Received: 7 August 2021 | Revised: 27 October 2021 | Accepted: 3 December 2021

DOI: 10.1002/qua.26870

REVIEW

International Journal of QUANTUM CHEMISTRY WILEY

Molecular representations for machine learning applications in chemistry

Shampa Raghunathan[1] | U. Deva Priyakumar[2]

[1]École Centrale School of Engineering, Mahindra University, Hyderabad, India

[2]Center for Computational Natural Sciences and Bioinformatics, International Institute of Information Technology, Hyderabad, India

Correspondence
Shampa Raghunathan, École Centrale School of Engineering, Mahindra University, Hyderabad 500043, India.
Email: shampa.raghunathan@mahindrauniversity.edu.in

U. Deva Priyakumar, Center for Computational Natural Sciences and Bioinformatics, International Institute of Information Technology, Hyderabad 500032, India.
Email: deva@iiit.ac.in

Funding information
Department of Science and Technology, Ministry of Science and Technology, India, Grant/Award Number: SR/WOS-A/CS-19/2018 (G); IHub-Data, IIIT Hyderabad

Abstract

Machine learning (ML) methods enable computers to address problems by learning from existing data. Such applications are becoming commonplace in molecular sciences. Interest in applying ML techniques across chemical compound space, from predicting properties to designing molecules and materials is in the surge. Especially, ML models have started to accelerate computational chemistry, and are often as accurate as state-of-the-art electronic/atomistic models. Being an integral part of the ML architecture, representation of a molecular entity, uniquely encoded, plays a crucial role to what extent an ML model would be accurately predicting the desired property. This review aims to demonstrate a hierarchy of representations which has been introduced, to capture all degrees of freedom of a molecule or an atom the best, to map the quantum mechanical properties. We discuss their diverse applications how they have been instrumental in harnessing the growing field of ML accelerated computational modeling.

KEYWORDS
Kernel methods, machine learning potential, molecular descriptor, molecular featurization

1 | INTRODUCTION

Every single facet of information science is rapidly becoming data-intensive, ranging from technology, business, public administration, and so on. The volume and diversity of the types and sources of these data raise new opportunities, but also bring new challenges. For example, automated pattern-recognition algorithms could lead to relevant valuable knowledge and, change how we develop treatments, classify patients or study diseases [1]. In case of quantitative structure–activity and structure–property relationship (QSAR/QSPR), using a computational model, once a satisfactory correlation between structure/activity is established any number of chemical compounds, including those not yet characterized in vivo can be readily screened in silico, and identify a suitable candidate with the desired property [2,3]. However, researchers often face challenges regarding quantity and quality of the data, even when they are well annotated/labeled so that algorithm can distinguish features and categorize pattern.

Machine learning (ML) methods consisting of sophisticated algorithms assist computers with the ability to learn functional relationships from raw data not necessarily providing them a priori [4,5]. For decades now, conventional ML model required careful engineering and significant domain expertise to design a feature extractor which would transform the raw data into a suitable representation or feature vector, from which, the learning subsystem would be able to detect the input pattern. Representation learning uses raw data and automatically discover the representations required for detection or classification [6]. Finally, deep learning (DL) models composing of representation-learning algorithms with multiple levels of representations, starting with simple objects like, edges, are combined to form complex objects like, faces [7]. DL models have been gaining superiority over existing state-of-the-art ML algorithms across several fields. In addition to revolutionizing image classification and speech recognition, DL methods have shown major advances in fields, as diverse as, high-energy physics [8], computational chemistry [9–11], biology [1,12], medicines [13], translation among written languages [14], and so forth.

Within this novel paradigm, the choice of data representation in a machine-readable form (so-called descriptors) is highly problem-specific. ML methods are often employed to predict the activities of chemical substances. A molecular structure can be represented in terms of a labeled

graph with nodes corresponding to atoms, and edges corresponding to bonds between these atoms [15]. To obtain a significant structure/activity correlation, it is crucial that appropriate descriptors must be employed, whether, they are theoretical, empirical or derived from readily available experimental characteristics of the structures [3,16]. For example, chemical descriptors based on elemental properties (for instance, atomic charges, electronegativity and ionization potentials in a compound) have been extensively used in the chemical and biological engineering domain to develop QSAR/QSPR model of compounds [17,18]. Similarly, molecular fingerprints (FP) [19] have been extensively applied to drug discovery, virtual screening and compound similarity search. ML models applied on data calculated using the quantum mechanical (QM) approaches have been actively contributing toward the accelerated discovery of chemicals with desired properties. Here, mostly molecules were represented as Coulomb matrices [20] and their variants [21,22], bag-of-bonds [23], fingerprints [24], smooth overlap of atomic positions [25], atom-centered symmetry functions (ACSFs) [26], interatomic many body expansions (MBE) [27–29], and so forth.

The more suitable the representation of the input data, the more accurately can one algorithm map it to the output data. Selecting how best to represent the data could require insight into both the underlying scientific problem and the operation of the learning algorithm, because it is not always obvious which choice of representation will give the best performance; this is an active topic of research for chemical systems [30,31]. For example, molecular graph in terms of dihedral angles might be advantageous over the pairwise distances while modeling a complex molecular transition process. Also a descriptor should be able to capture existing physical invariances which are basically prior knowledges of the task at hand. For example, reconstructing a potential energy surface (PES) smoothness assumption of a descriptor is desired in order to find the underlying regularity between various input structures. Basically similar inputs should result in similar outputs. So that we can extrapolate for the new sample from a limited number of training samples. In case of materials field, instead of the total energy of a system we are more interested in a unique local functional of single atomic energy or a bond for a given neighbor environment. While constructing a many-body descriptor with interatomic potentials, such as, bonds, angles, dihedrals addition of higher order contributions systematically should increase the similarity to the true potential energy of a system.

This is a review work mainly focused on molecular representations and their usage as an integrated part of ML models. This is an active research field growing and becoming voluminous daily. A review like this, cannot capture the entire extent of available computational applications in every facet of molecular sciences that utilizes molecular representation. Instead, here we present recent applications of ML models those made use of efficient molecular descriptors for advancing computational modeling of molecules and materials. The next section gives a very brief overview of ML and DL models. The Section 3 focuses on various popularly used molecular featurization techniques. The Section 4 discusses their applications and performances. The Section 5 summarizes with a brief outlook that discusses a few computationally challenging molecular problems which can possibly be taken up for future ML research.

2 | AN OVERVIEW OF ML AND DL MODELS

Understanding, how a specific problem is being addressed employing the ML approach requires an introduction of the machinery involved within. A canonical ML workflow typically involves the following steps: (i) data cleaning and preprocessing, (ii) feature extraction, (iii) building a model (training, validating and testing). In case of supervised learning, data must be labeled. Conventionally, one data sample is denoted as input x (usually a vector of numbers), and labeled with its property of interest, generally referred to as the target or output value y (usually a scalar). A supervised ML model aims to learn a function $f(x) = y$, from a list of training pairs (x_1, y_1), (x_2, y_2), … for which data are recorded. One typical application in chemistry is to predict QM properties of equilibrium molecular structures. The input feature (x) nuclear charges and positions $\{Z_I, \mathbf{R}_I\}$, which together with the property **P** (output label y) can be used to train a regression model (functional relationship f). Now, given a new $\{Z_I^*, \mathbf{R}_I^*\}$ (x^*), the learned function predicts its property (output label y^*) by calculating $f(x^*)$. While unsupervised ML approaches aim to discover patterns from the data samples x themselves, without the need for output labels y. Methods, such as, clustering, K-means clustering, principal component analysis and outliers detection are typical examples of unsupervised models [32].

A conventional deep neural network (DNN) inputs the raw data (x) at the lowest layer and transforms them into a representation at a higher and slightly more abstract level by composing outputs from the preceding layer, thus a complex function is learned in the process. For example, a DL model consisting of n no. of functions $f^{(i)}(x)$, where $i \in n$ determines the *depth* of the model. Consequently, the terminology *deep* NN was turned out.

3 | MOLECULAR FEATURIZATION

A molecular representation, also known as "descriptor" of feature vectors, encodes chemical identity of a molecular entity in terms of its chemical composition and atomic configuration. Only after the chemical identity is converted into a descriptor (an array of numbers), computer can efficiently process a large number of structures. Descriptor is a crucial ingredient for the development of predictive ML models over the input → output mapping surface. There are three key invariances of physical systems (schematics shown in Figure 1) those are preferably captured by a common molecular descriptor. Formally, these are as follows:

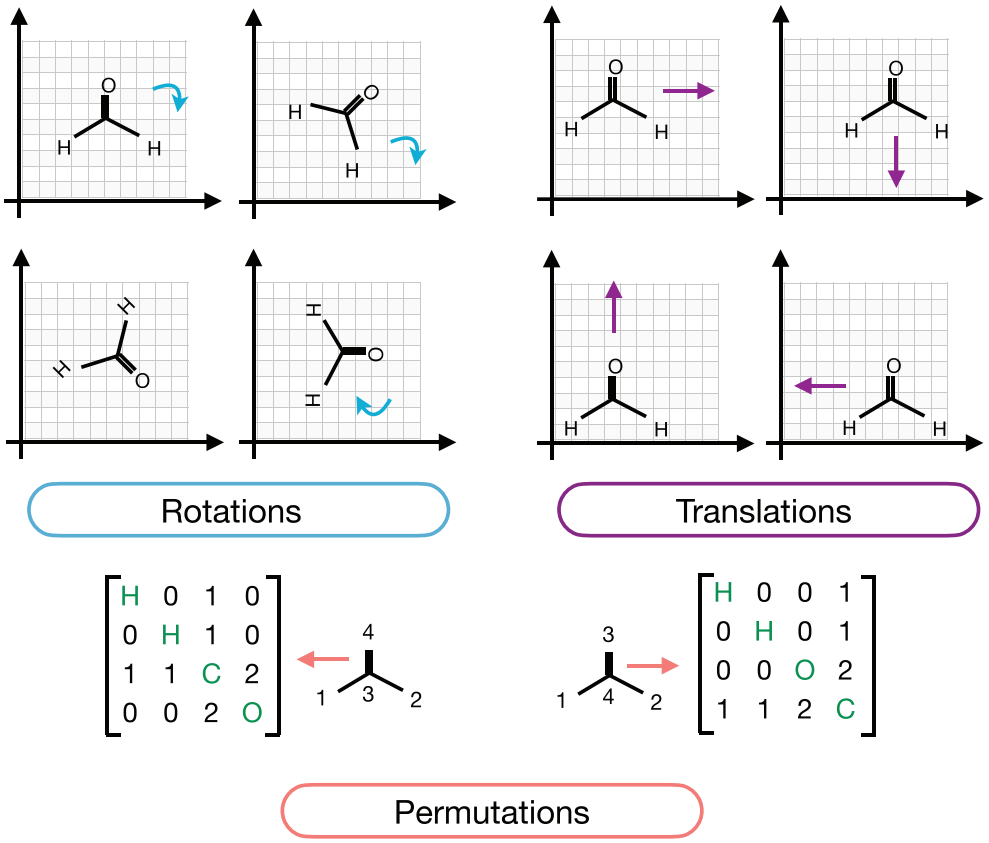


FIGURE 1 Formaldehyde molecule undergoing rotation and translation operations in a fixed reference grid. Permutation operation showing two equivalent formaldehyde representations. The atom connectivity matrix introduced by Spialter has atomic symbols (alternatively, atomic numbers) listed along the diagonal [33]. The off-diagonals represent interatomic connectivities, for example, bond order values 0, 1, and 2 for none, single and double bonds, respectively

1. Rotational invariance: Representation must be invariant upon a rotation operation.
2. Translational invariance: Representation must be unchanged upon a translation in space.
3. Permutation invariance: Representation must be unaltered due to the change in the particular order of the atoms.

We will start this section on how to represent (computationally) a chemical structure considering only the chemical bonding between atoms rather than their three-dimensional structure.

3.1 | Graph representation

Chemical structures are popularly represented as molecular graphs [34,35]. In mathematics, a graph can be an abstract structure consisting of nodes and connected by edges. In a molecular graph, atoms can be nodes and bonds can be edges (Figure 2A); often hydrogens are omitted. The nodes and edges have properties, for instance, atomic number or atom type may correspond to each node whereas, bond order to each edge. In a protein, the local environment of a central amino acid residue can be thought of as a graph consisting of residues as nodes and residue–residue interactions as edges [36].

3.2 | Simplified molecular input line entry system representation

There should exist means by which molecular graphs can be communicated to and from computers. Often connectivity table is used to do so. Alternatively, a linear notation—alphanumeric characters is used to encode the molecular structure. A linear notation that has found widespread acceptance is the simplified molecular input line entry system (SMILES) notation [37,38]. In SMILES, atomic symbols

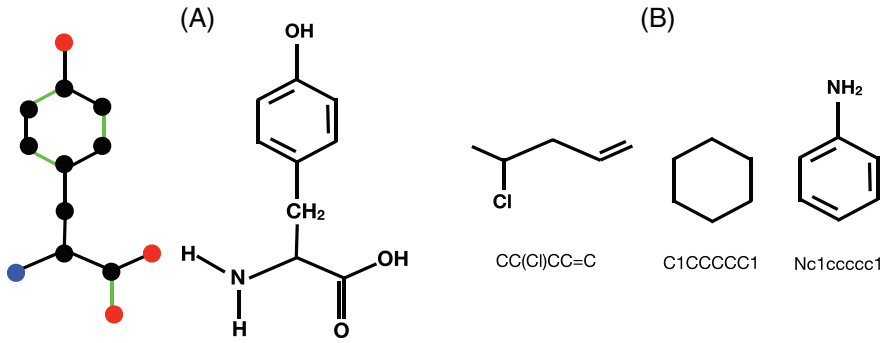
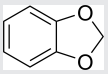


FIGURE 2 (A) Graph representation of a chemical structure, tyrosine. Carbon, nitrogen and oxygen atoms are colored in black, blue, and red circles, respectively. Single and double bonds are marked in black and green colors, respectively. (B) Simplified molecular input line entry system (SMILES) representation of 4-chloro-1-pentene, cyclohexane and aniline

TABLE 1 Different sequence-based molecular representations for 1,3-benzodioxole

2D structure	
IUPAC	1,3-Benzodioxole
SMILES	c1ccc2c(c1)OCO2
Canonical SMILES	c2ccc1OCOc1c2
InChI	InChI = 1S/C7H6O2/c1-2-4-7-6(3-1)8-5-9-7/h1-4H,5H2
InChIKey	InChIKey = FTNJQNQLEGKTGD-UHFFFAOYSA-N

Abbreviations: IUPAC, International Union of Pure and Applied Chemistry; SMILES, simplified molecular input line entry system.

are used to represent atoms. Aliphatic and aromatic atoms are represented by upper and lower case symbols, respectively. Normally hydrogen atoms are not explicitly shown, like the single bonds. Double and triple bonds are written as "=" and "#," respectively. Branches are represented by parenthesis "()". Rings are shown by allocating digits to the two connecting ring atoms. For further details on notation, see Leach and Gillet[39] (Table 1).

