MACHINE LEARNING
Science and Technology

PURPOSE-LED PUBLISHING™ Part of the MACHINE LEARNING Series

PAPER • OPEN ACCESS

Atomic orbits in molecules and materials for improving machine learning force fields

IOP Publishing Mach. Learn.: Sci. Technol. 6 (2025) 035005 https://doi.org/10.1088/2632-2153/adea0b

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
28 March 2025

REVISED
22 May 2025

ACCEPTED FOR PUBLICATION
30 June 2025

PUBLISHED
16 July 2025

PAPER

Atomic orbits in molecules and materials for improving machine learning force fields

Anton Charkin-Gorbulin[1,2], Artem Kokorin[1], Huziel E Sauceda[3,4], Stefan Chmiela[5,6], Claudio Quarti[2], David Beljonne[2], Alexandre Tkatchenko[1] and Igor Poltavsky[1,*]

[1] Department of Physics and Materials Science, University of Luxembourg, L-1511 Luxembourg, Luxembourg
[2] Laboratory for Chemistry of Novel Materials, University of Mons,B-7000 Mons, Belgium
[3] Instituto de Física Universidad Nacional Autónoma de México Cd de México C.P. 04510, Mexico
[4] BASLEARN, BASF-TU joint Lab, Technische Universität Berlin, 10587 Berlin, Germany
[5] Machine Learning Group, Technische Universität Berlin, 10587 Berlin, Germany
[6] BIFOLD—Berlin Institute for the Foundations of Learning and Data, 10587 Berlin, Germany
[*] Author to whom any correspondence should be addressed.

E-mail: igor.poltavskyi@uni.lu

Keywords: force fields, machine learning, atomistic simulations, molecular dynamics, symmetry search, perovskites, graphene

Supplementary material for this article is available online

Abstract

The accurate representation of atoms within their environment forms the backbone of any reliable machine learning force field (MLFF). While modern MLFFs treat atoms of the same type as indistinguishable, their identities can be further resolved by accounting for the composition of their chemical environment. This can improve the parameterization of the MLFF model in chemically diverse systems. In this work, we introduce a novel, data-driven approach designed to find permutation symmetries in isolated and periodic systems, delivering key insights that enable the identification of atomic 'orbits', atoms that share consistent chemical and structural environments throughout the dataset. We demonstrate the effectiveness of the orbit representation by incorporating it into the kernel-based symmetric gradient-domain ML (sGDML) model and the equivariant message-passing neural network, MACE. For sGDML, trained on ethanol, 1,8-naphthyridine, D-alanine, and D-histidine adsorbed on graphene, we establish a strong correlation between force prediction accuracy and chemical diversity, quantified by orbit count. The results for the Ac-Phe-Ala5-Lys molecule further underscore the critical role of orbits in force reconstruction across various MLFF architectures. Incorporating orbits into MACE enables us to reduce the model size by an order of magnitude while preserving predictive accuracy, as demonstrated for the $CsPbI_3$ perovskite slab and graphene with a pyridinic-N defect. Overall, our approach provides a scalable and efficient solution for modeling complex chemical systems with state-of-the-art MLFFs.

1. Introduction

Machine learning (ML) force fields (FFs), trained on high-quality reference data, have become indispensable tools for modeling increasingly complex and large-scale systems-enabling applications that were inconceivable just two decades ago. Early approaches relied primarily on Gaussian process models combined with atomic environment descriptors such as SOAP and ACSF, offering a balance between predictive reliability and computational feasibility [1–5]. These methods laid the groundwork for kernel-based models such as GDML and its symmetric extension symmetric gradient-domain ML (sGDML), which achieved substantial improvements in accuracy for small- to medium-sized molecular systems by explicitly incorporating symmetries [6–10]. More recently, the development of invariant and equivariant neural network (NN) architectures-such as SchNet [11], PaiNN [12], NequIP [13], Allegro [14, 15], PhysNet [16], ACE-based models [17–19], and SO3krates [20, 21] -has markedly advanced the accuracy and

generalizability of MLFFs, extending their applicability to chemically diverse and structurally complex systems. State-of-the-art machine-learning FFs increasingly adopt integrated strategies that combine scalable fragment-based training with task-specific neural architectures. Unke *et al* [22] demonstrated broad chemical coverage by training a quantum FF on an extensive fragment set, Kabylda *et al* [23] pretrained an equivariant model for biomolecular simulations, and the two MACE variants [24, 25] achieved transferable accuracy across both organic molecules and inorganic materials. Other architectures embed domain knowledge: the extensible ANI potential [26] attains DFT-level accuracy for organic chemistry, AIMNet2 [27] extends this paradigm to charged and hybrid systems, feNNix [28] fuses classical force-field terms with an equivariant network to capture atom-in-molecule properties, and OrbNet [29] leverages orbital features for accelerated quantum predictions, while the kernel-based FCHL19 representation [30] and its GPU implementation QML-lightning [31] enable ultrafast training without sacrificing accuracy. These advanced MLFFs also capture both short- and long-range interactions with high fidelity. For extended solids, ALIGNN-FF [32], elemental-SDNNFF [33], FIREANN [34] and the deep potential framework [35, 36] embed explicit long-range physics, whereas the moment tensor potential [37] and SNAP [38] provide systematically improvable local many-body expansions. Universal graph potentials such as M3GNet [39] and the survey by Isayev and co-workers [40] highlight strategies for balancing locality and transferability; active-learning schemes further refine accuracy on-the-fly, as shown by Vandermause *et al* [41] and the scalable SevenNet algorithm [42]. Building on these capabilities, on-the-fly MLFFs are already transforming demanding simulations: Jinnouchi *et al* obtained entropy-driven phase transitions for hybrid perovskites within weeks instead of years by combining DFT with on-the-fly ML scheme [43]. Using a symmetry-incorporated kernel-based model, Sauceda *et al* successfully performed highly-accurate path integral molecular dynamics simulations to obtain novel insights into the localization of benzene–graphene dynamics induced by nuclear quantum effects [9, 10]. Moreover, new developments in ML algorithms facilitate the automatic identification of molecular building blocks, allowing for the creation of coarse-grained (CG) FFs, reproducing long-range interactions, and determining reaction coordinates [44–46].

One of the essential elements that enabled the successes mentioned above is exploiting symmetries present in the systems under study [2, 47–49]. Symmetries are crucial for QC calculations and ML models, particularly for complex systems. Traditional MLFFs typically rely on full symmetry groups to model atomic environments. However, recent advances suggest that strategically restricting these symmetry groups can simplify learning tasks. For example, atom embeddings with restricted permutation symmetry, derived from fully permutation-invariant NN representations (e.g. SchNet, PaiNN), have been shown to improve prediction accuracy [50].

In modern MLFFs, atoms of the same chemical element are typically treated as identical, permitting intricate structural and chemical transformations. However, in complex systems, atoms of the same element often participate in various interaction patterns, such as forming single and double bonds with neighboring atoms of differing types. In these cases, restricting the full symmetry group by defining identical particles based on shared chemical neighborhoods may be advantageous, thus simplifying the learning task for ML models. This approach aligns with strategies used for decades in empirical FFs, where, for example, the bond order of carbon atoms determines their interaction parameters [51, 52]. However, to achieve broad applicability, it is crucial to go beyond identifying symmetries tied to specific chemical compositions and geometries. Instead, symmetries should be defined across the entire relevant chemical and configurational space.

This work presents a novel data-driven symmetry search method that can reveal all relevant permutations in molecules and materials present in a given dataset. These permutations are used to differentiate atoms based on their surrounding chemical environments, allowing us to improve the accuracy of MLFFs by adjusting the level of symmetry exercised by the model. We start from a cloud of disconnected vertices that represent atoms and establish connections between them for all representative configurations in a given dataset using the van der Waals radii and interatomic angles. Only the most frequent links are used to build the final graph. In the case of the multicomponent system, the symmetry independence of separate components is analyzed by a CG procedure, and the final graph is modified accordingly. The final graph serves as an input for the graph automorphism search algorithm [53–57]. This algorithm produces three primary outputs: (i) the total number of permutational symmetries, (ii) a minimal set of symmetry operations sufficient to generate all permitted permutations of vertices, and (iii) the orbits—equivalence classes of the graph's vertices under the permitted permutations. Additionally, the constructed graph representation facilitates the analysis of local atomic environments, particularly useful in cases where reliance on global symmetries leads to overly complex or inadequate descriptions of the system.

We demonstrate the applicability of the developed approach on examples of a kernel-based model (sGDML) and a message-passing equivariant NN architecture (MACE) [9, 18, 19]. The sGDML model employs a global representation that treats the system as a whole rather than partitioning it into localized

atomic contributions. The model ensures permutational invariance by explicitly symmetrizing its predictions through integration over the permutation group of a system. Consequently, incorporating permutational symmetries into the model requires obtaining a comprehensive list of these symmetries. The current implementation uses a spectral graph-matching algorithm limited to non-periodic systems. Our method extends this approach to periodic systems such as those featuring organic molecules interacting with a graphene surface [58–61]. In order to achieve this, a number of unique challenges presented by the periodic graph isomorphism problem have to be addressed.

We investigate atomic force reconstruction heterogeneity using results from the TEA Challenge 2023, revealing how employing orbits—atoms with persistent neighborhood chemical compositions-provides new insights into force prediction errors for the Ac-Phe-Ala5-Lys molecule [62]. This analysis, which yields consistent results across various deep learning architectures and kernel-based ML models, including MACE, SO3krates, and sGDML models, highlights the limitations of classifying atoms solely by their elemental identity [7, 18, 19, 21].

