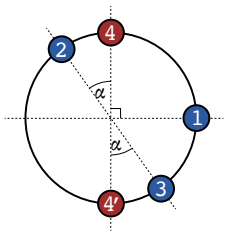
is injective: knowledge of the first N features $\langle n|\rho\rangle$ allows us to uniquely reconstruct the configuration (but not the index of the atoms). That is, this *minimal feature set* $\boldsymbol{\xi}(A)$ is indeed *geometrically complete*. It is not too difficult to construct similar complete 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* ν-correlations provides a complete set of features (after all, they are even complete in the sense of forming a complete linear basis), however, as we explained at the beginning of Sec. VI, this is not a practically useful property when considering nonlinear regression schemes. 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 ν-point correlations have a number of components that scales with N^ν. 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 symmetrized 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[29,122,200], it was not until recently that the connection to environment descriptors was made[124]. The fact that degenerate pairs of inequivalent environments mapping to the same descriptor exist for two-correlation (distance-angle) representations came as a surprise because so many "successful" models for potential energy surfaces have been published based on



	1	2	3	4
1	0	$\frac{\pi}{2}+\alpha$	$\frac{\pi}{2}-\alpha$	$\frac{\pi}{2}$
2		0	π	$\boldsymbol{\alpha}$
3			0	$\boldsymbol{\pi-\alpha}$
4				0

	1	2	3	4’
1	0	$\frac{\pi}{2}+\alpha$	$\frac{\pi}{2}-\alpha$	$\frac{\pi}{2}$
2		0	π	$\boldsymbol{\pi-\alpha}$
3			0	$\boldsymbol{\alpha}$
4’				0

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 α 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.

such descriptors in the past decade[14,33]. An example of such “degenerate pair” is given in Figure 13. The construction involves an environment with four neighbors on the unit circle, with the two structures corresponding to the labels $(1,2,3,4)$ and $(1,2,3,4')$ being different, but having the same unordered list of distances and angles. The total number of degrees of freedom for this layout is three (because one neighbor can be fixed on the x axis), and there is one degree of freedom in the construction of the degenerate pair (the angle labelled α in Fig. 13). Thus, this manifold of pairs of degenerate configurations has a *codimension* of two, i.e. it has a dimensionality that involves two fewer degrees of freedom than the total. A more general construction, that yields a family of 3D degenerate pairs including an arbitrary number of neighbors, is discussed in Ref. 124. The fact that the degenerate pairs form a manifold does not mean that there is a *degenerate manifold*, i.e. a manifold of configurations all mapping to the same descriptor. This type of degeneracy occurs between pairs of configurations which are typically far from one another, and so this degeneracy problem differs from that of assessing the sensitivity of a representation to small atomic displacements[201,202].

As was shown in Ref. 124, in order to break this degeneracy, the correlation order has to be increased. Three-correlations ($|\overline{\rho_i^{\otimes 3}}\rangle$, equivalent to the unordered set of central tetrahedra, and the bispectrum of the atomic density) indeed distinguish environments such as those in Fig. 13. It is however possible to build pairs of environments, composed of 7 or more neighbors, which are distinct but have the same three-correlations. This example, also discussed in Ref. 124, raises a number of open mathematical questions: (i) is the $\nu = 3$ descriptor complete for $N < 7$ neighbors, (ii) are all ν-correlations degenerate for sufficiently many neighbors, (iii) what is the codimension of the manifold of degenerate configurations for $\nu > 2$?

The concept of completeness applies both to representing entire structures and to atomic environments, 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 $\nu = 2$ correlations around a central atom contain information on the full set of interparticle distances between the neighbors, and so any pair of environments that are degenerate in terms of $|\overline{\rho_i^{\otimes 2}}\rangle$ is also (removing the particle at the origin) a pair of *structures* with a degenerate description in terms of distances.[203] Note that the problem of completeness for entire structures is exactly the same as the problem of reconstructing point sets[122].

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. (17). Following the above reasoning, a pair of environments that are degenerate in terms of the list of distances and angles are also (removing the central atom) 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.[124] 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 hold 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[204], 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[205,206]. See also Uhrin [207] for an excellent review connecting 3D signal processing and reconstruction of atomic configurations.

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[134], PIP[186], aPIP[95], ACE[126,208] and NICE[149] (see also Section V A). Any of these bases can be expressed in terms of the ν-correlations via a linear transformation, and vice-versa. Even for fixed maximum polynomial degree, these are enormous representations. Depending on how ν-correlation features are chosen their number might scale as rapidly as $\binom{q_{\max}+\nu}{\nu}$, where $q_{\max}$ is the number of one-particle features.

There is a class of much lower-dimensional descriptor maps based on the eigenspectra of overlap matrices[68,106] that lifts the degeneracy for the known examples, although their actual completeness is unknown. A simplified 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:

$$T_{jj'} = f_{\rm cut}(r_{ji}) f_{\rm cut}(r_{ij'}) t(r_{jj'}), \tag{88}$$

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) $\{\tau_k\}_k$ is an invariant descriptor of A_i. Due to eigenvalue crossings, the mapping $A_i \mapsto \{\tau_k\}_k$ is non-smooth, hence one may wish to project it on a smooth basis, e.g. polynomials,

$$\langle n | A_i ; \mathbf{T} \rangle := \sum_k (\tau_k)^n. \tag{89}$$

The spectral features (or, *fingerprints* as they are also called[106]) $\{ \langle n|\mathbf{T}\rangle\}_{n=1}^N$ correspond to the moments of the histogram of eigenvalues, and contain precisely the same information.

An alternative way to write $\langle n|T\rangle$ is

$$\langle n|\mathbf{T}\rangle = \operatorname{Tr} \mathbf{T}^n, \tag{90}$$

which is *not* computationally more efficient, but highlights the close connection between $\{ \langle n|\mathbf{T}\rangle\}_n$ and the body-ordered features we discussed in previous sections. From (90) we observe that

$$\begin{aligned}
\langle 1|\mathbf{T}\rangle &= N t(0), \\
\langle 2|\mathbf{T}\rangle &= \sum_{j_1,j_2} t(r_{j_1 j_2})^2 \cdot f_{\rm cut}(r_{ij_1}) f_{\rm cut}(r_{ij_2}) \\
\langle 3|\mathbf{T}\rangle &= \sum_{j_1,j_2,j_3} t(r_{j_1 j_2}) t(r_{j_2 j_3}) t(r_{j_3 j_1}) \cdot \prod_{\alpha=1}^{3} f_{\rm cut}(r_{ij_\alpha}),
\end{aligned} \tag{91}$$

and so forth. That is, $\langle n|\mathbf{T}\rangle$ contains the projection of the histogram of n-simplices onto a single basis function. In other words, for $n = 2$, the cutoff function $f_{\rm cut}$ and the overlap function t play the role of R_n and P_l in (37). More in general, $\langle n|\mathbf{T}\rangle$ describes n-neighbors correlations, and so it could be written, in principle, as a linear combination of a complete set of $|\overline{\rho_i^{\otimes n}}\rangle$ features. Thus, the $\langle n|\mathbf{T}\rangle$ provide invariant high body-order features at relatively low computational cost, even though each scalar overlap matrix $\mathbf{T}$ contains information on a single feature per body order.

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 (89) 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. 106. 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[127,149] 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 $\mathbf{r}_{ji}$, 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

$$T_{jj'} := \cos \theta_{ijj'}$$

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

Even for a general atomic environment, $T_{jj'} = r_{ji} r_{j'i} \cos\theta_{ijj'}$ is the Gram matrix of the interatomic distance vectors, which has at most three non-zero eigenvalues – and hence the collection $(\langle n|\mathbf{T}\rangle)_{n=1}^N$ contains at most three independent features even though *formally*, $\langle n|\mathbf{T}\rangle$ has body-order n. 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 $t(r_{jj'})$, and for the general case with an appropriate (albeit contrived) choice of $f_{\rm cut}$ 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.

D. Completeness: summary and open challenges

To conclude our discussion of *completeness of representations* we briefly review and contrast the two

key notions of completeness that we introduced and also mention a third concept that we implicitly encountered in Sec. VI A. In the following, let $\boldsymbol{\xi}(A) = \{\langle q|A\rangle\}_q$ again denote a finite or infinite collection of equivariant features of a configuration or environment A.

Complete linear basis: This is the correct notion of completeness of $\boldsymbol{\xi}$ for *linear models*, $\sum_q \langle y|q\rangle \langle q|A\rangle$, such as PIPs, aPIPs, MTP, ACE, NICE. It is now well-understood how to systematically generate such a complete linear basis in a variety of different ways. This is the strongest requirement one can make on a feature set.

Geometric completeness: This is the correct notion of $\boldsymbol{\xi}$ completeness for nonlinear models, $\tilde{y}(\boldsymbol{\xi}(A))$, i.e., it is the minimal requirement to ensure systematic convergence of such a model. Ensuring only injectivity of the mapping $A \mapsto \boldsymbol{\xi}(A)$, means it is a much weaker requirement than being a complete linear basis. We therefore expect that complete feature vectors are generally significantly sparser, which is important for the performance of nonlinear regression schemes. At present, there is no systematic construction of minimal geometrically complete feature sets.

Algebraic completeness: We say that $\boldsymbol{\xi}$ *is algebraically complete* if every element of a complete linear basis $\langle q|A;\overline{\rho_i^{\otimes\nu}}\rangle$ can be written as a polynomial of the entries of $\boldsymbol{\xi}$, $p_q(\boldsymbol{\xi}(A_i))$. This is precisely the concept we used to construct a geometrically complete feature set in the pedagogical example of Sec. VI A. The set of invariants used to construct PIP[186] and aPIP[95] potentials form a minimal algebraically complete descriptor. The concept was also proposed as part of the NICE framework[149] as a mechanism to reduce the size of descriptor set.

In general, algebraic completeness is strictly stronger than geometric completeness and an algebraically complete feature set will be larger than a minimal geometrically complete one. It is nevertheless an interesting and useful concept: (i) it provides a stepping stone towards a theoretical understanding of geometric completeness; (ii) for the purpose of effective regression schemes it may in fact prove to be more important since it preserves polynomials, while inverting a minimal geometrically complete descriptor is likely to introduce singularities. Indeed, reducing algebraic dependence is a common technique in the signal processing literature. Uhrin [207] reviews those techniques and modifies them for the construction of descriptors with relatively few entries, that can in principle be made complete.

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[210–221], to obtain low-dimensional projections that help visualize complex datasets[1,4,6,222–226], and more generally to describe the lie of the land in (free)energy landscapes and interpret structure-property relationships in complex systems[38,227–230]. 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[231], discrete secondary-structure categories[232,233], as well as sophisticated continuous fingerprints of secondary structure and backbone chirality[234,235]. Simulations of clusters and condensed-phase systems have often used more general indicators, such as Steinhardt order parameters[236], cubic harmonics[237,238], radial distribution functions (either directly[239,240] or in the form of entropy-inspired fingerprints[241]), histograms of coordination numbers[4], that can be seen as precursors of the atom-density correlation representations that we discuss in Section IV D. More broadly, general-purpose descriptors that can be understood, more or less transparently, as a special case of the density-correlation features $|\overline{\rho_i^{\otimes\nu}}\rangle$ 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.[138], the local order metric of Martelli et al.[242], the spectral representations of Sadeghi et al.[68], the Minkowski structure metric of Mickel et al.[243] (that closely resembles and anticipates the construction of the moment tensor potentials), and the use of SOAP features to analyze materials and molecules[69,70,244].

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[245]. 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 $\langle n_1 n_2 l|\overline{\rho_i^{\otimes 2}}\rangle$) to analyze a simulation of undercooled iron[209]. Atoms are colored according to a two-dimensional projection describing the associated environments, in this case obtained using a kernel principal component analysis[246] built on the feature vectors $\boldsymbol{\xi}(A_i)$. 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

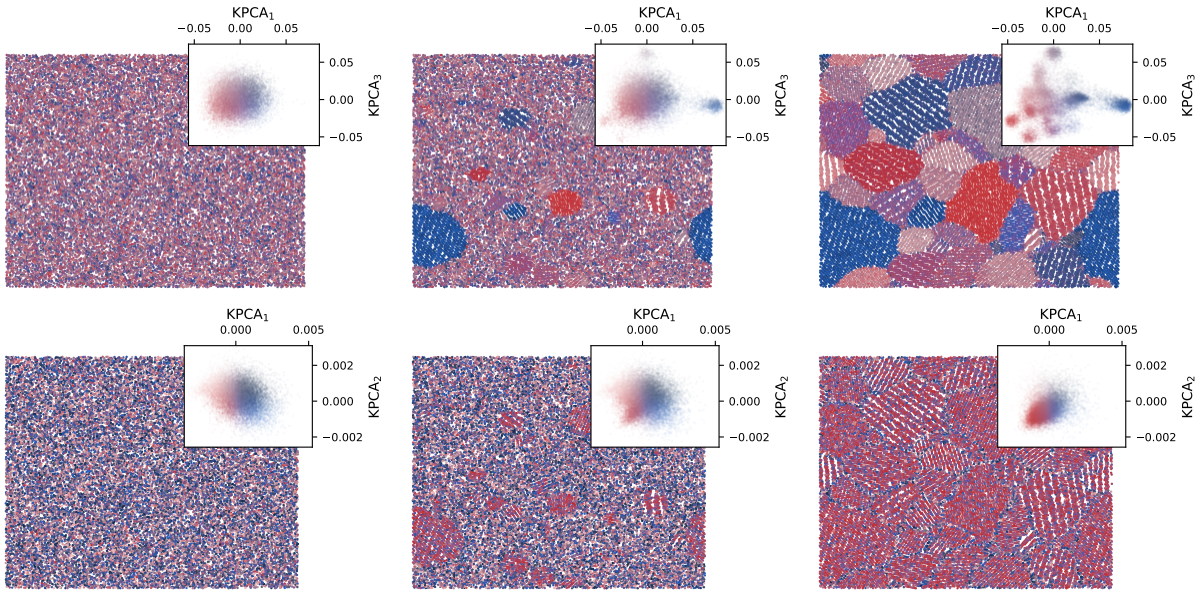


FIG. 14. Visualizing crystallization in a million-atoms simulation of undercooled iron (data from Ref. 209). The inset shows a KPCA map of the environments, and the atoms are color-coded following the same scheme. Top: map and coloring based on translationally-invariant $\langle nlm|\rho_i\rangle$. Bottom: map and coloring based on fully-invariant $\langle n_1 n_2 l|\overline{\rho_i^{\otimes 2}}\rangle$.

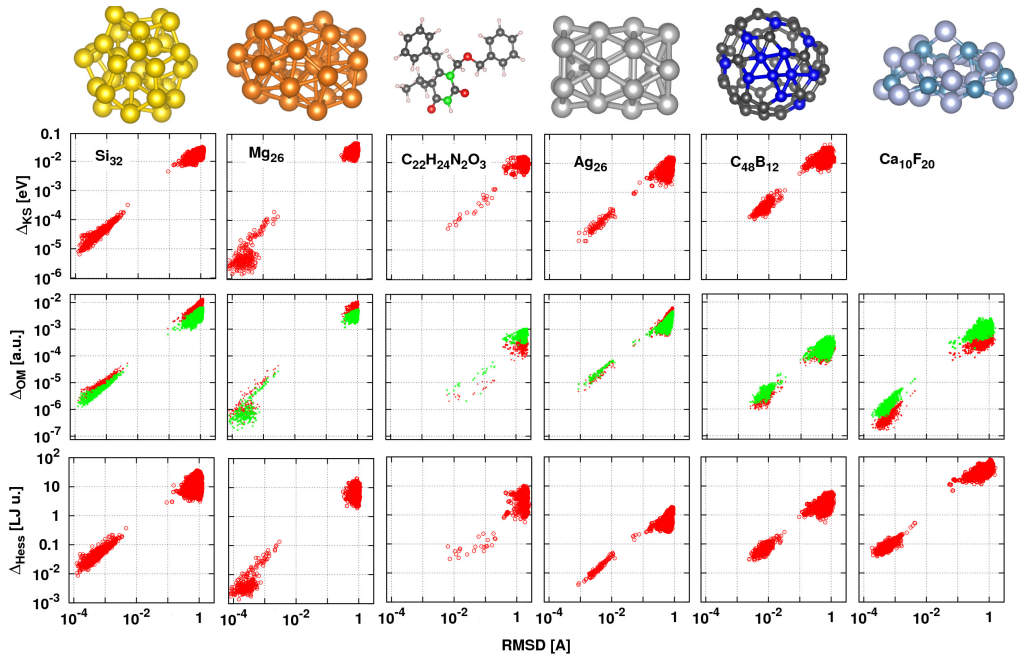


FIG. 15. Comparison of distances between local minimum-energy configurations of various clusters (rows) constructed based on the sorted eigenvalues of the Kohn-Sham Hamiltonian matrix (first row), the overlap matrix (second row), and the Lennard-Jones Hessian matrix, and plotted against a permutation-invariant RMSD. For the overlap matrix, results are shown for matrices based only on s-type orbitals (red) and both s and p orbitals (green). Details of the different systems and the fingerprint construction are discussed in Ref. 68. Reproduced with permission from Ref. 68. Copyright 2013 American Institute of Physics.

possible to recognize the disordered environments at the grain boundaries. This kind of analysis can also be used to elucidate the properties of different representations, investigating the effect of different choices on the unsupervised analysis of a well-understood system to better appreciate the relation between structure 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 $\boldsymbol{\xi}(A_i)$, 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 $\boldsymbol{\xi}$, it is possible to define a distance using e.g. a Euclidean metric, $\mathrm{d}(A_i, A'_{i'})^2 = \|\boldsymbol{\xi}(A_i) - \boldsymbol{\xi}(A'_{i'})\|^2$, and use as a kernel the scalar product $\mathrm{k}(A_i, A'_{i'}) = \boldsymbol{\xi}(A_i) \cdot \boldsymbol{\xi}(A'_{i'})$, or a non-linear function, e.g. an exponential of a squared distance $\mathrm{k}(A_i, A'_{i'}) = \exp -\gamma\, \mathrm{d}(A_i, A'_{i'})^2$.

The opposite is also true: for a given set of configurations M, and any (negative definite) distance or (positive definite) kernel[247] 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 $K_{ij} = \mathrm{k}(M_i, M_j)$, and find its eigenvalues and eigenvectors $\mathbf{K}\mathbf{u}^{(j)} = \lambda_j \mathbf{u}^{(j)}$. It is easy to see that the scalar product between the reproducing features

$$\phi^K_j(A) = \sum_{i \in M} \mathrm{k}(A, M_i) u_i^{(j)} / \sqrt{\lambda_j} \qquad (92)$$

computed for two members of the reference dataset yields exactly the value of the kernel function between the two configurations.[246] It is also possible to define a kernel-induced distance

$$\mathrm{d}(A, A')^2 = \mathrm{k}(A, A) + \mathrm{k}(A, A') - 2\,\mathrm{k}(A, A'). \qquad (93)$$

Even though different techniques may be formulated more naturally in terms of features, distances or kernels, it is always possible to translate – at least approximately – one description into another.

B. Measuring structural similarity

Most unsupervised learning algorithms rely on the definition of a metric to tell apart structures depending on their similarity. 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[248–252]. 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.[68]. 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. 68: a good metric needs to be invariant to rotations, translations, and permutations[253], and still be capable of telling distinct structures apart[106]. The comparison between the resolving power of different metrics has been often determined using distance-distance correlation maps[68,69,124,202], 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 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[124].

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

$$\mathrm{d}^2(A_i, A'_{i'}) \equiv \|\boldsymbol{\xi}(A_i) - \boldsymbol{\xi}(A'_{i'})\|^2 . \qquad (94)$$

Different ways of combining atom-centered representations to obtain a structure-level comparison are discussed and benchmarked in Ref. 69, 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,

$$\bar{\mathrm{d}}^2(A, A') = \frac{1}{N_A N_{A'}} \sum_{i \in A, i' \in A'} \mathrm{d}^2(A_i, A'_{i'}). \qquad (95)$$

Using the abstract notation $|A_i\rangle$ rather than $\boldsymbol{\xi}(A_i)$ to highlight the connection with the definition of the global representation $|A; \overline{\rho^{\otimes 2}}\rangle$ as the sum of environmental $|A; \rho_i\rangle$ (see Section IV C) it is easy to see that

$$\bar{\mathrm{d}}^2(A, A') = \frac{1}{N_A N_{A'}} \sum_{i \in A, i' \in A'} \| \, |A'_{i'}\rangle - |A_i\rangle \, \|^2 =$$
$$\left[\sum_{i \in A} \frac{|A_i\rangle}{N_A} - \sum_{i' \in A'} \frac{|A'_{i'}\rangle}{N_{A'}} \right]^2 \equiv \| \, |\overline{A'}\rangle - |\overline{A}\rangle \, \|^2, \qquad (96)$$

i.e. that the average environment distance $\bar{\mathrm{d}}^2(A, A')$ can be computed by taking the Euclidean distance between the mean of the environment's features in

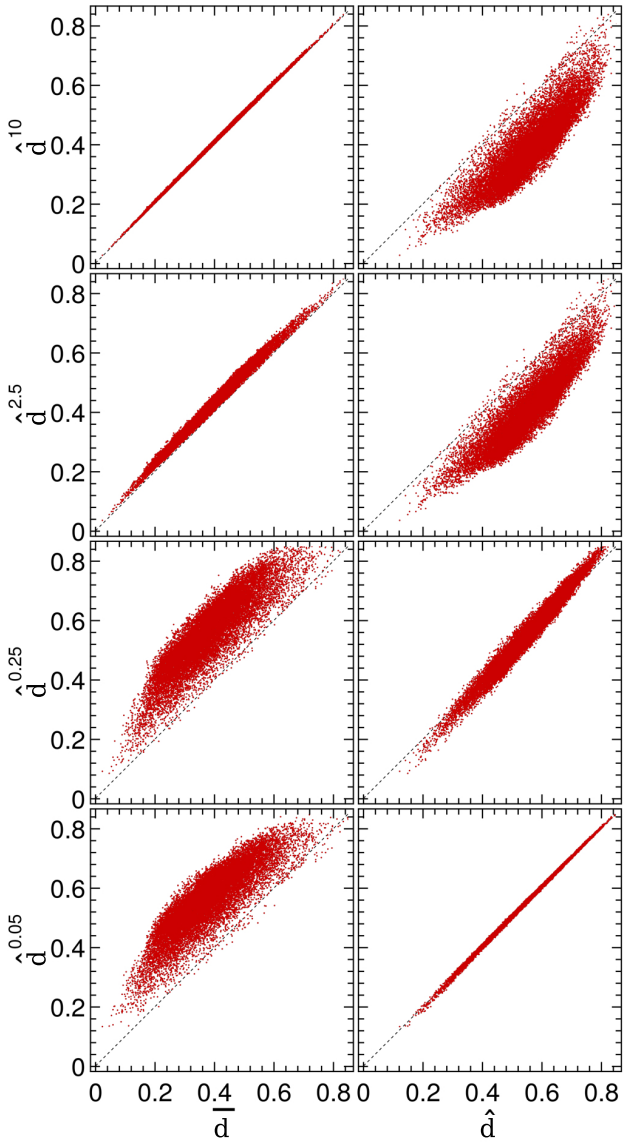


FIG. 16. Distance-distance correlation plots comparing the average environment distance (95) to the best-match (97) and REMatch (98) distances, with different values of the entropy regularization parameter γ. The reference structures are taken from the QM7b dataset of small organic molecules[254], and the environments are described by SOAP features $\langle a_1 n_1; a_2 n_2; l|\overline{\rho_i^{\otimes 2}}\rangle$. Reproduced with permission from Ref. 69. Copyright 2016 PCCP Owner Societies.