There may be many ways to build the SMILES strings for a given molecule; SMILES string maybe written, starting from any atom or following different sequences through the molecule. For example, in Figure 2B, for aniline, many possible SMILES strings can be written besides, Nc1ccccc1, such as, c1ccccc1N, c1cc(N)ccc1, and so forth. To solve this problem, using the "canonical representation" a specific chemical structure is given one unique notation, for example, canonical SMILES [38], and InChI (formerly IChI) [40] being developed by International Union of Pure and Applied Chemistry (IUPAC). As a note, an InChIKey with a 27 character fixed-length representation is a hashed version of an InChI. The former was designed for open-web searching of chemical databases. A variant of SMILES, designed for identifying reaction data sets contains specialized grammar rules useful inputs to neural network models [41,42]. Atoms-in-molecule-based fragment–*amon* SMILES representation is one of the latest additions to the active learning in transferable quantum machine learning (QML) model [43].

3.3 | Molecular fingerprinting

Molecular fingerprinting is a vectorized representation of molecules capturing precise details of atomic configurations within. Depending on the problem in hand, fingerprinting can be coarser or finer. For instance, if one has to predict the mechanical or electrical strength of materials, where prediction accuracy is less critical, coarser fingerprint may be defined. It can be of any general attributes of the atoms by which, the material is made of, or other potentially relevant properties (e.g., band gap). On the other hand, if accuracy in predictions is demanded, such as, total energies and atomic forces, solid state space groups that fingerprint has to be finer with atomic-level structural information [44]. A reaction fingerprint which is often made from concatenating few fingerprints of reactant and reagent molecules, can be the input for a neural network [41]. A popularly used fingerprint is extended connectivity fingerprints (ECFPs) [24]. During the featurization process, a molecule is decomposed into substructures (e.g., fragments) of a fixed length binary fingerprint assembled into an array whose each element is either 1 or 0 (see Figure 3). Fingerprinting is usually used in similarity searching, clustering and virtual screening.

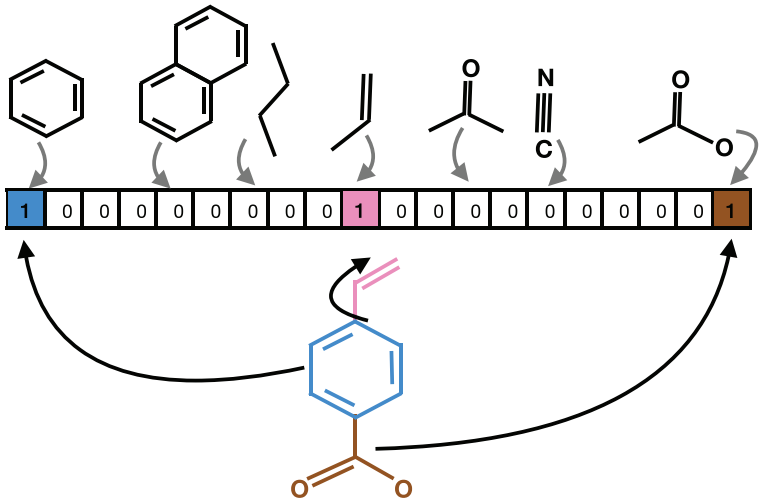


FIGURE 3 Representation of a molecular fingerprint encoded the presence (1) or absence (0) of certain substructures in a compound. This molecule is represented by a vector of length 20 consisting of binary numbers

3.4 | Coulomb matrix

In the "Coulomb matrix" (CM) representation of a molecule, Rupp et al. have used the same molecular information that enters in the Hamiltonian for an ab initio electronic structure calculation, namely, the set of Cartesian coordinates, $\{\mathbf{R}_I\}$ and nuclear charges, $\{Z_I\}$ [20]. The Coulomb matrix $\mathbf{M}$ is defined as

$$M_{IJ} = \begin{cases} 0.5Z_I^{2.4}, & \text{for } I = J \\ \dfrac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|}, & \text{for } I \neq J \end{cases} \tag{1}$$

Here, the diagonal elements correspond to a polynomial fit of the atomic energies to the nuclear charge Z_I. Whereas, off-diagonal elements encode the Coulomb repulsion between the nuclei I and J, which can be considered as scaled inverse distances. In order to represent an entire chemical compound space (CCS), an ML model needs to construct a nonlinear map between molecular characteristics and properties (e.g., atomization energies). This requires a molecular (dis)similarity measure which is invariant to translations, rotations, and the index ordering of atoms. Apart from being able to describe a complete set of pairwise distances of all atoms, the CM allows to encode information about the chemical composition of an entity without the introduction of additional descriptors. However, the CM is not invariant to permutations of the indices of atoms. Which means that many CMs can result to the same molecule in a dataset by just permuting rows and columns. To enforce such invariances, Montavon et al. proposed random Coulomb matrices by sorting: (i) eigenvalues of the CM, (ii) the rows and columns by their corresponding norm and (ii) randomly, rows and columns in order to generate a set of randomly sorted CM for each molecule [21,22]. To obtain fixed length feature vectors of a dataset, the size of the CM is given by the number of atoms of the largest molecule present in the dataset $d = d_{\text{max}}$, and feature vectors for smaller systems with number of atoms less than d_{max} are padded with zeros (see Figure 4).

3.5 | Bag-of-bonds

The "bag-of-bonds" (BoB) featurization, a variant of the Coulomb matrix, introduced by Hansen et al. becomes invariant to permutations of indices of atoms [23]. It is inspired by the so-called "bag of words" descriptor used in natural language processing where, a *bag* encodes the frequency of a particular word appearing in text. BoB follows a similar approach by making *bags* of different types of bonds (C—O, C—N, and so on), and order of the bond (single, double, triple) shown in Figure 5. Each *bag* is basically a vector. Each element of such vector is computed as $Z_I Z_J / |\mathbf{R}_I - \mathbf{R}_J|$, where Z_I and Z_J are the nuclear charges; $\mathbf{R}_I$ and $\mathbf{R}_J$ are the positions of the atoms participating in a given bond. All bags are concatenated in a specific order and, are also padded with zeros to make bags of equal sizes of all the molecules across the data set. This representation is invariant under rotation, translation and permutation. However, it fails to distinguish between so-called homometric configurations [45,46].

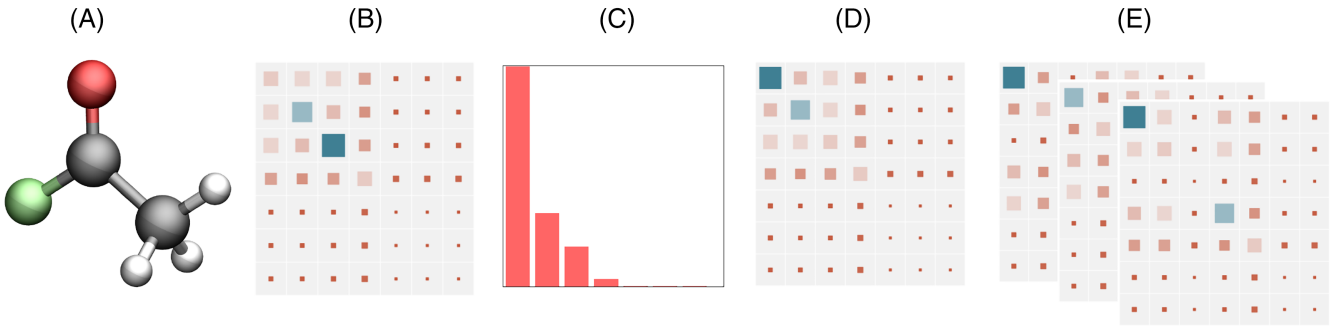


FIGURE 4 Coulomb matrix representation: (A) ball and stick representation of the CH_3COF molecule, (B) original (non-sorted) Coulomb matrix, (C) eigenvalues of the Coulomb matrix, (D) sorted Coulomb matrix, (E) set of randomly sorted Coulomb matrices

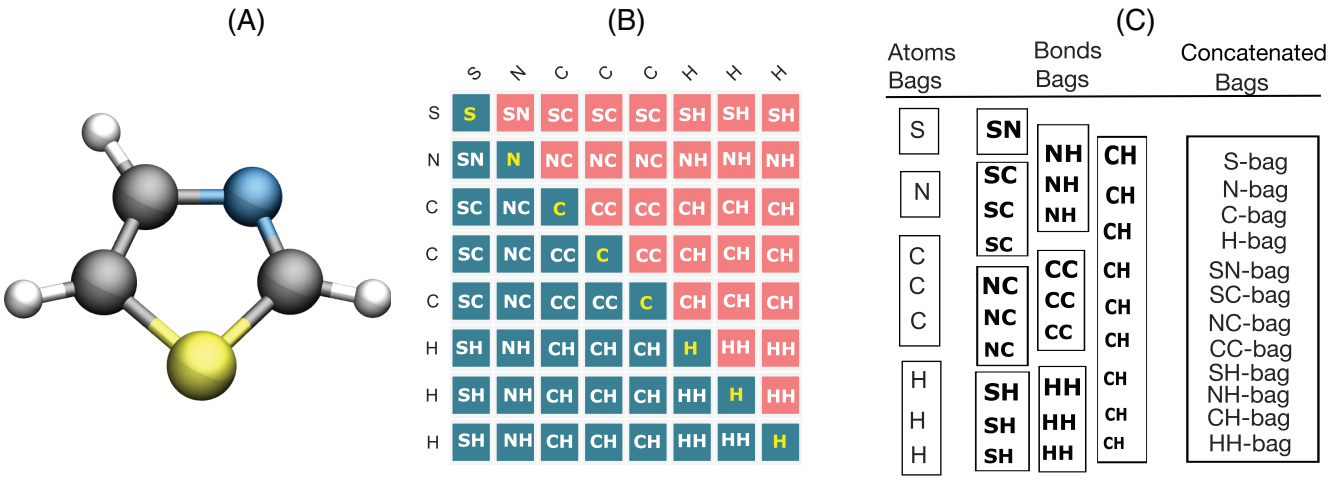


FIGURE 5 Schematic view of the bag-of-bonds (BoB) representation: (A) ball and stick representation of the thiazole (C_3H_3NS) molecule, (B) Coulomb matrix elements, (C) different Coulomb matrix entries (off-diagonal) are sorted into different bags. Here, concatenated bags are not padded with zeros for clarity

3.6 | Bonds-angles representation

In order to construct an improved representation over BoB, Huang and von Lilienfeld introduced bonds–angles (BA) representation. It relies on interatomic MBE, like, the bonding term of a force field, including bonding, angular, and higher order terms. They have simply used a hierarchy, consisting of bags of: (i) dressed atoms (M^D), (ii) atoms + bonds (M^B), (iii) atoms + bonds + angles (M^A), and (iv) atoms + bonds + angles + torsions (M^T) [27]. For the pair-wise bonding and nonbonding intramolecular terms, they used Morse and Lennard-Jones potentials, respectively. Sinusoidal functions were used for angles (3-body), and torsions (4-body) between covalently bonded atoms. Here, functional forms and parameters were chosen, such that, BA-representations correspond to the universal force field (UFF) [47] used for molecular mechanics and molecular dynamics simulations. Histograms of distances, angles, and dihedrals (HDAD) were introduced by manually generated bins in histograms for each feature, followed by condensing to a single value by summing up their individual contributions rendering into a fixed-size representation for each molecule [31]. Descriptors FCHL18 [28] and its variant FCHL19 [29] came up with the similar idea of using interatomic many-body interactions for atomic environment (considering atoms upto a cutoff value) of a molecule and/or condensed phase systems.

The FCHL19 [29] representation is composed of a two-body and a three-body term. The two-body term encodes radial distributions between a central atom and its neighboring atoms of a given element type. The log-normal radial basis function is

$$G^{2-\text{body}} = \zeta_2(R_{IJ}) f_c(R_{IJ}) \frac{1}{R_s \sigma(r_{ij}) \sqrt{2\pi}} \exp\left(-\frac{\left(\ln R_s - \mu(r_{ij})\right)^2}{2\sigma(r_{ij})^2}\right) \tag{2}$$

where R_s is the distance location of the grid point. $\mu(r_{ij})$ and $\sigma(r_{ij})$ parameters of the log-normal distribution depend on the interatomic distance R_{IJ}. The three-body term encodes the atom to atom-pair (neighboring) distances, angle between the triplet, and element types of the atomic environment.

$$G^{3-\text{body}} = \zeta_3 G^{\text{rad}} G^{\text{ang}} f_c(R_{IJ}) f_c(R_{IK}) f_c(R_{JK}) \tag{3}$$

The radial part G^{rad} is closely related to the ACSFs used in the Accurate NeurAl networK engINe for Molecular Energies (ANI)-1 neural network [48]. The angular part G^{ang} has two components which are cosine and sine terms. The FCHL19 representation is able to show state-of-the-art prediction accuracy on several datasets (QM7b, QM9, MD17) with much reduced computational cost. This is quite promising toward making transferable model that can be routinely trained and run molecular dynamics simulations efficiently throughout the CCS [29].