To improve the accuracy and efficiency of MACE, we extend its architecture by explicitly incorporating atomic orbits-groups of atoms identified by permutational symmetry analysis as sharing equivalent chemical environments-as distinct atomic types. As a demonstration, we select two systems: the cesium lead iodide perovskite slab and graphene with pyridinic-N defect. We demonstrate that the proposed symmetry search method notably enhances FF reconstruction for complex systems when combined with the message-passing graph NN architecture. This benefits systems where atoms of the same species can be classified into multiple groups with varying neighborhood chemical compositions. For example, for graphene with a pyridinic-N defect, the presence of the defect produces different carbon environments based on proximity to the defect site, while in the $CsPbI_3$ perovskite slab system, the presence of surfaces and internal layers leads to different atomic environments for the same atomic species.

The article's structure is the following: section 2 provides a comprehensive breakdown of each step within the proposed graph-based symmetry search algorithm. In section 3, we present the results for the sGDML model trained for four organic molecules adsorbed on a graphene substrate, followed by an analysis of atomic force reconstruction heterogeneity based on TEA Challenge 2023 results for the Ac-Phe-Ala5-Lys molecule. Lastly, in the section, we explore the performance of the MACE model incorporating permutational symmetries, applied to a graphene system featuring a pyridinic-N defect and the $CsPbI_3$ perovskite slab. Section 4 concludes our study.

2. Methodology

The representation of a system as a graph, where the vertices represent atoms and the edges represent the links between them, is widely used in chemistry and crystallography [63, 64]. The permutational symmetries of the system form the graph automorphism group, i.e. they preserve the edge-vertex connectivity. Consequently, the graph representation reduces the problem of finding symmetries in a system to finding elements of an automorphism group, a well-studied problem in the field of graph theory [65, 66]. However, building a representative graph encompassing the symmetries in a given relatively complex system along its finite temperature evolution is not straightforward. This work proposes a procedure for constructing such graph representation that describes an entire dataset of configurations rather than one particular structure.

The first step establishes edges between individual atoms at distances smaller or equal to their respective elements' van der Waals radii to account for all physically and chemically relevant neighbors for each individual configuration, as shown in figure 1(I). This provides a system representation resembling chemical bonds in which neighboring atoms become directly connected.

Some edges created during the first step form *sharp angles* that are not chemically or physically meaningful. Therefore, the second step is to remove the longer of the two edges that form such angles, as depicted in figure 1(II). The threshold defining sharp angles differs from system to system and should be set manually. We found that taking this threshold as half of the smallest angle formed by translation vectors of a super-cell in the case of a periodic system is an excellent choice to obtain a structural-informative graph. After performing the first two steps, a graph representation is formed for each system configuration at our disposal. In practice, we can use only a limited number of representative structures, making the method easily applicable for large datasets.

During the third step, we sum the adjacency matrices of the constructed graphs, obtaining the frequencies of edge formations. Only the edges whose frequencies are above some predefined threshold become a part of the candidate for the *final graph*, as illustrated in figure 1(III). This eliminates the edges that are not characteristic of most configurations. The eliminating threshold value varies from system to system and depends on both the softness of the system and the presence of transient states and should be set

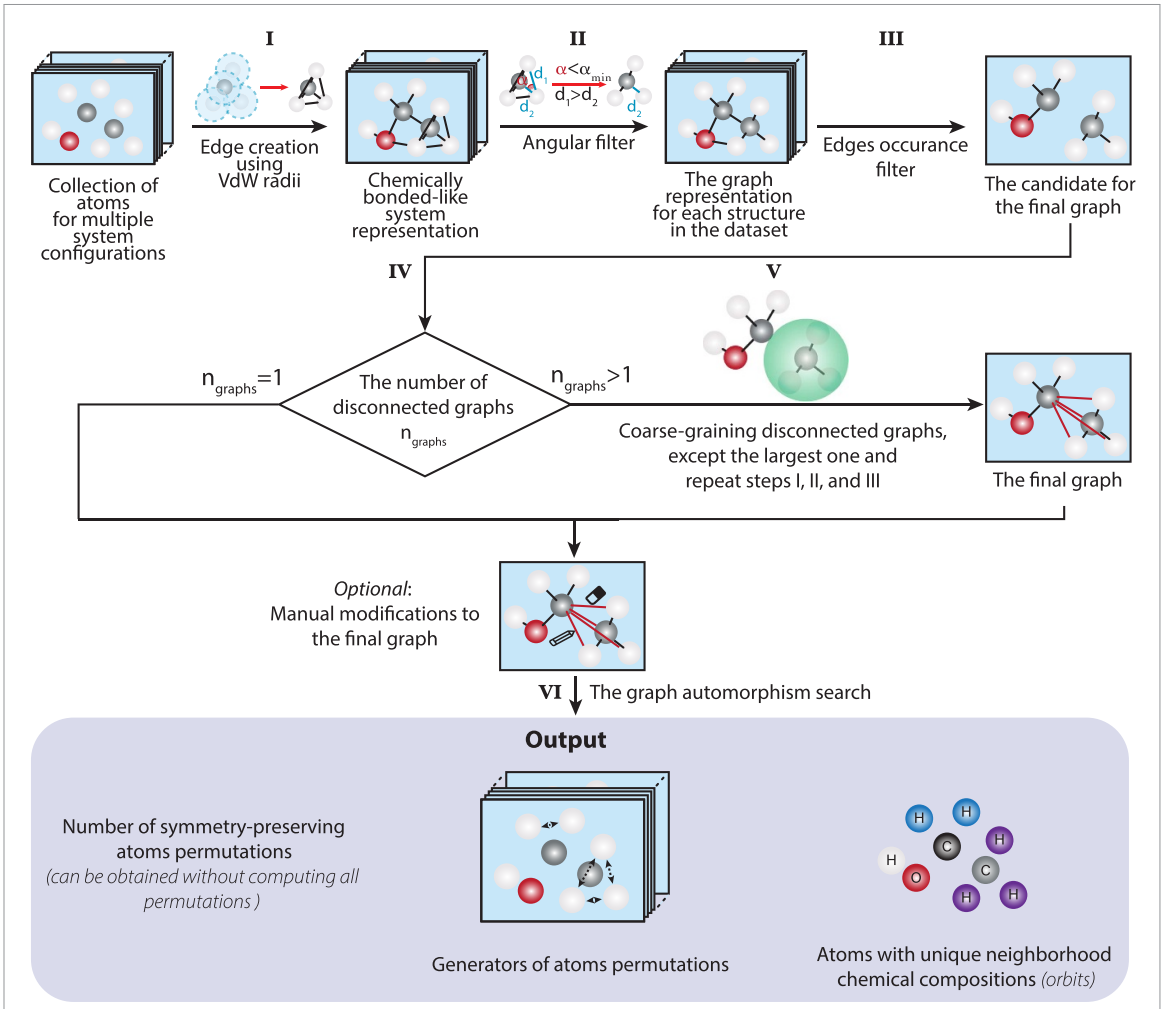

Figure 1. An illustrative description of the main components of the graph-based symmetry search procedure. The procedure consists of six steps: (I) creating of edges between atoms using van der Waals radii as cutoff; (II) filtering out the longest of two created edges if the formed angle between them is greater than the threshold angle; (III) filtering edges by occurrence frequency in the dataset; (IV) estimating the number of disconnected components; (V) coarse-graining the disconnected components; (VI) performing the graph automorphism search.

manually. A good choice is between 70% and 90%. After the threshold is applied, the candidate for the final graph is constructed.

As shown in figure 1(IV), the fourth step is determining the number of non-connected components in the candidate's final graph. It is common for these components to arise from subsystems either distant from one another or moving quickly enough to prevent them from forming stable edges. If the candidate for the final graph has no separate components, it is considered the final graph. At this point, one can proceed directly to the search for its automorphisms.

However, if the candidate for the final graph contains disconnected subgraphs, a fifth step, shown in figure 1(V), is necessary to analyze the relationships between them. This step begins by replacing all fragments-represented by disconnected subgraphs, except the largest-with CG particles positioned at each fragment's center of mass. The initial four steps are then repeated on this semi-CG system dataset to create a new graph with CG vertices corresponding to CG particles. The radius for edge creation around each CG particle is defined as the distance from its center of mass to its most distant atom plus the van der Waals radius of that atom. Once these steps are complete, each CG vertex is replaced by its original disconnected subgraph, with new connections added to each subgraph's vertices based on the CG vertex links. By the end of the fifth step, any previously missed interatomic connections are re-established, and the distinct fragments are identified.

Whether the final graph was obtained after the fourth or the fifth step, it can be modified manually before searching for symmetries. Manual modifications can offer advantages in systems like the chemisorbed Pd/MgO-surface, see figure 2. In this scenario, when a palladium atom becomes trapped in one of the numerous local minima, it can introduce bias into the final graph. Accordingly, the biased final graph does not have information associated with other local minimum configurations missing in the dataset. However,

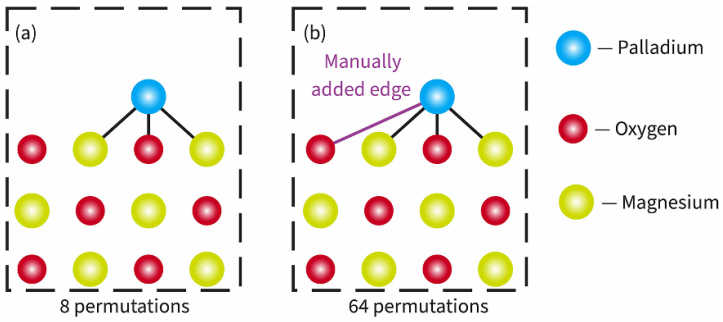

Figure 2. The example of manual modifications of the final graphs for the chemisorbed Pd/MgO-surface. All edges of the final graph that are not connected to the palladium atom are hidden in both (a) and (b). (a) Illustration of the final graph before the manual modifications, where permutational symmetries are biased toward single local minima. (b) Illustration of the final graph after the manual modifications, where permutational symmetries take into account other possible locations of a palladium atom.

manual modifications applied to the final graph can help to restore the absent information in the trajectory to train a robust MLFF model. Therefore, being an optional step, the manual modifications can both compensate for insufficient configurational space sampling and include physical and chemical intuitions.