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

$$\hat{\mathrm{d}}^2(A,A') = \underset{\mathbf{P}\in\mathbb{U}^{N_A\times N_{A'}}}{\operatorname{argmin}} \sum_{i\in A, i'\in A'} \mathrm{d}^2(A_i, A'_{i'})P_{ii'} \qquad (97)$$

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 $\mathbf{P}$ 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[257]. Much like the case of the use of sorted interatomic distances as a structural representation (Section III B), 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 $O(N_A^2)$ and that can be applied to the comparison of structures of different sizes, by introducing an entropy regularization in Eq. (97)

$$\hat{\mathrm{d}}^\gamma(A,A')^2 = \underset{\mathbf{P}\in\mathbb{U}^{N_A\times N_{A'}}}{\operatorname{argmin}} \sum_{i\in A, i'\in A'} P_{ii'}(\mathrm{d}^2(A_i, A'_{i'}) + \gamma \ln P_{ii'}), \qquad (98)$$

controlled by the magnitude of the parameter γ. This approach was introduced in Ref. 258 for the general problem of solving optimal transport problems and of evaluating the Wasserstein distance between probability distributions, and was first applied in Ref. 69 to atomistic problems in terms of regularized entropy match (REMatch) kernels. By introducing a non-additive combination of the environments, REMatch kernels and the associated distances offer an increased resolving power compared to the plain average distance (95), as demonstrated in Figure 16. The figure also shows that Eq. (98) interpolates between the average and the best-match metrics, to which it tends respectively for $\gamma \to \infty$ and $\gamma \to 0$.

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 a bias that will be visible in the end result of the analysis.[245]. 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_{cut}, density smearing σ_a, 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 a dataset that consists in different polymorphs of a family of molecular materials[255]. In the first panel, in particular, one can recognize a de-

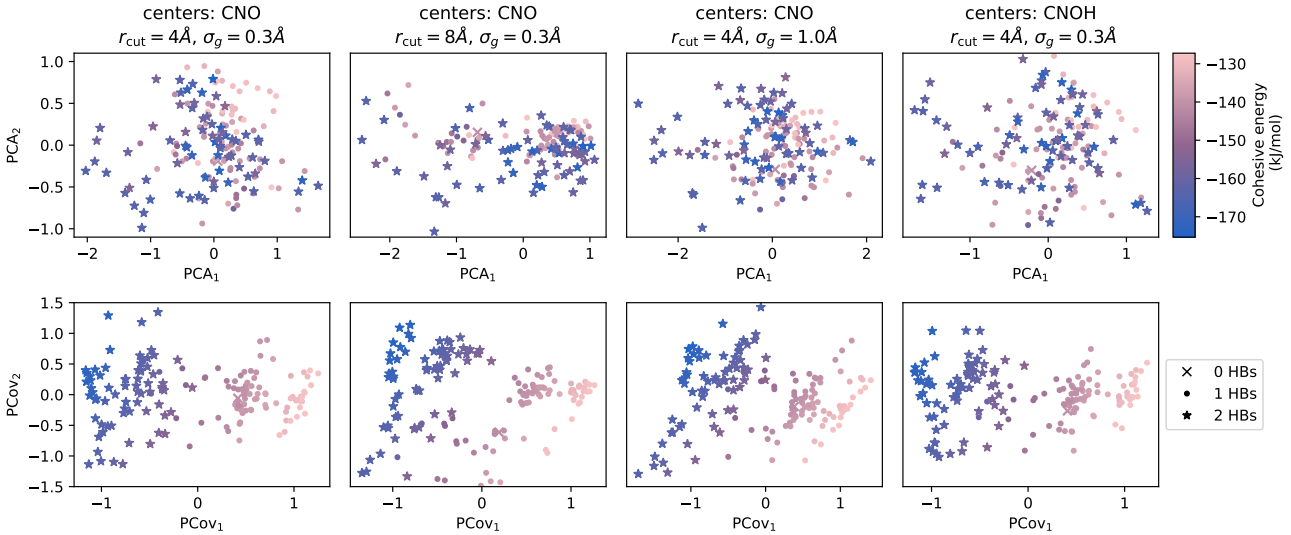


FIG. 17. Each map describes a set of 156 low-energy polymorphs of 21 different isomers of azaphenacene. The configurations are the same subset of the structures from Ref. 255 that was used in Ref. 256. Each point corresponds to a structure, color-coded based on its lattice energy, and with a symbol that indicates the number of hydrogen bonds per molecule, identified with a self-consistent definition[214]. The top row reports the first two principal components from a principal component analysis of SOAP $\overline{|\rho_i^{\otimes 2}\rangle}$ structures. The bottom row shows maps obtained using KPCovR[256]. Each column is computed using different SOAP hyperparameters, as indicated in the plot titles.

gree 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[259]. 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)[260]. In PCovR one minimizes a loss built as a mixture of a PCA and a linear regression loss, weighted by a mixing parameter α

$$\ell = \sum_i \alpha \left\| \boldsymbol{\Xi} - \boldsymbol{\Xi}\mathbf{P}_{\Xi T}\mathbf{P}_{T\Xi} \right\|^2 + (1-\alpha) \left\| \mathbf{Y} - \boldsymbol{\Xi}\mathbf{P}_{\Xi T}\mathbf{P}_{TY} \right\|^2 . \quad (99)$$

The matrix $\mathbf{P}_{\Xi T}$ projects from the feature space to a low-dimensional latent space, $\mathbf{P}_{T\Xi}$ reconstructs an approximation of the full-dimensional feature vector based on its latent-space embedding, and $\mathbf{P}_{TY}$ regresses the property matrix $\mathbf{Y}$ 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[256]) 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[256]. 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 by deliberate or accidental tuning of the hyperparameters of the representation.

D. Analyzing representations and datasets

The unsupervised 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 better understanding of the structure of feature space. For example, Fig. 18a

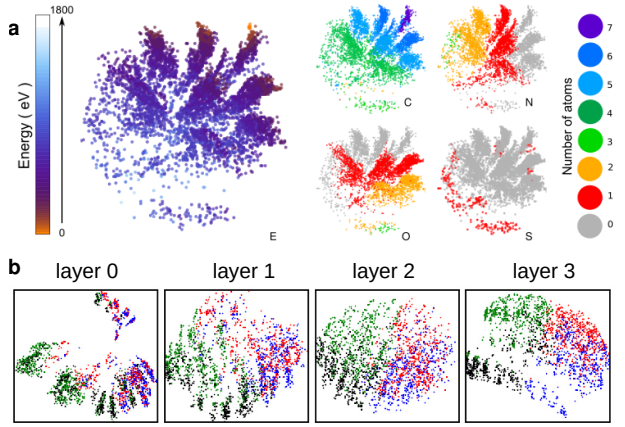


FIG. 18. (a) A sketch-map[4] representation of the QM7 molecular dataset[254] based on a SOAP kernel distance. Each point corresponds to one molecule. Left: points are colored according to the atomization energy; right: points are colored according to composition. Adapted with permission from Ref. 69. Copyright 2016 PCCP Owner Societies. (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. Reproduced from Ref. 254. Copyright 2013 American Chemical Society.

tells us less about the QM7 dataset[254] (that contains small organic molecules containing C, H, N, O, S, Cl) than about the SOAP features that underlie the representation: the unsupervised analysis shows that the chemical composition is the most clear-cut differentiating characteristic when looking at this dataset through SOAP lenses. Fig. 18b visualizes the same QM7 data using a different representation, based on the Coulomb matrix, and shows how successive layers of a neural network transform these features into nonlinear combinations that correlate 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 regression, and diagnose the inner workings of the deep neural 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 configurations it contains, but about its makeup, or the relationship 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

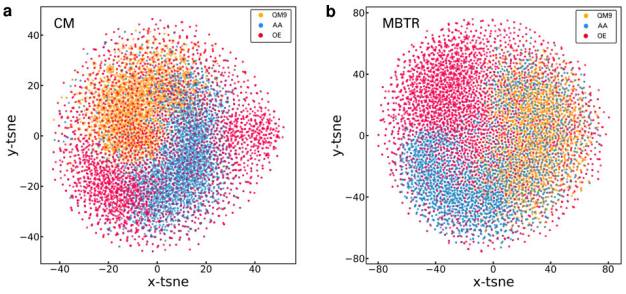


FIG. 19. 2D maps obtained applying the t-SNE dimensionality reduction algorithm[261] to three different molecular datasets – the systematic enumeration of 9-non-H-atoms molecules in QM9[262], the conformers of aminoacids in the Berlin aminoacid dataset AA[263] and the large molecules extracted from the Cambridge structural dataset of the OE dataset[264] . Panel (a) uses a Coulomb matrix representation, panel (b) uses the MBTR features[31]. Reproduced with permission from Ref. 115. Copyright 2019 American Institute of Physics.

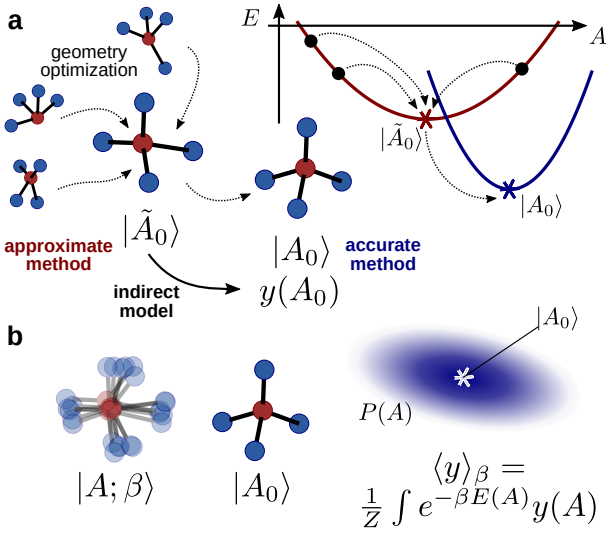


FIG. 20. A schematic overview of the process of using atomic structure representations to predict properties that are not directly associated with the starting structure. (a) prediction of the properties of the minimum-energy configuration of a structure; the problem can be made well-posed by using a cheap approximate method to optimize the structure, and taking the representation of this approximate structure as the input to regress accurate energy and geometry. (b) prediction of a property that is associated with a thermodynamic average; the minimum energy structure can be taken as a proxy for the ensemble, but a more formally precise "ensemble representation" is also possible.

sets. Other examples of this kind of analysis are discussed in Section IX.

E. Indirect structure-property relationships

The one-to-one mapping between an atomic structure and its representation is one of the key requirements to achieve accurate "surrogate quantum models" 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 configuration at hand. For example, consider the databases of molecular properties (e.g. the QM9 dataset[262]) 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 geometry has been optimized at the DFT level and use it to predict the DFT energy – an exercise that is manifestly of little practical utility. A more useful approach, instead, would be using a non-optimized structure to predict 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 structure $\tilde{A}_0$. Thus, the many-to-one mapping is realized by the local optimization procedure, and the corresponding representation $|\tilde{A}_0\rangle$ can be used to uniquely identify the entire basin of attraction of the local minimum. 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 relationship between $\tilde{A}_0$ and its high-quality counterpart is learned implicitly. This kind of "indirect" model has been used, for instance, in Ref. 30, where structures optimized at the semiempirical PM7[265] 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. The problem is very well understood in the context of cheminformatics, where molecular-graph descriptors can be thought as representing the entire set of molecular conformers. In the technique known as 4D-QSAR, "ensembles" of conformers are used to build fingerprints that encompass explicitly the structural variability of each compound[266]. These two approaches can also be applied while using the kind of representations discussed in the present review. 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 its $|\tilde{A}_0\rangle$ as a proxy of the thermodynamic state (Fig. 20b). The case in which the target property can be estimated as an ensemble average can be formulated very elegantly in the case of a linear model. Consider for instance the mean of a property y over the Boltzmann distribution at inverse temperature β, $P(A) = e^{-\beta E(A)}/Z$,

$$\langle y \rangle_\beta \equiv \frac{1}{Z} \int \mathrm{d}A \, e^{-\beta E(A)} y(A) \tag{100}$$

where $Z = \int \mathrm{d}A \, e^{-\beta E(A)}$ is the canonical partition function. Exploiting the linear nature of the representation one can define an "ensemble ket"

$$|A;\beta\rangle \equiv \frac{1}{Z} \int \mathrm{d}A \, e^{-\beta E(A)} \, |A\rangle \, . \tag{101}$$

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

$$\langle y \rangle_\beta \approx \sum_q \langle y|q\rangle \, \langle q|A;\beta\rangle \, , \tag{102}$$

which is convenient because it allows using properties of configurations and of ensembles on the same footings – and possibly combining them in a single training exercise. The same approach can also 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.

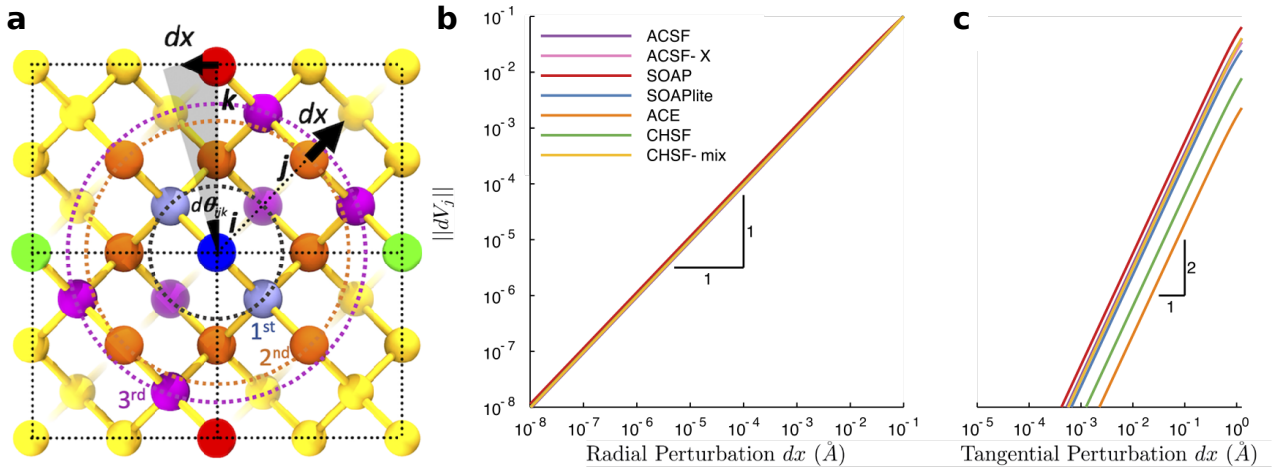


FIG. 21. (a) a $4 \times 4 \times 4$ Si cell, that is taken as the reference structure for radial and tangential perturbation of the neighbors of the central atom. (b,c) Norm of difference of atomic descriptors on atom i as a neighboring atom j is perturbed from its reference position. (b) A radial perturbation 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. Reproduced with permission from Ref. 201. Copyright 2020 American Institute of Physics.

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.[94,267,268]. 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[201], 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[201], or through an analysis of the Jacobian $J_{jk} = \partial \langle k|A_i\rangle / \partial \mathbf{r}_j$[202]. 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.[201] 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 201 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. Analizing the response of the features to perturbations in terms of the Jacobian, as in Ref. 202, 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. In some cases, as demonstrated by the finite-displacement deformation in Fig. 21b, 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, $\mathrm{d}(A_i, A'_{i'})$ 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 Eucledian distance between feature vectors, Eq. (94). Written as a function of the Cartesian coordinates of $A'_{i'}$, $\mathrm{d}(A_i, A'_{i'})$ is a scalar field which 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 $\mathrm{d}(A_i, A'_{i'}) = 0$, 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 $\mathrm{d}(A_i, A'_{i'}) < \epsilon$ where ϵ is a small number. Using a random search approach, the numerical sensitivity of the feature landscape has been analysed in Ref 29. Reference structures A_i were perturbed and then reconstructed by minimising the distance $\mathrm{d}(A_i, A'_{i'})$, and the optimised structures compared to the reference ones. For small numbers of neighbors in the reference environment,

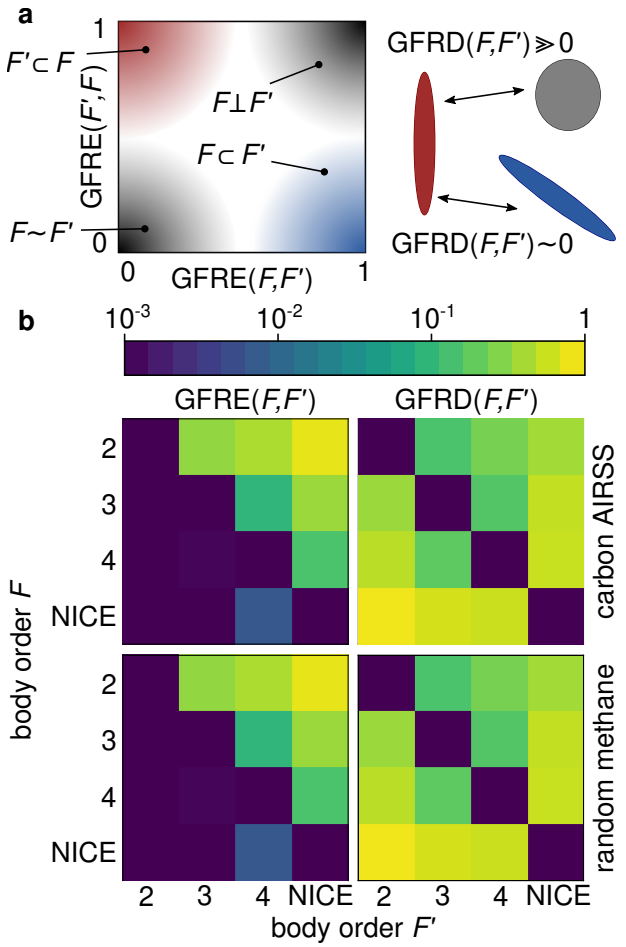


FIG. 22. (a) Schematic depiction of the interpretation of the error (GFRE, Eq. (103)) and the distortion (GFRD, Eq. (104)) that describe the relationship between two feature spaces. (b) Comparison between density correlation features of different order, as well as the NICE features[149] up to $\nu = 4$, computed in terms of the GFRE and GFRD, for a data set of random CH_4 configurations[269] and hypothetical carbon allotropes predicted by AIRSS[250,270]. Adapted from Ref. 271.

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

$$\mathrm{GFRE}(\boldsymbol{\Xi},\boldsymbol{\Xi}') = \min_{\mathbf{P}} \sqrt{\|\boldsymbol{\Xi}'_{\mathrm{test}} - \boldsymbol{\Xi}_{\mathrm{test}}\mathbf{P}\|^2/n_{\mathrm{test}}}. \tag{103}$$

$\mathbf{P}$ is a linear regression weight matrix obtained on a training subset of the rows of $\boldsymbol{\Xi}$ and $\boldsymbol{\Xi}'$, and both sets of features are assumed to be standardised.[271] The GFRE can be extended to also incorporate nonlinearity in the mapping, either by a locally-linear approach, or by using a kernelized version. Loosely speaking, it measures the relative amount of information encoded by the two feature spaces, and is not symmetric. $\mathrm{GFRE}(\boldsymbol{\Xi},\boldsymbol{\Xi}') \ll \mathrm{GFRE}(\boldsymbol{\Xi}',\boldsymbol{\Xi})$ indicates that the featurization underlying $\boldsymbol{\Xi}$ is more informative than that used to build $\boldsymbol{\Xi}'$, and vice versa. $\mathrm{GFRE}(\boldsymbol{\Xi},\boldsymbol{\Xi}') \approx \mathrm{GFRE}(\boldsymbol{\Xi}',\boldsymbol{\Xi}) \approx 0$ implies that the two featurizations contain similar information (Fig. 22a). A similar asymmetric measure of similarity between feature spaces can be defined by comparing the resolving power of the corresponding metrics[272], translating the information that is present in distance-distance correlation plots (Sec. VII B) into a quantitative measure of information content.

Having $\mathrm{GFRE}(\boldsymbol{\Xi},\boldsymbol{\Xi}') \approx \mathrm{GFRE}(\boldsymbol{\Xi}',\boldsymbol{\Xi}) \approx 0$ does not mean that $\boldsymbol{\Xi}$ and $\boldsymbol{\Xi}'$ they are equivalent and can be used interchangeably. 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 $\boldsymbol{\Xi}$ or $\boldsymbol{\Xi}'$, can be measured by defining a global feature space distortion (GFRD)

$$\mathrm{GFRD}(\boldsymbol{\Xi},\boldsymbol{\Xi}') = \min_{\mathbf{Q}\in\mathbb{U}} \sqrt{\|\boldsymbol{\Xi}_{\mathrm{test}}\mathbf{P} - \boldsymbol{\Xi}_{\mathrm{test}}\mathbf{Q}\|^2/n_{\mathrm{test}}}. \tag{104}$$

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

If both GFRE and GFRD are zero, then the linearly independent components of $\boldsymbol{\Xi}$ and $\boldsymbol{\Xi}'$ are related by a unitary transformation, which implies that distances and scalar products between feature vectors are equal in $\boldsymbol{\Xi}$ and in $\boldsymbol{\Xi}'$. Figure 22 demonstrates the use of these measures to compare $|\overline{\rho_i^{\otimes\nu}}\rangle$ 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-ν 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

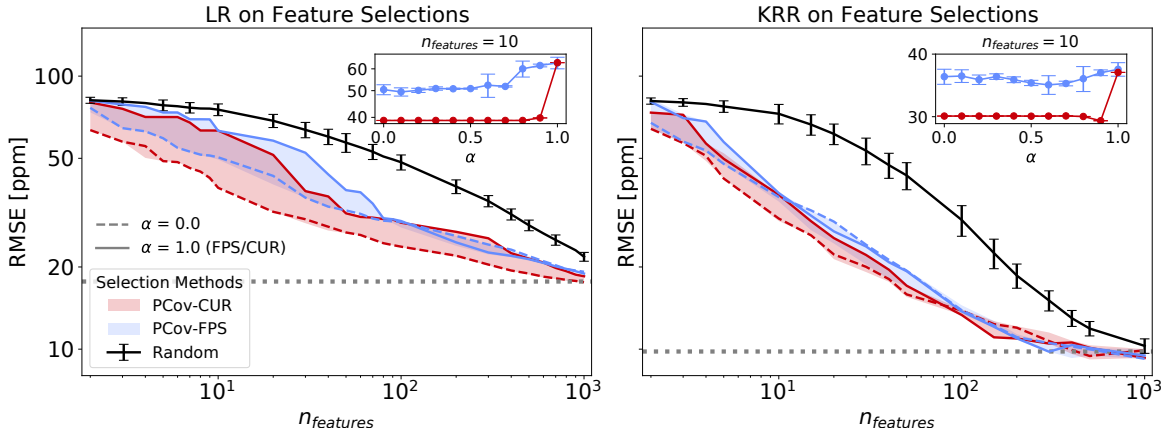


FIG. 23. Test-set error in the prediction of a linear regression (left) and kernel ridge regression (right) model of the nuclear chemical shieldings of atoms in a set of molecular materials, as a function of the number of features used in the model. Features are selected using the FPS and CUR methods (full lines) from a set of 2520 SOAP features. The shaded areas indicate the range of values obtained varying the mixing parameter α in a principal covariate-augmented version of the methods. Adapted from Ref. 273. Copyright 2021 IOP Publishing under Creative Commons Attribution 4.0 International License `https://creativecommons.org/licenses/by/4.0/`.