3.7 | Bonds-in-molecules

The "bonds in molecule" (BIM) [49] featurization proposed by Yao et al., is similar to the BoB and BAML representations. A molecule with the BIM representation, is broken down into overlapping bonds, such as, C—H, C—C, C—O, and so forth. Covalent bonds are included using simply with an element specific distance cutoff. The descriptor contains the bond length of the target bond, and the nearest-neighbor bonds and angles connected to it. Each type of bond then has its own branch that predicts the respective bond energy which are then summed to get the total energy of the molecule (see Figure 6). Though, the neural network was trained on the total molecular energy, it is able to make predictions of relative bond strengths which agree well with experimental and calculated values.

3.8 | Bonds–angles–nonbonds–dihedrals

Recently Laghuvarapu et al. introduced a transferable (to both equilibrium and nonequilibrium structures) and molecule size-independent ML model where, bonds (B), angles (A), nonbonded (N) interactions, and dihedrals (D) are incorporated in a molecular representation within a neural network potential (NNP) inspired by molecular mechanics force fields [50]. In the BAND molecular representation, each molecule is viewed as consisting of bonded pairs (adjacent atoms) and nonbonded pairs (non-adjacent atoms); a pair of consecutive bonds forming an angle, and three consecutive bonds forming a dihedral angle. An 8-length feature vector, concatenated from atom name (4-length) and the nearest-neighbor atom type (4-length) connected by a bond represents an atom, shown in Figure 7. Bonds, angles, nonbonded and dihedrals feature vectors of the

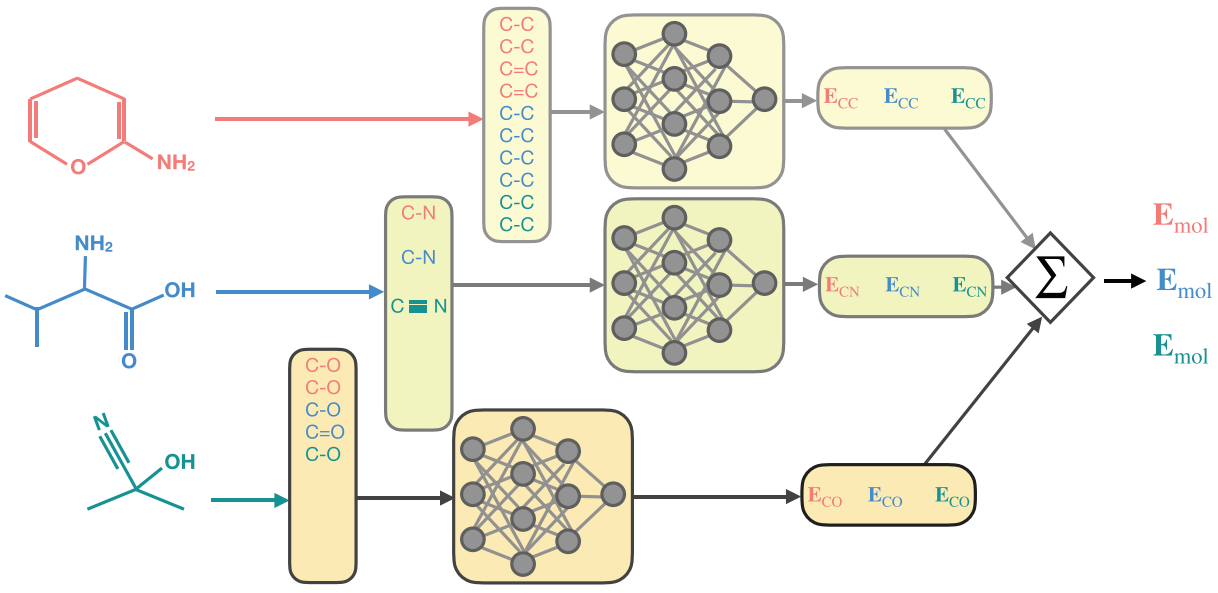


FIGURE 6 A scheme of how a bonds in molecule-neural network (BIM-NN) computational graph is trained. A batch of molecules decomposed into different bond-types. Within the batch, each bond-type is fed into a type-specific sub-network, and finally molecular energy is reassembled by a linear transformation

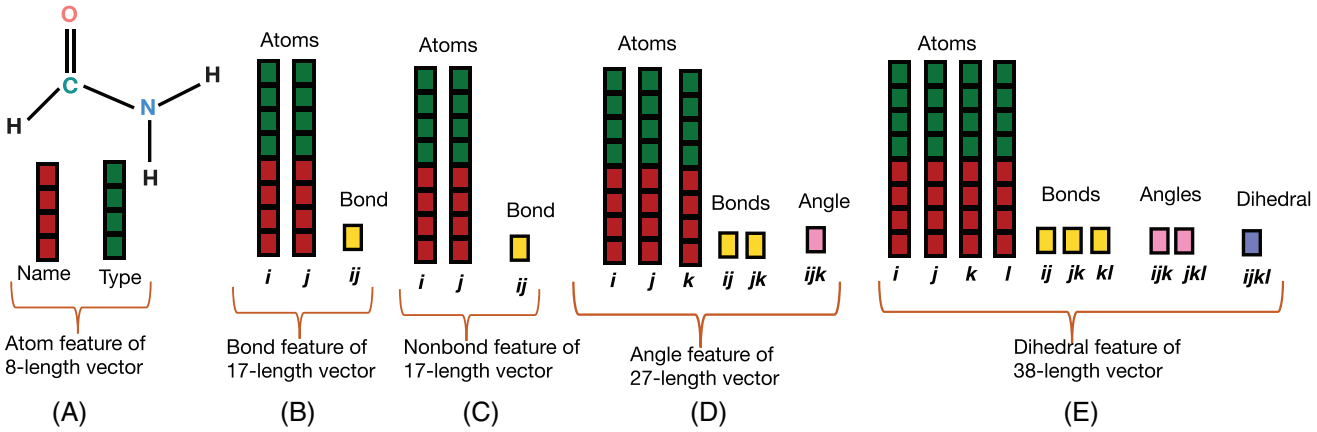


FIGURE 7 Schematic representation of the feature vectors of (A) atoms, (B) bonds, (C) nonbonds, (D) angles, and (E) dihedrals in BAND representation of a formamide molecule

participating atoms are then concatenated. In case of bonded and nonbonded terms, the feature vector is made of the atom vectors of the participating atom-pair, and the corresponding interatomic distance. The angle feature vector has three participating atoms, bond lengths of the participating bond-pair and the angle between them. The dihedral feature vector is consisted of four participating atoms, three bond lengths, two angles and the respective dihedral angle.

3.9 | Behler and Parrinello symmetry functions

Behler and Parrinello proposed ACSFs applicable to high-dimensional systems containing large numbers of atoms [26]. The sum of all local atomic energies E_I in the system of N atoms is equal to the total potential energy E,

$$E = \sum_{I=1}^{N} E_I \tag{4}$$

The local chemical environment of each atom is defined by a set of ACSFs. Interatomic interactions within a system are screened upto a cutoff radius R_c by using a cutoff function,

$$f_c(R_{IJ}) = \begin{cases} 0.5\left[\cos\left(\frac{\pi R_{IJ}}{R_c}\right)+1\right], & \text{for } R_{IJ} \le R_c \\ 0, & \text{for } R_{IJ} > R_c \end{cases} \tag{5}$$

where, R_{IJ} is the inter-atomic distance. This function essentially captures those interactions, as R_{IJ} to all neighboring atoms J, inside the cutoff radius and zero otherwise (Figure 8).

The positions of the atoms inside the R_c are then described by many-body atom-centered radial and angular symmetry functions which depend, consequently on the positions of all neighboring atoms. The ACSFs are: The "radial" function,

$$G_I^{\text{rad}} = \sum_{J \neq I}^{N} \exp\left[-\eta(R_{IJ} - R_s)^2\right] f_c(R_{IJ}) \tag{6}$$

and, the "angular" function,

$$G_I^{\text{ang}} = 2^{1-\zeta} \sum_{J \neq I}^{N} \sum_{K > J}^{N} (1 + \lambda\cos(\theta_{IJK}))^{\zeta} \times \exp\left[-\eta\left(R_{IJ}^2 + R_{IK}^2 + R_{JK}^2\right)\right] f_c(R_{IJ}) f_c(R_{IK}) f_c(R_{JK}) \tag{7}$$

The "radial" arrangement in Equation (6), is expressed by a product of a Gaussian and the cutoff function. The parameter η changes the width of the Gaussian distribution whereas, R_s is used to shift the center of the peak; and multiplication with the f_c ensures a smooth decay in value and

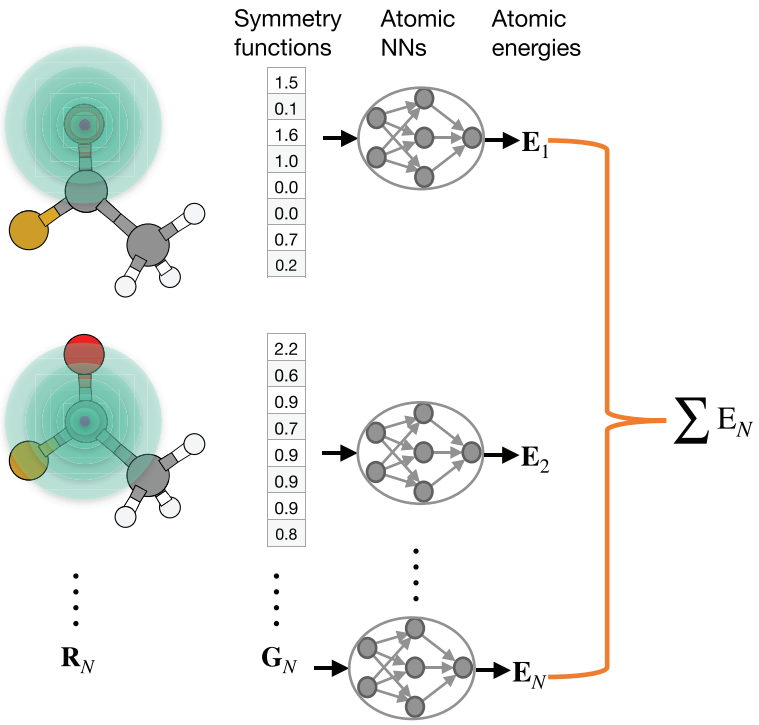


FIGURE 8 Structure of a high-dimensional neural network potential. First, Cartesian coordinates, $\mathbf{R}^I$ of each atom are transformed to a set of symmetry function values $\mathbf{G}_I$ depending also on the Cartesian coordinates of all atoms in the local environment. The symmetry function values represent the input vectors for atomic NNs yielding the atomic energy contributions E_I. The total energy E is the sum of all E_I

slope to zero at R_c. Typically a set of η values would entail a radial fingerprint of the atomic environment. The "angular" ACSF (Equation 7) gives an angular fingerprint of the atomic environment using the angles θ_{IJK} formed between the central atom I and with a pair of neighbors J and K; $\theta_{IJK} = \frac{R_{IJ}R_{IK}}{|R_{IJ}||R_{IK}|}$, where $\mathbf{R}_{IJ} = \mathbf{R}_I - \mathbf{R}_J$. Different angular regions are covered by adjusting the parameter ζ. While $\lambda = \pm 1$ shifts the positions of the maxima of the cosine function either to π or 2π. Typically, 50–100 symmetry functions with varying parameters (η, R_s, ζ, and λ) are used. Also, the cutoff R_c convergence is normally reached between 6 and 9 Å [51]. Since ACSFs depend on internal coordinates (R_{IJ} and θ_{IJK}), the rotational and translational invariances are implied/conserved. Further, due to the cumulation of all neighbors for a given atom I, ACSFs are invariant to any permutation of chemically equivalent atoms within the environment.

Apart from expressing the total energy as a sum of atomic energy contributions as given in Equation (4), it can also be shown as a sum of atom pair energies

$$E = \sum_{I=1}^{N} \sum_{J>I}^{N} E_{IJ} \tag{8}$$

like the Tersoff potential [52]. This form was also employed to construct machine learning potentials (MLPs) using the pair symmetry functions [53].

3.10 | ANI

The original BP symmetry functions are used to compute an atomic environment vector (AEV), $\overrightarrow{G_I^X} = \{G_1, G_2, G_3, \ldots G_M\}$, composed of elements, G_M which probe specific regions of an individual atom's radial and angular chemical environment. Equation (6) runs over a set of η and R_s parameters. Therefore, when probing with many small η parameters, vector elements can grow to very large values which are detrimental to the training of NNPs. An ANI potential uses only a single η to produce a thin Gaussian peak, and multiple R_s values are used to probe outward from the central atom: that is, for a set of $M = \{m_1, m_2, m_3, \ldots\} = \{(\eta_1, R_{s_1}), (\eta_2, R_{s_2}), (\eta_3, R_{s_3}) \ldots\}$ [48]. Each m with a constant η and multiple R_s parameters probe its own distinct region of an atom's radial environment. M consists of such a set of m.