The group theory-based graph automorphism search algorithm implemented in the `nauty` package efficiently determines the total size of the permutation group and its generators without requiring the explicit generation of all possible permutations [67]. Moreover, the information on orbits—the equivalence classes of the graph's vertices under the action of the automorphisms—can be obtained without additional computations. Different orbits represent atoms with distinct *global* neighborhood chemical compositions. Compared to the number that can be expected for a given system, a large number of different orbits reveals that the dataset is insufficiently representative or that a transition state is present, requiring careful construction of the dataset to capture its behavior accurately.

For the final multicomponent graph, the graph automorphism search is applied to each component separately to examine subsystem symmetries. As the number of components increases, the total number of symmetries grows exponentially and equals the product of the number of symmetries of each non-connected part. On the other hand, the total number of orbits shows a linear growth and is equal to the sum of the number of atoms with distinct neighborhood chemical compositions in each component. For instance, the system consisting of a 7×7 graphene sheet and 1,8-naphthyridine molecule has a total of 1176 permutational symmetries. Specifically, 588 of these symmetries come from graphene, while two come from the 1,8-naphthyridine molecule. However, this system has ten distinct neighborhood chemical compositions, consisting of one orbit from graphene and nine orbits from the 1,8-naphthyridine molecule.

The presence of local symmetry perturbations in the final graph representation reduces the number of permutational symmetries while inflating the number of orbits. Focusing on local neighborhood chemical compositions-rather than relying exclusively on global orbits-can be more advantageous. The extent of these environments is determined by the number of nearest neighbors considered, a choice guided by the range of relevant interactions in the system. In the subsequent section, we will showcase a graphene system with a pyridinic-N defect to illustrate this scenario.

The sGDML architecture enforces permutation invariance by explicitly enumerating all symmetry-equivalent atomic permutations provided by the proposed method during dataset preprocessing. Incorporating orbit information into an MPNN also requires no changes to the model architecture. The modification is applied at the data level: each atomic species label is replaced with its corresponding orbit identifier, effectively redefining species based on neighborhood chemical compositions.

Although the graph-based symmetry search procedure was discussed in the context of data produced by QC methods, it can be applied to any data represented by atomic species and their coordinates, including those sourced from an experiment.

3. Results

3.1. Organic molecules absorbed on graphene

sGDML is a global kernel ridge regression model designed for efficient FF reconstruction of extended systems [9]. It leverages permutational symmetries to enhance data efficiency and employs a periodic Coulomb descriptor to capture system periodicity accurately. To demonstrate the capabilities of the developed graph-based symmetry search algorithm in combination with global kernel model, we selected

Table 1. Accuracy of the sGDML model, along with the number of permutational symmetries and atoms with distinct neighborhood chemical compositions (orbits), for molecule/graphene systems. The models were trained on 400 configurations and tested on 12 000 configurations for each system.

		# Permutational symmetries		Force $\left(\frac{\text{kcal}}{\text{mol}\cdot\text{Å}}\right)$		Energy $\left(\frac{\text{kcal}}{\text{mol}}\right)$	
	# Orbits	Graphene	Molecule	MAE	RMSE	MAE	RMSE
Ethanol/graphene	7	588	6	0.31	0.54	0.53	0.67
1,8-naphthyridine/graphene	10	588	2	0.45	0.72	0.61	0.78
D-alanine/graphene	11	588	6	0.55	1.07	1.07	1.39
D-histidine/graphene	19	588	4	0.59	1.20	1.35	1.92

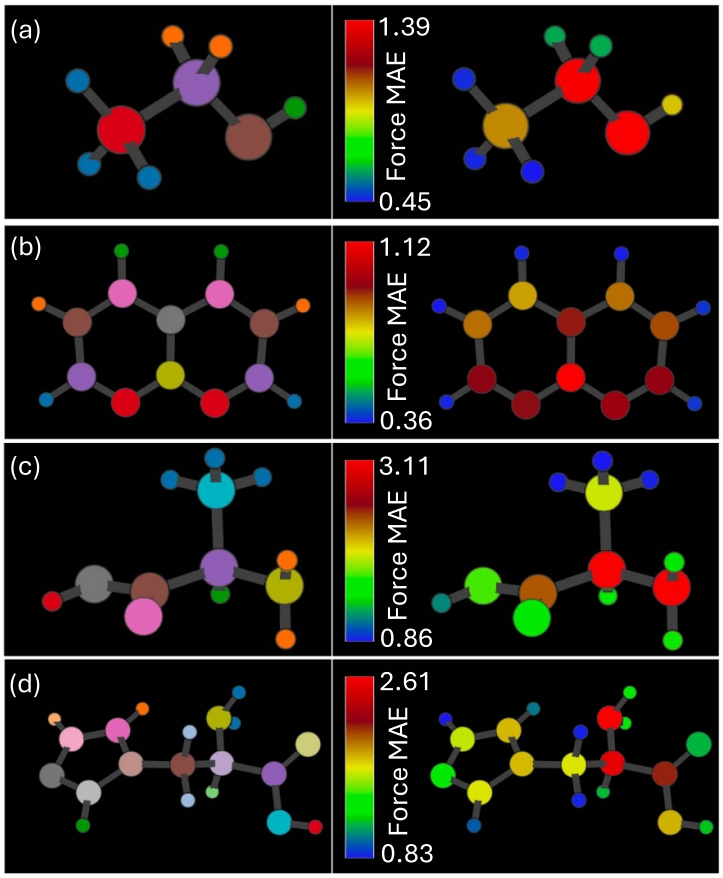

Figure 3. 3D visualizations of molecule/graphene systems, where the first column shows atoms colored according to orbit classifications, reflecting distinct neighborhood chemical compositions, and the second column shows atoms colored by their mean average force error in kcal $(\text{mol}\cdot\text{Å})^{-1}$ as predicted by the sGDML models. Figures correspond to: (a) ethanol, (b) 1,8-naphthyridine, (c) D-alanine, and (d) D-histidine molecules in the respective molecule/graphene systems.

four periodic systems in which a 7×7 graphene surface interacts with various organic molecules: ethanol, 1,8-naphthyridine, D-alanine, and D-histidine (details provided in the supplementary information). For each system, we extracted 400 configurations from DFT-generated molecular dynamics trajectories to identify permutational symmetries and integrate them into the sGDML model in closed form training process [68–70].

Table 1 presents the mean absolute error (MAE) and root mean squared error (RMSE) values for the sGDML models applied to molecular systems on graphene, along with the number of orbits and permutational symmetries [9]. The results show that the force MAE remains below 1 kcal $(\text{mol}\cdot\text{Å})^{-1}$ across all systems. While both permutational symmetries and orbits influence the accuracy of sGDML models, the impact of orbits appears to be more significant. Notably, the MAE strongly correlates with the number of orbits, representing atoms in distinct neighborhood chemical compositions. Furthermore, the increasing disparity between RMSE and MAE for forces with a growing number of orbits suggests that atomic force prediction errors are highly heterogeneous, likely due to variations in interaction patterns for different orbits.

Figure 3 presents the molecules with atoms color-coded according to two criteria: their classification into orbits, representing distinct neighborhood chemical compositions, and their individual force MAE, averaged

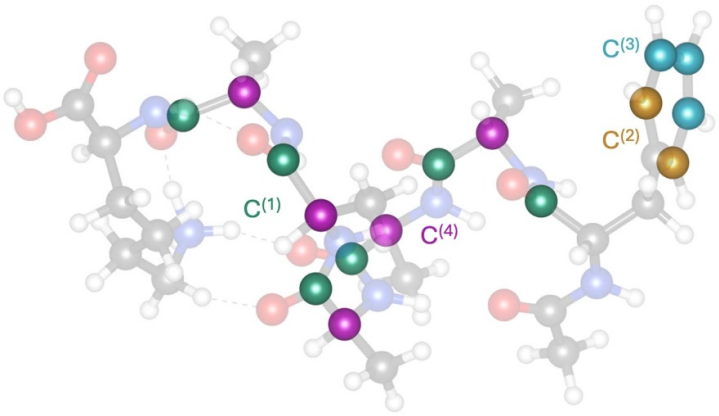

Figure 4. A snapshot of the Ac-Phe-Ala5-Lys molecule geometry. $C^{(1)}$–$C^{(4)}$ represent sets of carbon atoms belonging to four different orbits.

over the entire test set [71, 72]. As illustrated by the example of hydrogen atoms in ethanol shown in figure 3(a), three distinct types of hydrogen atoms can be identified based on their neighborhood chemical compositions. Hydrogen atoms in the R-CH_3 group share the same neighborhood chemical composition and thus belong to the same orbit (blue). In contrast, hydrogen atoms bonded to the adjacent carbon (orange) or to the oxygen atom (green) form distinct orbits, reflecting differences in their local chemical surroundings. Although these three types of hydrogen atoms exhibit intrinsically similar MAEs, their values differ depending on their specific neighborhood chemical compositions, as highlighted by the orbit-based classification. The impact of neighborhood chemical compositions on the distribution of errors is particularly evident in 1,8-naphthyridine, shown figure 3(b). Reflection symmetry inherent to the molecule ensures that symmetrically positioned atoms share identical neighborhood chemical compositions, resulting in similar MAEs. Consistent conclusions are observed when applying similar analyses to each atom in the other systems examined.

Overall, the analysis suggests that while the sGDML model accounts for symmetries by considering the complete set of all possible permutations in the system—without explicitly distinguishing orbits—integrating orbit-based approaches could enhance performance in architectures capable of leveraging this distinction.

