JCTC Journal of Chemical Theory and Computation

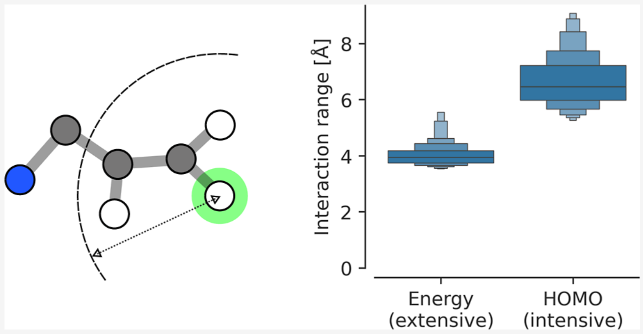
pubs.acs.org/JCTC Article

Analyzing Atomic Interactions in Molecules as Learned by Neural Networks

Malte Esders,* Thomas Schnake,∇ Jonas Lederer,∇ Adil Kabylda, Grégoire Montavon, Alexandre Tkatchenko,* and Klaus-Robert Müller*

Cite This: J. Chem. Theory Comput. 2025, 21, 714–729

ABSTRACT: While machine learning (ML) models have been able to achieve unprecedented accuracies across various prediction tasks in quantum chemistry, it is now apparent that accuracy on a test set alone is not a guarantee for robust chemical modeling such as stable molecular dynamics (MD). To go beyond accuracy, we use explainable artificial intelligence (XAI) techniques to develop a general analysis framework for atomic interactions and apply it to the SchNet and PaiNN neural network models. We compare these interactions with a set of fundamental chemical principles to understand how well the models have learned the underlying physicochemical concepts from the data. We focus on the strength of the interactions for different atomic species, how predictions for intensive and extensive quantum molecular properties are made, and analyze the decay and many-body nature of the interactions with interatomic distance. Models that deviate too far from known physical principles produce unstable MD trajectories, even when they have very high energy and force prediction accuracy. We also suggest further improvements to the ML architectures to better account for the polynomial decay of atomic interactions.



1. INTRODUCTION

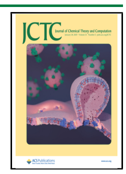
Methods for modeling atomistic systems range between computationally cheap but less precise (e.g., classical force fields), to computationally expensive but more precise [e.g., first-principles calculations based on density functional theory (DFT), coupled-cluster method with single, double and triple excitations (CCSD(T)), or quantum Monte Carlo techniques[1,2]]. Machine learning force fields (MLFFs) are an emerging technology that tries to favorably position itself by being computationally efficient while simultaneously approaching the more expensive methods in accuracy.[3]

Due to the many-body nature of the Schrödinger equation, the computational cost of accurate ab initio methods grows extremely fast (exponentially or steeply polynomially) with the number of particles in a system.[4,5] Conversely, approximate methods with a lower computational cost inevitably need to "cut corners" and therefore may not adequately represent the full complexity of a system under study.[6,7] As a result, numerous quantum-chemical approximation methods have been developed, each with its own trade-offs. The usefulness of these methods lies in the detailed understanding of their limitations, allowing one to choose the most appropriate method for the task at hand.

Despite the vast potential of MLFFs, they may ultimately only become trusted once their strengths and weaknesses are similarly understood. For instance, a common problem of ML models is that they do not extrapolate well beyond their training domain,[8] and MLFFs are no exception. Although research into transferable models that are trained on well-curated data sets that broadly cover chemical space is ongoing,[9−15] for the foreseeable future there likely will not be a one-size-fits-all model. This necessitates a deeper analysis of the underlying prediction strategy. The nonlinear nature of complex ML models complicates our understanding of how they form predictions, particularly when it comes to identifying potential shortcomings. The current study serves as a crucial step to address this issue: based on recent advances,[16−18] we present a method to uncover in detail the prediction strategies and learned representations of MLFFs. On the basis of four common chemical principles listed below, we examine to what extent they are embodied by learning models.

Recently, several studies highlighted the need to move beyond just the validation accuracy, because the validation accuracy was shown to be insufficient to predict MD stability. Therefore, the validation accuracy by itself is not a good

Received: October 22, 2024
Revised: December 30, 2024
Accepted: January 2, 2025
Published: January 10, 2025



https://doi.org/10.1021/acs.jctc.4c01424
J. Chem. Theory Comput. 2025, 21, 714−729