The ANI model sees two modifications to the original "angular" BP symmetry functions given in Equation (7). The first addition is θ_s, and the second is a modified exponential factor that allows addition of the R_s parameter. The θ_s allows arbitrary number of shifts in the angular environment whereas, R_s allows the angular environment to be considered within radial shells based on the average of the distance from the neighboring atoms. For a given triple *I*, *J*, and *K*, an angle θ_{IJK}, centered on the atom *I* is computed along with two distances R_{IJ} and R_{IK}. A single element, $G_m^{A_{\text{mod}}}$ of $\vec{G}_I^X$, to probe the angular environment of atom *I*, takes the form of a sum overall *J* and *K* neighboring atom pairs, of the product of a radial and an angular factor,

$$G_m^{A_{\text{mod}}} = 2^{1-\zeta} \sum_{J,K \neq I}^{\text{all}} (1+\cos(\theta_{IJK}-\theta_s))^{\zeta} \times \exp\left[-\eta\left(\frac{R_{IJ}+R_{IK}}{2}-R_s\right)^2\right] f_c(R_{IJ}) f_c(R_{IK}) \tag{9}$$

3.11 | Weighted atom-centered symmetric functions

As we discussed in the previous section, the symmetry functions introduced by Behler and Parrinello as pairs (radial ACSFs, Equation 6) and triples (angular ACSFs, Equation 7), in order to describe the arrangement of different chemical elements relative to the central atom. For instance, if a molecular system contains three types of elements, for example, H, C, and O, the environment of a carbon atom would be described by a set of radial ACSFs for the pairs C—C, C—H, C—O, and angular ACSFs for the triples C—C—C, C—C—H, C—C—O, C—H—H, C—H—O, C—O—O. For each of these elemental combinations, in order to probe the radial and angular fingerprint of the atomic environment sufficiently enough, several ACSFs are used while varying their parameters (η, R_s, λ, and ζ). A system with the number of elements, N_{elem}, in order to account for every possible pair and triple, there should be N_{elem} and $N_{\text{elem}}(N_{\text{elem}}+1)$ number of ACSFs, respectively. That is, size of the descriptor scales "quadratically" with the N_{elem} [54]. Not only this, further, considering several ACSFs for each of these combinations, one again amplifies the undesirable scaling with N_{elem}. This is really the bottleneck for this descriptor for developing high-dimensional neural network potentials (HDNNPs). By virtue, the growing size of the BP's ACSFs become less suited with an increasing number of different elements in chemical systems. Due to the very same reason as Smith et al. improved the "angular" symmetry functions [48], Gastegger et al. proposed a modification namely, weighted ACSFs [55], w-ACSF. Here, element-dependent weighting functions were introduced implicitly, instead of using separate functions for each elemental combination. The radial part takes the form as

$$W_I^{\text{rad}} = \sum_{J \neq I}^{N} g(Z_J) \exp\left[-\eta(R_{IJ}-R_s)^2\right] f_c(R_{IJ}) \tag{10}$$

and, the angular part is expressed as,

$$W_I^{\text{ang}} = 2^{1-\zeta} \sum_{J \neq I}^{N} \sum_{K \neq I,J}^{N} h(Z_J,Z_K)(1+\lambda\cos(\theta_{IJK}))^{\zeta} \times \exp\left[-\eta\left(\frac{R_{IJ}+R_{IK}+R_{JK}}{3}-R_s\right)^2\right] f_c(R_{IJ}) f_c(R_{IK}) f_c(R_{JK}) \tag{11}$$

Z_J, and Z_K are the atomic numbers of the nuclei *J* and *K*, respectively. $g(Z_J)$ and $h(Z_I,Z_K)$ are weighting functions, those modify the contribution of each radial and angular term based on the chemical element involved. It was found that simply setting $g(Z_J) = Z_J$ and $h(Z_I,Z_K) = Z_I,Z_K$ gives satisfying results without introducing additional parameters.

3.12 | Neighborhood density functions

Bartók and co-workers proposed an alternative approach to describe atomic environment for constructing MLPs for high-dimensional systems using Gaussian processes [56]. Here, for each atom *I*, a neighbor density $\rho_I(\mathbf{R})$ is built at each point of space **R**, and a δ function is placed upto the cutoff.

$$\rho_I(\mathbf{R}) = \delta(\mathbf{R}) + \sum_J f_c(|\mathbf{R}_{IJ}|)\delta(\mathbf{R}-\mathbf{R}_{IJ}) \tag{12}$$

Where, the cutoff function f_c can take a form as given in Equation (5) which ensures a smooth decay to zero at the cutoff distance R_c. Using the spherical and hyperspherical harmonics, the neighborhood density can be transformed into a bispectrum matrix (for further details on the derivation, see Albert et al. [25]) in order to achieve translational and rotational as well as permutation invariance. Using the bispectrum components,

Thompson et al. introduced spectral neighbor analysis potential (SNAP) [57]. Here, bispectrum components linearly depend on the atomic energy, unlike the Gaussian approximation potential (GAP).

Later, by Artrith et al. weights (w_J) were introduced to enable a discrimination of different elements in addition to atom-centered radial (bond length) and angular (bond angle) distribution functions [54]; this does not scale with the number of chemical species present in the system.

3.13 | Smooth overlap of atomic positions

As the neighbor density defined in Equation (12) employs δ-functions, even slight deviations of the atomic positions between two environments may lead to strong numerical changes. Alternatively, the Smooth Overlap of Atomic Positions (SOAP) approach was introduced [25]. Here, the δ-functions have been replaced by Gaussians centered on each of the neighbors, N_{env} of the central atom I, as well as on the central atom itself. The neighbor density in SOAP is then given by

$$\rho_{\text{SOAP}}(\mathbf{R}) = \sum_{I=1}^{N_{\text{env}}} \exp\left(-\alpha|\mathbf{R}-\mathbf{R}_I|^2\right) \tag{13}$$

The SOAP kernel is built as the rotationally averaged squared overlap of the two corresponding neighbor atom densities, that is, a similarity measure between atomic neighborhoods is constructed as,

$$k\left(\rho_{\text{SOAP}},\rho'_{\text{SOAP}}\right)(\mathbf{R}) = \int d\widehat{R}\left|\int \rho_{\rho_{\text{SOAP}}}(\mathbf{r})\ \ \rho_{\rho'_{\text{SOAP}}}\left(\widehat{\mathbf{R}}\mathbf{r}\right)dr\right|^2 \tag{14}$$

Further, in order to extract a single similarity measure from the matrix of pairwise environment similarities, regularized entropy match (REMatch) SOAP kernel was introduced [58]. This descriptor was very useful for describing similarities of both lone molecular and bulk periodic structures (Figure 9). Recently Unke and Meuwly have used the neighborhood density function as a product of a radial Gaussian function and spherical harmonics using neural networks to predict molecular energies [60].

3.14 | Gaussian potential

Brockherde and co-workers used Gaussian potentials to represent molecules while learning the density—potential and energy—density maps to predict electron densities and energies for a series of small molecules [61]. Either to model energy E^{ML} directly as a functional of external potential

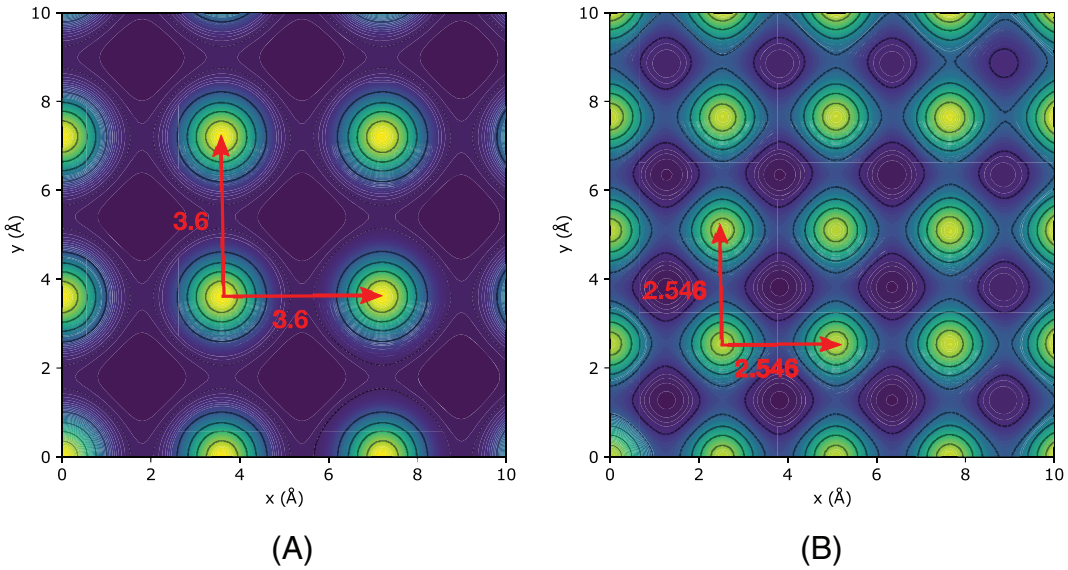


FIGURE 9 Smooth Overlap of Atomic Positions (SOAP) descriptions of periodic Cu/fcc lattices: (A) cubic and (B) orthorhombic; they are generated using the DScribe [59] package

$\nu(\mathbf{r})$ or to model density n^{ML}, require the characterization of the Hamiltonian by its external potential. Therefore, as a feature, they used artificial Gaussian potential of the form

$$\nu(\mathbf{r}) = \sum_{I=1}^{N} \mathbf{Z}_I \exp\left(\frac{-|\mathbf{r}-\mathbf{R}_I|^2}{2\gamma^2}\right) \tag{15}$$

where, $\mathbf{R}_I$ are the positions, $\mathbf{Z}_I$ are the atomic charges of the N atoms. The width γ is a hyper-parameter of the ML algorithm. The idea of using Gaussians to represent $\nu(\mathbf{r})$ was introduced earlier [56]. The Gaussian potential was discretized on a coarse grid of spacing 0.08. Then the discretized potential was used in Gaussian kernel with the kernel ridge regression (KRR) model. Gaussians are a popular choice for ML applications because of its local nature controlled by the width and their smoothness.

3.15 | Machine learned features

Kernel-based ML methods while use the hand engineered features discussed so far, like, CM, BoB, and so forth, descriptor-based neural network models, like, ANI [48] and TensorMol [62] make use of ACSFs. In contrast, an end-to-end NN architecture uses atom types and Cartesian coordinates as inputs, and learns an appropriate representation from the data. Many popular end-to-end NN models, such as, deep tensor neural network (DTNN) [63], SchNet [64], HIP-NN [65], and PhysNet [66] were inspired by the graph neural network model of Scarselli et al. [67] which was casted later in a message passing neural network (MPNN) by Gilmer et al. [68].

4 | ACCELERATING COMPUTATIONAL MODELING: ML APPLICATIONS

Applications of ML in science are manifold. This section reviews particularly, a few selected areas of chemistry where ML has contributed toward advancing the research field, that is otherwise dominated by quantum chemical methods. A significant effort has been devoted to addressing the challenges that faces when chemistry meets ML. It is identified that ML is already boosting computational chemistry at various levels. Quantum chemical calculations are now becoming popular source of new molecular descriptors, which can, in principle determine all of the geometric and electronic properties of molecules and their interactions [69]. Study of condensed phase chemical systems nowadays is becoming feasible with reasonable accuracy at lower cost employing the density functional theory (DFT)-based approaches in combination with ML models [70,71]. In the past few years, different ML models have been introduced, in order to understand various aspects of chemistry. Here, we focus in summarizing the developments related to molecular descriptors.

4.1 | QM calculated properties

The QM-based molecular modeling studies, where ML methods have been employed successfully are reported here, from recent contributions. Those can be considered as complementary, and give us an overall highlight of the applications. These include various QM observables: enthalpies of formation [30,72,73], atomization energies [20], reorganization energies [74], chemical reactivity [75], polymer properties [76], optimizing transition state surface which divides reactants from products [77], electronic ground state properties [22,31,49], electronic excitation energies [78], electron transmission coefficients [79], NMR nuclear shifts [80], frontier orbital eigenvalues [81], atomic charges, dipole and quadrupole moments [82], thermochemical properties (besides enthalpies, entropies and heat capacities) [83], infrared spectra [84], polarizabilities [85,86], energies of methanol clusters [87], and transition metals [88]. ML has been playing potential roles in constructing approximate QM methods [89], as well as predicting electron densities from DFT [61,90], improving the density functionals [91], basis set effects [92], finding approximate density functionals [93,94], Møller–Plesset (MP) theory correction [95], prediction of coupled-cluster (CC) singles and doubles amplitudes from MP2-derived properties [96], also predicting MP2 and CC energies from Hartree–Fock orbitals [97–99].

The majority of DL algorithms currently relies on neural networks. DL methodologies have often been used physics-based calculations in order to determine properties of a given molecular system. This involves training the neural network to predict one of the key components/properties of the overall calculation. As a natural choice, neural networks were too used to predict DFT- and wavefunction-based energies [95,100–103] like conventional ML models often did. One of the DL applications includes to predict PES. PES fitting caught attention in the early 1980s by the work done by Wagner, Schatz and Bowman. They demonstrated modern computing perspectives of, how to gain maximum information on the surface while using the smaller portion of the surface [104,105]. Then, for almost two decades, neural network fits to PES pitched up with ample contribution to the field by various research groups [26,48,63,68,106–114].

While kernel-based ML models for fitting the potential were introduced in 1996 [115], over the following years, it started to surge in a variety of applications [51,56,57,116–119]. Few of these approaches were used to construct molecular force fields for molecular dynamics simulations [118,120,121]. For MD simulations, one calculates energies and forces for a large number of atomic configurations in order to have adequate sampling of the phase space. That can be obtained by electronic structure calculations, often using the DFT or ab initio MD calculations (DFT-based), therefore restricted to a few 100 s of atoms with shorter simulation times, due to their time-consuming nature.

In case of conventional ML utilizing KRR, von Lilienfeld and co-workers used "Coulomb matrix" as descriptor in a so-called QML approach within the CCS, and got very impressive success in predicting quantum chemical properties without solving the Schrödinger equation. A randomized variant of the "Coulomb matrix" representation was later used to predict atomization energies [122], static polarizabilities, frontier orbital eigenvalues (HOMO, LUMO), ionization potential, and electron affinity [22]. The randomized variant was seen to improve the accuracy greatly. DL models have achieved quite satisfactory chemical accuracy in predicting PES using ACSFs and their modified versions. Neural network potential of protonated water clusters were recently studied by Behler and co-workers [113] with the CC accuracy, which is known as gold-standard wavefunction-based QM method. They have employed ACSFs as descriptors. However, SOAP-based ML model unifies predicting CC atomization energy of molecules, stability of molecular conformers, such as, conformers of glucose [123], receptor-ligand binding, ground state structures of silicon surfaces [70]. While calculating the excitation spectra, the CM representation was used with the conventional ML [78] as well as DL [124] models to predict the spectra of the same dataset. DL model gained similar accuracy to the ML approach, and Gghosh et al. stated therein DL will be able to work with the smaller data set also. In a gradient-domain machine learning (GDML) [118] approach to construct accurate molecular force fields using a restricted number of samples from ab initio molecular dynamics trajectories, Chmiela et al. used the reciprocal of the Euclidean distance of an atom-pair, inspired by the CM descriptor in order to define Cartesian geometries that are physically equivalent. Further, in a symmetrized GDML version, the same group used adjacency matrices (similar to CM) as representation to build molecular graphs [125] and constructed ML FF at CCSD(T) level of accuracy [121] for MD simulations. Being simple and efficient, the CM representation contributed in learning various components from QM calculations, also for constructing molecular force fields for MD simulations either using the conventional ML or DL methodologies.