3.2. Ac-Phe-Ala5-Lys

To further explore the influence of orbits on atomic force reconstruction across various chemical systems and MLFF architectures, we analyze the results from the TEA Challenge 2023 [62]. The TEA Challenge 2023 is a benchmarking initiative that evaluates MLFF performance across diverse molecular systems, systematically assessing models with various architectures such as MACE, SO3krates, sGDML, SOAP/GAP, and FCHL19* [1, 6–8, 18–21, 30]. The test systems included two biomolecular structures (alanine tetrapeptide and N-acetylphenylalanyl-pentaalanyl-lysine), a 1,8-naphthyridine/graphene interface, and a methylammonium lead iodide perovskite. The challenge focused on key tasks, including the reconstruction of potential energy surfaces, the handling of incomplete reference data, the modeling of multi-component systems, and the simulation of complex periodic structures. The results revealed pronounced heterogeneity in atomic force prediction errors, with worst-to-best atomistic MAE ratios ranging from 5:1 to 6:1, consistently observed across different architectures. Here, we selected the Ac-Phe-Ala5-Lys molecule (100 atoms) to investigate the origin of this heterogeneity using orbits identified through the developed permutational symmetry search method.

Since orbits are defined as sets of vertices that remain indistinguishable under all graph automorphisms, applying graph automorphism analysis directly to complex molecules can lead to an excessive number of chemically distinct environments, even for atoms that should be treated identically based on chemical intuition. Distinguishing atomic species based on their local neighborhood chemical compositions is more effective than relying solely on global permutational symmetry in such systems. Therefore, while orbits traditionally refer to global permutations, hereafter, we extend the term to include atoms characterized by distinct local environments.

Figure 4 presents a molecular snapshot highlighting four sets of carbon atoms that form orbits with identical neighborhood chemical compositions up to the second nearest neighbors, considering only orbits containing multiple atoms. Table 2 reports the MAEs for atomic force reconstruction using two MPNN architectures, MACE and SO3krates, along with the kernel-based sGDML model. The errors are evaluated on

Table 2. Normalized atomistic force MAE (in %) for Ac-Phe-Ala5-Lys molecule. Orbits correspond to carbon atoms highlighted in figure 4. For details on the MLFF architectures and different test sets (complete (com), incomplete (incom), and unknown (unkn)), refer to reference [62].

	MACE			SO3krates			sGDML		
Orbit	com	incom	unkn	com	incom	unkn	com	incom	unkn
$C^{(1)}$	0.85	0.74	1.08	2.22	2.10	2.79	8.40	7.44	13.42
	0.73	0.63	0.99	1.98	1.80	2.59	8.11	7.07	13.35
	0.72	0.63	0.94	1.87	1.72	2.41	8.24	7.21	13.78
	0.68	0.61	0.82	1.74	1.64	2.05	7.92	7.14	12.03
	0.71	0.64	0.90	1.90	1.73	2.37	8.04	7.10	13.37
	0.83	0.72	1.11	2.28	2.13	2.89	8.30	7.33	13.38
$C^{(2)}$	0.50	0.45	0.60	1.25	1.22	1.46	5.55	4.91	8.70
	0.51	0.45	0.59	1.24	1.21	1.45	5.53	4.89	8.62
$C^{(3)}$	0.40	0.35	0.44	0.98	0.97	1.09	5.06	4.46	7.70
	0.39	0.34	0.44	0.96	0.93	1.10	5.00	4.37	7.84
	0.38	0.33	0.41	0.88	0.85	0.93	4.95	4.41	7.48
$C^{(4)}$	0.65	0.57	0.79	1.73	1.62	2.13	8.62	7.72	14.05
	0.63	0.55	0.76	1.66	1.56	2.00	8.45	7.47	14.40
	0.58	0.52	0.68	1.51	1.43	1.73	8.16	7.27	13.08
	0.61	0.55	0.74	1.62	1.53	1.87	8.43	7.39	14.05
	0.69	0.61	0.84	1.86	1.74	2.27	8.51	7.47	13.90

the same test set and model prediction data used to generate table 1 in [62]. To account for variations in force magnitudes across the molecule, MAEs are expressed as percentages of the averaged norm of the true force acting on an atom. Carbon atoms within the same orbit are grouped together, revealing a clear pattern: relative atomic force MAEs remain highly consistent within each orbit but differ significantly across different orbits. Similar results can be observed for absolute atomic force MAE values, as well as for other chemical elements and systems benchmarked in the TEA Challenge 2023.

This trend aligns with our findings for smaller molecules on graphene in the previous subsection, further emphasizing the limitations of treating all atoms of a given chemical element identically in ML models, despite their diverse bonding environments.

3.3. Graphene with pyridinic-N defect

Beyond providing insight into the origin of force reconstruction heterogeneity observed across all ML models and chemical systems analyzed, orbits can also be directly leveraged to improve the performance of state-of-the-art MLFF architectures. As a demonstration, we considered a 15×15 graphene lattice containing a pyridinic-N defect, shown in figure 5(a), which disrupts the symmetry among carbon atoms. For this study, we choose the MACE model, an equivariant MPNN, and employ orbits (atoms with persistent neighborhood chemical compositions) extracted using the developed graph-based symmetry search algorithm to enhance training efficiency and predictive performance. [18, 19] The standard MACE architecture has three hyperparameters that are critical for accuracy and ultimately define the model's size:

- The maximum order of irreducible representations (irreps) that determine the complexity of the internal feature space;
- The number of channels, i.e. features belonging to each hidden irreps, indicates the dimensionality of the internal representation at a given symmetry level.
- The design of the multilayer perceptron (MLP) that defines the radial resolution of the model.

The default configuration for the MACE model incorporates hidden representations up to the second order, with each order containing 128 features (`128x0e + 128x1o'). Additionally, the radial MLP comprises three fully connected layers, each with 64 neurons. However, for experiments with models' sizes that were changed with by varying channel counts, the radial MLP consists of three fully connected layers, each with only 8 neurons, since that was enough to obtain error below 1 kcal $(\text{mol}\cdot\text{Å})^{-1}$. All MACE models were trained under identical conditions, employing early stopping on the validation loss with a patience of 2048 epochs to ensure convergence.

As an equivariant model, MACE captures symmetries by translating discrete permutation symmetry into a continuous representation governed by rotational equivariance. This approach enables efficient parameterization through the use of irreps of the rotation group $SO(3)$. However, this conversion also introduces a limitation: the model is constrained to operate within the full rotation group and cannot

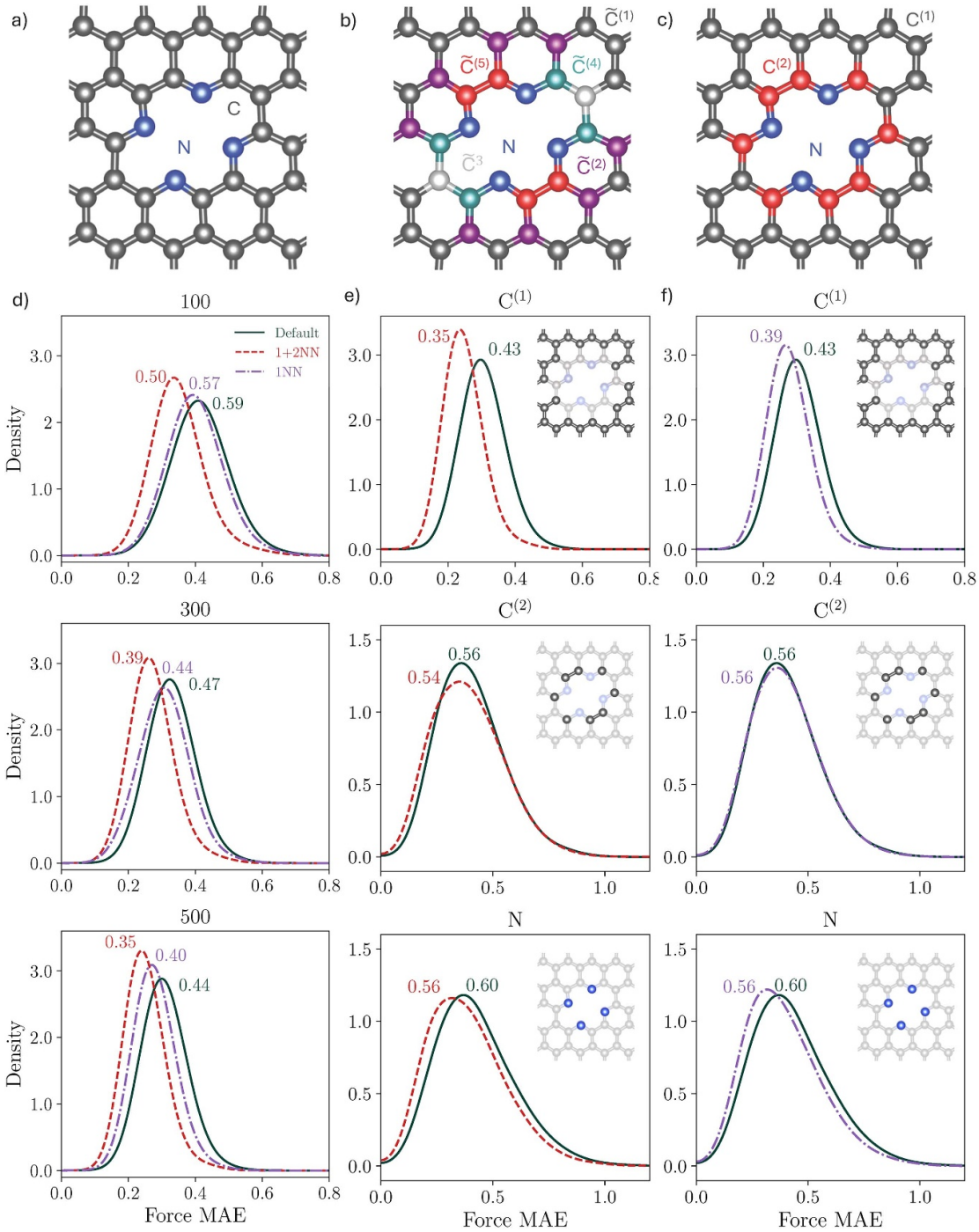

Figure 5. Graphene with a pyridinic-N defect system is represented using three approaches: (a) default, (b) first- and second-neighbor (1+2NN), and (c) first-neighbor structural encoding (1NN). (d)–(f) Distributions of mean absolute error (MAE) in force predictions, in kcal $(\text{mol}\cdot\text{Å})^{-1}$, for MACE models applied to the graphene system with pyridine-type defect. Solid green lines represent the default MACE architecture, red dashed lines indicate the first and second-neighbor-based model, and purple dashed-dotted line denote the first-neighbor-based model. (d) Force MAE distributions calculated for all atoms in the system, for the MACE models trained on 100, 300, and 500 reference geometries. (e) and (f) Force MAE distributions specifically for nitrogen, carbons bonded to nitrogen, and other carbon atoms in the system using MACE models trained on 500 configurations, where default MACE architecture compared with (e) the first and second-neighbor-based MACE model, and with (f) the first-neighbor-based MACE model.

explicitly emphasize specific symmetry subgroups, such as particular atom permutations. By preserving a direct encoding of permutation symmetry, it becomes possible to treat both discrete permutation and continuous rotational symmetries explicitly and independently, offering greater flexibility and potentially improved inductive bias in the model. By maintaining an explicit connection to permutation symmetry-specifically by partitioning atoms into equivalence classes based on their local chemical environments-the model can separately and directly encode both discrete permutation and continuous rotational symmetries.