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 the full $\nu = 3$ features. The GFRD is rather large between all featurizations, indicating 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 interatomic potentials.[15,180,183,195]

B. Feature selection

Numerical feature vectors $\boldsymbol{\xi}(A_i)$ are the result of a basis set expansion of the abstract atom-centered representations, which are, for practical purposes, truncated. A concrete discretization of the symmetrized ν-correlations is obtained by choosing a finite subset from the set of all possible features,

$$\mathbf{q} \subset \mathbf{q}_{\text{total}} := \big\{(n_\alpha l_\alpha)_{\alpha=1}^{\nu}, \quad \nu \in \mathbb{N}\big\}. \tag{105}$$

(A choice of $(n_\alpha, l_\alpha)_\alpha$ naturally induces a choice of m_α 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 to be *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 ν to achieve a desired accuracy, e.g. on the order $O(10'000)$ features for single-species models. The additional cost in training and evaluating the features is of course offset 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 ν (curse of dimensionality). For example, a total-degree D discretisation,

$$\mathbf{q}(\nu^{\max}, D) = \big\{(n_\alpha l_\alpha m_\alpha)_{\alpha=1}^{\nu} \;:\; \nu \le \nu^{\max} \quad \sum_\alpha n_\alpha + l_\alpha \le D \big\} \tag{106}$$

was used by Bachmayr et al.[127], while closely related a priori sparsifications were used by Braams and Bowman[11], Shapeev[134] 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[127,149]. 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.

A simple example of a feature selection strategy is the farthest point sampling (FPS) technique[274]. One chooses an initial column $\boldsymbol{\chi}_{c_0}$ (indexed by c_0) of the feature matrix, and then iterates selecting the columns that maximize the Haussdorf distance to the previously selected columns

$$c_{m+1} = \text{argmax}_j \left\{ \min_{i \in \mathbf{c}_m} \left\| \boldsymbol{\chi}_i - \boldsymbol{\chi}_j \right\| \right\}, \tag{107}$$

effectively identifying the indices $\mathbf{c}$ of the features that have the most diverse values across the data set. FPS has also been used in a similar manner, but on the rows of $\boldsymbol{\Xi}$ in order to select a representative set of data points.[30,69,275] The CUR matrix decomposition[276], instead, generates a low-rank approximation of the feature matrix $\boldsymbol{\Xi}$, in the form

$$\boldsymbol{\Xi} \approx \mathbf{CUR}. \tag{108}$$

Unlike singular value decomposition, CUR uses the actual columns ($\mathbf{C}$) and rows ($\mathbf{R}$) of $\boldsymbol{\Xi}$. To make the selection, a *leverage score* is associated with each feature c

$$\pi_c = \frac{1}{k} \sum_{i=1}^{k} (\mathbf{v}_i)_c^2, \tag{109}$$

based on the right singular vectors $\mathbf{v}_i$ of the singular value decomposition of $\boldsymbol{\Xi}$. k is usually taken to be the approximate rank of $\boldsymbol{\Xi}$. Features may be selected in a probabilistic procedure or simply based on their score. Imbalzano *et al.*[277] 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. 277 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 $\boldsymbol{\Xi}$ 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 π_c, Eq. (109).

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 with the structures,[273] as in Eq. (99). 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.*[277] used a reference data set on liquid water[278] 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[24,279,280]. A more systematic investigation of the effectiveness of feature selection for many commonly used atomic descriptors has been recently reported by Onat et al.[201], 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 D, 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[29,117,195], 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 heuristics[141]. An optimization of the hyperparameters by gradient descent has also been proposed[281] to

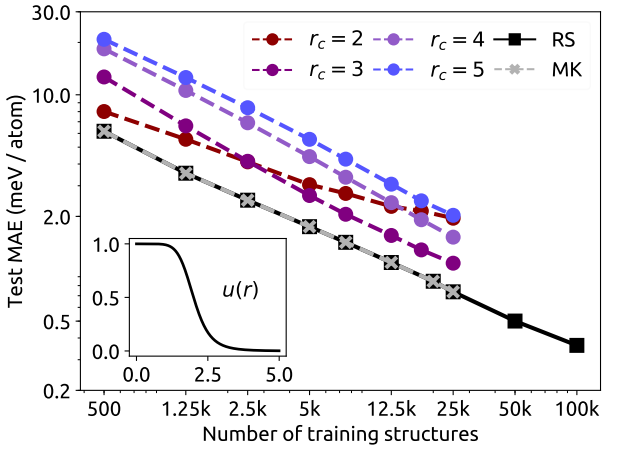


FIG. 24. Learning curves for the atomization energy of molecules in the QM9 data set[262]. Four of the lines show the MAE on the test set for kernel regression models based on SOAP ($\langle\overline{|\rho_i^{\otimes 2}\rangle}$) features with different cutoff radii (dashed lines graduating from red to blue). The other lines show the MAE on the test set for the optimal radially-scaled (RS) and multiple-kernel (MK) SOAP models (black and grey lines respectively). In every model, the features were constructed with very converged hyperparameters, $n_{\max} = 12$ and $l_{\max} = 9$. The inset shows the radial-scaling function $u(r)$ from $r = 0$Å to $r = 5$Å with the parameters that were found to minimize the tenfold cross validation MAE on the optimization set through a grid search, $r_0 = 2$Å and $m = 7$. The multiple-kernel model combines the $r_{\rm cut} = 2, 3, 4$ and RS kernels in the ratio 100'000 : 1 : 2 : 10'000, and the learning curve agrees with the RS result to within graphical accuracy. Error bars are omitted because they are as small as the data point markers. Note that errors are expressed on a per-atom basis. Error per molecule expressed in kcal/mol can be obtained approximately by multiplying the scale by 0.4147, that is computed based on the average size of a molecule in the QM9 database. Reproduced with permission from Ref. 125. Copyright 2018 PCCP Owner Societies.

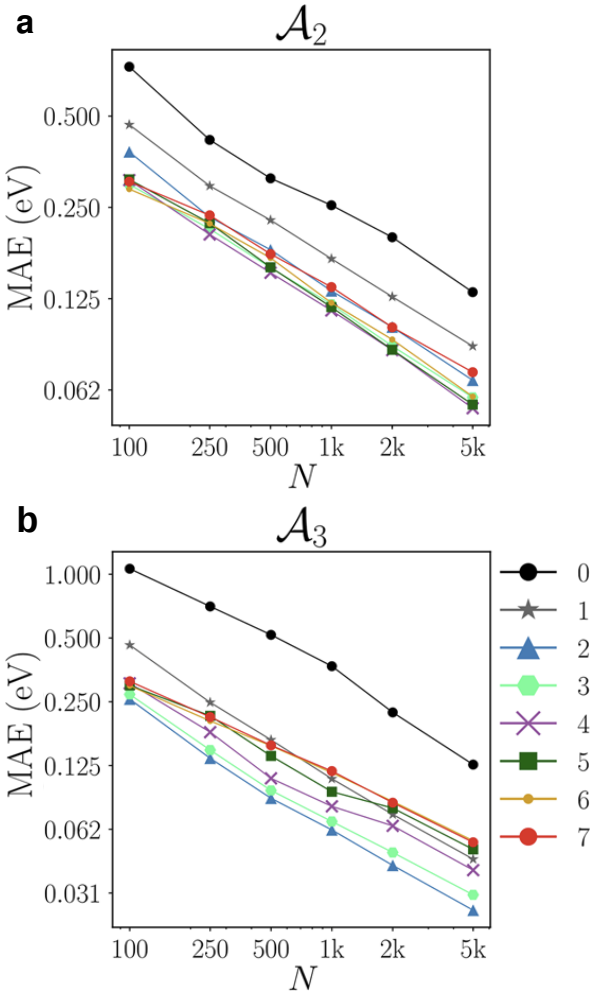


FIG. 25. Optimization of the exponents in scaling power laws. a) Out-of-sample MAE for atomization/formation energy predictions as a function of training set size on the QM9 dataset. Learning curves are generated using KRR with a 2-body FCHL representation. The legends indicate the exponent n_2 used in the scaling power law, $\xi_2(d)$. Leftb) Out-of-sample MAE for atomization/formation energy predictions as a function of training set size on the QM9 dataset. Learning curves are generated using KRR with a 3-body FCHL representation. The legends indicate the exponent n_3 used in the scaling power law, $\xi_3(d)$. In order to compare results to Fig. 24, the ordinates must be divided by 18. Adapted from Ref. 116.

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 VIII B. 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, affects 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 predictions[123]. 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 $r_{\rm cut}$ relates to the typical decay length of interactions, as mentioned in Section III 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 $r_{\rm cut}$. A very large cutoff $r_{\rm cut} = 5$Å does not yield the best performance, despite providing information on a wider range of distances. In fact, one observes the need to balance the

complexity of the model and the available data: a very short-range $r_{\rm cut} = 2$Å yields the most effective description in the data-poor regime, but the accuracy of the corresponding model saturates due to lack of information on non-covalent interactions. Combining multiple representations in a "multi-kernel" model (which is effectively equivalent to concatenating multiple feature vectors, each scaled separately) yields consistently better performances[22,30]. The weighting of different components – that can be optimized by cross-validation – indicates the relative importance of correlations on various length-scales. The fact that large-$r_{\rm cut}$ features carry low weight in the optimal combination suggests that an improvement of performance can be obtained by calibrating the distance-dependent contributions of neighbors to the environment description. This can be achieved by introducing a radial scaling function (indicated as $u(r)$ in Ref. 125, as $f(r, r_j, r_{\rm cut})$ in Ref.[191] and as $\xi_\nu(d)$ in Ref. 116) that downweights the contributions of atoms in the far field. As shown in Fig. 25 for the FCHL representation[116], the choice of the form of this scaling can change the accuracy of the model by more than a factor of 2. A similar effect is seen in Fig. 24 for the case of SOAP features. It is also worth noting that the cutoff function usually adopted in Behler-Parrinello-like frameworks decays rapidly well before reaching $r_{\rm cut}$ – suggesting that a similar optimization is implicitly at play.[152] 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. 69, that is also a core component of the FCHL framework[116], but has been implemented in different forms in the context of atom-centered symmetry functions[283–285] and of generic atom-density correlation features[125]. In general terms, the idea is to achieve a reduction of the dimensionality of the chemical space, writing formally a (linear) projection of the elemental features

$$\langle \tilde{a}| = \sum_a \langle \tilde{a}|a\rangle \, \langle a| \,, \qquad (110)$$

where the coefficients $\langle \tilde{a}|a\rangle$ enact the projection between the elemental and the "alchemical" basis. The reduction in the dimensionality of the feature space can be substantial: for powerspectrum ($\nu = 2$) features, the number of components scales quadratically with the number of species, and so even just halving the dimension of the chemical space reduces the

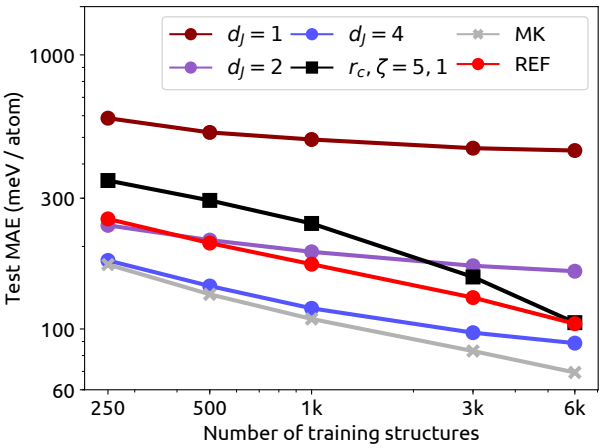


FIG. 26. Learning curves for a model of the cohesive energy of a database of elpasolite structures, each containing a random selection of four elements chosen among 39 main group elements.[282] The standard SOAP curve is shown in black, the best curve from Ref. 282 is shown in bright red (REF) and the curves obtained with an alchemical model with reduced dimensionality d_J are shown in dark red ($d_J = 1$), purple ($d_J = 2$) and blue ($d_J = 4$). The multiple-kernel model (shown in grey) combines three standard SOAP kernels with different cutoff and one alchemically optimized kernel with $d_J = 4$. Reproduced with permission from Ref. 125. Copyright 2018 PCCP Owner Societies.

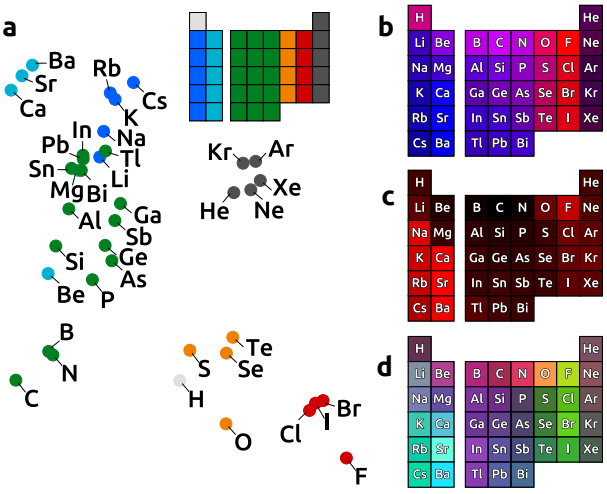


FIG. 27. Data-driven representations of the chemical space. (a) A 2D map of the elements contained in the elpasolite data set,[282] with the coordinates corresponding to $\langle 1|a\rangle$ and $\langle 2|a\rangle$ for the case with $d_J = 2$ (see also Fig. 26). Points are colored according to the group. (b) A periodic table colored according to the coordinates in the 2D chemical space. $\langle 1|a\rangle$ corresponds to the red channel and $\langle 2|a\rangle$ to the blue channel. (c) A periodic table colored according to $\langle 1|a\rangle$ (red channel) for a 1D chemical space. (d) A periodic table colored according to 4D chemical coordinates ($\langle 1|a\rangle$: red channel, $\langle 2|a\rangle$: green channel, $\langle 3|a\rangle$: blue channel, $\langle 4|a\rangle$: hatches opacity) Reproduced with permission from Ref. 125. Copyright 2018 PCCP Owner Societies.

number of powerspectrum features by 75%:

$$\langle \tilde{a}_1 n_1; \tilde{a}_2 n_2; l|\overline{\rho_i^{\otimes 2}}\rangle = \sum_{a_1 a_2} \langle \tilde{a}_1|a_1\rangle \, \langle \tilde{a}_2|a_2\rangle \, \langle a_1 n_1; a_2 n_2; l|\overline{\rho_i^{\otimes 2}}\rangle \,. \quad (111)$$

Figure 26 demonstrates how reducing the dimensionality of chemical space helps achieving a transferable, accurate model with a small number of training structures. By comparing learning curves with different degrees of compression, one sees that there is a similar data/complexity interplay as observed for radial correlations. A low-dimensional alchemical space is beneficial in the data-poor regime, as it allows the model to make an educated guess about the interactions of pairs of elements that are not represented in the training set. Learning curves with low chemical complexity, however, saturate in the limit of large training set, because they generate features that are not sufficiently flexible, and cannot describe the differences between elements.

The optimization of both geometrical and compositional components of the density-correlation features can be construed as a linear transformation of the kets (or, when seen in terms of the linear kernels built on such features, as the action of a Hermitian operator[109]). The requirement that such transformations do not affect the symmetry properties of the features restricts form they can take – for instance, they cannot mix different l or m dependent channels. These observations imply that (1) *linear* feature optimizations do not change the nature of the representations, and can be applied equally well to any implementation of $|\overline{\rho_i^{\otimes \nu}}\rangle$ features, (2) as long as the linear transformation is full rank, there is no loss of information, which means that the observed change in performance is linked to the details of the regression scheme, such as regularization in linear or kernel models.

As a final remark, let us mention that a critical analysis of a feature-optimization effort often reveals insights into the physical-chemical properties of the system being studied and the target properties. For instance comparing models of the energy using different $r_{\rm cut}$ can be used to infer relationships between the length and energy scales[30], and the inspection of the chemical mapping coefficients in Eq. (110) can be used to construct a data-driven periodic table of the elements (see Fig. 27). The use of interpretable, physics-inspired features can also be used to provide intuitive chemical insights by the construction of *knock-out models*[286] in which for instance correlations are restricted to 2-bodies, the cutoff reduced to first or second neighbors. The impact of these artificial restrictions on the features information content, and therefore on the asymptotic performance of the model, indicates how important 3 or higher-body order interactions are, or how much long-range effects are relevant to determine the value of the target property.[230]

D. Efficient implementation

Despite encoding similar information content, the differences in formulation of competing structural representations may lead to large variations in implementation and performance. A first fundamental divide is between evaluation of features by summing over clusters of ν neighbors and computing ν tensor products of atomic densities (see Section IV F). Consider the case of evaluating "SOAP-like" ACSF by cluster sum (cost ${n_{\rm max}}^2 l_{\rm max} n_{\rm neigh}^2$) and by density expansion (cost $n_{\rm neigh} n_{\rm max} {l_{\rm max}}^2$ for the density, and ${n_{\rm max}}^2 {l_{\rm max}}^2$ for the SOAP evaluation). Despite the adverse scaling of ACSF computed as a sum over clusters of neighbors, these representations can be implemented efficiently[94,268,288] by relying on a careful selection of the features (discussed in Section VIII B), reuse of parts of the computations, parallelism and GPU acceleration.[289–291] In fact, when computing a linear model which is explicitly equivalent to a $(\nu+1)$-body order potential, the low-order terms can be more efficiently evaluated as a sum over neighbors[195,292]

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 usually built on a *smooth* atom density,[109,125,149] that generalize the construction of the SOAP powerspectrum and bispectrum,[29] and those that are usually computed in a way that corresponds to a δ-like density, such as ACE[126,150] and MTP.[134,293] Indeed, both families of representations rely on three steps: (i) expansion of the local atom density on a suitable basis, e.g. Eq. (24), (ii) computation of ν tensor products of the expansion, and then (iii) contraction over the correlations to obtain equivariant features (Fig. 28). While these three steps have been implemented in different ways, their efficient implementation relies on similar considerations.

Atomic density expansion Equation (20) provides the blueprints for a broad class of $(\nu+1)$-body atom-density representation. Practical implementations differ by the type of localized function used to construct the local atom density (see Eq. (16)), 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*[134,243]

$$\mathbf{M}_n^{\otimes \nu}(\rho_i) = \sum_{j \in A_i} P_{n,\nu}(r_{ji}) \underbrace{\hat{\mathbf{r}}_{ji} \otimes \hat{\mathbf{r}}_{ji} \ldots \otimes \hat{\mathbf{r}}_{ji}}_{\nu \text{ times}} \quad (112)$$

where $P_{\mu,\nu}$ 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. 127. The performance

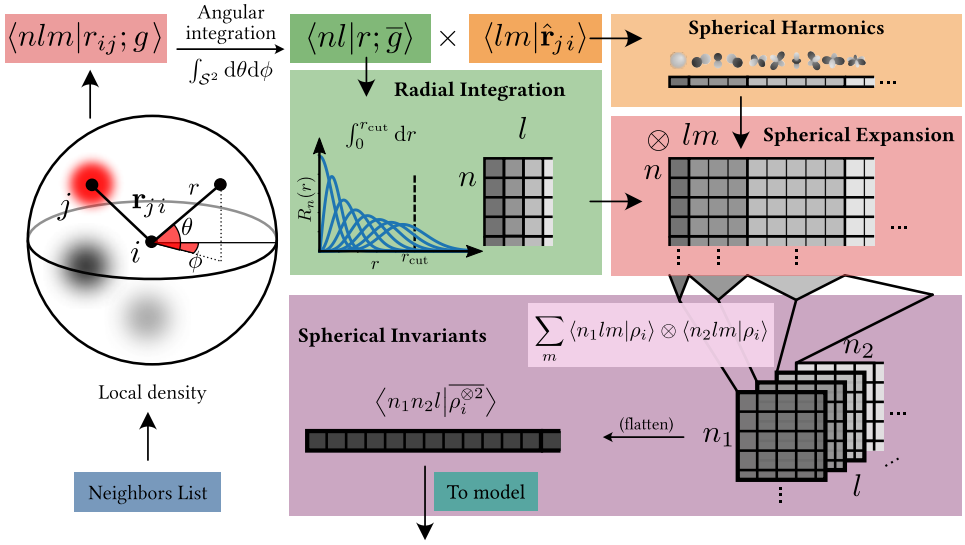


FIG. 28. Schematic overview of the process of expanding the density in a radial and angular basis set, and recombining those to form spherical invariants (or covariants). Reproduced with permission from Ref. 287. Copyright 2021 American Institute of Physics.

of the MTP representation,[268] which relies on an efficient recursive evaluation of the basis functions[134], are a testament to the effectiveness of this basis choice.

As shown in Section IV E, the choice of an angular basis of spherical harmonics simplifies greatly the evaluation of Eq. (20), that can be written in terms of contractions of density coefficients $\langle nlm|\rho_i\rangle$ (cf. Eq. (24)). If the environment-centered density is written in terms of a sum of density functions $g(\mathbf{x}-\mathbf{r}_{ji}) \equiv \langle \mathbf{x}|\mathbf{r}_{ji};g\rangle$, peaked at the neighbors positions, the expansion coefficients can be written as the accumulation

$$\langle nlm|\rho_i\rangle = \sum_{j\in i} f_{\mathrm{cut}}(r_{ji})\,\langle nlm|\mathbf{r}_{ji};g\rangle \tag{113}$$

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

$$\langle nlm|\mathbf{r}_{ji};g\rangle = \int \mathrm{d}\mathbf{x}\ \langle nl|x\rangle\ \langle lm|\hat{\mathbf{x}}\rangle\ \langle\mathbf{x}|\mathbf{r}_{ji};g\rangle\,. \tag{114}$$

In the $g\to\delta$ limit, the contribution from the j-th neighbor amounts simply to a product of the radial and angular functions evaluated at $\mathbf{r}_{ji}$,

$$\langle nlm|\mathbf{r}_{ji};\delta\rangle = \langle nl|r_{ji}\rangle\ \langle lm|\hat{\mathbf{r}}_{ji}\rangle\,. \tag{115}$$

Similar to the 1D case discussed in Sec. V B, *for a given choice of radial basis* the smearing of the density can be achieved by a mollification of the basis:

$$\begin{aligned}
&\langle nlm|\mathbf{r}_{ji};g\rangle = \\
&\int \mathrm{d}\mathbf{x}\ \langle nl|x\rangle\ \langle lm|\hat{\mathbf{x}}\rangle \int \mathrm{d}\mathbf{x}'\, g(\mathbf{x}-\mathbf{x}')\,\langle\mathbf{x}'|\mathbf{r}_{ji};\delta\rangle \\
&= \int \mathrm{d}\mathbf{x}'\ \langle\mathbf{x}'|\mathbf{r}_{ji};\delta\rangle \int \mathrm{d}\mathbf{x}\ \langle nl|x\rangle\ \langle lm|\hat{\mathbf{x}}\rangle\, g(\mathbf{x}-\mathbf{x}') \\
&= \int \mathrm{d}\mathbf{x}'\ \langle\mathbf{x}'|\mathbf{r}_{ji};\delta\rangle\ \langle lm|\hat{\mathbf{x}}'\rangle\ \langle nl;g|x'\rangle \\
&\equiv \langle nlm;g|\mathbf{r}_{ji};\delta\rangle\,,
\end{aligned} \tag{116}$$

where we use $\langle nl;g|$ to indicate the radial term that results from the Gaussian convolution. 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.[126,127,294]

It might appear that using a smooth atom density g complicates substantially the evaluation of Eq. (114). However when g is a spherical Gaussian with standard deviation σ, the integral over $\mathrm{d}\hat{\mathbf{x}}$ can be computed analytically[295]

$$\int \mathrm{d}\hat{\mathbf{x}}\ \langle lm|\hat{\mathbf{x}}\rangle\ \langle x\hat{\mathbf{x}}|\mathbf{r}_{ji};g\rangle = \langle x;l;|r_{ji};g\rangle\ \langle lm|\hat{\mathbf{r}}_{ji}\rangle\,. \tag{117}$$

where the radial integral reads

$$\langle x;l;|r;g\rangle = 4\pi e^{-r^2/2\sigma^2}x^2e^{-x^2/2\sigma^2}\mathrm{i}_l\big(xr/\sigma^2\big)\,, \tag{118}$$

so one gets

$$\langle nlm|\mathbf{r}_{ji};g\rangle = \langle lm|\hat{\mathbf{r}}_{ji}\rangle \int \mathrm{d}x\ \langle nl|x\rangle\ \langle l;x|r_{ji};g\rangle\,. \tag{119}$$

The radial part of the integral

$$\int \mathrm{d}x\ \langle nl|x\rangle\ \langle l;x|r_{ji};g\rangle = \langle nl|r_{ji};g\rangle \tag{120}$$

can be computed numerically for any form of the radial basis resulting in $n_{\max}l_{\max}n_{grid}$ evaluations of special functions. For instance, the original implementation of the SOAP representation uses a numerically orthogonalized, equispaced Gaussian basis.[29] Alternatively, this integral might also be performed analytically by using Gaussian type orbitals (GTO) as the radial basis[159,296], $\langle x|nl;\mathrm{GTO}\rangle$. This choice makes it possible to compute the coefficients of the smeared density as easily as for the $g \to \delta$ case

$$\langle nlm;\mathrm{GTO}|\mathbf{r}_{ji};g\rangle = \langle lm|\hat{\mathbf{r}}_{ji}\rangle\ \langle nl;\mathrm{GTO}|r_{ji};g\rangle\,, \tag{121}$$

where the only overhead comes from having to compute $\mathcal{O}(n_{\max}l_{\max})$ terms for the radial part and its orthonormalization. This is asymptotically cheaper than combining radial and angular terms – which requires $\mathcal{O}\left(n_{\max}{l_{\max}}^2\right)$ multiplications per neighbor – but can be substantial in practical cases, because the analytical integrals in Eqs. (117) and (120) yield non-standard special functions.