4.2 | Computational materials

The fourth paradigm of science based on data-driven approaches now playing a big role in materials research [126]. The application of ML to address challenges in this emerging field has been extensively reviewed in prior publications [2,51,127,128]. Computational materials design early in 1998 [129] predicting of magnetic and optoelectronic materials, creating large high-throughput DFT databases [130–132], finding new ternary oxides [133,134], semiconductors [135,136] from known compounds, predicting new 15 compounds from naturally occurring 98 elements from the periodic table consisting of database upto five-component systems to enable new functionality [137], and almost naturally, QM-calculated properties (data) have been extended to accelerate the design and realization of advanced functional materials, often denoted as QSPR modeling [3,76,138–141]. In an inverse-design, very similar to molecular chemistry [142] approach, atomic configuration of an alloy in a combinatorial space could be realized from a target electronic structure property [143].

Many representations, developed until now, have been used for ML modeling of materials, such as, extended CM and sine matrix for predicting formation energies using a data set of 3938 crystal structures obtained from the materials project [144], crystal representation by Faber, Lindmaa, von Lilienfeld, Armiento (FLLA) [141] for predicting formation energies of elpasolite crystal, partial radial distribution functions [145] for predicting density of states on crystals from the inorganic crystal structure database [146], SOAP representations for Si_n [25] and Ta_n [147] clusters for predicting energy and forces, Voronoi tessellation [148], similarity function [149], property-labeled materials fragment descriptors [150], bond-order-potential [112], and so forth. Within a Gaussian process regression with input features consisting of the largest 10 principal components of fragment-based fingerprints, using the ML, a screening of pathways for the reaction of syngas ($CO + H_2$) on Rh(111) [151], at lower computational cost was achieved [152]. The binding energies of NO and CO_2 on small-crystal slabs were computed using the DFT and, then compared that with the larger RhAu [153] and NiGa [154] alloys, respectively applying ML approaches. The SOAP [25] kernels were used here. SOAP-based descriptors were also used to represent the entire configuration space of a complex material, such as, crystalline and amorphous Silicon [58], tungsten [155] and Ni_{19} nanoclusters [156,157] using DFT energies. ACSFs were successfully utilized in DFT-based NN potentials for predicting complex PESs [158] of multi-component systems, such as, ZnO supported copper clusters [159], bimetallic AuCu nanoparticles [160,161] and finding the global minimum of Au_{58} nanocluster [162], for studying global optimizations and phase transitions in Na_{20} to Na_{40} sodium clusters [163]. A new topology-based descriptor came up for energy prediction of metal nanoclusters [164]. Generation of materials with desired properties using conditional variational autoencoders with one-hot-encoding of the material composition was reported, recently [165]. Some of the open ML software being developed are listed in Table 2 which enlists sources of databases for materials as well.

Features based on molecular structure → property mapping have been useful for efficiently predicting QM-based properties. However, no such study exists, where prediction errors have been investigated to an extent that reader can chose a descriptor as a blackbox. It ought to be problem-specific, systematic as well [188].

TABLE 2 Machine learning codes and databases for materials

References	Name	Links
Packages for ML-based simulations of materials		
Khorshidi and Peterson [166]	AMP	https://amp.readthedocs.io/en/latest https://singroup.github.io/dscribe
Bartók and Csányi [167]	GAP	https://github.com/libAtoms/QUIP
Thompson et al. [57]	SNAP	https://lammps.sandia.gov/doc/pair_snap.html, https://github.com/materialsvirtuallab/snap
Artrith and Urban [168]	AENET	http://ann.atomistic.net
Huan et al. [169]	AGNI	https://lammps.sandia.gov/doc/pair_agni.html
Kolb et al. [170]	PROPhet	https://github.com/biklooost/PROPhet
Yao et al. [62]	TensorMol	https://github.com/jparkhill/TensorMol
Smith et al. [48]	ANI	https://github.com/isayev/ASE_ANI
Smith et al. [171]	COMP6	https://github.com/isayev/COMP6
Wang et al. [172]	DeePMD-kit95	https://github.com/deepmodeling/deepmd-kit
Mardt et al. [173]	VAMPnet	https://github.com/markovmodel/deeptime
Xie and Grossman [174]	CGCNN	https://github.com/txie-93/cgcnn
Jha et al. [175]	ElemNet	https://github.com/dipendra009/ElemNet
Jha et al. [176]	OQMD-SC	https://github.com/dipendra009/ElemNet
Data mining libraries of materials		
Curtarolo et al. [177]	AFLOWLIB	http://aflow.org/aflow-ml
Jain et al. [178]	Materials project	http://materialsproject.org
Kirklin et al. [179]	OQMD	http://oqmd.org
Bartók et al. [180]	libAtoms.org	http://www.libatoms.org/Home/DataRepository
Ward et al. [181]	Matminer	https://github.com/markovmodel/deeptime
Gómez-Bombarelli et al. [182]	Chemical VAE	https://github.com/aspuru-guzik-group/chemical_vae
Choudhary et al. [183]	JARVIS-DFT	https://github.com/usnistgov/jarvis
Pun et al. [112]	DScribe, descriptors	https://github.com/SINGROUP/dscribe,
Draxl and Scheffler [184]	NOMAD	https://analytics-toolkit.nomad-coe.eu
Olsthoorn et al. [185]	OMDB	https://omdb.mathub.io/dataset
Chapman et al. [186]	Khazana	https://khazana.gatech.edu
Kayastha and Ramakrishnan [187]	MolDis	https://moldis.tifrh.res.in/index.html

4.3 | Synthesis planning

Synthesis planning can be summed up into three different components: (i) Retrosynthesis, where the known-product to be synthesized from unknown-reactants, (ii) Reaction prediction, where known-reactants to be transformed into unknown-products, and (iii) Reaction optimization, where known-reactants/products of a specific reaction, would try to maximize the yield/efficiency. There exists a long history in computer-assisted chemical synthesis planning [189–193]. In early 80s Jorgensen and co-workers introduced computer assisted mechanistic evaluation of organic reactions [194]. The field is advancing rapidly with the major developments in ML approaches and the availability of millions of tabulated reaction examples. Optimization of a given reaction, and reaction prediction successfully utilized ML approaches [41,75,195,196]. To find retro-synthetic routes, often DL approaches have been employed [42,197–203]. ML models for MD simulations were also useful for predicting rate, yield of chemical reactions with mechanistic details [204]. ML has been making enormous strides accelerating computational retrosynthesis. However, in vivo evaluation of those predictions poses even greater challenges [205]. Often studies in this field suffer human biases, such as, imposing reaction-selectivity rules and heuristics that are suitable for interpolating known chemistry [206].

One of the early works that applied DL to reaction prediction involved DNNs with molecular fingerprints [41]. Very recently Coley et al. used GCNN to predict the products of organic reactions given their reactants, reagents, and solvent(s) [196]. The model shows an accuracy of predicting the major product correctly over 85%, which is significantly higher than the previous ML approaches [207,208], and performs on par with expert chemists with years of formal training. Here, a graph-based representation of reactant species introduced earlier [208], was used to

propose changes in bond order. Even, in 80s, graph theoretical approaches were applied in analyzing various aspects of chemical systems, like, definition, enumeration, and systematic coding or nomenclature of constitutional or steric isomers, valence isomers (especially of annulenes), and condensed polycyclic aromatic hydrocarbons (PAH) [34], and also in representing organic reactions [35]. A sequence-to-sequence (seq2seq) model that converts a product SMILES to reactant(s) SMILES was proposed in a template-free retrosynthesis study [200]. Segler and Waller used a neural network to score on the molecular fingerprints which are the relevant templates [197]; later, they extended this approach to Monte Carlo Tree Search with DNN policy for chemical synthesis planning [198]. Coley et al. used similarity-based scoring on ECFPs between the target and all known products, to extract a highly generalized template [203]. This method outperforms the template-free seq2seq model [200] and also extends to complete route planning when applied recursively. In summary, graph-based, molecular fingerprints and SMILES representations of the chemical structure are widely implemented descriptors in synthesis planning research.

Two main-stream research areas with biological relevance have been clearly boosted by the rapid development of ML algorithms and computer architectures [36,209–213]. They are namely, drug discovery and protein structure prediction. Each one consisting of various sub-areas holds a high esteem to the research field. A Review work by Yang et al. gives current status on various approaches which are popularly used in computer-assisted drug discovery [214]. The AlphaFold [215,216] model leveraging neural network algorithms took possibly the biggest leap in solving three-dimensional structure of a protein from its amino acid sequence posing a new challenge to biology.

4.4 | Performance of molecular representation

This section starts with giving details of standard datasets which are popularly used for QM/ML methods followed by the performance of molecular representation combined with the kernel-based ML and DL methods.

The performance of an ML model undeniably depends on the size and quality of the data to be used for learning. The majority of the ML modeling of the QM-calculated properties are using the QM7 [20], QM7b [22], QM8 [78], QM9 [217], QM7-X [218], and ANI-1 (variants) [48,219] datasets. They are created from the parent dataset GDB upto 11 [220] or 13 [221] or 17 [222] heavy atoms. QM7 (QM7b) is a collection of 7165 (7221) molecules with, upto seven heavy atoms. Whereas, QM9 is setup with 133 885 molecules with upto nine heavy atoms. The QM9 exists with geometric, energetic, electronic, and thermodynamic properties computed at DFT/B3LYP/6-31G(2df,p) level of theory. Due to its robustness, the QM9 dataset has become a classical benchmark for ML. PC9—a new QM9 equivalent dataset (only H, C, N, O and F and up to nine heavy atoms) of the PubChemQC project was published very recently [223]. Though, PC9 has wider chemical diversity, and only 18% of PC9 is common with QM9. Nevertheless, the overall accuracy in total energy is higher in case of QM9. Whereas, it was found that models trained on PC9 showed a better ability to predict energies of the QM9 dataset.

MoleculeNet [224] provides a uniform platform for performing molecular ML. It offers data-loading framework for various public datasets, data-splitting methods for comparison and evaluation, and high quality open-source implementations of multiple molecular featurizations and learning algorithms. Here, it has been emphasized that for QM dataset, proper choice of featurization appears critical. Especially for representing a molecule, atomic charges and positions, and atom-pair distances by which an interaction potential is formed, might as closely as, be the input for solving the Schrödinger equation.

Table 3 is composed of recently published articles which compares accuracy of the ML model employing various standard datasets. Mean absolute error (MAE) atomization energy predictions are considered here. From Table 3, it is apparent that, using the QM7 and QM9 datasets, a well-established KRR model almost reaches to the *chemical accuracy* (∼1 kcal/mol) using both BoB [27] and BAML [31], also a recent addition–*amon* [43] descriptors. Further, for QM9 dataset, KRR again with a many-body interactions to represent the structural and chemical environment of an atom in a compound outperforms the BoB, and BAML descriptors [28]. FCHL19 [29] shows the best performance in terms of accuracy for the QM9 dataset compared to previously mentioned representations. KRR model with the SOAP-based descriptor increased the accuracy level as MAE reaches 0.14 kcal/mol [226].

Alongside with kernel methods, NN architectures have been used the QM9 dataset for training and testing. A MPNN model proposed by Gilmer et al., uses graph structured data, additionally the interatomic distances [68]. Using the QM9 benchmark, it predicted within *chemical accuracy* (1 kcal/mol on total energies and 0.1 eV for orbital energies) for 11 out of 13 properties [68]. The ANI model utilizes a Behler and Parrinello symmetry functions to construct single-atom AEV [48]. While the ANI model (with much larger training size, a subset of the GDB-11 database), showed the accuracy <1.5 kcal/mol in predicting the conformational energies, recently a refined version–transfer learning ANI-1ccx model approaches the CCSD(T)/CBS accuracy [227]. Schütt et al. on a DTNN model with the QM9 dataset reached MAE of 0.32 kcal/mol for U_0; 0.04 and 0.03 eV for HOMO and LUMO energies, respectively [64]. Unke and Meuwly proposed a complex NN model, called PhysNet [66], predicted MAE of 0.14 kcal/mol for total energies; this is the best performing neural network model for the QM9 dataset.

As recent DL-based and regression-based ML approaches try to directly capture the electronic degrees of freedom either by learning the wavefunction or densities of the ground state, that lie at the heart of quantum chemistry, physics-based learning foresees exciting avenues toward new approximate QM methods with high-level quantum chemical accuracy at low computational cost.