In this framework, permutation invariance within each equivalence class is enforced *a priori*, allowing the network to concentrate its capacity on learning rotational equivariance within each class. This approach offers finer control over symmetry treatment and improves both the accuracy and efficiency of MACE. By organizing atoms into classes with stable neighborhood chemical environments, the model reduces hypothesis space complexity, accelerates convergence, and lowers generalization error. Crucially, it focuses the network on learning meaningful geometric correlations rather than rediscovering redundant permutation symmetries.

To further explore how these equivalence classes influence model performance, we assessed the impact of different orbit-based structural encodings on MACE's performance by constructing three datasets containing 200, 600, and 1000 configurations each. For every dataset, we trained three types of models using standard hyperparameter settings: one with default atomic species (formally classified as a 0NN model) and two incorporating 1NN and $1+2$NN structural encodings. Training process followed a 50/50 split between training and validation datasets.

The default structural encoding of the system shown in figure 5(a) distinguishes only carbon and nitrogen atoms. It is then refined by adjusting the number of nearest neighbors that define each atom's neighborhood chemical composition, capturing a more detailed view of local symmetries. For example, including first and second neighbors into consideration ($1+2$NN) yields four additional inequivalent orbits for carbon. Specifically, $\tilde{C}^{(1)}$ is unaffected by the defect, $\tilde{C}^{(2)}$ has one nitrogen atom among its second neighbors, $\tilde{C}^{(3)}$ has two nitrogen atoms among its second neighbors, $\tilde{C}^{(4)}$ has one nitrogen atom as a first neighbor, and $\tilde{C}^{(5)}$ has one nitrogen atom as a first neighbor plus another as a second neighbor. In contrast, figure 5(c), which considers only first-neighbor structural encoding (1NN), introduces just two inequivalent orbits for carbon: $C^{(1)} = \tilde{C}^{(1)} \cup \tilde{C}^{(2)} \cup \tilde{C}^{(3)}$, unaffected by the defect, and $C^{(2)} = \tilde{C}^{(4)} \cup \tilde{C}^{(5)}$, which has one nitrogen atom as a first neighbor.

The MACE models that were trained on default (0NN), 1NN, and $1+2$NN structural encodings use a 5 Å cutoff and two interaction blocks, effectively increasing their effective interaction range to 10 Å. In the considered defective graphene system, every atom's third-nearest neighbors lie within 4 Å. Thus, orbit partitioning does not extend the model's spatial reach under any encoding, but simply reorganizes the same local information into symmetry-informed classes, letting the network focus on geometric features rather than relearning permutation invariance.

Figure 5(d) illustrates the force MAE distributions predicted for 0NN, 1NN, and 1+2NN models, each trained on 100, 300, and 500 structures, respectively. The force MAE distributions for graphene with a pyridinic-N defect display a single Gaussian-like peak across all dataset sizes. For 100 training points, the default model yields a force MAE distribution peaking around 0.40 kcal $(\text{mol}\cdot\text{Å})^{-1}$ with RMSE of 0.59 kcal $(\text{mol}\cdot\text{Å})^{-1}$. In contrast, the 1NN and $1+2$NN models produce noticeably narrower distributions that peak near 0.38 and 0.30 kcal $(\text{mol}\cdot\text{Å})^{-1}$, with corresponding RMSE values of 0.57 and 0.50 kcal $(\text{mol}\cdot\text{Å})^{-1}$, respectively. As the number of training points increases, the relative performance ranking of the models remains consistent, while the error distributions shift toward smaller values. With 500 training points, the RMSEs decrease to 0.35, 0.40, and 0.44 kcal $(\text{mol}\cdot\text{Å})^{-1}$ for the $1+2$NN, 1NN, and default MACE models, respectively.

To go deeper into the effects of different structural encodings on specific subsets of atoms, we separately analyze the precision of the predicted forces for nitrogen atoms (N), carbons bonded to nitrogen ($C^{(2)}$), and all other carbons ($C^{(1)}$). Figures 5(e) and (f) show the force MAE distributions for those atoms, illustrating the influence of the local pyridinic-N defect on the performance of MACE models when trained with 500 configurations. By examining these subsets individually, one finds that orbits-based models outperform the default one most noticeably for $C^{(1)}$—the largest subset of atoms. For $C^{(1)}$, the MAE distributions of the models narrow and shift toward smaller errors, reducing the RMSE from 0.43 (0NN) to 0.39 (1NN) and 0.35 ($1+2$NN) kcal $(\text{mol}\cdot\text{Å})^{-1}$.

However, the MAE distributions and RMSE for the $C^{(2)}$ subset remain largely unchanged in both orbit-based models, likely due to its role as a 'force outlier' for carbon atoms. By employing orbits, the larger $C^{(1)}$ subset is effectively decoupled from smaller and chemically distinct $C^{(2)}$, allowing orbit-based models to significantly enhance force reconstruction for $C^{(1)}$. Similarly, nitrogen atoms also exhibit improved force predictions when orbit-based models are used. This can be attributed to the fact that, in the default MACE model, the nitrogen subset initially has the highest error, 0.60 kcal $(\text{mol}\cdot\text{Å})^{-1}$, making it more responsive to orbit-based encoding. As a result, its MAE distributions improve, ultimately aligning the nitrogen atoms and $C^{(2)}$ RMSEs at 0.56 kcal $(\text{mol}\cdot\text{Å})^{-1}$.

These findings indicate that orbits with the greatest impact on the loss function-typically those containing the largest subset of atoms or exhibiting the highest errors-experience the most substantial improvements.

To validate that the $1+2$NN orbit-based model accurately captures key physical properties, we compare the Γ-point phonon spectrum of the system with that obtained from the default MACE model. Although the

default model has approximately 1.6 times more parameters, both models achieve nearly identical force RMSE and MAE. As shown in figure SI 2, the phonon frequencies at the Γ-point closely align across all vibrational modes, demonstrating that the orbit-based model reproduces vibrational behavior with accuracy comparable to the default, despite its significantly reduced complexity.

3.4. $CsPbI_3$ perovskite slab

To validate our conclusions regarding the advantages of incorporating orbits in MLFFs across different system sizes, we extend our analysis to a larger system-a $CsPbI_3$ perovskite slab with 216 atoms per unit cell. As shown in figure 6(a), the $CsPbI_3$ perovskite slab exhibits only reflective symmetry along the z-axis, resulting in three distinct, non-equivalent layers of PbI_2 and CsI: the outer layer, the in-between layer, and the middle layer. Consequently, despite containing only three chemical elements, the system features 12 distinct orbits defined by global permutational symmetry-six for iodine, three for cesium, and three for lead-providing a more granular structural representation for MACE models. By enforcing these orbits, the orbit-based variant effectively extends its receptive field beyond the nominal 13 Å with 6.5 Å cutoff, allowing it to capture longer-range structural information that the default model cannot access.

To assess the impact of orbits on data efficiency and model accuracy, we assembled three DFT-based datasets, comprising 600, 1000, and 2000 configurations each [73–79]. Subsequently, two MACE models with standard hyperparameter settings were trained for each dataset—one utilizing default atomic species and the other incorporating orbits. The training process used a 50/50 split for training and validation subsets. For experiments on model size, four distinct models were trained for each model type and channel count using a cross-validation procedure on 2000 configurations, with a 50/50 split between training and validation datasets. Additionally, early stopping on the validation loss with a 2048-epoch patience was used to guarantee full convergence for all models. All obtained ML models were tested on the test sets consisting of 8000 configurations, ensuring the stability and reliability of our results.

Figure 6(b) illustrates the distribution of the MAE of forces predicted on the same test set by the MACE models trained on 300, 500, and 1000 structures, respectively. For all training set sizes except 1000, two distinct peaks are shown in the distribution. The left and right peaks in the low and high MAE regimes highlight the existence of properly sampled and under-sampled regions of the configuration phase space, respectively. For 300 training points, the right-side peak is centered at 0.54 kcal $(\text{mol}\cdot\text{Å})^{-1}$ for the model with default species and thus has slightly better accuracy for those under-sampled regions than the model with orbits, which shows a broader right-side peak centered at 0.58 kcal $(\text{mol}\cdot\text{Å})^{-1}$. The difference between the two models is highlighted in well-sampled regions of configurational space. For the left-side peak, the model with orbits has higher accuracy than the default model with a peak centered around 0.23 kcal $(\text{mol}\cdot\text{Å})^{-1}$ as opposed to 0.30 kcal $(\text{mol}\cdot\text{Å})^{-1}$. In total, the two models present similar overall accuracies, with the inclusion of orbits favoring the well-sampled regions.