To reduce this overhead, one can choose a form of the atomic density that is symmetric about $\mathbf{r}_i$ instead of $\mathbf{r}_{ji}$[191]

$$\langle \mathbf{x}|\mathbf{r}_{ji};\hat{g}\rangle = \exp\left[-\frac{(x-r_{ji})^2}{2\sigma_r^2} - \frac{r_{ji}^2}{\sigma_\perp^2}(1-\hat{\mathbf{r}}_{ji}\cdot\hat{\mathbf{x}})\right]. \tag{122}$$

Together with a choice of radial functions that do not depend explicitly on l, this allows factorizing the radial integral (120) as

$$\int \mathrm{d}x\ \langle n|x\rangle\ \langle x;l|r_{ji};\hat{g}\rangle = \langle n|r_{ji};\hat{g}\rangle\ \langle l|r_{ji};\hat{g}\rangle\,. \tag{123}$$

Coupled with the polynomial basis proposed in Ref. 29, 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 $\langle nl|r;g\rangle$ 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.[150,294] Several well-established techniques can be used to speed up the calculation of Y_l^m, 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 $\hat{\mathbf{r}}_{ji}$.

Symmetrized n-body correlations The density coefficients $\langle anlm|\rho_i\rangle$ 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 ν sets of density coefficients to yield density

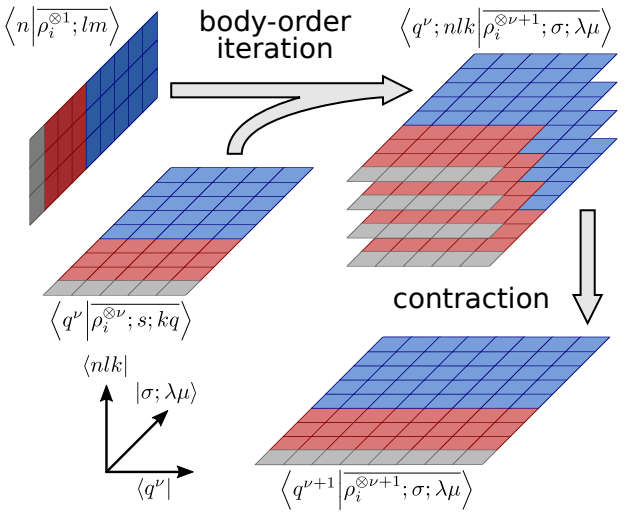


FIG. 29. A schematic representation of the NICE framework. A hierarchy of N-body equivariant features is built by iterative combination with the atom density coefficients, and the exponential increase in feature space size is kept at bay by successive contractions. Reproduced with permission from Ref. 149. Copyright 2020 American Institute of Physics.

correlations in the uncoupled basis $\langle(a_i n_i l_i m_i)_{i=1}^{\nu}|$, and then a contraction along the m_i indices, that generates the equivariant features expressed in the coupled basis $\langle(a_i n_i l_i k_i)_{i=1}^{\nu}|$. 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 $-l \le m \le l$. Depending on the hardware architecture, it might be beneficial to store the coefficients in a regular $(l_{\max}+1)\times(2l_{\max}+1)$ 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.[126]) increase exponentially with ν. Even though the exponential scaling is related to the expansion parameters $l_{\max}$ and $n_{\max}$, 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 ν atoms (see Sec. IV F and V A), it makes the enumeration of a complete linear basis prohibitively expensive. The recurrence relations[149] of Eq. (46) (or the equivalent ones for the invariant features proposed in Ref.[127]) make it possible to evaluate individual equivariant features with a cost that scales only linearly with ν. To beat completely the exponential scaling, these recursive expressions should be combined with feature selection schemes such as those discussed in Section VIII B. For example, the n-body iterative contraction of equivariant (NICE) features incorporates a selection/contraction step at each level of the itera-

tion. For each equivariant component $\langle q'|\overline{\rho_i^{\otimes\nu};\sigma;\lambda\mu}\rangle$, 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

$$\langle q^{\nu;\sigma\lambda}|\overline{\rho_i^{\otimes\nu};\sigma;\lambda\mu}\rangle = \sum_{q'} U^{\nu;\sigma\lambda}_{qq'} \langle q'|\overline{\rho_i^{\otimes\nu};\sigma;\lambda\mu}\rangle\,. \quad (124)$$

Given that this operation only mixes features with the same equivariant behavior, it is then possible to perform an iteration equivalent to Eq. (46) to increase the body order further

$$\begin{aligned}\langle q|\overline{\rho_i^{\otimes(\nu+1)};\sigma;\lambda\mu}\rangle \equiv \langle q^{\nu;\tau k};nlk|\overline{\rho_i^{\otimes(\nu+1)};\sigma;\lambda\mu}\rangle = \\ \delta_{\sigma(\tau(-1)^{l+k+\lambda})}\sum_m \langle lm;k(\mu-m)|\lambda\mu\rangle \\ \times\ \langle n|\overline{\rho_i^{\otimes 1};lm}\rangle\ \langle q^{\nu;\tau\lambda}|\overline{\rho_i^{\otimes\nu};\tau;k(\mu-m)}\rangle\,. \quad (125)\end{aligned}$$

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 $y(A_i)$ in terms of the *unsymmetrized* correlations,

$$\begin{aligned}y(A_i) \approx \sum_{n_1l_1m_1\ldots n_\nu l_\nu m_\nu} \langle y|n_1l_1m_1\ldots n_\nu l_\nu m_\nu\rangle \\ \times\ \langle n_1l_1m_1\ldots n_\nu l_\nu m_\nu|\rho_i^{\otimes\nu}\rangle\,,\end{aligned}$$

with the desired symmetries imposed through constraints on the coefficients $\langle y|n_1l_1m_1\cdots n_\nu l_\nu m_\nu\rangle$. 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

$$\langle n_1l_1m_1;\ldots n_\nu l_\nu m_\nu|\rho_i^{\otimes\nu}\rangle = \prod_{\alpha=1}^{\nu}\langle n_\alpha l_\alpha m_\alpha|\rho_i\rangle\,,$$

which is relatively straightforward to construct,[127] the key challenge being to retain only the $(n_\alpha, l_\alpha, m_\alpha)_\alpha$ features that give rise to non-zero coefficients.

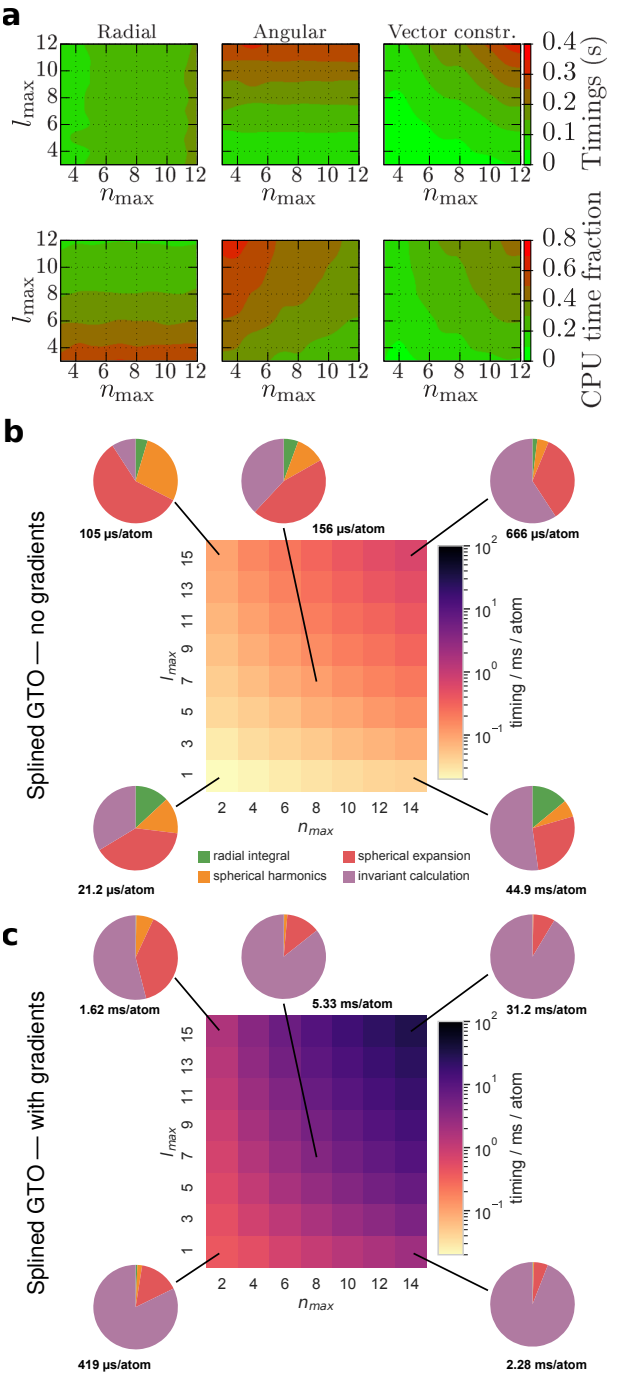


FIG. 30. (a) Single-core timings for the evaluation of radial expansion, angular expansion, and SOAP vector construction for an atomic structure containing 10'000 randomly-placed atoms, using the implementation discussed in Ref. 191 and as a function of $(n_{\max}, l_{\max})$. Reproduced with permission from Ref. 191. Copyright 2019 American Physical Society. (b) Single-core timings for the evaluation of SOAP features for a dataset of molecular crystals[172], using the implementation in librascal[297], as a function of $(n_{\max}, l_{\max})$; to compare with panel (a), consider that the presence of 4 distinct chemical elements corresponds roughly to a fourfold increase of $n_{\max}$. The breakdown of the total timing in the different steps of the calculation is shown for a few representative sizes of the expansion. (c) As in (b), including also the calculations of the gradients of the features with respect to atomic positions. Reproduced with permission from Ref. 287. Copyright 2021 American Institute of Physics.

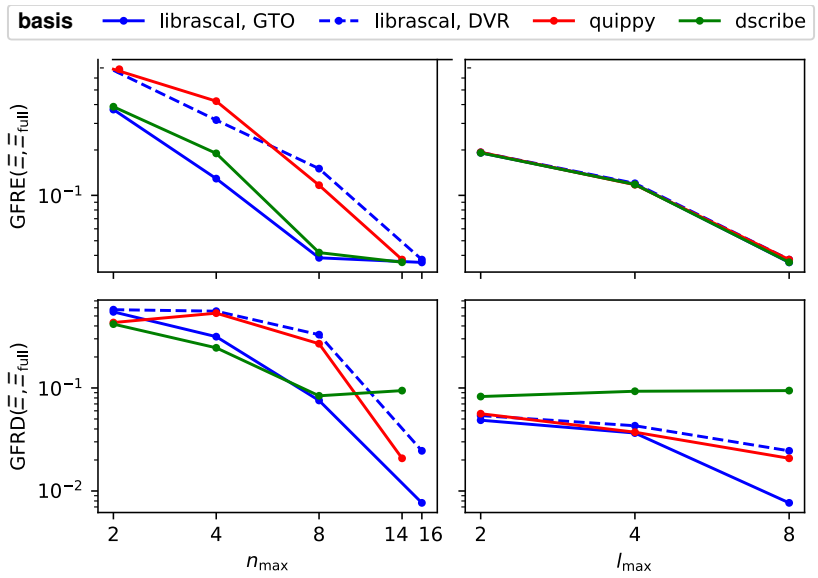


FIG. 31. The panels demonstrate the convergence of SOAP features as computed for 10'000 random CH_4 configurations[269] using the radial bases implemented in different codes, and measured in terms of the error one incurs when linearly predicting fully-converged features $\mathbf{\Xi}_{\text{full}}$ ($n_{\text{max}} = 24, l_{\text{max}} = 12$, computed with the GTO implementation in librascal) using features with $l_{\text{max}} = 8$ and growing values of n_{max} (left) and with $n_{\text{max}} = 14$ (16 for librascal) and growing values of l_{max} (right). Top panels show the linear reconstruction error (GFRE) which measures the amount of information that cannot be linearly decoded from the coarser features. Bottom panels show the reconstruction distortion (GFRD) which measures the additional error one makes when limiting the reconstruction to an orthogonal transformation.[271]

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[298], dscribe[296] and librascal[287,297], 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[299], and atomic structures are assumed to be stored in the variable `structures`, an instance of ASE's `Atoms` object. In all of these implementations, the descriptor vectors are returned as `numpy.array` objects, from which kernel values may be obtained by computing the dot products between descriptor vectors. We also do not discuss the computational efficiency of the different codes, which is still the subject of very active development. Fig. 30 provides some representative timings from librascal, and from a recent implementation of SOAP that uses non-Gaussian atomic densities[191]. The wildly different breakdown of the computational effort as a function of the basis set size, and the large overhead associated with the evaluation of the gradients of the features highlight some of the implementation challenges.

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.

```
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 `Descriptor` object is initialised using a string containing the kernel parameters in a `key=value` format, with some keys being mandatory.

The `dscribe` package[296] implements multiple descriptors, including SOAP, MBTR[31] 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 `quippy` and `dscribe` are the choice of radial basis functions, which are spherical primitive Gaussian Type Orbitals (GTOs), orthogonalised using the method suggested by Löwdin[300]. 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.

```
from dscribe.descriptors import SOAP
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 the python object providing the descriptor is constructed from the class `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 $|\overline{\rho_i^{\otimes\nu}}\rangle$ features, librascal implements two kinds of radial functions, namely a family of GTO-like radial functions[157] 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,[301] 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 $(n_{\max}, l_{\max})$ are increased. Figure 31 demonstrates the convergence of the GFRE and GFRD (see Section VIII A and Ref. 271) between small-$(n_{\max}, l_{\max})$ features and a highly converged $\boldsymbol{\Xi}_{\text{full}}$ featurization. In all cases we consider GFRE$(\boldsymbol{\Xi}_{\text{full}}, \boldsymbol{\Xi})$ is at least one order of magnitude smaller than GFRE$(\boldsymbol{\Xi}, \boldsymbol{\Xi}_{\text{full}})$. One sees that, reassuringly, in all cases the feature reconstruction errors converge towards zero. For $n_{\max} = 16$ 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 $l_{\max}$ 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 (although they are not fully equivalent, as they are parameterized differently), 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 $(n_{\max}, l_{\max})$ increase beyond the values that are typically used in practice.

IX. APPLICATIONS AND CURRENT TRENDS

In this Section we 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.

A. Best match kernels for ligand binding

Contrary to the problem of predicting interatomic potentials, or other extensive properties, 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 32 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)[51]. 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

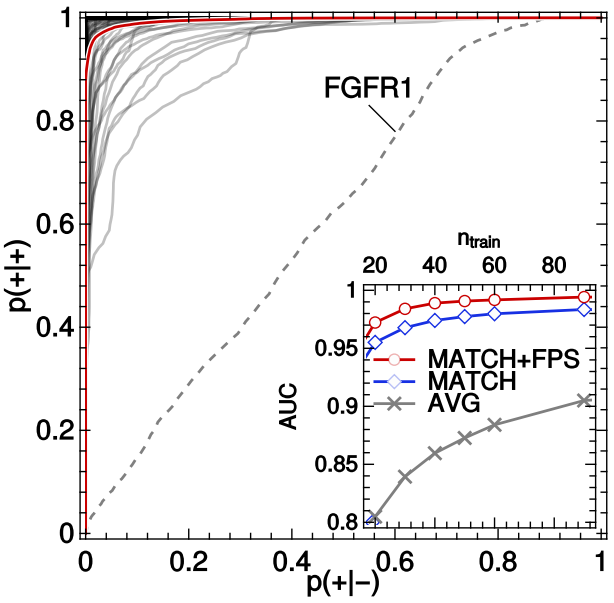


FIG. 32. ROCs of binary classifiers based on a SOAP kernel, applied to the prediction of the binding behavior of ligands and decoys taken from the DUD-E[51], trained on 60 examples. Each ROC corresponds to one specific protein receptor, and plots the fraction of true positives $p(+|+)$ against the fraction of false negatives $p(+|-)$. 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[30]. 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). Reprinted with permission from Ref. 30. © The Authors, some rights reserved; exclusive licensee AAAS. Distributed under a Creative Commons Attribution License 4.0 (CC BY-NC).

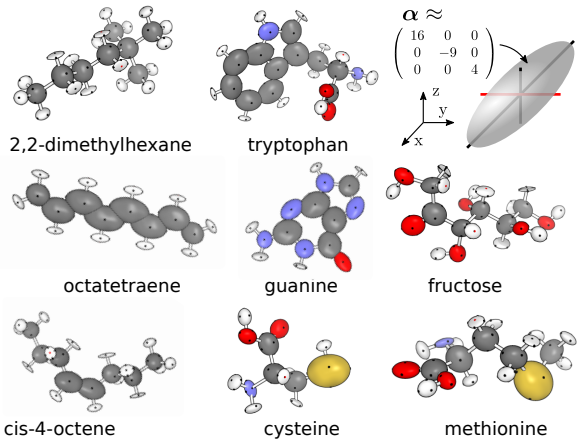


FIG. 33. Predicted atomic contributions to the total CCSD polarizability tensor for a selection of the showcase dataset as reported in Refs. 24,302. The ellipsoids are aligned along the principal axes of $\boldsymbol{\alpha}_i$, and their extent is proportional to the square root of the corresponding eigenvalue. The principal axes are shown, and are colored based on whether the corresponding eigenvalues are positive (black) or negative (red). Reproduced with permission from Ref. 24. Copyright 2019 National Academy of Sciences.

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. 32 shows that a model based on an average metric – that describes each molecule as the average of its environments, Eq. (95) – 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. 97, and implemented in practice as the small-γ limit of the REMatch kernel[69]) 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.

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 $\boldsymbol{\mu}$ (and the equivalent bulk quantity, polarization), polarizability $\boldsymbol{\alpha}$ (and the closely-related electronic dielectric constant) as well as higher-order terms, such as the first hyperpolarizability $\boldsymbol{\beta}$. We discuss the case of the static dipole polarizability $\boldsymbol{\alpha}$ as a representative case that highlights many of the current ideas and applications. In its Cartesian form, $\boldsymbol{\alpha}$ 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, $\alpha_0^{(0)}$, corresponds to the trace of the tensor, while the 5 anisotropic components, $\alpha^{(2)}_{\{-2,-1,0,+1,+2\}}$ transform collectively as $\lambda = 2$ spherical harmonics, and can be computed using recursive relationships that are explicitly reported in Ref. 158. A clear advantage of this construction is that, unlike the components of the Cartesian tensor, the two ISCs of $\boldsymbol{\alpha}$ 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. (40).[24,157,159]

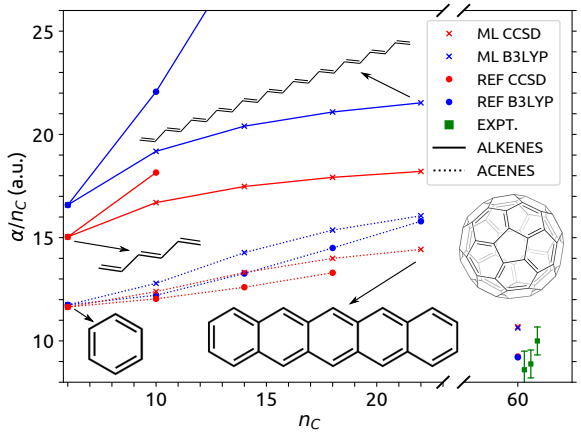


FIG. 34. Polarizability per carbon atom (α/n_C) vs. number of carbons (n_C) for the series of s-*trans* alkenes (from C_6H_8 to $C_{22}H_{24}$, full line) and acenes (from benzene to pentacene, dotted line), as well as fullerene (C_{60}). The green squares (and error bars) indicate the experimental measurements for C_{60}[303]. Results are provided from DFT (blue) and CCSD (red) calculations, as well as the corresponding AlphaML models. Reproduced with permission from Ref. 24. Copyright 2019 National Academy of Sciences.

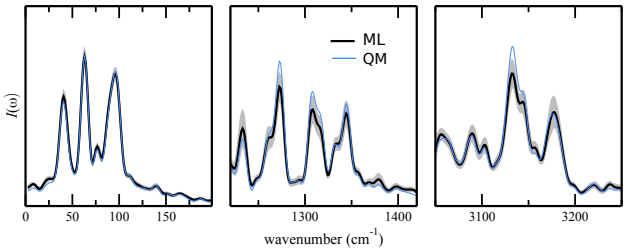


FIG. 35. (*black line*) Raman spectrum prediction of paracetamol form-I averaged over 16 different training models. Each training model is obtained by a random subselection of 2000 configurations over a total of 2500. (*shaded area*) Standard deviation of the predicted spectra over the 16 models, calibrated with a likelihood maximization procedure described in Ref. 172. (*blue line*) Reference *ab initio* Raman spectrum. Adapted from Ref. 160. Copyright 2019 IOP Publishing under Creative Commons Attribution 4.0 International License `https://creativecommons.org/licenses/by/4.0/`.