TABLE 3 Mean absolute errors for atomization energies U_0 (kcal/mol), HOMO and LUMO energies for several models kernel ridge regression (KRR), elastic net (EN), Gaussian process regression (GPR), and neural networks (NN) reported in the literature (from oldest to most recent)

References	ML method/descriptors	Dataset	U_0	HOMO (eV)	LUMO (eV)
Rupp et al. [20]	KRR/CM	QM7	9.9	–	–
Montavon et al. [21]	Multilayer NN/random CM	QM7	3.5	–	–
Montavon et al. [22]	Multitask NN/random CM	QM7b	3.7	0.15	0.12
Hansen et al. [122]	KRR/random CM	QM7	3.0	–	–
Hansen et al. [23]	KRR/BoB	QM7	1.5	–	–
Huang and von Lilienfeld [27]	KRR/BoB	QM7b	1.8	0.15	0.16
Huang and von Lilienfeld [27]	KRR/BAML	QM7b	1.2	0.10	0.11
De et al. [58]	KRR/REMatch-SOAP	QM7b	0.92	0.11	0.08
Faber et al. [31]	EN/CM	QM9	21.0	0.34	0.63
Faber et al. [31]	EN/BoB	QM9	13.9	0.28	0.52
Faber et al. [31]	KRR/CM	QM9	3.0	0.13	0.18
Faber et al. [31]	KRR/BoB	QM9	1.5	0.09	0.12
Faber et al. [31]	KRR/BAML	QM9	1.2	0.09	0.12
Faber et al. [31]	KRR/HDAD	QM9	0.6	0.07	0.08
Huo and Rupp [225]	KRR/MBTR	QM7b	0.6	–	–
Bartók et al. [70]	GPR/SOAP-GAP	QM7b	0.40	–	–
Willatt et al. [226]	KRR/SOAP multi-kernel	QM9	0.14	–	–
Christensen et al. [29]	KRR/FCHL19	QM9	0.25		
Huang and von Lilienfeld [43]	AML/amon	QM9	<1.0		
References	**DL method/descriptors**	**Dataset**	U_0	**HOMO (eV)**	**LUMO (eV)**
Gilmer et al. [68]	MPNN	QM9	0.45	0.99	0.87
Smith et al. [48,171,227]	ANI-1[a] NN/ACSF	COMP6[b]	<1.5	–	–
Gastegger et al. [55]	HDNN/w-ACSF	QM9	<1.8	–	–
Hou et al. [228]	Multitask NN/CM	QM9	44.0	0.38	0.63
Schütt et al. [64]	SchNet NN	QM9	0.32	0.04	0.03
Lubbers et al. [65]	HIP-NN	QM9	0.26	–	–
Unke and Meuwly [60]	HDNN	QM9	0.41	–	–
Gómez-Bombarelli et al. [182]	RNN/VAE	QM9	–	0.16	0.16
Unke and Meuwly [66]	PhysNet NN	QM9	0.14	–	–
Schütt et al. [114]	SchNOrb NN	MD17[c]	<0.046	<0.02	<0.1

[a]Variants of ANI-1;
[b]GDB-07to09 benchmark, conformation energy;
[c]Chmiela et al. [118], reference calculations were performed at Hartree–Fock and DFT level.

5 | SUMMARY AND OUTLOOK

Applications of ML for molecular/material science are progressing rapidly beyond the conventional approach. ML molecular descriptors are being popularly used till today, whereas DL machine learned features are utilized commonly with high rate of success as well. However, underlying fundamental principles behind DL models, perhaps one of the prime aspects are yet to be uncovered. Hence, limiting its application toward molecular science where physical and chemical complex variables play the basic role. For example, electronic to atomic principals within a molecular entity will differ with its varying state specificity from gas phase to condensed phase, rigid inorganic materials to flexible bio-materials, and so forth. Crafting an ML representation of a molecule with $\{(Z_i, \mathbf{r_i})\}_{i=1}^{N}$, where Z_i and $\mathbf{r_i}$ are atomic number and three-dimensional position vector of the atom i, respectively is possible. However, in case of solids atomic environment descriptor together with cutoff distance is used to limit the number of neighboring atoms to compute the representation. Another key factor is to exploit invariances of the task at hand. For example, molecules can generally be translated and rotated in space without affecting their properties, that is, predictions are unchanged. Other well-known challenges of ML science are scalability, transferability, data-size-extensivity [229].

In computational chemistry where scientists still face challenges, like, solving chemical reaction dynamics problems where bond-breaking and bond-formation occurring hence, the ambiguity behind applying the force field-based approach or the wave function-based approach. Solving ground electronic states are highly influenced by electron correlation methods or by the exchange-correlation functionals used within the Kohn-Sham DFT. Calculation of electronic excited states beyond the linear response regime within the adiabatic approximation of the TDDFT raises subtle questions to be answered. Within the multiscale modeling approach, like, QM/MM where polarization effects are included either in the QM Hamiltonian or in the MM force fields as an additional contribution. Considering applications of theoretical methods lie in the core of the problem of interest, ML model built with a specific molecular representation therefore, performs the best, so far. Nevertheless, overcoming the problem of computational cost, applications and developments of new ML algorithms with a specific molecular descriptor or a DL learned features are positively assisting molecular science, and are going to witness significant research activities in future.

ACKNOWLEDGMENTS

S.R. thanks the WOS-A/DST Grant, No. SR/WOS-A/CS-19/2018 (G) for the financial support. We also thank IHub-Data, IIIT Hyderabad for funding.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

AUTHOR CONTRIBUTIONS

Shampa Raghunathan: Conceptualization; data curation; formal analysis; funding acquisition; software; visualization; writing – original draft; writing – review and editing. **U. Deva Priyakumar:** Conceptualization; funding acquisition; writing – review and editing.

DATA AVAILABILITY STATEMENT

Data available on request from the authors

ORCID

Shampa Raghunathan https://orcid.org/0000-0002-8872-4411

REFERENCES

[1] S. Webb, *Nature* **2018**, *554*(7693), 555.
[2] V. Tu Le, C. Epa, F. R. Burden, D. A. Winkler, *Chem. Rev.* **2012**, *112*(5), 2889.
[3] M. Karelson, V. S. Lobanov, A. R. Katritzky, *Chem. Rev.* **1996**, *96*(3), 1027.
[4] D. Ruppert, *J. Am. Stat. Assoc.* **2004**, *99*, 567.
[5] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, MIT Press, Cambridge, MA **2012**.
[6] Y. LeCun, Y. Bengio, G. Hinton, *Nature* **2015**, *521*(7553), 436.
[7] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, MIT Press, Cambridge, MA **2016**.
[8] P. Baldi, P. Sadowski, D. Whiteson, *Nat. Commun.* **2014**, *5*, 4308.
[9] A. C. Mater, M. L. Coote, *J. Chem. Inf. Model.* **2019**, *59*(6), 2545.
[10] G. B. Goh, N. O. Hodas, A. Vishnu, *J. Comput. Chem.* **2017**, *38*(16), 1291.
[11] T. F. G. G. Cova, A. A. C. C. Pais, *Front. Chem.* **2019**, *7*, 809.
[12] C. Angermueller, T. Pärnamaa, L. Parts, O. Stegle, *Mol. Syst. Biol.* **2016**, *12*(7), 878.
[13] T. Ching, D. S. Himmelstein, B. K. Beaulieu-Jones, A. A. Kalinin, B. T. Do, G. P. Way, E. Ferrero, P.-M. Agapow, M. Zietz, M. M. Hoffman, W. Xie, G. L. Rosen, B. J. Lengerich, J. Israeli, J. Lanchantin, S. Woloszynek, A. E. Carpenter, A. Shrikumar, J. Xu, E. M. Cofer, C. A. Lavender, S. C. Turaga, A. M. Alexandari, Z. Lu, D. J. Harris, D. DeCaprio, Y. Qi, A. Kundaje, Y. Peng, L. K. Wiley, M. H. S. Segler, S. M. Boca, S. J. Swamidass, A. Huang, A. Gitter, C. S. Greene, *J. R. Soc. Interface* **2018**, *15*(141), 20170387.
[14] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Ł. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, J. Dean Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint* arXiv:1609.08144, **2016**.
[15] H. Jippo, T. Matsuo, R. Kikuchi, D. Fukuda, A. Matsuura, M. Ohfuchi, *Mol. Inf.* **2019**, *39*, 1800155.
[16] D. J. Durand, N. Fey, *Chem. Rev.* **2019**, *119*(11), 6561.
[17] R. Todeschini, V. Consonni, *Molecular Descriptors for Chemoinformatics: Volume I: Alphabetical Listing/Volume II: Appendices, References*, Vol. 41, John Wiley & Sons, Weinheim, Germany **2009**.
[18] G. Schneider, *Nat. Rev. Drug Discov.* **2010**, *9*(4), 273.
[19] S. K. Chakravarti, *ACS Omega* **2018**, *3*(3), 2825.
[20] M. Rupp, A. Tkatchenko, K.-R. Müller, O. A. von Lilienfeld, *Phys. Rev. Lett.* **2012**, *108*(5), 058301.
[21] G. Montavon, K. Hansen, S. Fazli, M. Rupp, F. Biegler, A. Ziehe, A. Tkatchenko, O. A. von Lilienfeld, K.-R. Müller, Learning invariant representations of molecules for atomization energy prediction. in *Advances in Neural Information Processing Systems 25* (Eds: F. Pereira, C. J. C. Burges, L. Bottou,

K. Q. Weinberger), Curran Associates, Inc., NY, USA **2012**, p. 440 http://papers.nips.cc/paper/4830-learning-invariant-representations-of-molecules-for-atomization-energy-prediction.pdf
[22] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, O. A. von Lilienfeld, *New J. Phys.* **2013**, *15*(9), 095003.
[23] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K.-R. Müller, A. Tkatchenko, *J. Phys. Chem. Lett.* **2015**, *6*(12), 2326.
[24] D. Rogers, M. Hahn, *J. Chem. Inf. Model.* **2010**, *50*(5), 742.
[25] P. Albert, R. K. Bartók, G. Csányi, *Phys. Rev. B* **2013**, *87*(18), 184115.
[26] J. Behler, M. Parrinello, *Phys. Rev. Lett.* **2007**, *98*(14), 146401.
[27] B. Huang, O. A. von Lilienfeld, *J. Chem. Phys.* **2016**, *145*(16), 161102.
[28] F. A. Faber, A. S. Christensen, B. Huang, O. A. von Lilienfeld, *J. Chem. Phys.* **2018**, *148*(24), 241717.
[29] A. S. Christensen, L. A. Bratholm, F. A. Faber, O. A. von Lilienfeld, *J. Chem. Phys.* **2020**, *152*(4), 044107.
[30] O. A. von Lilienfeld, R. Ramakrishnan, M. Rupp, A. Knoll, *Int. J. Quantum Chem.* **2015**, *115*(16), 1084.
[31] F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley, O. A. von Lilienfeld, *J. Chem. Theory Comput.* **2017**, *13*(11), 5255.
[32] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, NY, USA **2006**.
[33] L. Spialter, *J. Am. Chem. Soc.* **1963**, *85*(13), 2012.
[34] A. T. Balaban, *J. Chem. Info. Comput. Sci.* **1985**, *25*(3), 334.
[35] S. Fujita, *J. Chem. Info. Comput. Sci.* **1986**, *26*(4), 205.
[36] Y. B. L. Samaga, S. Raghunathan, U. D. Priyakumar, *J. Phys. Chem. B* **2021**, *125*(38), 10657.
[37] D. Weininger, *J. Chem. Info. Comput. Sci.* **1988**, *28*(1), 31.
[38] D. Weininger, A. Weininger, J. L. Weininger, *J. Chem. Info. Comput. Sci.* **1989**, *29*(2), 97.
[39] A. R. Leach, V. J. Gillet, Representation and manipulation of 2D molecular structures. in *An Introduction to Chemoinformatics*, Springer, Dordrecht, Netherlands **2007**, p. 1.
[40] S. Heller, A. McNaught, S. Stein, D. Tchekhovskoi, I. Pletnev, *Aust. J. Chem.* **2013**, *5*(1), 7.
[41] J. N. Wei, D. Duvenaud, A. Aspuru-Guzik, *ACS Cent. Sci.* **2016**, *2*(10), 725.
[42] C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green, K. F. Jensen, *ACS Cent. Sci.* **2017**, *3*(5), 434.
[43] B. Huang, O. A. von Lilienfeld., *Nat. Chem.* **2020**, *12*(10), 945.
[44] R. Ramprasad, R. Batra, G. Pilania, *NPJ Comput. Mater.* **2017**, *3*(1), 1.
[45] A. L. Patterson, *Nature* **1939**, *143*(3631), 939.
[46] A. Sadeghi, S. A. Ghasemi, B. Schaefer, S. Mohr, M. A. Lill, S. Goedecker, *J. Chem. Phys.* **2013**, *139*(18), 184118.
[47] A. K. Rappé, C. J. Casewit, K. S. Colwell, W. A. Goddard III., W. M. Skiff., *J. Am. Chem. Soc.* **1992**, *114*(25), 10024.
[48] J. S. Smith, O. Isayev, A. E. Roitberg, *Chem. Sci.* **2017**, *8*(4), 3192.
[49] K. Yao, J. E. Herr, S. N. Brown, J. Parkhill, *J. Phys. Chem. Lett.* **2017**, *8*(12), 2689.
[50] S. Laghuvarapu, Y. Pathak, U. D. Priyakumar, *J. Comput. Chem.* **2020**, *41*(8), 790.
[51] J. Behler, *J. Chem. Phys.* **2016**, *145*(17), 170901.
[52] J. Tersoff, *Phys. Rev. Lett.* **1986**, *56*(6), 632.
[53] K. V. J. Jose, N. Artrith, J. Behler, *J. Chem. Phys.* **2012**, *136*(19), 194111.
[54] N. Artrith, A. Urban, G. Ceder, *Phys. Rev. B* **2017**, *96*(1), 014112.
[55] M. Gastegger, L. Schwiedrzik, M. Bittermann, F. Berzsenyi, P. Marquetand, *J. Chem. Phys.* **2018**, *148*(24), 241709.
[56] A. P. Bartók, M. C. Payne, R. Kondor, G. Csányi, *Phys. Rev. Lett.* **2010**, *104*(13), 136403.
[57] A. P. Thompson, L. P. Swiler, C. R. Trott, S. M. Foiles, G. J. Tucker, *J. Comput. Phys.* **2015**, *285*, 316.
[58] S. De, A. P. Bartók, G. Csányi, M. Ceriotti, *Phys. Chem. Chem. Phys.* **2016**, *18*(20), 13754.
[59] L. Himanen, M. O. J. Jäger, E. V. Morooka, F. F. Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke, A. S. Foster, *Comput. Phys. Commun.* **2020**, *247*, 106949.
[60] O. T. Unke, M. Meuwly, *J. Chem. Phys.* **2018**, *148*(24), 241708.
[61] F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, K. Burke, K.-R. Müller, *Nat. Commun.* **2017**, *8*(1), 1.
[62] K. Yao, J. E. Herr, D. W. Toth, R. Mckintyre, J. Parkhill, *Chem. Sci.* **2018**, *9*(8), 2261.
[63] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, A. Tkatchenko, *Nat. Commun.* **2017**, *8*(1), 1.
[64] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, K.-R. Müller, *J. Chem. Phys.* **2018**, *148*(24), 241722.
[65] N. Lubbers, J. S. Smith, K. Barros, *J. Chem. Phys.* **2018**, *148*(24), 241715.
[66] O. T. Unke, M. Meuwly, *J. Chem. Theory Comput.* **2019**, *15*(6), 3678.
[67] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, G. Monfardini, *IEEE Trans. Neural Netw.* **2008**, *20*(1), 61.
[68] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, G. E. Dahl, Neural message passing for quantum chemistry. in *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70, Sydney, Australia, JMLR. org, **2017**, p. 1263.
[69] A. V. Sinitskiy, V. S. Pande. *Physical Machine Learning Outperforms "Human Learning" in Quantum Chemistry*, **2019**, arXiv preprint arXiv:1908.00971.
[70] A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi, M. Ceriotti, *Sci. Adv.* **2017**, *3*(12), e1701816.
[71] P. Pattnaik, S. Raghunathan, T. Kalluri, P. Bhimalapuram, C. V. Jawahar, U. D. Priyakumar, *J. Phys. Chem. A* **2020**, *124*(34), 6954.
[72] L. H. Hu, X. J. Wang, L. H. Wong, G. H. Chen, *J. Chem. Phys.* **2003**, *119*(22), 11501.
[73] J. Sun, J. Wu, T. Song, L. H. Hu, K. L. Shan, G. H. Chen, *J. Phys. Chem. A* **2014**, *118*(39), 9120.
[74] M. Misra, D. Andrienko, B. Baumeier, J.-L. Faulon, O. A. von Lilienfeld., *J. Chem. Theory Comput.* **2011**, *7*(8), 2549.
[75] M. A. Kayala, P. Baldi, *J. Chem. Inf. Model.* **2012**, *52*(10), 2526.
[76] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, R. Ramprasad, *Sci. Rep.* **2013**, *3*(1), 1.
[77] Z. D. Pozun, K. Hansen, D. Sheppard, M. Rupp, K.-R. Müller, G. Henkelman, *J. Chem. Phys.* **2012**, *136*(17), 174101.
[78] R. Ramakrishnan, M. Hartmann, E. Tapavicza, O. A. von Lilienfeld, *J. Chem. Phys.* **2015**, *143*(8), 084111.
[79] A. Lopez-Bezanilla, O. A. von Lilienfeld, *Phys. Rev. B* **2014**, *89*(23), 235411.
[80] M. Rupp, R. Ramakrishnan, O. A. von Lilienfeld, *J. Phys. Chem. Lett.* **2015**, *6*(16), 3309.