With the increasing number of training points, the right-side peak vanishes faster for the models with orbits than for those with default species, pushing the overall accuracy in the orbits' favor. At 1000 training points, both models demonstrate a single peak distribution. The MACE model with orbits is narrower, with a peak of around 0.08 kcal $(\text{mol}\cdot\text{Å})^{-1}$ compared to the MACE model with default species presenting a broader distribution and a peak at around 0.18 kcal $(\text{mol}\cdot\text{Å})^{-1}$.

Figure 6(c) depicts the distributions of the MAE across configurations for three lead orbits separately. This analysis pertains to both types of models trained on 500 configurations, thereby showcasing the individual contribution of each lead's orbit to the overall force MAE distribution. Including orbits in the model yields more significant improvements in accuracy for both the outer and in-between layers than for the middle layer. This difference is attributed to the middle layer having two times fewer lead instances than other layers. Other chemical elements within the system show similar trends and distributions.

In MPNN models, the increase in atom types correlates with a rise in trainable parameters. However, our findings demonstrate that the observed improvement of MACE models within the proposed approach is more than just a result of an increased number of parameters. Figure 7(a) illustrates the accuracy of these models as a function of the number of channels, which represent the dimensionality of the internal representation at a specific symmetry level. Each model type and channel count underwent training and testing on four distinct datasets, with the resulting variability in accuracy depicted as error bars. Graphs for all model types exhibit linear trends in the logarithmic scale, where the models treating all atoms as the same species have the lowest accuracy. In contrast, orbit models display the highest accuracy for the same number of channels. Notably, the difference in accuracy among these models remained nearly constant when plotted on a logarithmic scale. This suggests that increasing the number of parameters for a given tensor representation can provide sufficient capacity to capture the neighborhood chemical composition, even

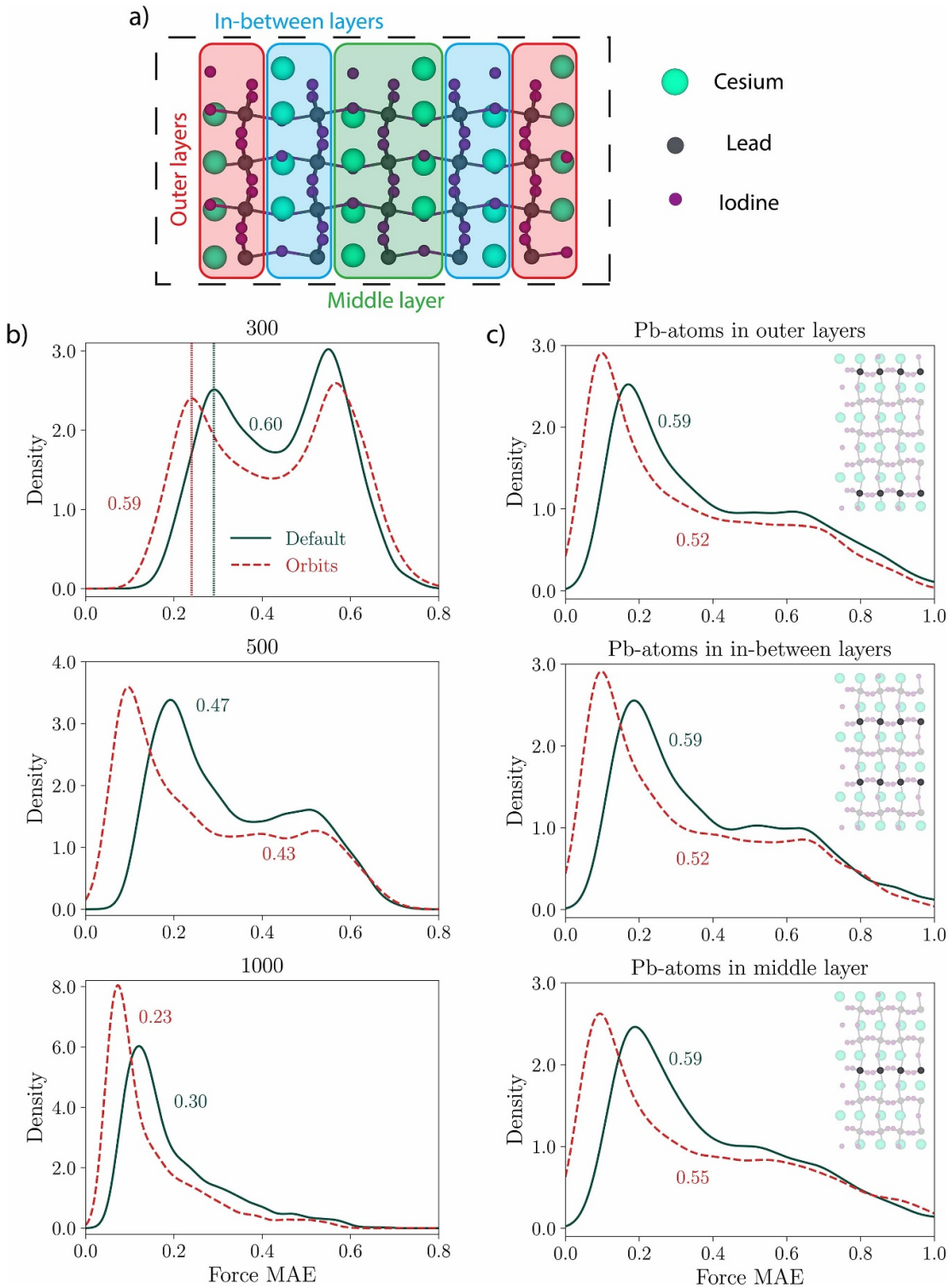

Figure 6. (a) The perovskite $CsPbI_3$ slab system in the zx projection. (b) and (c) Distributions of mean absolute error (MAE) in force predictions, in kcal $(mol \cdot Å)^{-1}$, for MACE models applied to the $CsPbI_3$ perovskite slab system. Solid green lines represent the default MACE architecture, while red dashed lines indicate the orbit-based MACE models. (b) Force MAE distributions calculated for all atoms in the system, for the MACE models trained on 300, 500, and 1000 reference slab geometries. (c) Force MAE distributions specifically for lead atoms, analyzed across different slab layers, using MACE models trained on 500 configurations.

without considering specific atom types. This is particularly true in systems where the neighborhood chemical composition remains consistent during the system's evolution.

Any MPNN, including MACE, is invariant under permutation of the same chemical species, using the same function for force predictions. Employing orbits expands the number of different chemical species in the models, increasing the number of distinct functions used. To benefit from such an increase, the training

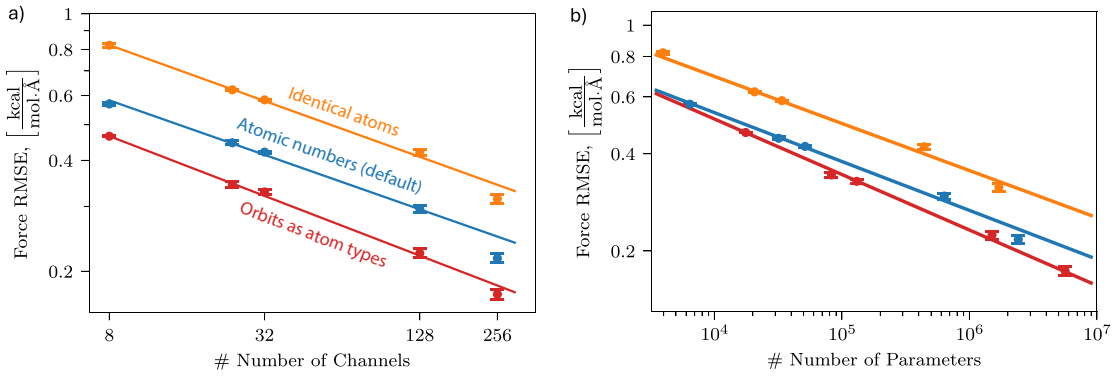

Figure 7. Force prediction accuracy of MACE models in the $CsPbI_3$ slab system as a function of the models' size. The root mean square error (RMSE) is evaluated across three model architectures: the default RMSE architecture (blue), an orbit-based MACE architecture (red), and a MACE architecture treating all atoms as identical (orange). Uncertainty was quantified using cross-validation testing. (a) The RMSE as a function of the number of channels in hidden irreducible representations, with other model parameters held constant. (b) The RMSE as a function of the total number of trainable parameters.

set needs to be large enough for every orbit to be well sampled; otherwise, the parameters of those functions cannot be accurately trained. As a demonstration, figure 7(b) shows a relationship between the accuracy of different types of the MACE models and the total number of trainable parameters they have, where the variability in accuracy is depicted as error bars. The figure corresponds precisely to the same models as figure 7(a). Notably, orbit-based MACE models achieve comparable accuracy with significantly fewer parameters than the default MACE models. For example, to reach a target accuracy of 0.4 kcal $(\text{mol}\cdot\text{Å})^{-1}$, an orbit-based MACE model requires approximately 9×10^4 parameters. In contrast, the default MACE model, which relies on atomic species, requires around 7×10^5 parameters-an order of magnitude more.

Our results show that while increasing the number of trainable parameters in MPNN models generally improves prediction accuracy, the proposed orbit-based approach achieves comparable or even superior accuracy with significantly fewer parameters. This suggests that explicitly incorporating chemically distinct atomic environments through orbits enables a more efficient parameterization, reducing model complexity without sacrificing accuracy.

4. Conclusion

We have developed a graph-based method for identifying permutational symmetries, applicable to both periodic and non-periodic chemical systems. Our approach systematically uncovers symmetries across the entire configurational space by analyzing atomic positions from a dataset of configurations. It outputs the size and generators of atomic permutation sets, along with a classification of atoms into chemically distinct environments (orbits). Notably, this information is determined prior to explicitly computing permutations. Our method has been successfully integrated with the kernel-based sGDML framework and the equivariant MPNN MACE model.