The inherent locality of the model means that the tensor prediction can be broken down in the sum of individual atomic contributions $\boldsymbol{\alpha} = \sum_i \boldsymbol{\alpha}_i$. 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 model[24,304] to fit against coupled-clusters (CCSD) reference values, computed on small organic molecules from the QM7b dataset[254,302], 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 $\boldsymbol{\alpha}$, which are represented in Fig. 33, in terms of ellipsoids aligned along the principal axes of $\boldsymbol{\alpha}(A_i)$. Even though these components do not have to be physically meaningful – given that the only training target is given by *total* polarizabilities – the local $\boldsymbol{\alpha}(A_i)$ 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. Due to the spatial nearsightedness of the representation, the model breaks down when asked to predict the polarizability of large polyenes and polyacenes based on the information learned on simpler and smaller molecular units. This is well represented in Fig. 34, where the prediction of $\boldsymbol{\alpha}$ is tested for conjugated carbon-based molecules of increasing size, including fullerene[24].

The prediction of $\boldsymbol{\alpha}$ 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. 160, 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. 35, given the local 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[305] – a task that until very recently required enormous computational effort[306].

C. Long-range and non-local responses

The clear breakdown of a ML model based on local features that is apparent in Fig. 34 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 36. 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. 36, adapted

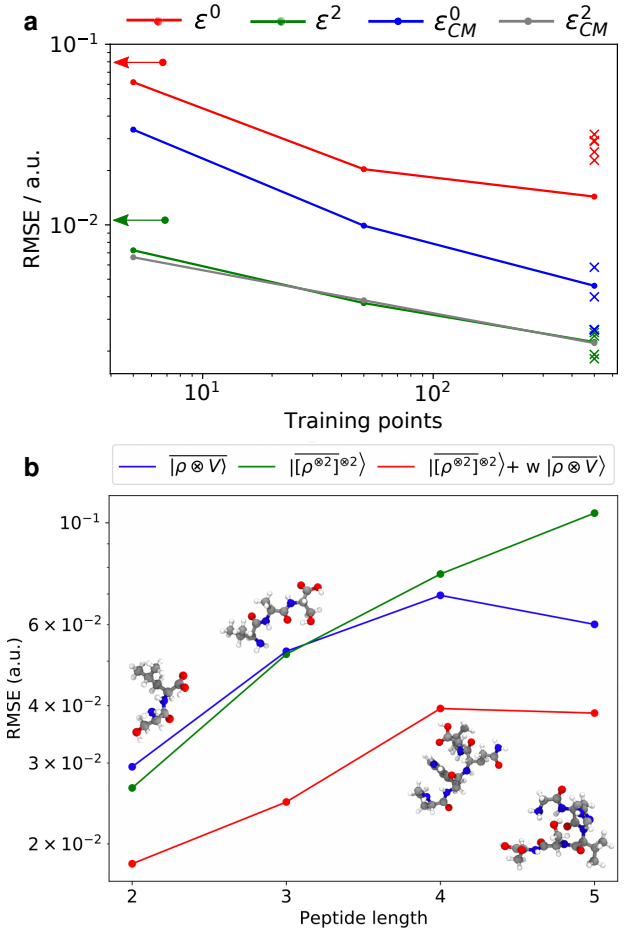


FIG. 36. (a) Learning curves of the $\lambda = 0$ and $\lambda = 2$ components of the dielectric response tensor $\boldsymbol{\epsilon}_\infty$ of water, through direct learning (red and green lines, respectively) and indirect learning going through the Clausius-Mossotti relation (blue and gray). The 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. Adapted with permission from Ref. 159. Copyright 2018 American Physical Society (b) 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 with permission from Ref. 163. Copyright 2020 Royal Society of Chemistry.

from Ref. 159, is an example of this approach. The electronic dielectric response $\boldsymbol{\epsilon}_\infty$ 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, $\boldsymbol{\alpha} = V(\boldsymbol{\epsilon} - 1)/(\boldsymbol{\epsilon} + 2)$, that links $\boldsymbol{\epsilon}_\infty$ to an effective molecular polarizability. This effective $\boldsymbol{\alpha}$ is more readily learnable by a local model, leading to better accuracy and transferability in predicting $\boldsymbol{\epsilon}_\infty$.

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 (Section IV H and V C) 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. 36, 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 peptides 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[166] – where however the splitting between local and long-range physics was achieved by combining different regression models rather than by different choices of features.

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 $\tilde{\rho}(\mathbf{r})$ of an atomic structure.[307] 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*[308,309]. 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 $\tilde{\rho}(\mathbf{r})$ as a sum of atom-centered contributions, $\tilde{\rho}(\mathbf{r}) = \sum_i \tilde{\rho}(A_i;\mathbf{r})$. These atom-centered terms can then be conveniently decomposed as a sum of local functions, at the price of adopting a multi-centered non-orthogonal basis for the expansion i.e.,

$$\tilde{\rho}(A_i,\mathbf{r}) = \sum_{n\lambda\mu} \tilde{c}_{n\lambda\mu}(A_i) R_n(|\mathbf{r}-\mathbf{r}_i|) Y^\lambda_\mu\left(\widehat{\mathbf{r}-\mathbf{r}_i}\right) \tag{126}$$

where R_n represent some suitably optimized radial functions (for instance those used in resolution of the identity methods in quantum chemistry[310]) and $\tilde{c}_{n\lambda\mu}(A_i)$ 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 λ-component can be independently predicted using equivariant features of the corresponding order – for instance with a linear model

$$\tilde{c}_{n\lambda\mu}(A_i) \approx \sum_q \langle \tilde{c}_{n\lambda} | Q \rangle \, \langle Q | A_i; \overline{\rho_i^{\otimes\nu}}; \sigma; \lambda\mu \rangle \,, \quad (127)$$

even though current implementations use a kernel regression scheme[23,311]. The non-orthogonality of the basis used to represent $\tilde{\rho}(\boldsymbol{r})$, implies that the learning phase has now to be performed considering all the different density components at the same time[23,311]. 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. 37, where the electron density prediction of C(8) hydrocarbons[23] 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[311], and to the prediction of other scalar fields such as the on-top density[312].

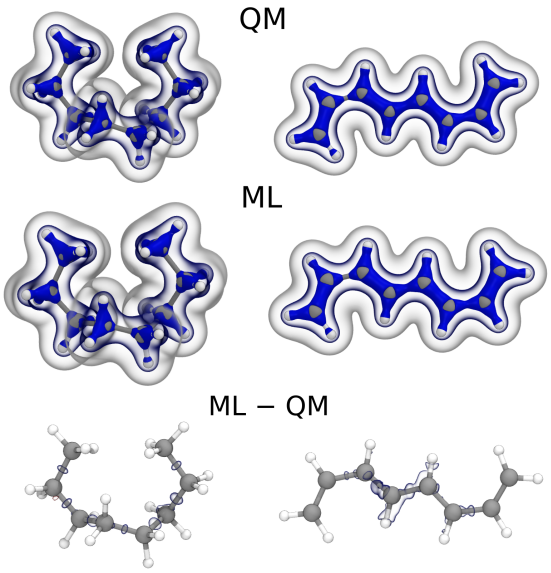


FIG. 37. Extrapolation results for the valence electron density of one octane (*left*) and one octatetraene (*right*) conformer, using a model trained on butadiene and butane. (*top*) DFT/PBE density isosurface at 0.25, 0.1, 0.01 Bohr^{-3}, (*middle*) machine-learning prediction isosurface at 0.25, 0.1, 0.01 Bohr^{-3}, (*bottom*) machine-learning error, red and blue isosurfaces refer to ± 0.005 Bohr^{-3} respectively. Reproduced from Ref. 23. Copyright 2018 American Chemical Society.

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 turns 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 38, adapted from Ref. 313 shows a classification of snapshots taken from simulations of different phases of water, based on Steinhardt order parameters[236], which are closely related to $|\overline{\rho_i^{\otimes 2}}\rangle$ features, and make it possible to partly differentiate between phases. In the same study it is shown how a neural network based on atom-centered symmetry functions can be trained to achieve near-perfect classification accuracy. 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. 218, using permutation-invariant vectors[316] 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[1]. In combination with a dimensionality-reduction scheme[4], and with a generalized convex hull construction that attempts to estimate the synthesizability of materials by considering jointly their pre-

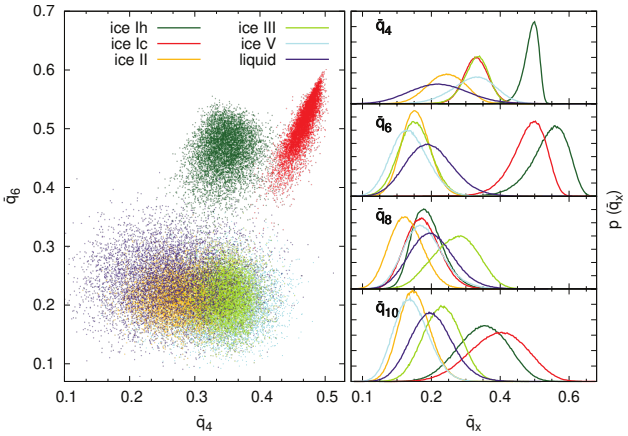


FIG. 38. Scatter plot and histograms based on Steinhardt order parameters[236] q_n computed for simulations of liquid water and different phases of ice. Reproduced with permission from Ref. 313. Copyright 2013 American Institute of Physics.

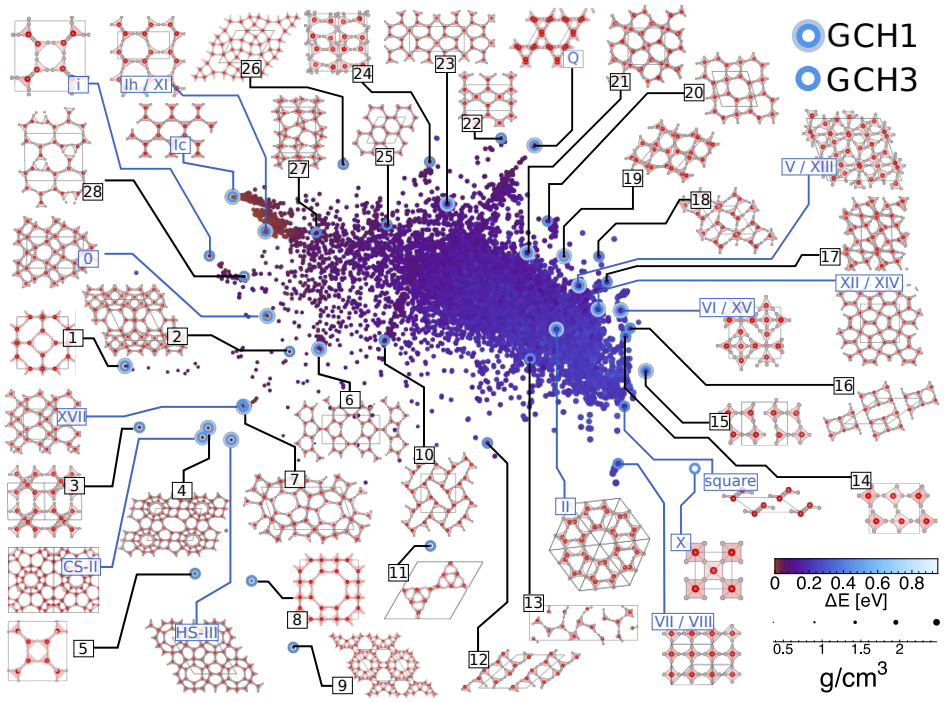


FIG. 39. Sketch map of the structural similarity of 15,869 distinct PBE-DFT geometry-optimised ice structures, as computed by the Euclidean distance in SOAP space. The sketch-map coordinates are obtained by minimizing the error in reproducing this similarity in terms of distances between points on a 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[314] construction. Reproduced from Ref. 315. Copyright 2018 Springer Nature under Creative Commons Attribution 4.0 International License `https://creativecommons.org/licenses/by/4.0/`.

dicted stability, and the structural similarity to other potential candidates[314], 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[315] (see Fig. 39).

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[244,318], molecular materials[255,319,320], inorganic perovskites[321], corrosion inhibitors[322]; the identification and characterization of defects in solids[323–325] and self-assembled polymers[326,327]; the classification of secondary-structure patterns in polypeptides[259] and the building blocks of zeolites[230,328] and porous materials[329]; the classification of different phases in multi-phase materials[330,331]; the characterization of amorphous systems[18,332–335]; the search of stable phases of materials[336]; the determination of the convergence of microsolvation studies of the hydration free energy[337].

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. 98 a PCA map was used to understand the ef-

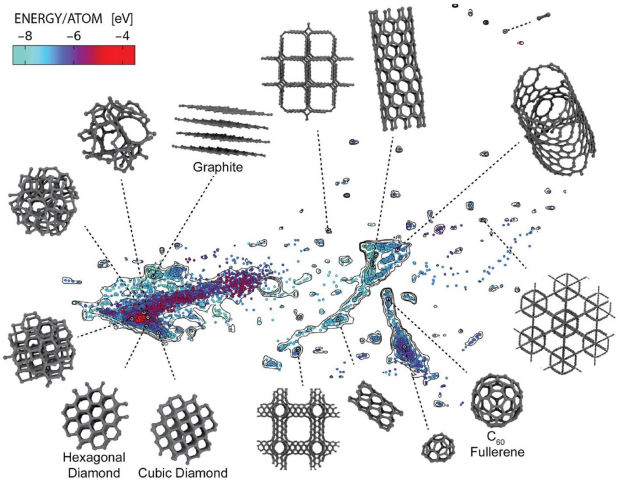


FIG. 40. Sketch-map describing the makeup of the structures included in the train set of an accurate and transferable potential for carbon. Selected structures are identified for graphite, diamond, hexagonal diamond (Lonsdaleite), amorphous carbon and fullerenes. Points are coloured according to their energy, while contours indicate the density of the database population in a particular region. Adapted with permission from Ref. 317. Copyright 2020 American Institute of Physics.

fect 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. 69, 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. 94, 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.[30,69,114,115,280,338–340] An example of such map is given in Fig. 40, showing the diversity of the structures used to train a transferable machine-learning potential for carbon.[317] 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 and reaction predictions

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 III, 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.[341] From a conceptual point of view, their coarsness 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,[342,343] as well as for data analytics approaches for materials informatics[344]. 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.[345,346] As we briefly discuss in Section VII E, 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[347], the computational screening for antiviral drugs[348], and the prediction of enantioselectivity of organocatalysts[349].

G. Descriptors from electronic-structure theory

Another growing trend that is worth a brief mention involves the use of information from electronic-structure calculations in the construction of structural representations. The idea has been applied in different forms. At the simplest level, electronic-structure-based indicators of chemical similarity, obtained for bulk elements, have been used in the construction of elemental similarity kernels,[69] to obtain models that are more predictive across chemical space[350]. Alternatively, electronic-structure indicators, such as the local density of states, can be used side-by-side with purely structural representations, yielding a substantial improvement of the accuracy of the model[351,352].

Elements of an electronic structure calculation, such as the charge density[353], the electron density of states[354] or the elements of the Fock matrix[355] can be used directly as the basis for a molecular representation. This approach requires an electronic structure calculation in order to make predictions for each new structure, which implies a substantial overhead in comparison with methods using as inputs only the atomic positions. However, the increase in the transferability of the models may well justified the greater computational effort, particularly when using descriptors based on low levels of quantum mechanical theory to predict high-end, accurate molecular properties.[356,357]

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 criteria 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 the statistical physics of liquids (ν-point density correlations) and of alloys (the cluster expansion), as well as with the construction of potential energy surfaces for molecules and the condensed phase.

A formal treatment of symmetries 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. Even though it is possible to build systematically a complete basis to expand in a linear fashion any structure-property relation, it is not clear how to build a *minimal* set of features that guarantees an injective mapping when used as the input of a general *non-linear* model, or how to reduce in an effective manner the size of a complete linear basis. A better understanding of the mathematical properties of representations is likely to lead, in the near future, to more robust and better performing implementations, and might also help design better "deep" models, by identifying the algebraic manipulations that increase most effectively the expressive power of the features used as inputs.

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. The quantitative assessment of the mutual information content of alternative descriptors, of their sensitivity to structural deformations, and to the degree to which they correlate with the target properties may serve as a guide to strike a balance between these conflicting goals, and to make better informed choices between alternative frameworks.

One of the most recent research directions aims at extending even further the reach of the class of descriptors we discuss in this review, 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 corresponding to 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 connections 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. A definitive answer to this question is still lacking, but we believe that the general principles that we have summarized in this review may indicate the direction to follow, and provide some guidance to the practitioners who seek to make an informed choice among the

ever increasing number of representations for atomic-scale modeling.

ACKNOWLEDGMENTS

The authors would like to thank Yasushi Shibuta for providing the structures used in Fig. 14, and Stefan Goedecker 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. 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 inito 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 Modelling 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. 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.

REFERENCES

* michele.ceriotti@epfl.ch

[1] Isayev, O.; Fourches, D.; Muratov, E. N.; Oses, C.; Rasch, K.; Tropsha, A.; Curtarolo, S. Materials Cartography: Representing and Mining Materials Space Using Structural and Electronic Fingerprints. *Chem. Mater.* **2015**, *27*, 735–743.

[2] Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering. *Science* **2018**, *361*, 360–365.

[3] Das, P.; Moll, M.; Stamati, H.; Kavraki, L. E.; Clementi, C. Low-Dimensional, Free-Energy Landscapes of Protein-Folding Reactions by Nonlinear Dimensionality Reduction. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 9885–9890.

[4] Ceriotti, M.; Tribello, G. A.; Parrinello, M. Simplifying the Representation of Complex Free-Energy Landscapes Using Sketch-Map. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 13023–13028.

[5] Spiwok, V.; Králová, B. Metadynamics in the Conformational Space Nonlinearly Dimensionally Reduced by Isomap. *J. Chem. Phys.* **2011**, *135*, 224504.

[6] Rohrdanz, M. A.; Zheng, W.; Clementi, C. Discovering Mountain Passes via Torchlight: Methods for the Definition of Reaction Coordinates and Pathways in Complex Macromolecular Reactions. *Annu. Rev. Phys. Chem.* **2013**, *64*, 295–316.

[7] Kanekal, K. H.; Bereau, T. Resolution limit of data-driven coarse-grained models spanning chemical space. *Journal of Chemical Physics* **2019**, *151*.

[8] Jackson, N. E.; Bowen, A. S.; Antony, L. W.; Webb, M. A.; Vishwanath, V.; de Pablo, J. J. Elec-

tronic structure at coarse-grained resolutions from supervised machine learning. *Science Advances* **2019**, *5*, eaav1190.

[9] Wang, J.; Olsson, S.; Wehmeyer, C.; Pérez, A.; Charron, N. E.; De Fabritiis, G.; Noé, F.; Clementi, C. Machine Learning of Coarse-Grained Molecular Dynamics Force Fields. *ACS Central Science* **2019**, *5*, 755–767.

[10] Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.

[11] Braams, B. J.; Bowman, J. M. Permutationally Invariant Potential Energy Surfaces in High Dimensionality. *Int. Rev. Phys. Chem.* **2009**, *28*, 577–606.

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

[13] Sosso, G. C.; Miceli, G.; Caravati, S.; Behler, J.; Bernasconi, M. Neural Network Interatomic Potential for the Phase Change Material GeTe. *Phys. Rev. B* **2012**, *85*, 174103.

[14] Behler, J. Perspective: Machine Learning Potentials for Atomistic Simulations. *J. Chem. Phys.* **2016**, *145*, 170901.

[15] Deringer, V. L.; Csányi, G. Machine Learning Based Interatomic Potential for Amorphous Carbon. *Phys. Rev. B* **2017**, *95*, 094203.

[16] Dragoni, D.; Daff, T. D.; Csányi, G.; Marzari, N. Achieving DFT Accuracy with a Machine-Learning Interatomic Potential: Thermomechanics and Defects in Bcc Ferromagnetic Iron. *Phys. Rev. Materials* **2018**, *2*, 013808.

[17] Cheng, B.; Mazzola, G.; Pickard, C. J.; Ceriotti, M. Evidence for Supercritical Behaviour of High-Pressure Liquid Hydrogen. *Nature* **2020**, *585*, 217–220.

[18] Deringer, V. L.; Bernstein, N.; Csányi, G.; Ben Mahmoud, C.; Ceriotti, M.; Wilson, M.; Drabold, D. A.; Elliott, S. R. Origins of Structural and Electronic Transitions in Disordered Silicon. *Nature* **2021**, *589*, 59–64.

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

[20] Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: An Extensible Neural Network Potential with DFT Accuracy at Force Field Computational Cost. *Chem. Sci.* **2017**, *8*, 3192–3203.

[21] Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet – A Deep Learning Architecture for Molecules and Materials. *J. Chem. Phys.* **2018**, *148*, 241722.

[22] Paruzzo, F. M.; Hofstetter, A.; Musil, F.; De, S.; Ceriotti, M.; Emsley, L. Chemical Shifts in Molecular Solids by Machine Learning. *Nat. Commun.* **2018**, *9*, 4501.

[23] Grisafi, A.; Fabrizio, A.; Meyer, B.; Wilkins, D. M.; Corminboeuf, C.; Ceriotti, M. Transferable Machine-Learning Model of the Electron Density. *ACS Cent. Sci.* **2019**, *5*, 57–64.

[24] Wilkins, D. M.; Grisafi, A.; Yang, Y.; Lao, K. U.; DiStasio, R. A.; Ceriotti, M. Accurate Molecular Polarizabilities with Coupled Cluster Theory and Machine Learning. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 3401–3406.

[25] Schütt, K. T.; Gastegger, M.; Tkatchenko, A.; Müller, K.-R.; Maurer, R. J. Unifying Machine Learning and Quantum Chemistry with a Deep Neural Network for Molecular Wavefunctions. *Nat Commun* **2019**, *10*, 5024.

[26] Kalita, B.; Li, L.; McCarty, R. J.; Burke, K. Learning to Approximate Density Functionals. *Acc. Chem. Res.* **2021**, acs.accounts.0c00742.

[27] Ouyang, R.; Curtarolo, S.; Ahmetcik, E.; Scheffler, M.; Ghiringhelli, L. M. SISSO: A Compressed-Sensing Method for Identifying the Best Low-Dimensional Descriptor in an Immensity of Offered Candidates. *Phys. Rev. Mater.* **2018**, *2*, 083802.

[28] Behler, J.; Lorenz, S.; Reuter, K. Representing Molecule-Surface Interactions with Symmetry-Adapted Neural Networks. *The Journal of Chemical Physics* **2007**, *127*, 014705.

[29] Bartók, A. P.; Kondor, R.; Csányi, G. On Representing Chemical Environments. *Phys. Rev. B* **2013**, *87*, 184115.

[30] Bartók, A. P.; De, S.; Poelking, C.; Bernstein, N.; Kermode, J. R.; Csányi, G.; Ceriotti, M. Machine Learning Unifies the Modeling of Materials and Molecules. *Sci. Adv.* **2017**, *3*, e1701816.

[31] Huo, H.; Rupp, M. Unified Representation for Machine Learning of Molecules and Crystals. *ArXiv Prepr. ArXiv170406439* **2017**, *13754*.

[32] Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S. P. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chem. Mater.* **2019**, *31*, 3564–3572.

