CHEMICAL REVIEWS

pubs.acs.org/CR

Review

Physics-Inspired Structural Representations for Molecules and Materials

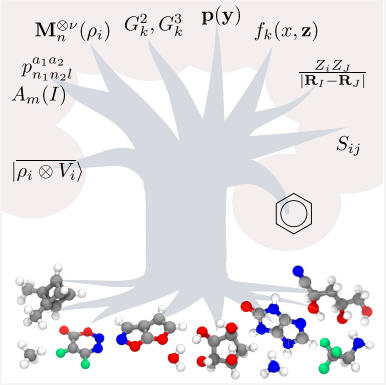
Felix Musil, Andrea Grisafi, Albert P. Bartók, Christoph Ortner, Gábor Csányi, and Michele Ceriotti*

Cite This: *Chem. Rev.* 2021, 121, 9759–9815

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: 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

Special Issue: Machine Learning at the Atomic Scale

Received: January 8, 2021
Published: July 26, 2021

1. INTRODUCTION

The past 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 dataset 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, and 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 we 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 realization, 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 $k(A, A')$ (e.g., the scalar product between feature vectors) or in terms of a distance between configurations $d(A, A')$ (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 is 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 multiscale nature of interactions, and 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). More complex, nonlinear 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

https://doi.org/10.1021/acs.chemrev.1c00021
Chem. Rev. 2021, 121, 9759−9815

dedicated to understanding and improving their properties. These efforts follow several directions. First, the scalable and parallel implementation of the construction of a given set of features is essential to ensure computational efficiency. Second, limiting the number of features that are 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 thematic 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.

2. 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 relationship (QSPR) techniques, attempting to map[40] descriptors of molecular structure—based on chemoinformatics 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 describe 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] have 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 became 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 predictions 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 6. 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 8.7.

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

https://doi.org/10.1021/acs.chemrev.1c00021
Chem. Rev. 2021, 121, 9759−9815

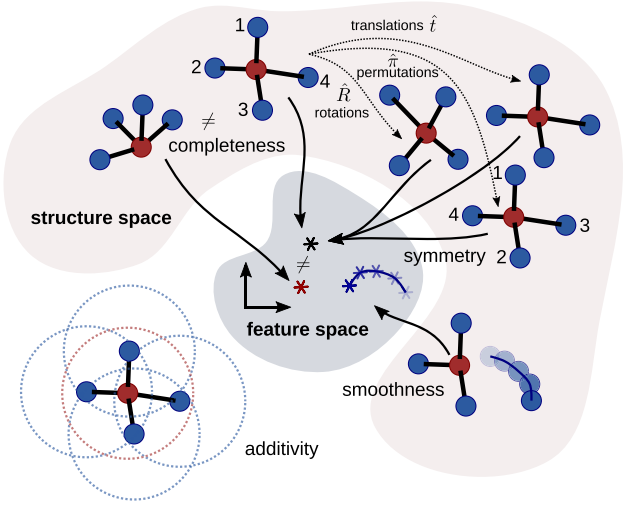


Figure 1. Schematic overview of the requirements for an effective structural representation. The mapping between structures and feature space should obey fundamental physical symmetries (equivalent structures should be mapped to the same features), should be complete (inequivalent structures should be mapped to distinct features), and 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.

system or the labeling of identical atoms is modified; be smooth, so that models built on the features inherit a regular behavior with changing atomic coordinates; and be complete, so that fundamentally distinct configurations are never mapped to the same set of features. Furthermore, many machine-learning tasks benefit greatly from being based on *local* features, which describe atoms or groups of atoms. Even though this is a less stringent requirement and, as we discuss below, global descriptors have been used very successfully, representations based on local environments are usually associated with higher transferability, reflecting a "divide and conquer" approach to materials modeling.[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 7.