The practical utility of our approach was demonstrated by training accurate machine-learning FFs for four organic molecules: ethanol, 1,8-naphthyridine, D-alanine, and D-histidine adsorbed on a pristine graphene substrate using the sGDML architecture. We established that the accuracy of atomic force predictions is strongly influenced by the number of orbits in a system. Greater chemical diversity (resulting in more orbits) leads to increased force reconstruction heterogeneity and larger MAE and RMSE values. This trend was further validated across multiple MLFF architectures, including MACE, SO3krates, and sGDML, for the N-acetylphenylalanyl-pentaalanyl-lysine using the results of the TEA Challenge 2023 benchmark.

By incorporating orbit information into the MACE training process, we observed significant improvements over the default MACE architecture, which relies solely on atomic types. The tests were run for the graphene with pyridinic-N defect and $CsPbI_3$ perovskite slab. For example, to achieve a target accuracy of 0.4 kcal $(\text{mol}\cdot\text{Å})^{-1}$ for the $CsPbI_3$ slab (sufficient for multiple practical applications), an orbit-based MACE model required approximately 9×10^4 parameters-an order of magnitude fewer than the 7×10^5 parameters needed for the default MACE model. The advantages of using orbits become more pronounced as system complexity and dataset size increase. Our results consistently demonstrate that orbit-based models outperform standard approaches in force prediction accuracy, owing to their more efficient parameterization.

In this work, we have focused on a structurally well-defined benchmark system to demonstrate the accuracy improvements enabled by explicit permutation-symmetry encoding. Extending orbit-based symmetry methods to fully dynamic settings, where atoms may alter their local chemical environments through processes such as defect migration, bond breaking and formation, or other structural transformations, will require the development of dynamic orbit assignments. In particular, future MLFF architectures should smoothly embed atomic neighborhoods to allow orbit labels to evolve consistently throughout MD trajectories. We consider this a critical and promising direction for future research.

In summary, the proposed graph-based symmetry search method represents a significant step forward in modeling large, complex, and chemically diverse systems using state-of-the-art MLFFs. By leveraging orbits, our approach enhances predictive accuracy and computational efficiency, making it a valuable tool for future applications in computational chemistry and materials science.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://zenodo.org/records/15041256.

Acknowledgment

The work at the University of Luxembourg was performed with funding from the European Research Council Executive Agency (ERCEA) under Project 101054629 (FITMOL) and from the Luxembourg National Research Fund (FNR) under the CORE project C19/MS/13718694/QML-FLEX. The work at the University of Mons and the University of Luxembourg was carried out within the framework of the M-ERA.NET project PHANTASTIC, supported by the Luxembourg National Research Fund (FNR) (INTER/MERA22/16521502/PHANTASTIC) and by the Belgian National Fund for Scientific Research (F.R.S.-FNRS) under Grant R.8003.22. H.E.S. acknowledges support from DGAPA-UNAM Project PAPIIT No. IA106023 and CONAHCyT project CF-2023-I-468. S.C. acknowledges support by the German Federal Ministry of Education and Research (BMBF) for BIFOLD (01IS18037A). C.Q. is a F.R.S.-FNRS Research Associate, and D.B. is a F.R.S.-FNRS Research Director.

The simulations were performed on the Luxembourg national supercomputer MeluXina. The authors gratefully acknowledge the LuxProvide teams for their expert support. Computational resources were provided by the Consortium des Équipements de Calcul Intensif (CÉCI), funded by the F.R.S.-FNRS under Grant 2.5020.11. The present research also benefited from access to Lucia, the Tier-1 supercomputer of the Walloon Region, an infrastructure funded by the Walloon Region under Grant Agreement No. 1910247.

We thank Prof. Oleg Prezhdo for his helpful comments.

Code and data availability

The datasets used in this manuscript are available at https://doi.org/10.5281/zenodo.15 041 256. The proposed algorithm, integrated into the FFAST package as a separate brunch, can be accessed at GitHub/fonsecag/FFAST.

References

[1] Bartók A P, Payne M C, Kondor R and Csányi G 2010 Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons *Phys. Rev. Lett.* **104** 136403
[2] Bartók A P, Kondor R and Csányi G 2013 On representing chemical environments *Phys. Rev.* B **87** 184115
[3] Deringer V L, Bartók A P, Bernstein N, Wilkins D M, Ceriotti M and Csányi G 2021 Gaussian process regression for materials and molecules *Chem. Rev.* **121** 10073–141
[4] Bartók A P and Csányi G 2015 Gaussian approximation potentials: a brief tutorial introduction *Int. J. Quantum Chem.* **115** 1051–7
[5] Behler J 2011 Atom-centered symmetry functions for constructing high-dimensional neural network potentials *J. Chem. Phys.* **134** 074106
[6] Chmiela S, Tkatchenko A, Sauceda H E, Poltavsky I, Schütt K T and Müller K-R 2017 Machine learning of accurate energy-conserving molecular force fields *Sci. Adv.* **3** e1603015
[7] Chmiela S, Sauceda H E, Poltavsky I, Müller K-R and Tkatchenko A 2019 sGDML: constructing accurate and data efficient molecular force fields using machine learning *Comput. Phys. Commun.* **240** 38–45
[8] Chmiela S, Sauceda H E, Müller K-R and Tkatchenko A 2018 Towards exact molecular dynamics simulations with machine-learned force fields *Nat. Commun.* **9** 3887
[9] Sauceda H E, Gálvez-González L E, Chmiela S, Paz-Borbón L O, Müller K-R and Tkatchenko A 2022 BIGDML—towards accurate quantum machine learning force fields for materials *Nat. Commun.* **13** 3733

[10] Chmiela S, Vassilev-Galindo V, Unke O T, Kabylda A, Sauceda H E, Tkatchenko A and Müller K-R 2023 Accurate global machine learning force fields for molecules with hundreds of atoms *Sci. Adv.* **9** eadf0873
[11] Schütt K T, Kindermans P-J, Sauceda H E, Chmiela S, Tkatchenko A and Müller K-R 2017 SchNet: a continuous-filter convolutional neural network for modeling quantum interactions *Proc. 31st Int. Conf. on Neural Information Processing Systems NIPS'17* pp 992–1002 (Curran Associates Inc.)
[12] Schütt K T, Unke O T and Gastegger M 2021 Equivariant message passing for the prediction of tensorial properties and molecular spectra *Int. Conf. on Machine Learning* (PMLR) pp 9377–88
[13] Batzner S, Musaelian A, Sun L, Geiger M, Mailoa J P, Kornbluth M, Molinari N, Smidt T E and Kozinsky B 2022 E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials *Nat. Commun.* **13** 2453
[14] Musaelian A, Batzner S, Johansson A, Sun L, Owen C J, Kornbluth M and Kozinsky B 2023 Learning local equivariant representations for large-scale atomistic dynamics *Nat. Commun.* **14** 579
[15] Musaelian A, Johansson A, Batzner S, and Kozinsky B 2023 Scaling the leading accuracy of deep equivariant models to biomolecular simulations of realistic size (arXiv:2304.10061)
[16] Unke O T and Meuwly M 2019 Physnet: a neural network for predicting energies, forces, dipole moments and partial charges *J. Chem. Theory Comput.* **15** 3678–93
[17] Drautz R 2019 Atomic cluster expansion for accurate and transferable interatomic potentials *Phys. Rev.* B **99** 014104
[18] Batatia I, Batzner S, Kovács D P, Musaelian A, Simm G N C, Drautz R, Ortner C, Kozinsky B and Csányi G 2022 The design space of E(3)-equivariant atom-centered interatomic potentials (arXiv:2205.06643)
[19] Batatia I, Kovacs D P, Simm G N C, Ortner C and Csanyi G 2022 MACE: higher order equivariant message passing neural networks for fast and accurate force fields *Advances in Neural Information Processing Systems* vol 35 ed S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho and A Oh (Curran Associates, Inc.) pp11423–36
[20] Frank J T, Unke O T and Müller K-R 2022 So3krates: equivariant attention for interactions on arbitrary length-scales in molecular systems *Advances in Neural Information Processing Systems* vol 35 S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho and A Oh (Curran Associates, Inc.) pp11423–36
[21] Frank J T, Unke O T, Müller K-R and Chmiela S 2024 A Euclidean transformer for fast and stable machine learned force fields *Nat. Commun.* **15** 6539
[22] Unke O T *et al* 2024 Biomolecular dynamics with machine-learned quantum-mechanical force fields trained on diverse chemical fragments *Sci. Adv.* **10** eadn4397
[23] Kabylda A, Frank J T, Dou S S, Khabibrakhmanov A, Sandonas L M, Unke O T, Chmiela S, Müller K-R and Tkatchenko A 2024 Molecular simulations with a pretrained neural network and universal pairwise force fields *Chemxiv Preprint* (https://doi.org/10.26434/chemrxiv-2024-bdfr0) (accessed 08 October 2024)
[24] Kovács D P, Moore J H, Browning N J, Batatia I, Horton J T, Kapil V, Witt W C, Magdău I-B, Cole D J and Csányi G 2023 MACE-OFF23: transferable machine learning force fields for organic molecules (arXiv:2312.15211)
[25] Batatia I *et al* 2024 A foundation model for atomistic materials chemistry (arXiv:2401.00096)
[26] Smith J S, Isayev O and Roitberg A E 2017 ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost *Chem. Sci.* **8** 3192–203
[27] Anstine D, Zubatyuk R and Isayev O 2024 AIMNet2: a neural network potential to meet your neutral, charged, organic, and elemental-organic needs *Chem. Sci.* **16** 10228–44
[28] Plé T, Lagardère L and Piquemal J-P 2023 Force-field-enhanced neural network interactions: from local equivariant embedding to atom-in-molecule properties and long-range effects *Chem. Sci.* **14** 12554–69
[29] Qiao Z, Christensen A S, Welborn M, Manby F R, Anandkumar A and Miller T F III 2022 Informing geometric deep learning with electronic interactions to accelerate quantum chemistry *PNAS* **119** e2205221119
[30] Christensen A S, Bratholm L A, Faber F A, Anatole von Lilienfeld O and von Lilienfeld O A 2020 FCHL revisited: faster and more accurate quantum machine learning *J. Chem. Phys.* **152** 044107
[31] Browning N J, Faber F A, Anatole von Lilienfeld O and 2022 GPU-accelerated approximate kernel method for quantum machine learning *J. Chem. Phys. O* **157** 214801
[32] Choudhary K, DeCost B, Major L, Butler K, Thiyagalingam J and Tavazza F 2023 Unified graph neural network force-field for the periodic table: solid state applications *Digit. Discov.* **2** 346–55
[33] Rodriguez A *et al* 2023 Million-scale data integrated deep neural network for phonon properties of heuslers spanning the periodic table *npj Comput. Mater.* **9** 20
[34] Zhang Y and Jiang B 2023 Universal machine learning for the response of atomistic systems to external fields *Nat. Commun.* **14** 6424
[35] Zeng J *et al* 2023 DeePMD-kit v2: a software package for deep potential models *J. Chem. Phys.* **159** 054801
[36] Wang H, Zhang L, Han J and Weinan E 2018 Deepmd-kit: a deep learning package for many-body potential energy representation and molecular dynamics *Comput. Phys. Commun.* **228** 178–84
[37] Shapeev A V 2016 Moment tensor potentials: a class of systematically improvable interatomic potentials *Multiscale Model. Sim.* **14** 1153–73
[38] Thompson A P, Swiler L P, Trott C R, Foiles S M and Tucker G J 2015 Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials *J. Comput. Phys.* **285** 316–30
[39] Chen C and Ong S P 2022 A universal graph deep learning interatomic potential for the periodic table *Nat. Comput. Sci.* **2** 718–28
[40] Gokcan H and Isayev O 2022 Learning molecular potentials with neural networks *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **12** e1564
[41] Vandermause J, Torrisi S B, Batzner S, Xie Y, Sun L, Kolpak A M and Kozinsky B 2020 On-the-fly active learning of interpretable Bayesian force fields for atomistic rare events *npj Comput. Mater.* **6** 20
[42] Park Y, Kim J, Hwang S and Han S 2024 Scalable parallel algorithm for graph neural network interatomic potentials in molecular dynamics simulations *J. Chem. Theory Comput.* **20** 4857–68
[43] Jinnouchi R, Lahnsteiner J, Karsai F, Kresse G and Bokdam M 2019 Phase transitions of hybrid perovskites simulated by machine-learning force fields trained on-the-fly with Bayesian inference *Phys. Rev. Lett.* **122** 225701
[44] Unke O T, Chmiela S, Gastegger M, Schütt K T, Sauceda H E and Müller K-R 2021 SpookyNet: learning force fields with electronic degrees of freedom and nonlocal effects *Nat. Commun.* **12** 7273
[45] Lederer J, Gastegger M, Schütt K T, Kampffmeyer M, Müller K-R and Unke O T 2023 Automatic identification of chemical moieties *Phys. Chem. Chem. Phys.* **25** 26370–9
[46] Majewski M, Pérez A, Thölke P, Doerr S, Charron N E, Giorgino T, Husic B E, Clementi C, Noé F and De Fabritiis G 2023 Machine learning coarse-grained potentials of protein thermodynamics *Nat. Commun.* **14** 5739