[33] Deringer, V. L.; Bartók, A. P.; Noam Bernstein, D. M. W.; Ceriotti, M.; Csányi, G. Gaussian Process Regression for Materials and Molecules. *Chem. Rev. (under review)* **2021**,

[34] Unke, O. T.; Chmiela, S.; Sauceda, H. E.; Gastegger, M.; Poltavsky, I.; Schütt, K. T.; Tkatchenko, A.; Müller, K.-R. Machine Learning Force Fields. *Chem. Rev.* **2021**,

[35] Manzhos, S.; Carrington, T. Neural Network Potential Energy Surfaces for Small Molecules and Reactions. *Chem. Rev.* **2020**,

[36] Behler, J. Four Generations of High-Dimensional Neural Network Potentials. *Chem. Rev.* **2021**,

[37] Westermayr, J.; Marquetand, P. Machine Learning for Electronically Excited States of Molecules. *Chem. Rev.* **2020**,

[38] Glielmo, A.; Husic, B. E.; Rodriguez, A.; Clementi, C.; Noé, F.; Laio, A. Unsupervised Learning Methods for Molecular Simulation Data. *Chem. Rev.* **2021**,

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

[40] Karelson, M. *Molecular Descriptors in QSAR/QSPR*; Wiley-Interscience: New York, 2000.

[41] Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28*, 31–36.

[42] Todeschini, R.; Consonni, V. *Molecular Descriptors*

for Chemoinformatics; Methods and Principles in Medicinal Chemistry; Wiley, 2010; Vol. 2; pp 1–252.

[43] Wills, T. J.; Polshakov, D. A.; Robinson, M. C.; Lee, A. A. Impact of Chemist-In-The-Loop Molecular Representations on Machine Learning Outcomes. *J. Chem. Inf. Model.* **2020**, *60*, 4449–4456.

[44] Schneider, G.; Fechner, U. Computer-Based de Novo Design of Drug-like Molecules. *Nat Rev Drug Discov* **2005**, *4*, 649–663.

[45] Karelson, M.; Lobanov, V. S.; Katritzky, A. R. Quantum-Chemical Descriptors in QSAR/QSPR Studies. *Chem. Rev.* **1996**, *96*, 1027–1044.

[46] Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research* **2012**, *40*, D1100–D1107.

[47] Sterling, T.; Irwin, J. J. ZINC 15 – Ligand Discovery for Everyone. *Journal of Chemical Information and Modeling* **2015**, *55*, 2324–2337.

[48] Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem substance and compound databases. *Nucleic Acids Research* **2016**, *44*, D1202–D1213.

[49] Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chemical Science* **2018**, *9*, 513–530.

[50] Obrezanova, O.; Csányi, G.; Gola, J. M. R.; Segall, M. D. Gaussian Processes: A Method for Automatic QSAR Modeling of ADME Properties. *J. Chem. Inf. Model.* **2007**, *47*, 1847–1857.

[51] Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking. *Journal of Medicinal Chemistry* **2012**, *55*, 6582–6594.

[52] Jones, J. E. On the Determination of Molecular Fields. —II. From the Equation of State of a Gas. *Proc. R. Soc. Lond. A* **1924**, *106*, 463–477.

[53] Alder, B. J.; Gass, D. M.; Wainwright, T. E. Studies in molecular dynamics. VIII. The transport coefficients for a hard-sphere fluid. *The Journal of Chemical Physics* **1970**, *53*, 3813–3826.

[54] Stillinger, F. H.; Rahman, A. 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. Chem. Phys. Gen. Method J. Chem. Phys. J. Chem. Phys. J. Chem. Phys.* **1974**, *601*, 1545–926.

[55] Car, R.; Parrinello, M. Unified Aproach for Molecular Dynamics and Density-Functional Theory. R. Car and M. Parrinello.Pdf. *Phys. Rev. Lett.* **1985**, *55*, 2471–2474.

[56] Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford University Press, USA, 1990.

[57] Frenkel, D.; Smit, B. *Understanding Molecular Simulation*, 2nd ed.; Academic Press: London, 2002.

[58] Parr, R. G.; Yang, W. *Density-Functional Theory of Atoms and Molecules*, 1st ed.; International Series of Monographs on Chemistry 16; Oxford Univ. Press [u.a.]: New York, NY, 1994.

[59] Burke, K. Perspective on Density Functional Theory. *J. Chem. Phys.* **2012**, *136*, 150901.

[60] Booth, G. H.; Grüneis, A.; Kresse, G.; Alavi, A. Towards an Exact Description of Electronic Wavefunctions in Real Solids. *Nature* **2013**, *493*, 365–370.

[61] Partridge, H.; Schwenke, D. W. The Determination of an Accurate Isotope Dependent Potential Energy Surface for Water from Extensive Ab Initio Calculations and Experimental Data. *J. Chem. Phys.* **1997**, *106*, 4618.

[62] Huang, X.; Braams, B. J.; Bowman, J. M. Ab Initio Potential Energy and Dipole Moment Surfaces for $H_5O_2\hat{+}$. *J. Chem. Phys.* **2005**, *122*, 44308.

[63] Russell, C. L.; Manolopoulos, D. E. How to Observe the Elusive Resonances in F + H2 Reactive Scattering. *Chemical Physics Letters* **1996**, *256*, 465–473.

[64] Kim, J. B.; Weichman, M. L.; Sjolander, T. F.; Neumark, D. M.; K os, J.; Alexander, M. H.; Manolopoulos, D. E. Spectroscopic Observation of Resonances in the F + H2 Reaction. *Science* **2015**, *349*, 510–513.

[65] Tuckerman, M. *Statistical Mechanics and Molecular Simulations*; Oxford University Press, 2008.

[66] Feynman, R. P.; Hibbs, A. R. *Quantum Mechanics and Path Integrals*; McGraw-Hill: New York, 1964.

[67] Markland, T. E.; Ceriotti, M. Nuclear Quantum Effects Enter the Mainstream. *Nat. Rev. Chem.* **2018**, *2*, 0109.

[68] Sadeghi, A.; Ghasemi, S. A.; Schaefer, B.; Mohr, S.; Lill, M. A.; Goedecker, S. Metrics for Measuring Distances in Configuration Spaces. *J. Chem. Phys.* **2013**, *139*, 184118.

[69] De, S.; Bartók, A. P.; Csányi, G.; Ceriotti, M. Comparing Molecules and Solids across Structural and Alchemical Space. *Phys. Chem. Chem. Phys.* **2016**, *18*, 13754–13769.

[70] Bernstein, N.; Bhattarai, B.; Csányi, G.; Drabold, D. A.; Elliott, S. R.; Deringer, V. L. Quantifying Chemical Structure and Machine-Learned Atomic Energies in Amorphous and Liquid Silicon. *Angew. Chem. Int. Ed.* **2019**, *58*, 7057–7061.

[71] Yang, W. Direct Calculation of Electron Density in Density-Functional Theory. *Phys. Rev. Lett.* **1991**, *66*, 1438–1441.

[72] Galli, G.; Parrinello, M. Large Scale Electronic Structure Calculations. *Phys. Rev. Lett.* **1992**, *69*, 3547–3550.

[73] Pulay, P.; Fogarasi, G.; Pang, F.; Boggs, J. E. Systematic Ab Initio Gradient Calculation of Molecular Geometries, Force Constants, and Dipole Moment Derivatives. *J. Am. Chem. Soc.* **1979**, *101*, 2550–2560.

[74] Mayo, S. L.; Olafson, B. D.; Goddard, W. A. DREIDING: A Generic Force Field for Molecular Simulations. *J. Phys. Chem.* **1990**, *94*, 8897–8909.

[75] Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell, A. D. CHARMM General Force Field: A Force Field for Drug-like Molecules Compatible with the CHARMM All-Atom Additive Biological Force Fields. *J. Comput. Chem.* **2009**, NA–NA.

[76] Halgren, T. A. Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490–519.

[77] Damm, W.; Frontera, A.; Tirado–Rives, J.; Jorgensen, W. L. OPLS All-Atom Force Field for Carbohydrates. *J. Comput. Chem.* **1997**, *18*, 1955–1970.

[78] Baker, J.; Hehre, W. J. Geometry Optimization in Cartesian Coordinates: The End of theZ-Matrix? *J. Comput. Chem.* **1991**, *12*, 606–610.

[79] Baker, J.; Chan, F. The Location of Transition States: A Comparison of Cartesian, Z-Matrix, and Natural Internal Coordinates. *J. Comput. Chem.* **1996**, *17*, 888–904.

[80] Harrison, J. A.; Schall, J. D.; Maskey, S.; Mikulski, P. T.; Knippenberg, M. T.; Morrow, B. H. Review of force fields and intermolecular potentials used in atomistic computational materials research. *Applied Physics Reviews* **2018**, *5*, 031104.

[81] Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. Advances in Neural Information Processing Systems. 2015; pp 2224–2232.

[82] Kamerlin, S. C. L.; Warshel, A. The Empirical Valence Bond Model: Theory and Applications. *WIREs Comput Mol Sci* **2011**, *1*, 30–45.

[83] Brown, A.; McCoy, A. B.; Braams, B. J.; Jin, Z.; Bowman, J. M. Quantum and Classical Studies of Vibrational Motion of CH5 on a Global Potential Energy Surface Obtained from a Novel Ab Initio Direct Dynamics Approach. *J. Chem. Phys.* **2004**, *121*, 4105–4116.

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

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

[86] Jiang, B.; Guo, H. Permutation Invariant Polynomial Neural Network Approach to Fitting Potential Energy Surfaces. *The Journal of Chemical Physics* **2013**, *139*, 054112.

[87] Collins, M. A.; Parsons, D. F. Implications of Rotation–Inversion–Permutation Invariance for Analytic Molecular Potential Energy Surfaces. *The Journal of Chemical Physics* **1993**, *99*, 6756–6772.

[88] Blank, T. B.; Brown, S. D.; Calhoun, A. W.; Doren, D. J. Neural Network Models of Potential Energy Surfaces. *J. Chem. Phys.* **1995**, *103*, 4129.

[89] Gassner, H.; Probst, M.; Lauenstein, A.; Hermansson, K. Representation of Intermolecular Potential Functions by Neural Networks. *J. Phys. Chem. A* **1998**, *102*, 4596–4605.

[90] Sanchez, J.; Ducastelle, F.; Gratias, D. Generalized Cluster Description of Multicomponent Systems. *Physica A: Statistical Mechanics and its Applications* **1984**, *128*, 334–350.

[91] Pettifor, D. New Many-Body Potential for the Bond Order. *Phys. Rev. Lett.* **1989**, *63*, 2480–2483.

[92] Horsfield, A. P.; Bratkovsky, A. M.; Fearn, M.; Pettifor, D. G.; Aoki, M. Bond-Order Potentials: Theory and Implementation. *Phys. Rev. B* **1996**, *53*, 12694–12712.

[93] Ozaki, T.; Aoki, M.; Pettifor, D. G. Block Bond-Order Potential as a Convergent Moments-Based Method. *Phys. Rev. B* **2000**, *61*, 7972–7988.

[94] Nguyen, T. T.; Székely, E.; Imbalzano, G.; Behler, J.; Csányi, G.; Ceriotti, M.; Götz, A. W.; Paesani, F. Comparison of Permutationally Invariant Polynomials, Neural Networks, and Gaussian Approximation Potentials in Representing Water Interactions through Many-Body Expansions. *J. Chem. Phys.* **2018**, *148*, 241725.

[95] van der Oord, C.; Dusson, G.; Csányi, G.; Ortner, C. Regularised Atomic Body-Ordered Permutation-Invariant Polynomials for the Construction of Interatomic Potentials. *Mach. Learn. Sci. Technol.* **2020**, *1*, 015004.

[96] Moussa, J. E. Comment on "Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning". *Phys. Rev. Lett.* **2012**, *109*, 059801.

[97] Montavon, G.; Hansen, K.; Fazli, S.; Rupp, M.; Biegler, F.; Ziehe, A.; Tkatchenko, A.; Lilienfeld, A. V.; Müller, K.-R. In *Advances in Neural Information Processing Systems 25*; Pereira, F., Burges, C. J. C., Bottou, L., Weinberger, K. Q., Eds.; Curran Associates, Inc., 2012; pp 440–448.

[98] Hansen, K.; Montavon, G.; Biegler, F.; Fazli, S.; Rupp, M.; Scheffler, M.; von Lilienfeld, O. A.; Tkatchenko, A.; Müller, K.-R. Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies. *J. Chem. Theory Comput.* **2013**, *9*, 3404–3419.

[99] Rupp, M.; Ramakrishnan, R.; von Lilienfeld, O. A. Machine Learning for Quantum Mechanical Properties of Atoms in Molecules. *J. Phys. Chem. Lett.* **2015**, *6*, 3309–3313.

[100] Pipolo, S.; Salanne, M.; Ferlat, G.; Klotz, S.; Saitta, A. \. e.; Pietrucci, F. Navigating at Will on the Water Phase Diagram. *Phys. Rev. Lett.* **2017**, *119*.

[101] Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; Von Lilienfeld, O. A.; Müller, K. R.; Tkatchenko, A. Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *J. Phys. Chem. Lett.* **2015**, *6*, 2326–2331.

[102] Vilhelmsen, L. B.; Hammer, B. Systematic Study of Au 6 to Au 12 Gold Clusters on MgO(100) F Centers Using Density-Functional Theory. *Phys. Rev. Lett.* **2012**, *108*, 126101.

[103] Vilhelmsen, L. B.; Hammer, B. A Genetic Algorithm for First Principles Global Structure Optimization of Supported Nano Structures. *The Journal of Chemical Physics* **2014**, *141*, 044711.

[104] Chen, X.; Jørgensen, M. S.; Li, J.; Hammer, B. Atomic Energies from a Convolutional Neural Network. *J. Chem. Theory Comput.* **2018**, *14*, 3933–3942.

[105] Pietrucci, F.; Andreoni, W. Graph Theory Meets Ab Initio Molecular Dynamics: Atomic Structures and Transformations at the Nanoscale. *Phys. Rev. Lett.* **2011**, *107*, 085504.

[106] Zhu, L.; Amsler, M.; Fuhrer, T.; Schaefer, B.;

Faraji, S.; Rostami, S.; Ghasemi, S. A.; Sadeghi, A.; Grauzinyte, M.; Wolverton, C.; Goedecker, S. A Fingerprint Based Metric for Measuring Similarities of Crystalline Structures. *J. Chem. Phys.* **2016**, *144*, 034203.

[107] Wang, H.; Zhang, L.; Han, J.; E, W. DeePMD-Kit: A Deep Learning Package for Many-Body Potential Energy Representation and Molecular Dynamics. *Computer Physics Communications* **2018**, *228*, 178–184.

[108] Çaylak, O.; von Lilienfeld, A.; Baumeier, B. Wasserstein Metric for Improved Quantum Machine Learning with Adjacency Matrix Representations. *Mach. Learn.: Sci. Technol.* **2020**,

[109] Willatt, M. J.; Musil, F.; Ceriotti, M. Atom-Density Representations for Machine Learning. *J. Chem. Phys.* **2019**, *150*, 154110.

[110] Ischtwan, J.; Collins, M. A. Molecular Potential Energy Surfaces by Interpolation. *The Journal of Chemical Physics* **1994**, *100*, 8080–8088.

[111] Ho, T.-S.; Rabitz, H. A General Method for Constructing Multidimensional Molecular Potential Energy Surfaces from *Ab Initio* Calculations. *The Journal of Chemical Physics* **1996**, *104*, 2584–2597.

[112] Prodan, E.; Kohn, W. Nearsightedness of Electronic Matter. *Proc. Natl. Acad. Sci.* **2005**, *102*, 11635–11638.

[113] Jung, H.; Stocker, S.; Kunkel, C.; Oberhofer, H.; Han, B.; Reuter, K.; Margraf, J. T. Size-Extensive Molecular Machine Learning with Global Representations. *ChemSystemsChem* **2020**, *2*.

[114] Faber, F.; Lindmaa, A.; von Lilienfeld, O. A.; Armiento, R. Crystal Structure Representations for Machine Learning Models of Formation Energies. *Int. J. Quantum Chem.* **2015**, n/a–n/a.

[115] Stuke, A.; Todorović, M.; Rupp, M.; Kunkel, C.; Ghosh, K.; Himanen, L.; Rinke, P. Chemical Diversity in Molecular Orbital Energy Predictions with Kernel Ridge Regression. *J. Chem. Phys.* **2019**, *150*, 204121.

[116] Faber, F. A.; Christensen, A. S.; Huang, B.; Von Lilienfeld, O. A. Alchemical and Structural Distribution Based Representation for Universal Quantum Machine Learning. *J. Chem. Phys.* **2018**, *148*, 241717.

[117] Christensen, A. S.; Bratholm, L. A.; Faber, F. A.; Anatole von Lilienfeld, O. FCHL Revisited: Faster and More Accurate Quantum Machine Learning. *J. Chem. Phys.* **2020**, *152*, 044107.

[118] Ramakrishnan, R.; von Lilienfeld, O. A. Many Molecular Properties from One Kernel in Chemical Space. *Chim. Int. J. Chem.* **2015**, *69*, 182–186.

[119] Barros, K.; Kato, Y. Efficient Langevin Simulation of Coupled Classical Fields and Fermions. *Phys. Rev. B* **2013**, *88*, 235101.

[120] Zheng, S.; Yan, X.; Yang, Y.; Xu, J. Identifying Structure–Property Relationships through SMILES Syntax Analysis with Self-Attention Mechanism. *J. Chem. Inf. Model.* **2019**, *59*, 914–923.

[121] Shin, B.; Park, S.; Kang, K.; Ho, J. C. Self-attention based molecule representation for predicting drug-target interaction. Machine Learning for Healthcare Conference. 2019; pp 230–248.

[122] Boutin, M.; Kemper, G. On Reconstructing N-Point Configurations from the Distribution of Distances or Areas. *Advances in Applied Mathematics* **2004**, *32*, 709–735.

[123] Huang, B.; von Lilienfeld, O. A. Communication: Understanding Molecular Representations in Machine Learning: The Role of Uniqueness and Target Similarity. *The Journal of Chemical Physics* **2016**, *145*, 161102.

[124] Pozdnyakov, S. N.; Willatt, M. J.; Bartók, A. P.; Ortner, C.; Csányi, G.; Ceriotti, M. Incompleteness of Atomic Structure Representations. *Phys. Rev. Lett.* **2020**, *125*, 166001.

[125] Willatt, M. J.; Musil, F.; Ceriotti, M. Feature Optimization for Atomistic Machine Learning Yields a Data-Driven Construction of the Periodic Table of the Elements. *Phys. Chem. Chem. Phys.* **2018**, *20*, 29661–29668.

[126] Drautz, R. Atomic Cluster Expansion for Accurate and Transferable Interatomic Potentials. *Phys. Rev. B* **2019**, *99*, 014104.

[127] Bachmayr, M.; Csanyi, G.; Drautz, R.; Dusson, G.; Etter, S.; van der Oord, C.; Ortner, C. Approximation of Atomic Interactions with Spherical Harmonics. **2019**,

[128] Langer, M. F.; Goeßmann, A.; Rupp, M. Representations of molecules and materials for interpolation of quantum-mechanical simulations via machine learning. *arxiv:2003.12081* **2020**,

[129] To facilitate the use of this notation in LaTeXdocuments, we provide a set of macros at `https://github.com/cosmo-epfl/cosmo-tools/tree/master/tex/dirac-rep`.

[130] Gieres, F. Mathematical Surprises and Dirac's Formalism in Quantum Mechanics. *Rep. Prog. Phys.* **2000**, *63*, 1893–1931.

[131] 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. The notation should be treated with some care when translating it into a practical implementation if the basis used is not orthonormal.

[132] Aronszajn, N. Theory of Reproducing Kernels. *Trans. Amer. Math. Soc.* **1950**, *68*, 337–337.

[133] Nachbin, L. *The Haar integral*; R. E. Krieger Pub. Co., 1976.

[134] Shapeev, A. V. Moment Tensor Potentials: A Class of Systematically Improvable Interatomic Potentials. *Multiscale Model. Simul.* **2016**, *14*, 1153–1173.

[135] Kajita, S.; Ohba, N.; Jinnouchi, R.; Asahi, R. A Universal 3D Voxel Descriptor for Solid-State Material Informatics with Deep Convolutional Neural Networks. *Sci. Rep.* **2017**, *7*, 1–9.

[136] Noh, J.; Kim, J.; Stein, H. S.; Sanchez-Lengeling, B.; Gregoire, J. M.; Aspuru-Guzik, A.; Jung, Y. Inverse Design of Solid-State Materials via a Continuous Representation. *Matter* **2019**, *1*, 1370–1384.

[137] Christiansen, M.-P. V.; Mortensen, H. L.; Meldgaard, S. A.; Hammer, B. Gaussian Representation for Image Recognition and Reinforcement Learning of Atomistic Structure. *J. Chem. Phys.* **2020**, *153*, 044107.

[138] Ziletti, A.; Kumar, D.; Scheffler, M.; Ghir-

inghelli, L. M. Insightful Classification of Crystal Structures Using Deep Learning. *Nat. Commun.* **2018**, *9*, 2775.

[139] Andersen, H. C.; Chandler, D. Optimized Cluster Expansions for Classical Fluids. I. General Theory and Variational Formulation of the Mean Spherical Model and Hard Sphere Percus-Yevick Equations. *J. Chem. Phys.* **1972**, *57*, 1918–1929.

[140] Chandler, D. *Introduction to Modern Statistical Mechanics*; Oxford University Press: New York, 1987.

[141] Behler, J. Atom-Centered Symmetry Functions for Constructing High-Dimensional Neural Network Potentials. *The Journal of Chemical Physics* **2011**, *134*, 074106.

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

[143] Mavračić, J.; Mocanu, F. C.; Deringer, V. L.; Csányi, G.; Elliott, S. R. Similarity Between Amorphous and Crystalline Phases: The Case of TiO_2. *J. Phys. Chem. Lett.* **2018**, *9*, 2985–2990.

[144] Anderson, B.; Hy, T. S.; Kondor, R. Cormorant: Covariant Molecular Neural Networks. NeurIPS. 2019; p 10.

[145] Thompson, W. J. W. J. *Angular momentum : an illustrated guide to rotational symmetries for physical systems*; Wiley-VCH, 2004; p 461.

[146] Kazhdan, M.; Funkhouser, T.; Rusinkiewicz, S. Rotation Invariant Spherical Harmonic Representation of 3D Shape Descriptors. Proceedings of the 2003 Eurographics/ACM SIGGRAPH Symposium on Geometry Processing. Goslar, DEU, 2003; p 156–164.

[147] Thompson, A.; Swiler, L.; Trott, C.; Foiles, S.; Tucker, G. Spectral Neighbor Analysis Method for Automated Generation of Quantum-Accurate Interatomic Potentials. *Journal of Computational Physics* **2015**, *285*, 316–330.

[148] Wood, M. A.; Thompson, A. P. Extending the Accuracy of the SNAP Interatomic Potential Form. *The Journal of Chemical Physics* **2018**, *148*, 241721.

[149] Nigam, J.; Pozdnyakov, S.; Ceriotti, M. Recursive Evaluation and Iterative Contraction of N -Body Equivariant Features. *J. Chem. Phys.* **2020**, *153*, 121101.

[150] Drautz, R. Atomic Cluster Expansion of Scalar, Vectorial, and Tensorial Properties Including Magnetism and Charge Transfer. *Phys. Rev. B* **2020**, *102*, 024104.

[151] Biedenharn, L. C.; Louck, J. D. *The Racah-Wigner Algebra in Quantum Theory*, 1st ed.; Cambridge University Press, 1984.

[152] Behler, J. Neural Network Potential-Energy Surfaces in Chemistry: A Tool for Large-Scale Simulations. *Phys. Chem. Chem. Phys. PCCP* **2011**, *13*, 17930–55.