2.1. 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 as 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

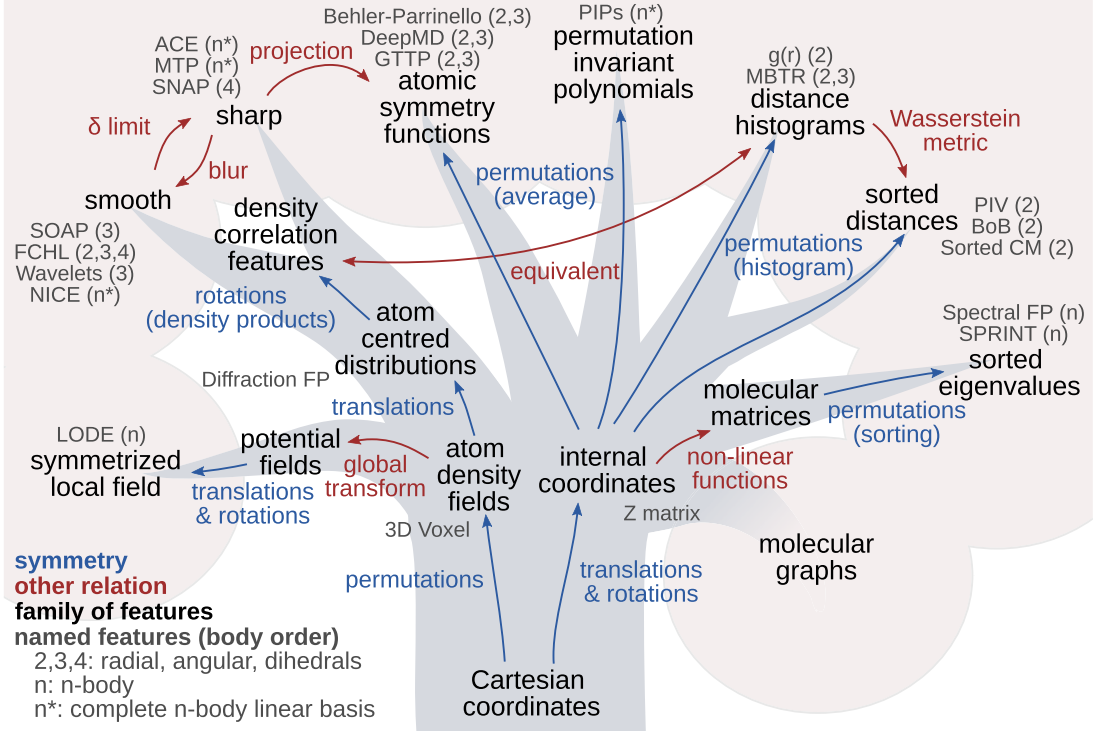


Figure 2. Phylogenetic tree of structural representations for materials and molecules. Arrows indicate the relationship between different groups of features. Lists of names, in gray, indicate the most common implementations for each class. Classes that appear as "leaves" of the tree are fully symmetric.

https://doi.org/10.1021/acs.chemrev.1c00021
Chem. Rev. 2021, 121, 9759−9815

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, and torsions) have been proposed, based on chemical intuition, as invariant descriptors of molecular geometry, and most of the molecular force fields 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 of 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 (ACSFs), which rely on a local expansion of the energy and on the construction of a symmetric description of atomic environments. Similarly to PIPs, ACSFs 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 ACSFs is kept under control by restricting the range of interactions (which we discuss further in section 2.3) 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 section 2.2.

The phylogenetic tree in Figure 2 shows that a large number of existing representations follow 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 does one proceed to symmetrize it over translations and rotations. We discuss in great detail this second approach in section 3. 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.

2.2. 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 3. Internal coordinates are also usually smooth, but the process of manipulating them to achieve a

https://doi.org/10.1021/acs.chemrev.1c00021
Chem. Rev. 2021, 121, 9759−9815

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] and "bag of bonds" (BoB) features.[101] Similar descriptors based on sorted distances have also been used to identify recurring structures in geometry 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 as 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

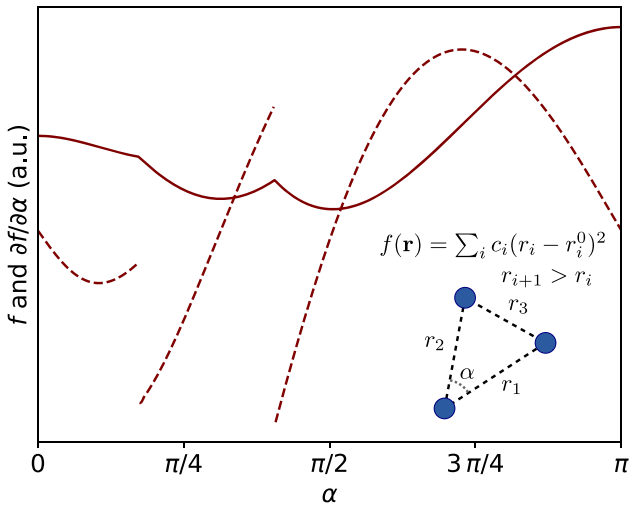


Figure 3. Toy model demonstrating a non-smooth property (solid line) and its discontinuous derivative (dashed line) that are defined as functions of the ordered list of interatomic distances for a three-atom cluster.