[81] E. O. Pyzer-Knapp, K. Li, A. Aspuru-Guzik, *Adv. Funct. Mater.* **2015**, *25*(41), 6495.
[82] T. Bereau, D. Andrienko, O. A. von Lilienfeld, *J. Chem. Theory Comput.* **2015**, *11*(7), 3225.
[83] C. A. Grambow, Y.-P. Li, W. H. Green, *J. Phys. Chem. A* **2019**, *123*(27), 5826.
[84] M. Gastegger, J. Behler, P. Marquetand, *Chem. Sci.* **2017**, *8*(10), 6924.
[85] A. Grisafi, D. M. Wilkins, G. Csányi, M. Ceriotti, *Phys. Rev. Lett.* **2018**, *120*(3), 036002.
[86] D. M. Wilkins, A. Grisafi, Y. Yang, K. U. Lao, R. A. DiStasio Jr., M. Ceriotti, *Proc. Natl. Acad. Sci. USA* **2019**, *116*(9), 3401.
[87] K. Yao, J. E. Herr, J. Parkhill, *J. Chem. Phys.* **2017**, *146*(1), 014106.
[88] J. P. Janet, H. J. Kulik, *Chem. Sci.* **2017**, *8*(7), 5137.
[89] H. Li, C. Collins, M. Tanha, G. J. Gordon, D. J. Yaron, *J. Chem. Theory Comput.* **2018**, *14*(11), 5764.
[90] K. Ryczko, D. A. Strubbe, I. Tamblyn, *Phys. Rev. A* **2019**, *100*(2), 022512.
[91] X. Zheng, L. H. Hu, X. J. Wang, G. H. Chen, *Chem. Phys. Lett.* **2004**, *390*(1–3), 186.
[92] R. M. Balabin, E. I. Lomakina, *Phys. Chem. Chem. Phys.* **2011**, *13*(24), 11710.
[93] J. C. Snyder, M. Rupp, K. Hansen, K.-R. Müller, K. Burke, *Phys. Rev. Lett.* **2012**, *108*(25), 253002.
[94] S. Dick, M. Fernandez-Serra, *Nat. Commun.* **2020**, *11*(1), 1.
[95] R. T. McGibbon, A. G. Taube, A. G. Donchev, K. Siva, F. Hernández, C. Hargus, K.-H. Law, J. L. Klepeis, D. E. Shaw, *J. Chem. Phys.* **2017**, *147*(16), 161725.
[96] J. Townsend, K. D. Vogiatzis, *J. Phys. Chem. Lett.* **2019**, *10*(14), 4129.
[97] M. Welborn, L. Cheng, T. F. Miller III., *J. Chem. Theory Comput.* **2018**, *14*(9), 4772.
[98] L. Cheng, N. B. Kovachki, M. Welborn, T. F. Miller III., *J. Chem. Theory Comput.* **2019**, *15*(12), 6668.
[99] L. Cheng, M. Welborn, A. S. Christensen, T. F. Miller III., *J. Chem. Phys.* **2019**, *150*(13), 131103.
[100] K. Yao, J. Parkhill, *J. Chem. Theory Comput.* **2016**, *12*(3), 1139.
[101] K. Mills, M. Spanner, I. Tamblyn, *Phys. Rev. A* **2017**, *96*(4), 042113.
[102] J. Hermann, Z. Schätzle, F. Noé, *Nat. Chem.* **2020**, *12*(10), 891.
[103] D. Pfau, J. S. Spencer, A. G. D. G. Matthews, W. M. C. Foulkes, *Phys. Rev. Res.* **2020**, *2*(3), 033429.
[104] A. F. Wagner, G. C. Schatz, J. M. Bowman, *J. Chem. Phys.* **1981**, *74*(9), 4960.
[105] G. C. Schatz, *Rev. Mod. Phys.* **1989**, *61*(3), 669.
[106] B. G. Sumpter, D. W. Noid, *Chem. Phys. Lett.* **1992**, *192*(5–6), 455.
[107] S. Lorenz, A. Groß, M. Scheffler, *Chem. Phys. Lett.* **2004**, *395*(4–6), 210.
[108] S. Manzhos, T. Carrington Jr., *J. Chem. Phys.* **2006**, *125*(8), 084109.
[109] S. Manzhos, R. Dawes, T. Carrington Jr., *Int. J. Quantum Chem.* **2015**, *115*(16), 1012.
[110] J. Behler, *Int. J. Quantum Chem.* **2015**, *115*(16), 1032.
[111] K. T. Schütt, P. Kessel, M. Gastegger, K. A. Nicoli, A. Tkatchenko, K.-R. Muller, *J. Chem. Theory Comput.* **2018**, *15*(1), 448.
[112] G. P. Purja Pun, R. Batra, R. Ramprasad, Y. Mishin, *Nat. Commun.* **2019**, *10*(1), 1.
[113] C. Schran, J. Behler, D. Marx, *J. Chem. Theory Comput.* **2019**, *16*, 88.
[114] K. T. Schütt, M. Gastegger, A. Tkatchenko, K.-R. Müller, and R. J. Maurer., *Nat. Commun.* **2019**, *10*(1), 1.
[115] T.-S. Ho, H. Rabitz, *J. Chem. Phys.* **1996**, *104*(7), 2584.
[116] Z. Li, J. R. Kermode, A. De Vita, *Phys. Rev. Lett.* **2015**, *114*(9), 096405.
[117] V. Botu, R. Ramprasad, *Int. J. Quantum Chem.* **2015**, *115*(16), 1074.
[118] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, K.-R. Müller, *Sci. Adv.* **2017**, *3*(5), e1603015.
[119] A. Owens, S. N. Yurchenko, A. Yachmenev, W. Thiel, *J. Chem. Phys.* **2015**, *143*(24), 244317.
[120] L. Zhang, J. Han, H. Wang, R. Car, E. Weinan, *Phys. Rev. Lett.* **2018**, *120*(14), 143001.
[121] S. Chmiela, H. E. Sauceda, K.-R. Müller, A. Tkatchenko, *Nat. Commun.* **2018**, *9*(1), 1.
[122] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. von Lilienfeld, A. Tkatchenko, K.-R. Müller, *J. Chem. Theory Comput.* **2013**, *9*(8), 3404.
[123] M. Marianski, A. Supady, T. Ingram, M. Schneider, C. Baldauf, *J. Chem. Theory Comput.* **2016**, *12*(12), 6157.
[124] K. Ghosh, A. Stuke, M. Todorović, P. B. Jørgensen, M. N. Schmidt, A. Vehtari, P. Rinke, *Adv. Sci.* **2019**, *6*(9), 1801367.
[125] S. Umeyama, *IEEE Trans. Pattern Anal. Mach. Intell.* **1988**, *10*(5), 695.
[126] A. Agrawal, A. Choudhary, *APL Mater.* **2016**, *4*(5), 053208.
[127] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, *Nature* **2018**, *559*(7715), 547.
[128] T. Mueller, A. Hernandez, C. Wang, *J. Chem. Phys.* **2020**, *152*(5), 050902.
[129] N. N. Kiselyova, V. P. Gladun, N. D. Vashchenko, *J. Alloys Compd.* **1998**, *279*(1), 8.
[130] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, C. Wolverton, *JOM* **2013**, *65*(11), 1501.
[131] R. K. Vasudevan, K. Choudhary, A. Mehta, R. Smith, G. Kusne, F. Tavazza, L. Vlcek, M. Ziatdinov, S. V. Kalinin, J. Hattrick-Simpers, *MRS Commun.* **2019**, *9*(3), 821.
[132] A. Jain, G. Hautier, S. P. Ong, K. Persson, *J. Mater. Res.* **2016**, *31*(8), 977.
[133] G. Hautier, C. C. Fischer, A. Jain, T. Mueller, G. Ceder, *Chem. Mater.* **2010**, *22*(12), 3762.
[134] T. Moot, O. Isayev, R. W. Call, S. M. McCullough, M. Zemaitis, R. Lopez, J. F. Cahoon, A. Tropsha, *Mater. Discov.* **2016**, *6*, 9.
[135] D. W. Davies, K. T. Butler, A. J. Jackson, A. Morris, J. M. Frost, J. M. Skelton, A. Walsh, *Chem* **2016**, *1*(4), 617.
[136] C. Schober, K. Reuter, H. Oberhofer, *J. Phys. Chem. Lett.* **2016**, *7*(19), 3973.
[137] A. Walsh, *Nat. Chem.* **2015**, *7*(4), 274.
[138] A. Pulido, L. Chen, T. Kaczorowski, D. Holden, M. A. Little, S. Y. Chong, B. J. Slater, D. P. McMahon, B. Bonillo, C. J. Stackhouse, A. Stephenson, C. M. Kane, R. Clowes, T. Hasell, A. I. Cooper, G. M. Day, *Nature* **2017**, *543*(7647), 657.
[139] D. C. Elton, Z. Boukouvalas, M. S. Butrico, M. D. Fuge, P. W. Chung, *Sci. Rep.* **2018**, *8*(1), 1.
[140] A. Mannodi-Kanakkithodi, G. Pilania, T. D. Huan, T. Lookman, R. Ramprasad, *Sci. Rep.* **2016**, *6*, 20952.
[141] F. A. Faber, A. Lindmaa, O. A. von Lilienfeld, R. Armiento, *Phys. Rev. Lett.* **2016**, *117*(13), 135502.