[153] Handley, C. M.; Popelier, P. L. A. Dynamically Polarizable Water Potential Based on Multipole Moments Trained by Machine Learning. *J. Chem. Theory Comput.* **2009**, *5*, 1474–1489.

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

[155] Liang, C.; Tocci, G.; Wilkins, D. M.; Grisafi, A.; Roke, S.; Ceriotti, M. Solvent Fluctuations and Nuclear Quantum Effects Modulate the Molecular Hyperpolarizability of Water. *Phys. Rev. B* **2017**, *96*, 041407.

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

[157] Grisafi, A.; Wilkins, D. M.; Willatt, M. J.; Ceriotti, M. In *Machine Learning in Chemistry*; Pyzer-Knapp, E. O., Laino, T., Eds.; American Chemical Society: Washington, DC, 2019; Vol. 1326; pp 1–21.

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

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

[160] Raimbault, N.; Grisafi, A.; Ceriotti, M.; Rossi, M. Using Gaussian Process Regression to Simulate the Vibrational Raman Spectra of Molecular Crystals. *New J. Phys.* **2019**, *21*, 105001.

[161] Glielmo, A.; Sollich, P.; De Vita, A. Accurate Interatomic Force Fields via Machine Learning with Covariant Kernels. *Phys. Rev. B* **2017**, *95*, 214302.

[162] Thomas, N.; Smidt, T.; Kearnes, S.; Yang, L.; Li, L.; Kohlhoff, K.; Riley, P. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv:1802.08219* **2018**,

[163] Grisafi, A.; Nigam, J.; Ceriotti, M. Multi-Scale Approach for the Prediction of Atomic Scale Properties. *Chem. Sci.* **2021**, *12*, 2078–2090.

[164] Yu, Q.; Bowman, J. M. Classical, Thermostated Ring Polymer, and Quantum VSCF/VCI Calculations of IR Spectra of $H_7O_3^+$ and $H_9O_4^+$ (Eigen) and Comparison with Experiment. *J. Phys. Chem. A* **2019**, *123*, 1399–1409.

[165] Gastegger, M.; Behler, J.; Marquetand, P. Machine Learning Molecular Dynamics for the Simulation of Infrared Spectra. *Chem. Sci.* **2017**, *8*, 6924–6935.

[166] Veit, M.; Wilkins, D. M.; Yang, Y.; DiStasio, R. A.; Ceriotti, M. Predicting Molecular Dipole Moments by Combining Atomic Partial Charges and Atomic Dipoles. *J. Chem. Phys.* **2020**, *153*, 024113.

[167] Christensen, A. S.; Faber, F. A.; von Lilienfeld, O. A. Operators in Quantum Machine Learning: Response Properties in Chemical Space. *J. Chem. Phys.* **2019**, *150*, 064105.

[168] Zhang, Y.; Ye, S.; Zhang, J.; Hu, C.; Jiang, J.; Jiang, B. Efficient and Accurate Simulations of Vibrational and Electronic Spectra with Symmetry-Preserving Neural Network Models for Tensorial Properties. *J. Phys. Chem. B* **2020**, *124*, 7284–7290.

[169] Batra, R.; Tran, H. D.; Kim, C.; Chapman, J.; Chen, L.; Chandrasekaran, A.; Ramprasad, R. General Atomic Neighborhood Fingerprint for Machine Learning-Based Methods. *J. Phys. Chem. C* **2019**, *123*, 15859–15866.

[170] Zhang, L.; Chen, M.; Wu, X.; Wang, H.; E, W.; Car, R. Deep Neural Network for the Dielectric Response of Insulators. *Phys. Rev. B* **2020**, *102*, 041121.

[171] Burns, L. A.; Faver, J. C.; Zheng, Z.; Marshall, M. S.; Smith, D. G. A.; Vanommeslaeghe, K.; MacKerell, A. D.; Merz, K. M.; Sherrill, C. D. The BioFragment Database (BFDb): An Open-Data Platform for Computational Chemistry Analysis of Noncovalent Interactions. *The Journal of Chemical Physics* **2017**, *147*, 161727.

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

[173] Yue, S.; Muniz, M. C.; Calegari Andrade, M. F.; Zhang, L.; Car, R.; Panagiotopoulos, A. Z. When Do Short-Range Atomistic Machine-Learning Models Fall Short? *J. Chem. Phys.* **2021**, *154*, 034111.

[174] Ambrosetti, A.; Ferri, N.; DiStasio, R. A.; Tkatchenko, A. Wavelike Charge Density Fluctuations and van Der Waals Interactions at the Nanoscale. *Science* **2016**, *351*, 1171–1176.

[175] Medders, G. R.; Babin, V.; Paesani, F. Development of a "First-Principles" Water Potential with Flexible Monomers. III. Liquid Phase Properties. *J. Chem. Theory Comput.* **2014**, *10*, 2906–2910.

[176] Medders, G. R.; Götz, A. W.; Morales, M. A.; Bajaj, P.; Paesani, F. On the Representation of Many-Body Interactions in Water. *J. Chem. Phys.* **2015**, *143*, 104102.

[177] Artrith, N.; Morawietz, T.; Behler, J. High-Dimensional Neural-Network Potentials for Multicomponent Systems: Applications to Zinc Oxide. *Phys. Rev. B* **2011**, *83*, 153101.

[178] Ghasemi, S. A.; Hofstetter, A.; Saha, S.; Goedecker, S. Interatomic Potentials for Ionic Systems with Density Functional Accuracy Based on Charge Densities Obtained by a Neural Network. *Phys. Rev. B* **2015**, *92*, 045131.

[179] Bereau, T.; DiStasio, R. A.; Tkatchenko, A.; von Lilienfeld, O. A. Non-Covalent Interactions across Organic and Biological Subsets of Chemical Space: Physics-Based Potentials Parametrized from Machine Learning. *The Journal of Chemical Physics* **2018**, *148*, 241706.

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

[181] Metcalf, D. P.; Koutsoukas, A.; Spronk, S. A.; Claus, B. L.; Loughney, D. A.; Johnson, S. R.; Cheney, D. L.; Sherrill, C. D. Approaches for Machine Learning Intermolecular Interaction Energies and Application to Energy Components from Symmetry Adapted Perturbation Theory. *J. Chem. Phys.* **2020**, *152*, 074103.

[182] Eickenberg, M.; Exarchakis, G.; Hirn, M.; Mallat, S. Solid Harmonic Wavelet Scattering: Predicting Quantum Molecular Energy from Invariant Descriptors of 3D Electronic Densities. *Adv. Neural Inf. Process. Syst.* **2017**, *2017-Decem*, 6541–6550.

[183] Bartók, A. P.; Csányi, G. Gaussian Approximation Potentials: A Brief Tutorial Introduction. *Int. J. Quantum Chem.* **2015**, *115*, 1051–1057.

[184] Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. International Convention Centre, Sydney, Australia, 2017; pp 1263–1272.

[185] Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K. R. Machine learning of accurate energy-conserving molecular force fields. *Science Advances* **2017**, *3*, e1603015.

[186] Bowman, G. R.; Beauchamp, K. A.; Boxer, G.; Pande, V. S. Progress and Challenges in the Automated Construction of Markov State Models for Full Protein Systems. *J. Chem. Phys.* **2009**, *131*, 124101.

[187] Richard, R. M.; Herbert, J. M. A Generalized Many-Body Expansion and a Unified View of Fragment-Based Methods in Electronic Structure Theory. *The Journal of Chemical Physics* **2012**, *137*, 064113.

[188] Jinnouchi, R.; Karsai, F.; Verdi, C.; Asahi, R.; Kresse, G. Descriptors Representing Two- and Three-Body Atomic Distributions and Their Effects on the Accuracy of Machine-Learned Inter-Atomic Potentials. *J. Chem. Phys.* **2020**, *152*, 234102.

[189] Bader, R. F. W. *Atoms in Molecules: A Quantum Theory*; The International Series of Monographs on Chemistry 22; Clarendon Press ; Oxford University Press: Oxford [England] : New York, 1994.

[190] Deringer, V. L.; Pickard, C. J.; Csányi, G. Data-Driven Learning of Total and Local Energies in Elemental Boron. *Phys. Rev. Lett.* **2018**, *120*, 156001.

[191] Caro, M. A. Optimizing Many-Body Atomic Descriptors for Enhanced Computational Performance of Machine Learning Based Interatomic Potentials. *Phys. Rev. B* **2019**, *100*, 024112.

[192] Natarajan, S. K.; Caro, M. A. Particle Swarm Based Hyper-Parameter Optimization for Machine Learned Interatomic Potentials. *arxiv:2101.00049* **2020**,

[193] Mills, K.; Ryczko, K.; Luchak, I.; Domurad, A.; Beeler, C.; Tamblyn, I. Extensive deep neural networks for transferring small scale learning to large scale systems. *Chemical Science* **2019**, *10*, 4129–4140.

[194] Kondor, R. N-body Networks: a Covariant Hierarchical Neural Network Architecture for Learning Atomic Potentials. **2018**,

[195] Glielmo, A.; Zeni, C.; De Vita, A. Efficient Nonparametric n -Body Force Fields from Machine Learning. *Phys. Rev. B* **2018**, *97*, 184307.

[196] Botu, V.; Ramprasad, R. Learning scheme to predict atomic forces and accelerate materials simulations. *Physical Review B - Condensed Matter and Materials Physics* **2015**, *92*, 094306.

[197] Saunders, C.; Gammerman, A.; Vovk, V. Ridge Regression Learning Algorithm in Dual Variables. *Proceedings of the 15th International Conference on Machine Learning* **1998**, 515–521.

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

[199] Macdonald, I. G. *Symmetric Functions and Hall Polynomials*, reprinted in paperback ed.; Oxford Classic Texts in the Physical Sciences; Clarendon Press: Oxford, 2015.

[200] von Lilienfeld, O. A.; Ramakrishnan, R.; Rupp, M.; Knoll, A. Fourier Series of Atomic Radial Distribution Functions: A Molecular Fingerprint for Machine Learning Models of Quantum Chemical Properties. *Int. J. Quantum Chem.* **2015**, *115*, 1084–1093.

[201] Onat, B.; Ortner, C.; Kermode, J. R. Sensitivity and Dimensionality of Atomic Environment Representations Used for Machine Learning Interatomic Potentials. *J. Chem. Phys.* **2020**, *153*, 144106.

[202] Parsaeifard, B.; De, D. S.; Christensen, A. S.; Faber, F. A.; Kocer, E.; De, S.; Behler, J.; von Lilienfeld, A.; Goedecker, S. An Assessment of the Structural Resolution of Various Fingerprints Commonly Used in Machine Learning. *Mach. Learn.: Sci. Technol.* **2020**,

[203] While the opposite is generally not true, the usual example of the pair of degenerate tetrahedra, as well as any pair of degenerate structures that can be inscribed in a sphere, correspond to a pair of environment placed at the center of such a sphere that is degenerate for $\nu = 2$ correlations.

[204] Yellott, J. I.; Iverson, G. J. Uniqueness properties of higher-order autocorrelation functions. *J. Opt. Soc. Am. A, JOSAA* **1992**, *9*, 388–404.

[205] Kakarala, R. The Bispectrum as a Source of Phase-Sensitive Invariants for Fourier Descriptors: A Group-Theoretic Approach. *J. Math. Imaging Vis.* **2012**, *44*, 341–353.

[206] Kakarala, R. The Bispectrum as a Source of Phase-Sensitive Invariants for Fourier Descriptors: A Group-Theoretic Approach. *J Math Imaging Vis* **2012**, *44*, 341–353.

[207] Uhrin, M. Through the eyes of a descriptor: Constructing complete, invertible, descriptions of atomic environments. *ArXiV e-prints* **2021**, *2104.09319*.

[208] Seko, A.; Togo, A.; Tanaka, I. Group-Theoretical High-Order Rotational Invariants for Structural Representations: Application to Linearized Machine Learning Interatomic Potential. *Phys. Rev. B* **2019**, *99*, 214108.

[209] Shibuta, Y.; Sakane, S.; Takaki, T.; Ohno, M. Submicrometer-Scale Molecular Dynamics Simulation of Nucleation and Solidification from Undercooled Melt: Linkage between Empirical Interpretation and Atomistic Nature. *Acta Materialia* **2016**, *105*, 328–337.

[210] MacQueen, J. Some Methods for Classification and Analysis of Multivariate Observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics. Berkeley, Calif., 1967; pp 281–297.

[211] Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. KDD 1996. 1996.

[212] Caflisch, A. Network and Graph Analyses of Folding Free Energy Surfaces. *Curr. Opin. Struct. Biol.* **2006**, *16*, 71–78.

[213] Ruppert, J.; Welch, W.; Jain, A. N. Automatic Identification and Representation of Protein Binding Sites for Molecular Docking. *Protein Sci.* **2008**, *6*, 524–533.

[214] Gasparotto, P.; Ceriotti, M. Recognizing Molecular Patterns by Machine Learning: An Agnostic Structural Definition of the Hydrogen Bond. *J. Chem. Phys.* **2014**, *141*, 174110.

[215] Rodriguez, A.; Laio, A. Clustering by Fast Search and Find of Density Peaks. *Science* **2014**, *344*, 1492–1496.

[216] Murtagh, F.; Contreras, P. Algorithms for Hierarchical Clustering: An Overview, II. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2017**, *7*, e1219.

[217] Gasparotto, P.; Meißner, R. H.; Ceriotti, M. Recognizing Local and Global Structural Motifs at the Atomic Scale. *J. Chem. Theory Comput.* **2018**, *14*, 486–498.

[218] Pietrucci, F.; Martoňák, R. Systematic Comparison of Crystalline and Amorphous Phases: Charting the Landscape of Water Structures and Transformations. *J. Chem. Phys.* **2015**, *142*, 104704.

[219] Piaggi, P. M.; Parrinello, M. Predicting Polymorphism in Molecular Crystals Using Orientational Entropy. *Proc. Natl. Acad. Sci.* **2018**, *115*, 10251–10256.

[220] Kahle, L.; Marcolongo, A.; Marzari, N. Modeling Lithium-Ion Solid-State Electrolytes with a Pinball Model. *Phys. Rev. Mater.* **2018**, *2*, 065405.

[221] Cheng, B.; Griffiths, R.-R.; Wengert, S.; Kunkel, C.; Stenczel, T.; Zhu, B.; Deringer, V. L.; Bernstein, N.; Margraf, J. T.; Reuter, K.; Csanyi, G. Mapping Materials and Molecules. *Acc. Chem. Res.* **2020**, *53*, 1981–1991.

[222] Coifman, R. R.; Lafon, S.; Lee, A. B.; Maggioni, M.; Nadler, B.; Warner, F.; Zucker, S. W. Geometric Diffusions as a Tool for Harmonic Analysis and Structure Definition of Data: Diffusion Maps. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 7426–7431.

[223] Ferguson, A. L.; Panagiotopoulos, A. Z.; Debenedetti, P. G.; Kevrekidis, I. G. Systematic Determination of Order Parameters for Chain Dynamics Using Diffusion Maps. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 13597–602.

[224] Rohrdanz, M. A.; Zheng, W.; Maggioni, M.; Clementi, C. Determination of Reaction Coordinates via Locally Scaled Diffusion Map. *J. Chem. Phys.* **2011**, *134*, 124116.

[225] Ramprasad, R.; Batra, R.; Pilania, G.; Mannodi-Kanakkithodi, A.; Kim, C. Machine Learning in Materials Informatics: Recent Applications and Prospects. *Npj Comput. Mater.* **2017**, *3*, 54.

[226] Lemke, T.; Peter, C. EncoderMap: Dimensionality Reduction and Generation of Molecule Conformations. *J. Chem. Theory Comput.* **2019**, *15*, 1209–1215.

[227] Wales, D. *Energy Landscapes: Applications to Clusters, Biomolecules and Glasses*; Cambridge University Press, 2003.

[228] Karpen, M. E.; Tobias, D. J.; Brooks, C. L. Statistical Clustering Techniques for the Analysis of Long Molecular Dynamics Trajectories: Analysis of 2.2-Ns Trajectories of YPGDV. *Biochemistry* **1993**, *32*, 412–420.

[229] Torda, A. E.; van Gunsteren, W. F. Algorithms for Clustering Molecular Dynamics Configurations. *J. Comput. Chem.* **1994**, *15*, 1331–1340.

[230] Helfrecht, B. A.; Semino, R.; Pireddu, G.; Auerbach, S. M.; Ceriotti, M. A New Kind of Atlas of Zeolite Building Blocks. *J. Chem. Phys.* **2019**, *151*, 154112.

[231] Ramachandran, G.; Ramakrishnan, C.; Sasisekharan, V. Stereochemistry of Polypeptide Chain Configurations. *Journal of Molecular Biology* **1963**, *7*, 95–99.

[232] Frishman, D.; Argos, P. Incorporation of Non-Local Interactions in Protein Secondary Structure Prediction from the Amino Acid Sequence. *Protein Eng Des Sel* **1996**, *9*, 133–142.

[233] Kabsch, W.; Sander, C. Dictionary of Protein Sec-

ondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* **1983**, *22*, 2577–2637.

[234] Pietrucci, F.; Laio, A. A Collective Variable for the Efficient Exploration of Protein Beta-Sheet Structures: Application to SH3 and GB1. *J. Chem. Theory Comput.* **2009**, *5*, 2197–2201.

[235] Pietropaolo, A.; Branduardi, D.; Bonomi, M.; Parrinello, M. A Chirality-Based Metrics for Free-Energy Calculations in Biomolecular Systems. *J. Comput. Chem.* **2011**, *32*, 2627–2637.

[236] Steinhardt, P. J.; Nelson, D. R.; Ronchetti, M. Bond-Orientational Order in Liquids and Glasses. *Phys. Rev. B* **1983**, *28*, 784–805.

[237] Angioletti-Uberti, S.; Ceriotti, M.; Lee, P. D.; Finnis, M. W. Solid-Liquid Interface Free Energy through Metadynamics Simulations. *Phys. Rev. B - Condens. Matter Mater. Phys.* **2010**, *81*, 125416.

[238] Carignano, M. A.; Saeed, Y.; Aravindh, S. A.; Roqan, I. S.; Even, J.; Katan, C. A Close Examination of the Structure and Dynamics of HC(NH $_2$) $_2$ PbI $_3$ by MD Simulations and Group Theory. *Phys. Chem. Chem. Phys.* **2016**, *18*, 27109–27118.

[239] Oganov, A. R.; Valle, M. How to Quantify Energy Landscapes of Solids. *The Journal of Chemical Physics* **2009**, *130*, 104504.

[240] Valle, M.; Oganov, A. R. Crystal Fingerprint Space – a Novel Paradigm for Studying Crystal-Structure Sets. *Acta Crystallogr A Found Crystallogr* **2010**, *66*, 507–517.

[241] Piaggi, P. M.; Parrinello, M. Entropy Based Fingerprint for Local Crystalline Order. *The Journal of Chemical Physics* **2017**, *147*, 114112.

[242] Martelli, F.; Ko, H.-Y.; Oğuz, E. C.; Car, R. Local-Order Metric for Condensed-Phase Environments. *Phys. Rev. B* **2018**, *97*, 064105.

[243] Mickel, W.; Kapfer, S. C.; Schröder-Turk, G. E.; Mecke, K. Shortcomings of the Bond Orientational Order Parameters for the Analysis of Disordered Particulate Matter. *The Journal of Chemical Physics* **2013**, *138*, 044501.

[244] De, S.; Musil, F.; Ingram, T.; Baldauf, C.; Ceriotti, M. Mapping and Classifying Molecules from a High-Throughput Structural Database. *J. Cheminformatics* **2017**, *9*, 1–14.

[245] Ceriotti, M. Unsupervised Machine Learning in Atomistic Simulations, between Predictions and Understanding. *J. Chem. Phys.* **2019**, *150*, 150901.

[246] Schölkopf, B.; Smola, A.; Müller, K.-R. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Comput.* **1998**, *10*, 1299–1319.

[247] Cuturi, M. Positive Definite Kernels in Machine Learning. *ArXiv Prepr. ArXiv09115367* **2009**,

[248] Oganov, A. R.; Glass, C. W. Crystal Structure Prediction Using Ab Initio Evolutionary Techniques: Principles and Applications. *J. Chem. Phys.* **2006**, *124*, 244704.

[249] Amsler, M.; Goedecker, S. Crystal Structure Prediction Using the Minima Hopping Method. *J. Chem. Phys.* **2010**, *133*, 224104.

[250] Pickard, C. J.; Needs, R. J. Ab Initio Random Structure Searching. *J. Phys. Condens. Matter* **2011**, *23*, 053201.

[251] Curtis, F.; Li, X.; Rose, T.; Vázquez-Mayagoitia, Á.; Bhattacharya, S.; Ghiringhelli, L. M.; Marom, N. GAtor: A First-Principles Genetic Algorithm for Molecular Crystal Structure Prediction. *J. Chem. Theory Comput.* **2018**, *14*, 2246–2264.

[252] Oganov, A. R.; Pickard, C. J.; Zhu, Q.; Needs, R. J. Structure Prediction Drives Materials Discovery. *Nat Rev Mater* **2019**, *4*, 331–348.

[253] Ferré, G.; Maillet, J.-B.; Stoltz, G. Permutation-Invariant Distance between Atomic Configurations. *The Journal of Chemical Physics* **2015**, *143*, 104114.

[254] Montavon, G.; Rupp, M.; Gobre, V.; Vazquez-Mayagoitia, A.; Hansen, K.; Tkatchenko, A.; Müller, K. R.; Anatole Von Lilienfeld, O. Machine Learning of Molecular Electronic Properties in Chemical Compound Space. *New J. Phys.* **2013**, *15*, 095003.

[255] Yang, J.; De, S.; Campbell, J. E.; Li, S.; Ceriotti, M.; Day, G. M. Large-Scale Computational Screening of Molecular Organic Semiconductors Using Crystal Structure Prediction. *Chem. Mater.* **2018**, *30*, 4361–4371.

[256] Helfrecht, B. A.; Cersonsky, R. K.; Fraux, G.; Ceriotti, M. Structure-Property Maps with Kernel Principal Covariates Regression. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 045021.

[257] Kuhn, H. W. The Hungarian Method for the Assignment Problem. *Nav. Res. Logist. Q.* **1955**, *2*, 83–97.

[258] Cuturi, M. In *Advances in Neural Information Processing Systems 26*; Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K. Q., Eds.; Curran Associates, Inc., 2013; pp 2292–2300.

[259] Helfrecht, B. A.; Gasparotto, P.; Giberti, F.; Ceriotti, M. Atomic Motif Recognition in (Bio)Polymers: Benchmarks from the Protein Data Bank. *Front. Mol. Biosci.* **2019**, *6*, 1–14.

[260] de Jong, S.; Kiers, H. A. Principal Covariates Regression. *Chemometrics and Intelligent Laboratory Systems* **1992**, *14*, 155–164.

[261] van der Maaten, L.; Hinton, G. Visualizing Data Using T-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.

[262] Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Sci. Data* **2014**, *1*, 1–7.

[263] Ropo, M.; Schneider, M.; Baldauf, C.; Blum, V. First-Principles Data Set of 45,892 Isolated and Cation-Coordinated Conformers of 20 Proteinogenic Amino Acids. *Sci. Data* **2016**, *3*, 160009.

[264] Schober, C.; Reuter, K.; Oberhofer, H. Virtual Screening for High Carrier Mobility in Organic Semiconductors. *J. Phys. Chem. Lett.* **2016**, *7*, 3973–3977.