a system of three atoms that is uniquely defined by the three interatomic distances r_i, where the index i denotes the position of the interatomic distance r_i in the ordered list of distances. We define a smooth function of the sorted distances, $f = \sum_i c_i (r_i - r_i^0)^2$ 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.

2.3. 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 designed 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] not only had 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 section 7.3. 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 5.2. 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.

2.4. Completeness

The requirements of symmetry, smoothness, and locality can be seen as geared toward reducing the complexity of the

https://doi.org/10.1021/acs.chemrev.1c00021
Chem. Rev. 2021, 121, 9759−9815

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 5.2 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, nonlinear models can often provide good results even when using a representation that involves excessive smoothing or a highly truncated version of a family of features. The interplay between model and features is discussed in more detail in section 4, and the (largely open) problem of completeness, in section 5.

3. 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 hinting to the connection between this class of features and a linear mapping between structure and properties, which is discussed in more detail in section 4.

3.1. Dirac Notation for Atomic Representations

We formalize a notation that extends the one introduced in refs 109 and 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 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 representations but occasionally make a link to the different notations that have become established to describe specific frameworks.

3.1.1. Representations in Bra-Ket Notation. We use a ket $|A\rangle$ to indicate an abstract feature vector associated 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 Figure 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.,

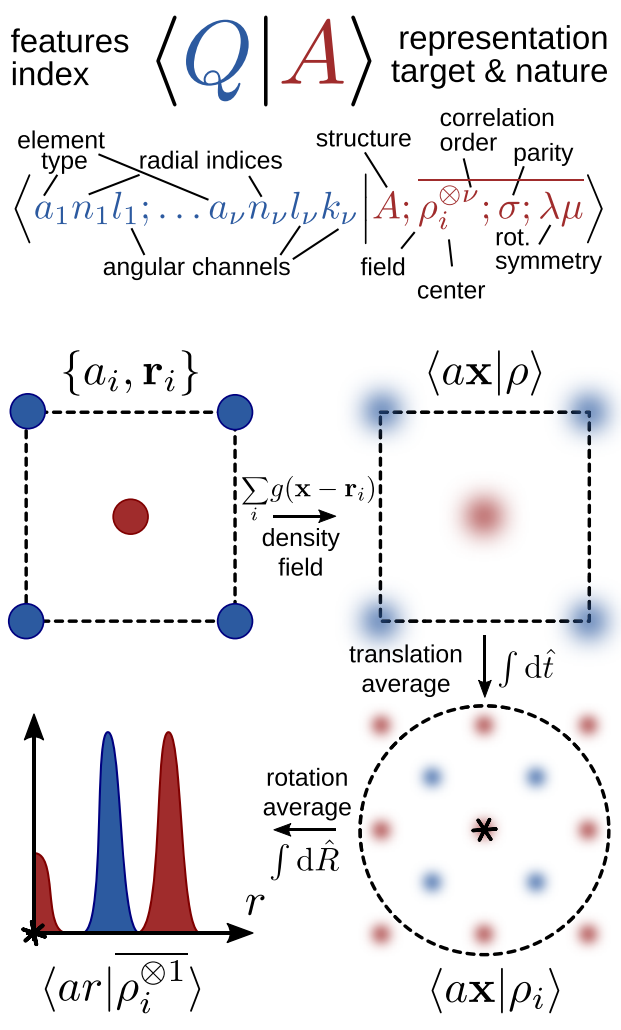


Figure 4. Top: overview of the notation we use to indicate the features that represent an atomistic structure. Bottom: summary of the steps in a symmetrized field construction.

https://doi.org/10.1021/acs.chemrev.1c00021
Chem. Rev. 2021, 121, 9759−9815