[142] C. Kuhn, D. N. Beratan, *J. Phys. Chem.* **1996**, *100*(25), 10595.
[143] A. Franceschetti, A. Zunger, *Nature* **1999**, *402*(6757), 60.
[144] F. A. Faber, A. Lindmaa, O. A. von Lilienfeld, R. Armiento, *Int. J. Quantum Chem.* **2015**, *115*(16), 1094.
[145] K. T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K.-R. Müller, E. K. U. Gross, *Phys. Rev. B* **2014**, *89*(20), 205118.
[146] FIZ Karlsruhe. *Inorganic Crystal Structure Database (ICSD)*, FIZ Karlsruhe GmbH, Karlsruhe **2011**.
[147] M. A. Wood, A. P. Thompson, *J. Chem. Phys.* **2018**, *148*(24), 241721.
[148] L. Ward, R. Liu, A. Krishna, V. I. Hegde, A. Agrawal, A. Choudhary, C. Wolverton, *Phys. Rev. B* **2017**, *96*(2), 024104.
[149] L. Yang, S. Dacek, G. Ceder, *Phys. Rev. B* **2014**, *90*(5), 054102.
[150] O. Isayev, C. Oses, C. Toher, E. Gossett, S. Curtarolo, A. Tropsha, *Nat. Commun.* **2017**, *8*(1), 1.
[151] N. Yang, A. J. Medford, X. Liu, F. Studt, T. Bligaard, S. F. Bent, J. K. Nørskov, *J. Am. Chem. Soc.* **2016**, *138*(11), 3705.
[152] Z. W. Ulissi, A. J. Medford, T. Bligaard, J. K. Nørskov, *Nat. Commun.* **2017**, *8*(1), 1.
[153] R. Jinnouchi, R. Asahi, *J. Phys. Chem. Lett.* **2017**, *8*(17), 4279.
[154] Z. W. Ulissi, M. T. Tang, J. Xiao, X. Liu, D. A. Torelli, M. Karamad, K. Cummins, C. Hahn, N. S. Lewis, T. F. Jaramillo, K. Chan, J. K. Nørskov, *ACS Catal.* **2017**, *7*(10), 6600.
[155] W. J. Szlachta, A. P. Bartók, G. Csányi, *Phys. Rev. B* **2014**, *90*(10), 104108.
[156] C. Zeni, K. Rossi, A. Glielmo, Á. Fekete, N. Gaston, F. Baletto, A. De Vita, *J. Chem. Phys.* **2018**, *148*(24), 241739.
[157] A. Glielmo, C. Zeni, A. De Vita, *Phys. Rev. B* **2018**, *97*(18), 184307.
[158] J. Behler, *J. Phys. Condens. Matter* **2014**, *26*(18), 183001.
[159] N. Artrith, B. Hiller, J. Behler, *Phys. Status Solidi B* **2013**, *250*(6), 1191.
[160] N. Artrith, A. M. Kolpak, *Nano Lett.* **2014**, *14*(5), 2670.
[161] N. Artrith, A. M. Kolpak, *Comput. Mater. Sci.* **2015**, *110*, 20.
[162] R. Ouyang, Y. Xie, D. Jiang, *Nanoscale* **2015**, *7*(36), 14817.
[163] S. Chiriki, S. S. Bulusu, *Chem. Phys. Lett.* **2016**, *652*, 130.
[164] R. Modee, S. Agarwal, A. Verma, K. Joshi, U. D. Priyakumar, *Phys. Chem. Chem. Phys.* **2021**, *23*, 21995.
[165] Y. Pathak, K. S. Juneja, G. Varma, M. Ehara, U. D. Priyakumar, *Phys. Chem. Chem. Phys.* **2020**, *22*(46), 26935.
[166] A. Khorshidi, A. A. Peterson, *Comput. Phys. Commun.* **2016**, *207*, 310.
[167] A. P. Bartók, G. Csányi, *Int. J. Quantum Chem.* **2015**, *115*(16), 1051.
[168] N. Artrith, A. Urban, *Comput. Mater. Sci.* **2016**, *114*, 135.
[169] T. D. Huan, R. Batra, J. Chapman, S. Krishnan, L. Chen, R. Ramprasad, *NPJ Comput. Mater.* **2017**, *3*(1), 1.
[170] B. Kolb, L. C. Lentz, A. M. Kolpak, *Sci. Rep.* **2017**, *7*(1), 1.
[171] J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev, A. E. Roitberg, *J. Chem. Phys.* **2018**, *148*(24), 241733.
[172] H. Wang, L. Zhang, J. Han, E. Weinan, *Comput. Phys. Commun.* **2018**, *228*, 178.
[173] A. Mardt, L. Pasquali, W. Hao, F. Noé, *Nat. Commun.* **2018**, *9*(1), 1.
[174] T. Xie, J. C. Grossman, *Phys. Rev. Lett.* **2018**, *120*(14), 145301.
[175] D. Jha, L. Ward, A. Paul, W.-k. Liao, A. Choudhary, C. Wolverton, A. Agrawal, *Sci. Rep.* **2018**, *8*(1), 1.
[176] D. Jha, K. Choudhary, F. Tavazza, W.-k. Liao, A. Choudhary, C. Campbell, A. Agrawal, *Nat. Commun.* **2019**, *10*(1), 1.
[177] S. Curtarolo, W. Setyawan, G. L. W. Hart, M. Jahnatek, R. V. Chepulskii, R. H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, M. J. Mehl, H. T. Stokes, D. O. Demchenko, D. Morgan, *Comput. Mater. Sci.* **2012**, *58*, 218.
[178] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K. A. Persson, *APL Mater.* **2013**, *1*(1), 011002.
[179] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, C. Wolverton, *NPJ Comput. Mater.* **2015**, *1*(1), 1.
[180] A. P. Bartók, J. Kermode, N. Bernstein, G. Csányi, *Phys. Rev.* **2018**, *8*(4), 041048.
[181] L. Ward, A. Dunn, A. Faghaninia, N. E. R. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla, K. Chard, M. Asta, K. A. Persson, G. J. Snyder, I. Foster, A. Jain, *Comput. Mater. Sci.* **2018**, *152*, 60.
[182] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, A. Aspuru-Guzik, *ACS Cent. Sci.* **2018**, *4*(2), 268.
[183] K. Choudhary, B. DeCost, F. Tavazza, *Phys. Rev. Mater.* **2018**, *2*(8), 083801.
[184] C. Draxl, M. Scheffler, *J. Phys. Mater.* **2019**, *2*(3), 036001.
[185] B. Olsthoorn, R. M. Geilhufe, S. S. Borysov, A. V. Balatsky, *Adv. Quantum Technol.* **2019**, *2*(7–8), 1900023.
[186] J. Chapman, R. Batra, R. Ramprasad, *Comput. Mater. Sci.* **2020**, *174*, 109483.
[187] P. Kayastha, R. Ramakrishnan, *J. Chem. Phys.* **2021**, *154*(6), 061102.
[188] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, M. Scheffler, *Phys. Rev. Lett.* **2015**, *114*(10), 105503.
[189] E. J. Corey, *Q. Rev. Chem. Soc.* **1971**, *25*(4), 455.
[190] E. J. Corey, W. T. Wipke, R. D. Cramer III., W. J. Howe, *J. Am. Chem. Soc.* **1972**, *94*(2), 431.
[191] E. J. Corey, W. J. Howe, H. W. Orf, D. A. Pensak, G. Petersson, *J. Am. Chem. Soc.* **1975**, *97*(21), 6116.
[192] E. J. Corey, A. K. Long, T. W. Greene, J. W. Miller, *J. Org. Chem.* **1985**, *50*(11), 1920.
[193] E. J. Corey, W. T. Wipke, R. D. Cramer III., W. J. Howe, *J. Am. Chem. Soc.* **1972**, *94*(2), 421.
[194] W. L. Jorgensen, E. R. Laird, A. J. Gushurst, J. M. Fleischer, S. A. Gothe, H. E. Helson, G. D. Paderes, S. Sinclair, *Pure Appl. Chem.* **1990**, *62*(10), 1921.
[195] D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher, A. G. Doyle, *Science* **2018**, *360*(6385), 186.
[196] C. W. Coley, W. Jin, L. Rogers, T. F. Jamison, T. S. Jaakkola, W. H. Green, R. Barzilay, K. F. Jensen, *Chem. Sci.* **2019**, *10*(2), 370.
[197] M. H. S. Segler, M. P. Waller, *Chem. A Eur. J.* **2017**, *23*(25), 5966.
[198] M. H. S. Segler, M. Preuss, M. P. Waller, *Nature* **2018**, *555*(7698), 604.
[199] B. A. Grzybowski, S. Szymkuć, E. P. Gajewska, K. Molga, P. Dittwald, A. Wołos, T. Klucznik, *Chem* **2018**, *4*(3), 390.
[200] B. Liu, B. Ramsundar, P. Kawthekar, J. Shi, J. Gomes, Q. L. Nguyen, S. Ho, J. Sloane, P. Wender, V. Pande, *ACS Cent. Sci.* **2017**, *3*(10), 1103.
[201] H. Gao, C. W. Coley, T. Struble, L. Li, Y. Qian, W. H. Green, K. F. Jensen, *React. Chem. Eng.* **2020**, *5*, 367.

[202] H. Gao, T. J. Struble, C. W. Coley, Y. Wang, W. H. Green, K. F. Jensen, *ACS Cent. Sci.* **2018**, *4*(11), 1465.
[203] C. W. Coley, W. H. Green, K. F. Jensen, *Acc. Chem. Res.* **2018**, *51*(5), 1281.
[204] F. Häse, I. F. Galván, A. Aspuru-Guzik, R. Lindh, M. Vacher, *Chem. Sci.* **2019**, *10*(8), 2298.
[205] D. Caramelli, J. M. Granda, S. H. M. Mehr, D. Cambié A. B. Henson, L. Cronin, *ACS Cent. Sci.* **2021**, *7*(11), 1821.
[206] T. J. Struble, J. C. Alvarez, S. Brown, M. Chytil, J. Cisar, R. DesJarlais, O. Engkvist, S. A. Frank, D. R. Greve, D. J. Griffin, X. Hou, J. W. Johannes, C. Kreatsoulas, B. Lahue, M. Mathea, G. Mogk, C. A. Nicolaou, A. D. Palmer, D. J. Price, R. I. Robinson, S. Salentin, L. Xing, T. Jaakkola, W. H. Green, R. Barzilay, C. W. Coley, K. F. Jensen, *J. Med. Chem.* **2020**, *63*, 8667.
[207] P. Schwaller, T. Gaudin, D. Lanyi, C. Bekas, T. Laino, *Chem. Sci.* **2018**, *9*(28), 6091.
[208] W. Jin, C. Coley, R. Barzilay, T. Jaakkola, Predicting organic reaction outcomes with Weisfeiler-Lehman network. in *Advances in Neural Information Processing Systems*. Curran Associates, Inc., NY, USA **2017**, p. 2607.
[209] S. Mehta, S. Laghuvarapu, Y. Pathak, A. Sethi, M. Alvala, U. D. Priyakumar, *Chem. Sci.* **2021**, *12*(35), 11710.
[210] M. Goel, S. Raghunathan, S. Laghuvarapu, U. D. Priyakumar, *J. Chem. Inf. Model.* **2021**, *12*, 5815–5826.
[211] V. Chelur, U. D. Priyakumar, *ChemRxiv*. Cambridge **2021**. https://doi.org/10.33774/chemrxiv-2021-013gn
[212] V. Bagal, R. Aggarwal, P. K. Vinod, U. D. Priyakumar, *J. Chem. Inf. Model.* **2021**. https://doi.org/10.1021/acs.jcim.1c00600
[213] R. Aggarwal, A. Gupta, C. Vineeth, C. V. Jawahar, U. D. Priyakumar, *J. Chem. Inf. Model.* **2021**. https://doi.org/10.1021/acs.jcim.1c00799
[214] X. Yang, Y. Wang, R. Byrne, G. Schneider, S. Yang, *Chem. Rev.* **2019**, *119*(18), 10520.
[215] A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. R. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D. T. Jones, D. Silver, K. Kavukcuoglu, D. Hassabis, *Nature* **2020**, *577*(7792), 706.
[216] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, *Nature* **2021**, *596*, 583.
[217] R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. von Lilienfeld., *Sci. Data* **2014**, *1*, 140022.
[218] J. Hoja, L. M. Sandonas, B. G. Ernst, A. Vazquez-Mayagoitia, R. A. DiStasio Jr., A. Tkatchenko, *Sci. Data* **2021**, *8*(1), 1.
[219] J. S. Smith, R. Zubatyuk, B. Nebgen, N. Lubbers, K. Barros, A. E. Roitberg, O. Isayev, S. Tretiak, *Sci. Data* **2020**, *7*(1), 1.
[220] T. Fink, J.-L. Reymond, *J. Chem. Inf. Model.* **2007**, *47*(2), 342.
[221] L. C. Blum, J.-L. Reymond, *J. Am. Chem. Soc.* **2009**, *131*(25), 8732.
[222] L. Ruddigkeit, R. Van Deursen, L. C. Blum, J.-L. Reymond, *J. Chem. Inf. Model.* **2012**, *52*(11), 2864.
[223] M. Glavatskikh, J. Leguy, G. Hunault, T. Cauchy, B. Da Mota, *Aust. J. Chem.* **2019**, *11*(1), 69.
[224] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, V. Pande, *Chem. Sci.* **2018**, *9*(2), 513.
[225] H. Huo, M. Rupp. Unified representation for machine learning of molecules and crystals. *arXiv preprint* arXiv:1704.06439, **2017**.
[226] M. J. Willatt, F. Musil, M. Ceriotti, *Phys. Chem. Chem. Phys.* **2018**, *20*(47), 29661.
[227] J. S. Smith, B. T. Nebgen, R. Zubatyuk, N. Lubbers, C. Devereux, K. Barros, S. Tretiak, O. Isayev, A. E. Roitberg, *Nat. Commun.* **2019**, *10*(1), 1.
[228] F. Hou, W. Zhenyao, H. Zheng, Z. Xiao, L. Wang, X. Zhang, G. Li, *J. Phys. Chem. A* **2018**, *122*(46), 9128.
[229] O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko, K.-R. Müller, *Chem. Rev.* **2021**, *121*(16), 10142.

AUTHOR BIOGRAPHIES

Shampa Raghunathan obtained her Ph.D. in 2010 under the supervision of Prof. Guntram Rauhut at the University of Stuttgart. She then moved to Technical University of Munich as a post-doctoral fellow with Dr. Mathias Nest. Followed by another move to University of Basel in the group of Prof. Markus Meuwly, she returned to India and worked as a research scientist at International Institute of Information Technology, Hyderabad. Currently, she is an Assistant Prof. at Mahindra University. She is a recipient of the National Post Doctoral Fellowship (N-PDF), and Women Scientists Scheme-A (WOS-A) from the Department of Science and Technology, India. Her primary area of research is multiscale modeling and simulations of physicochemical processes in biology and chemistry applying composite computational chemistry and machine learning methods.

U. Deva Priyakumar received his Ph.D. degree from Pondicherry University/Indian Institute of Chemical Technology followed by a postdoctoral fellowship at University of Maryland at Baltimore. He is currently a Professor at International Institute of Information Technology, Hyderabad, where he heads the Center for Computational Natural Sciences and Bioinformatics. He currently is the Academic Head of IHub-Data, a Technology Innovation Hub on data driven technologies. His research interests are using computational chemistry tools to investigate chemical and biological systems/processes, and applications of modern artificial intelligence/machine learning techniques for molecular/drug design and healthcare. He has been the recipient of awards such as the Chemical Research Society of India Medal, Indian National Science Academy Young Scientist Medal, JSPS invitation Fellowship, and Innovative Young Biotechnologist Award.

How to cite this article: S. Raghunathan, U. D. Priyakumar, *Int. J. Quantum Chem.* **2022**, *122*(7), e26870. https://doi.org/10.1002/qua.26870