[265] Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods VI: More Modifications to the NDDO Approximations and Re-Optimization of Parameters. *J Mol Model* **2013**, *19*, 1–32.

[266] Andrade, C. H.; Pasqualoto, K. F. M.; Ferreira, E. I.; Hopfinger, A. J. 4D-QSAR: Perspectives in Drug Design. *Molecules* **2010**, *15*, 3281–3294.

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

[268] Zuo, Y.; Chen, C.; Li, X.; Deng, Z.; Chen, Y.; Behler, J.; Csányi, G.; Shapeev, A. V.; Thompson, A. P.; Wood, M. A.; Ong, S. P. Performance and Cost Assessment of Machine Learning Interatomic Potentials. *J. Phys. Chem. A* **2020**, acs.jpca.9b08723.

[269] Pozdnyakov, S.; Willatt, M.; Ceriotti, M. Dataset: Randomly-Displaced Methane Configurations. `https://archive.materialscloud.org/record/2020.110`, 2020; (accessed 2020-11-05).

[270] Zamani, M.; Imbalzano, G.; Tappy, N.; Alexander, D. T. L.; Martí-Sánchez, S.; Ghisalberti, L.; Ramasse, Q. M.; Friedl, M.; Tütüncüoglu, G.; Francaviglia, L.; Bienvenue, S.; Hébert, C.; Arbiol, J.; Ceriotti, M.; Fontcuberta I Morral, A. Dataset: 3D Ordering at the Liquid–Solid Polar Interface of Nanowires. `https://archive.materialscloud.org/record/2020.141`, 2020; (accessed 2021-01-07).

[271] Goscinski, A.; Fraux, G.; Imbalzano, G.; Ceriotti, M. The Role of Feature Space in Atomistic Learning. *Mach. Learn.: Sci. Technol.* **2021**, *2*, 025028.

[272] Glielmo, A.; Zeni, C.; Cheng, B.; Csanyi, G.; Laio, A. Ranking the information content of distance measures. *arxiv:2104.15079* **2021**,

[273] Cersonsky, R. K.; Helfrecht, B.; Engel, E. A.; Kliavinek, S.; Ceriotti, M. Improving Sample and Feature Selection with Principal Covariates Regression. *Mach. Learn.: Sci. Technol.* **2021**,

[274] Eldar, Y.; Lindenbaum, M.; Porat, M.; Zeevi, Y. Y. The Farthest Point Strategy for Progressive Image Sampling. *IEEE Trans. Image Process. Publ. IEEE Signal Process. Soc.* **1997**, *6*, 1305–15.

[275] Ceriotti, M.; Tribello, G. A.; Parrinello, M. Demonstrating the Transferability and the Descriptive Power of Sketch-Map. *J. Chem. Theory Comput.* **2013**, *9*, 1521–1532.

[276] Mahoney, M. W.; Drineas, P. CUR Matrix Decompositions for Improved Data Analysis. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 697–702.

[277] Imbalzano, G.; Anelli, A.; Giofré, D.; Klees, S.; Behler, J.; Ceriotti, M. Automatic Selection of Atomic Fingerprints and Reference Configurations for Machine-Learning Potentials. *J. Chem. Phys.* **2018**, *148*, 241730.

[278] Morales, M. A.; Pierleoni, C.; Schwegler, E.; Ceperley, D. M. Evidence for a First-Order Liquid-Liquid Transition in High-Pressure Hydrogen from Ab Initio Simulations. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 12799–12803.

[279] Engel, E. A.; Anelli, A.; Hofstetter, A.; Paruzzo, F.; Emsley, L.; Ceriotti, M. A Bayesian Approach to NMR Crystal Structure Determination. *Phys. Chem. Chem. Phys.* **2019**, *21*, 23385–23400.

[280] Ben Mahmoud, C.; Anelli, A.; Csányi, G.; Ceriotti, M. Learning the Electronic Density of States in Condensed Matter. *Phys. Rev. B* **2020**, *102*, 235130.

[281] Gao, H.; Wang, J.; Sun, J. Improve the Performance of Machine-Learning Potentials by Optimizing Descriptors. *J. Chem. Phys.* **2019**, *150*, 244110.

[282] Faber, F. A.; Lindmaa, A.; Von Lilienfeld, O. A.; Armiento, R. Machine Learning Energies of 2 Million Elpasolite (ABC2D6) Crystals. *Phys. Rev. Lett.* **2016**, *117*, 135502.

[283] Artrith, N.; Urban, A.; Ceder, G. Efficient and Accurate Machine-Learning Interpolation of Atomic Energies in Compositions with Many Species. *Phys. Rev. B* **2017**, *96*, 014112.

[284] Gastegger, M.; Schwiedrzik, L.; Bittermann, M.; Berzsenyi, F.; Marquetand, P. wACSF—Weighted Atom-Centered Symmetry Functions as Descriptors in Machine Learning Potentials. *J. Chem. Phys.* **2018**, *148*, 241709.

[285] Rostami, S.; Amsler, M.; Ghasemi, S. A. Optimized Symmetry Functions for Machine-Learning Interatomic Potentials of Multicomponent Systems. *The Journal of Chemical Physics* **2018**, *149*, 124106.

[286] 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).

[287] Musil, F.; Veit, M.; Goscinski, A.; Fraux, G.; Willatt, M. J.; Stricker, M.; Ceriotti, M. Efficient Implementation of Atom-Density Representations. *J. Chem. Phys.* **2021**, *154*, 114109.

[288] Kamath, A.; Vargas-Hernández, R. A.; Krems, R. V.; Carrington, T.; Manzhos, S. 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* **2018**, *148*, 241702.

[289] Singraber, A.; Morawietz, T.; Behler, J.; Dellago, C. Parallel Multistream Training of High-Dimensional Neural Network Potentials. *J. Chem. Theory Comput.* **2019**, *15*, 3075–3092.

[290] Gao, X.; Ramezanghorbani, F.; Isayev, O.; Smith, J. S.; Roitberg, A. E. TorchANI: A Free and Open Source PyTorch-Based Deep Learning Implementation of the ANI Neural Network Potentials. *Journal of chemical information and modeling* **2020**, *60*, 3408–3415.

[291] Lu, D.; Wang, H.; Chen, M.; Liu, J.; Lin, L.; Car, R.; E, W.; Jia, W.; Zhang, L. 86 PFLOPS Deep Potential Molecular Dynamics simulation of 100 million atoms with ab initio accuracy. **2020**,

[292] Pozdnyakov, S.; Oganov, A. R.; Mazitov, A.; Kruglov, I.; Mazhnik, E. Fast general two- and three-body interatomic potential. *arxiv:1910.07513* **2019**,

[293] Novikov, I. S.; Gubaev, K.; Podryabinkin, E. V.; Shapeev, A. V. The MLIP Package: Moment Tensor Potentials with MPI and Active Learning. *Mach. Learn.: Sci. Technol.* **2021**, *2*, 025002.

[294] Limpanuparb, T.; Milthorpe, J. Associated Legendre Polynomials and Spherical Harmonics Computation for Chemistry Applications. **2014**,

[295] Kaufmann, K.; Baumeister, W. Single-Centre Expansion of Gaussian Basis Functions and the Angular Decomposition of Their Overlap Integrals. *J. Phys. B At. Mol. Opt. Phys.* **1989**, *22*, 1–12.

[296] Himanen, L.; Jäger, M. O. J.; Morooka, E. V.; Federici Canova, F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; Foster, A. S. DScribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications* **2020**, *247*, 106949.

[297] Musil, F.; Veit, M.; Junge, T.; Stricker, M.; Goscinki, A.; Fraux, G.; Ceriotti, M. LIBRASCAL. `https://github.com/cosmo-epfl/librascal`, 2020; `https://github.com/cosmo-epfl/librascal`.

[298] Kermode, J. R.; Bartók, A. P.; Csányi, G. QUIP. `http://www.libatoms.org/`, `http://www.libatoms.org/`.

[299] Hjorth Larsen, A. et al. The Atomic Simulation Environment—a Python Library for Working with Atoms. *J. Phys.: Condens. Matter* **2017**, *29*, 273002.

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

[301] Lejaeghere, K. et al. Reproducibility in Density Functional Theory Calculations of Solids. *Science* **2016**, *351*, aad3000–aad3000.

[302] Yang, Y.; Lao, K.-U.; Wilkins, D. M.; Grisafi, A.; Ceriotti, M. Quantum Mechanical Static Dipole Polarizabilities in the QM7b and AlphaML Showcase Databases. *Sci Data* **2019**, *6*, 152.

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

[304] Wilkins, D. M.; Grisafi, A.; Yang, Y.; Lao, K. U.; A. DiStasio Jr., R.; Ceriotti, M. AlphaML Website. `http://alphaml.org`, 2018; `http://alphaml.org`.

[305] Kapil, V.; Wilkins, D. M.; Lan, J.; Ceriotti, M. Inexpensive Modeling of Quantum Dynamics Using Path Integral Generalized Langevin Equation Thermostats. *J. Chem. Phys.* **2020**, *152*, 124104.

[306] Marsalek, O.; Markland, T. E. Quantum Dynamics and Spectroscopy of Ab Initio Liquid Water: The Interplay of Nuclear and Electronic Quantum Effects. *J. Phys. Chem. Lett.* **2017**, *8*, 1545–1551.

[307] We use $\tilde{\rho}$ and $\tilde{c}$ to refer to electron density and its expansion coefficients, to distinguish them from the similar terms used for the atom density.

[308] Alred, J. M.; Bets, K. V.; Xie, Y.; Yakobson, B. I. Machine Learning Electron Density in Sulfur Crosslinked Carbon Nanotubes. *Composites Science and Technology* **2018**, *166*, 3–9.

[309] Chandrasekaran, A.; Kamal, D.; Batra, R.; Kim, C.; Chen, L.; Ramprasad, R. Solving the Electronic Structure Problem with Machine Learning. *npj Comput Mater* **2019**, *5*, 22.

[310] Whitten, J. L. Coulombic Potential Energy Integrals and Approximations. *The Journal of Chemical Physics* **1973**, *58*, 4496–4501.

[311] Fabrizio, A.; Grisafi, A.; Meyer, B.; Ceriotti, M.; Corminboeuf, C. Electron Density Learning of Non-Covalent Systems. *Chem. Sci.* **2019**, *10*, 9424.

[312] Fabrizio, A.; Briling, K. R.; Girardier, D. D.; Corminboeuf, C. Learning On-Top: Regressing the on-Top Pair Density for Real-Space Visualization of Electron Correlation. *J. Chem. Phys.* **2020**, *153*, 204111.

[313] Geiger, P.; Dellago, C. Neural Networks for Local Structure Detection in Polymorphic Systems. *The Journal of Chemical Physics* **2013**, *139*, 164105.

[314] Anelli, A.; Engel, E. A.; Pickard, C. J.; Ceriotti, M. Generalized Convex Hull Construction for Materials Discovery. *Phys. Rev. Mater.* **2018**, *2*, 103804.

[315] Engel, E. A.; Anelli, A.; Ceriotti, M.; Pickard, C. J.; Needs, R. J. Mapping Uncharted Territory in Ice from Zeolite Networks to Ice Structures. *Nat. Commun.* **2018**, *9*, 2173.

[316] Gallet, G. A.; Pietrucci, F. Structural Cluster Analysis of Chemical Reactions in Solution. *J Chem Phys* **2013**, *139*, 74101.

[317] Rowe, P.; Deringer, V. L.; Gasparotto, P.; Csányi, G.; Michaelides, A. An Accurate and Transferable Machine Learning Potential for Carbon. *J. Chem. Phys.* **2020**, *153*, 034702.

[318] Maksimov, D.; Baldauf, C.; Rossi, M. The Conformational Space of a Flexible Amino Acid at Metallic Surfaces. *Int J Quantum Chem* **2020**,

[319] Musil, F.; De, S.; Yang, J.; Campbell, J. E.; Day, G. M.; Ceriotti, M. Machine Learning for the Structure-Energy-Property Landscapes of Molecular Crystals. *Chem. Sci.* **2018**, *9*, 1289–1300.

[320] Gryn'ova, G.; Lin, K.-H.; Corminboeuf, C. Read between the Molecules: Computational Insights into Organic Semiconductors. *J. Am. Chem. Soc.* **2018**, *140*, 16370–16386.

[321] Yang, J. Mapping Temperature-Dependent Energy–Structure–Property Relationships for Solid Solutions of Inorganic Halide Perovskites. *J. Mater. Chem. C* **2020**, 10.1039.D0TC04515B.

[322] Würger, T.; Feiler, C.; Musil, F.; Feldbauer, G. B. V.; Höche, D.; Lamaka, S. V.; Zheludkevich, M. L.; Meißner, R. H. Data Science Based Mg Corrosion Engineering. *Front. Mater.* **2019**, *6*, 53.

[323] Sharp, T. A.; Thomas, S. L.; Cubuk, E. D.; Schoenholz, S. S.; Srolovitz, D. J.; Liu, A. J. Machine Learning Determination of Atomic Dynamics at Grain Boundaries. *Proc Natl Acad Sci USA* **2018**, *115*, 10943–10947.

[324] Priedeman, J. L.; Rosenbrock, C. W.; Johnson, O. K.; Homer, E. R. Quantifying and Connecting Atomic and Crystallographic Grain Boundary Structure Using Local Environment Representation and Dimensionality Reduction Techniques. *Acta Materialia* **2018**, *161*, 431–443.

[325] Homer, E. R.; Hensley, D. M.; Rosenbrock, C. W.; Nguyen, A. H.; Hart, G. L. W. Machine-Learning Informed Representations for Grain Boundary Structures. *Front. Mater.* **2019**, *6*, 168.

[326] Gasparotto, P.; Bochicchio, D.; Ceriotti, M.; Pavan, G. M. Identifying and Tracking Defects in Dynamic Supramolecular Polymers. *J. Phys. Chem. B* **2020**, *124*, 589–599.

[327] Capelli, R.; Gardin, A.; Empereur-mot, C.; Doni, G.; Pavan, G. M. A Data-Driven Dimensionality Reduction Approach to Compare and Classify Lipid Force Fields. *ChemRxiv:14039834.v3* **2021**,

[328] Schwalbe-Koda, D.; Jensen, Z.; Olivetti, E.; Gómez-Bombarelli, R. Graph Similarity Drives Zeolite Diffusionless Transformations and Intergrowth. *Nat. Mater.* **2019**, *18*, 1177–1181.

[329] Nicholas, T. C.; Goodwin, A. L.; Deringer, V. L. Understanding the Geometric Diversity of Inorganic and Hybrid Frameworks through Structural Coarse-Graining. *Chem. Sci.* **2020**, *11*, 12580–12587.

[330] Dietz, C.; Kretz, T.; Thoma, M. H. Machine-Learning Approach for Local Classification of Crystalline Structures in Multiphase Systems. *Phys. Rev. E* **2017**, *96*, 011301.

[331] Fulford, M.; Salvalaglio, M.; Molteni, C. DeepIce: A Deep Neural Network Approach To Identify Ice and Water Molecules. *J. Chem. Inf. Model.* **2019**, *59*,

2141–2149.

[332] Deringer, V. L.; Caro, M. A.; Jana, R.; Aarva, A.; Elliott, S. R.; Laurila, T.; Csányi, G.; Pastewka, L. Computational Surface Chemistry of Tetrahedral Amorphous Carbon by Combining Machine Learning and Density Functional Theory. *Chem. Mater.* **2018**, *30*, 7438–7445.

[333] Zhou, Y.; Sun, L.; Zewdie, G. M.; Mazzarello, R.; Deringer, V. L.; Ma, E.; Zhang, W. Bonding Similarities and Differences between Y–Sb–Te and Sc–Sb–Te Phase-Change Memory Materials. *J. Mater. Chem. C* **2020**, *8*, 3646–3654.

[334] Huang, J.-X.; Csányi, G.; Zhao, J.-B.; Cheng, J.; Deringer, V. L. First-Principles Study of Alkali-Metal Intercalation in Disordered Carbon Anode Materials. *J. Mater. Chem. A* **2019**, *7*, 19070–19080.

[335] Caro, M. A.; Csányi, G.; Laurila, T.; Deringer, V. L. Machine Learning Driven Simulated Deposition of Carbon Films: From Low-Density to Diamondlike Amorphous Carbon. *Phys. Rev. B* **2020**, *102*, 174201.

[336] Reinhardt, A.; Pickard, C. J.; Cheng, B. Predicting the Phase Diagram of Titanium Dioxide with Random Search and Pattern Recognition. *Phys. Chem. Chem. Phys.* **2020**, *22*, 12697–12705.

[337] Basdogan, Y.; Groenenboom, M. C.; Henderson, E.; De, S.; Rempe, S. B.; Keith, J. A. Machine Learning-Guided Approach for Studying Solvation Environments. *J. Chem. Theory Comput.* **2020**, *16*, 633–642.

[338] Bernstein, N.; Csányi, G.; Deringer, V. L. De Novo Exploration and Self-Guided Learning of Potential-Energy Surfaces. *npj Comput Mater* **2019**, *5*, 99.

[339] Monserrat, B.; Brandenburg, J. G.; Engel, E. A.; Cheng, B. Liquid Water Contains the Building Blocks of Diverse Ice Phases. *Nat Commun* **2020**, *11*, 5757.

[340] Deringer, V. L.; Caro, M. A.; Csányi, G. A General-Purpose Machine-Learning Force Field for Bulk and Nanostructured Phosphorus. *Nat Commun* **2020**, *11*, 5461.

[341] Shen, C.; Ding, J.; Wang, Z.; Cao, D.; Ding, X.; Hou, T. From machine learning to deep learning: Advances in scoring functions for protein–ligand docking. *WIREs Computational Molecular Science* **2020**, *10*.

[342] Dastmalchi, S.; Hamzeh-Mivehroud, M.; Asadpour-Zeynali, K. Comparison of different 2D and 3D-QSAR methods on activity prediction of histamine H3 receptor antagonists. *Iranian Journal of Pharmaceutical Research* **2012**, *11*, 97–108.

[343] Jagiello, K.; Grzonkowska, M.; Swirog, M.; Ahmed, L.; Rasulev, B.; Avramopoulos, A.; Papadopoulos, M. G.; Leszczynski, J.; Puzyn, T. Advantages and limitations of classic and 3D QSAR approaches in nano-QSAR studies based on biological activity of fullerene derivatives. *Journal of Nanoparticle Research* **2016**, *18*, 256.

[344] Choudhary, K.; DeCost, B.; Tavazza, F. Machine Learning with Force-Field-Inspired Descriptors for Materials: Fast Screening and Mapping Energy Landscape. *Phys. Rev. Materials* **2018**, *2*, 083801.

[345] Kuz'min, V. E.; Artemenko, A. G.; Polischuk, P. G.; Muratov, E. N.; Hromov, A. I.; Liahovskiy, A. V.; Andronati, S. A.; Makan, S. Y. Hierarchic system of QSAR models (1D-4D) on the base of simplex representation of molecular structure. *Journal of Molecular Modeling* **2005**, *11*, 457–467.

[346] Zankov, D.; V.,; Matveieva, M.; Nikonenko, A.; Nugmanov, R.; Varnek, A.; Polishchuk, P.; Madzhidov, T. QSAR Modeling Based on Conformation Ensembles Using a Multi-Instance Learning Approach. **2020**,

[347] Weinreich, J.; Browning, N. J.; von Lilienfeld, O. A. Machine Learning of Free Energies in Chemical Compound Space Using Ensemble Representations: Reaching Experimental Uncertainty for Solvation. *J. Chem. Phys.* **2021**, *154*, 134113.

[348] Axelrod, S.; Gomez-Bombarelli, R. Molecular machine learning with conformer ensembles. **2020**,

[349] Gallarati, S.; Fabregat, R.; Laplaza, R.; Bhattacharjee, S.; Wodrich, M. D.; Corminboeuf, C. Reaction-Based Machine Learning Representations for Predicting the Enantioselectivity of Organocatalysts. *Chem. Sci.* **2021**, 10.1039.D1SC00482D.

[350] Jinnouchi, R.; Asahi, R. Predicting Catalytic Activity of Nanoparticles by a DFT-Aided Machine-Learning Algorithm. *J. Phys. Chem. Lett.* **2017**, *8*, 4279–4283.

[351] Caro, M. A.; Aarva, A.; Deringer, V. L.; Csányi, G.; Laurila, T. Reactivity of Amorphous Carbon Surfaces: Rationalizing the Role of Structural Motifs in Functionalization Using Machine Learning. *Chem. Mater.* **2018**, *30*, 7446–7455.

[352] Aarva, A.; Deringer, V. L.; Sainio, S.; Laurila, T.; Caro, M. A. Understanding X-Ray Spectroscopy of Carbonaceous Materials by Combining Experiments, Density Functional Theory, and Machine Learning. Part II: Quantitative Fitting of Spectra. *Chem. Mater.* **2019**, *31*, 9256–9267.

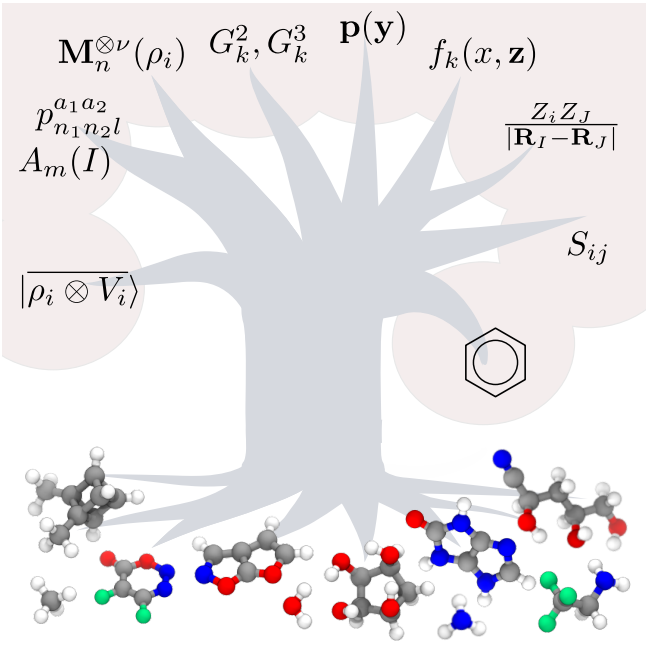
[353] Dick, S.; Fernandez-Serra, M. Learning from the Density to Correct Total Energy and Forces in First Principle Simulations. *J. Chem. Phys.* **2019**, *151*, 144102.

[354] Fung, V.; Hu, G.; Ganesh, P.; Sumpter, B. G. Machine Learned Features from Density of States for Accurate Adsorption Energy Prediction. *Nat Commun* **2021**, *12*, 88.

[355] Welborn, M.; Cheng, L.; Miller, T. F. Transferability in Machine Learning for Electronic Structure via the Molecular Orbital Basis. *J. Chem. Theory Comput.* **2018**, *14*, 4772–4779.

[356] Cheng, L.; Welborn, M.; Christensen, A. S.; Miller, T. F. A Universal Density Matrix Functional from Molecular Orbital-Based Machine Learning: Transferability across Organic Molecules. *J. Chem. Phys.* **2019**, *150*, 131103.

[357] Qiao, Z.; Welborn, M.; Anandkumar, A.; Manby, F. R.; Miller, T. F. OrbNet: Deep Learning for Quantum Chemistry Using Symmetry-Adapted Atomic-Orbital Features. *J. Chem. Phys.* **2020**, *153*, 124111.



TOC GRAPHICS