[47] Unke O T, Chmiela S, Sauceda H E, Gastegger M, Poltavsky I, Schütt K T, Tkatchenko A and Müller K-R 2021 Machine learning force fields *Chem. Rev.* **121** 10142–86
[48] Bergman D L 2020 *Symmetry Constrained Machine Learning* (*Intelligent Systems and Applications)* ed Y Bi, R Bhatia and S Kapoor (Springer) pp 501–12
[49] Grisafi A, Wilkins D M, Csányi G and Ceriotti M 2018 Symmetry-adapted machine learning for tensorial properties of atomistic systems *Phys. Rev. Lett.* **120** 036002
[50] Schmitz N F, Müller K-R and Chmiela S 2022 Algorithmic differentiation for automated modeling of machine learned force fields *J. Phys. Chem. Lett.* **13** 10183–9
[51] Ewig C S *et al* 2001 Derivation of classII force fields.VIII. Derivation of a general quantum mechanical force field for organic compounds *J. Comput. Chem.* **22** 1782–800
[52] Hwang M J, Stockfisch T P and Hagler A T 1994 Derivation of class II force fields. 2. Derivation and characterization of a class II force field, CFF93, for the alkyl functional group and alkane molecules *J. Am. Chem. Soc.* **116** 2515–25
[53] Frucht R 1949 Graphs of degree three with a given abstract group *Can. J. Math.* **1** 365–78
[54] Luks E M 1982 Isomorphism of graphs of bounded valence can be tested in polynomial time *J. Comput. Syst. Sci.* **25** 42–65
[55] Bohanec S and Perdih M 1993 Symmetry of chemical structures: a novel method of graph automorphism group determination *J. Chem. Inf. Comput. Sci.* **33** 719–26
[56] Faulon J-L 1998 Isomorphism, automorphism partitioning and canonical labeling can be solved in polynomial-time for molecular graphs *J. Chem. Inf. Comput. Sci.* **38** 432–44
[57] Merkys A, Vaitkus A, Grybauskas A, Konovalovas A, Quirós M and Gražulis S 2023 Graph isomorphism-based algorithm for cross-checking chemical and crystallographic descriptions *J. Cheminform.* **15** 25
[58] Novoselov K, Geim A, Morozov S, Jiang D, Zhang Y, Dubonos S, Grigorieva I and Firsov A 2004 Electric field in atomically thin carbon films *Science* **306** 666–9
[59] Schedin F, Geim A K, Morozov S V, Hill E W, Blake P, Katsnelson M I and Novoselov K S 2007 Detection of individual gas molecules adsorbed on graphene *Nat. Mater.* **6** 652–5
[60] Myung S, Yin P T, Kim C, Park J, Solanki A, Reyes P I, Lu Y, Kim K S and Lee K-B 2012 Label-free polypeptide-based enzyme detection using a graphene-nanoparticle hybrid sensor *Adv. Mater.* **24** 6081–7
[61] Lazar P, Karlický F, Jurečka P, Kocman M, Otyepková E, Mikuláš K and Otyepka M 2013 Adsorption of small organic molecules on graphene *J. Am. Chem. Soc.* **135** 6372–7
[62] Poltavsky I *et al* 2025 Crash testing machine learning force fields for molecules, materials and interfaces: model analysis in the tea challenge 2023 *Chem. Sci.* **16** 3720–37
[63] Bonchev D 1991 *Chemical Graph Theory: Introduction and Fundamentals* (*Chemical Graph Theory*) (Taylor & Francis)
[64] García-Domenech R, Gálvez J, de Julián-Ortiz J V and Pogliani L 2008 Some new trends in chemical graph theory *Chem. Rev.* **108** 1127–69
[65] Douglas B L and Wang J B 2008 A classical approach to the graph isomorphism problem using quantum walks *J. Phys. A: Math. Theor.* **41** 075303
[66] Lauri J and Scapellato R 2016 *Topics in Graph Automorphisms and Reconstruction* (*London Mathematical Society Lecture Note Series*) 2nd edn (Cambridge University Press)
[67] McKay B D and Piperno A 2014 Practical graph isomorphism, II *J. Symb. Comput.* **60** 94–112
[68] Perdew J P, Burke K and Ernzerhof M 1996 Generalized gradient approximation made simple *Phys. Rev. Lett.* **77** 3865–8
[69] Hermann J and Tkatchenko A 2020 Density functional model for van der Waals interactions: unifying many-body atomic approaches with nonlocal functionals *Phys. Rev. Lett.* **124** 146401
[70] Blum V, Gehrke R, Hanke F, Havu P, Havu V, Ren X, Reuter K and Scheffler M 2009 *Ab initio* molecular simulations with numeric atom-centered orbitals *Comput. Phys. Commun.* **180** 2175–96
[71] Fonseca G, Poltavsky I, Vassilev-Galindo V and Tkatchenko A 2021 Improving molecular force fields across configurational space by combining supervised and unsupervised machine learning *J. Chem. Phys.* **154** 124102
[72] Fonseca G, Poltavsky I and Tkatchenko A 2023 Force field analysis software and tools (FFAST): assessing machine learning force fields under the microscope *J. Chem. Theory Comput.* **19** 8706–17
[73] VandeVondele J, Krack M, Mohamed F, Parrinello M, Chassaing T and Hutter J 2005 Quickstep: fast and accurate density functional calculations using a mixed Gaussian and plane waves approach *Comput. Phys. Commun.* **167** 103–28
[74] Hutter J, Iannuzzi M, Schiffmann F and VandeVondele J 2014 CP2K: atomistic simulations of condensed matter systems *WIREs Comput. Mol. Sci.* **4** 15–25
[75] VandeVondele J and Hutter J 2007 Gaussian basis sets for accurate calculations on molecular systems in gas and condensed phases *J. Chem. Phys.* **127** 114105
[76] Goedecker S, Teter M and Hutter J 1996 Separable dual-space Gaussian pseudopotentials *Phys. Rev. B* **54** 1703–10
[77] Krack M 2005 Pseudopotentials for H to Kr optimized for gradient-corrected exchange-correlation functionals *Theor. Chem. Acc.* **114** 145–52
[78] Nosé S 1984 A unified formulation of the constant temperature molecular dynamics methods *J. Chem. Phys.* **81** 511–9
[79] Hoover W G 1986 Constant-pressure equations of motion *Phys. Rev. A* **34** 2499–500