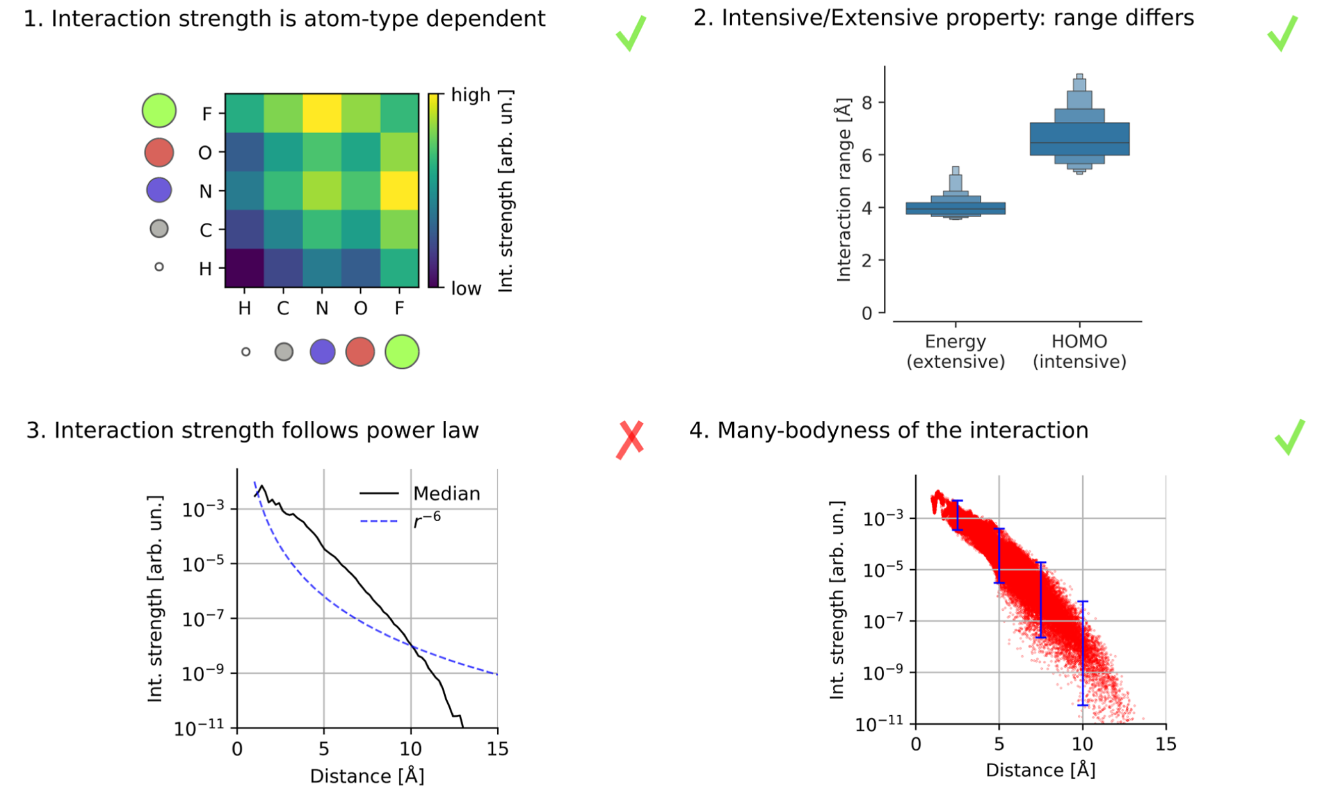


Figure 1. Using this study's explainability framework to inspect whether the models learned four common chemical principles from the data. Subfigure 1: mean interaction strengths for atom-pairs at a distance less than 3 Å on 1300 molecules from the QM9 data set. The color-scale is logarithmic. Subfigure 2: interaction range (eq 8) for a model trained on atomization energy (extensive property) and HOMO energy (intensive property) from the QM9 data set. Subfigure 3: median of the interaction strength across interatomic distance, compared to r^{-6}, a typical decay for the energy in London dispersion, e.g. as in the Lennard-Jones 12-6 potential (molecule: Ac-Ala3-NHMe from the MD22 data set). Subfigure 4: spread of the interaction strength at different distances (each dot in the scatter plot is one atom pair in one conformation of Ac-Ala3-NHMe); for selected distances, the maximum to minimum interaction strength is indicated by blue lines.

measure of the degree to which chemical principles were learned from the data.[19−23]

The FFAST software package[24] is an example of a tool designed for detailed analysis of MLFF prediction results, including visualization of per-atom prediction errors, force error densities, and challenging conformers. While such analysis can be invaluable, the current study aims to go beyond that by investigating the underlying GNN prediction strategy and understanding why prediction errors occur, rather than merely identifying whether and where they happen.

Training models is based on learning a mapping from atom positions and atomic numbers to properties like the atomization energy and the forces. It is generally hoped that models can learn the underlying physics purely from such data, but an analysis to which extent that is actually the case is so far lacking. In this study, we aim to fill this gap by proposing a way to systematically test the chemical plausibility of MLFF predictions. To this end, we posit the following four chemical principles:

I **The strength of interactions is atom-type and property dependent**: the relevance of atomic interactions predicted by MLFFs varies based on the atom types involved and the property being predicted. This atom-type and property dependence is particularly pronounced in bonded interactions, whereas at longer-range interactions, the dependence on the property becomes less prominent.

II **Different interaction range for intensive vs extensive properties**: extensive properties can be approximated by evaluating the property on parts of the whole, and summing these local contributions up to obtain the property for the entire system.[25] One could say the whole is the sum of the parts (at least up to a given accuracy). For intensive properties on the other hand, the entire system must be taken into consideration, and the whole is different from the sum of the parts. Therefore, one expects a higher interaction range when predicting intensive properties.

III **Decrease of interaction strength with distance follows a power law**: at higher distance ranges, forces within molecules often fall off with a power law.[26] For instance, forces between permanent dipoles fall off with r^{-4}, and London dispersion forces and dipole-induced dipoles fall off with r^{-7} (when using the pairwise approximation).

IV **Many-bodyness**: the interaction strength should be anisotropic, meaning in this case that the interaction strength for equally distant atom pairs should differ depending on other atoms in the neighborhood.[27,28] We call this property "many-bodyness", and contrast it with classical force fields, where interactions typically involve 4 or less directly bonded atoms. At higher distances, only 2-body terms are considered in widely used mechanistic force fields.[29,30]

An overview of these principles with some illustrative results can be found in Figure 1. We see in subfigure 1 that the

https://doi.org/10.1021/acs.jctc.4c01424
J. Chem. Theory Comput. 2025, 21, 714−729

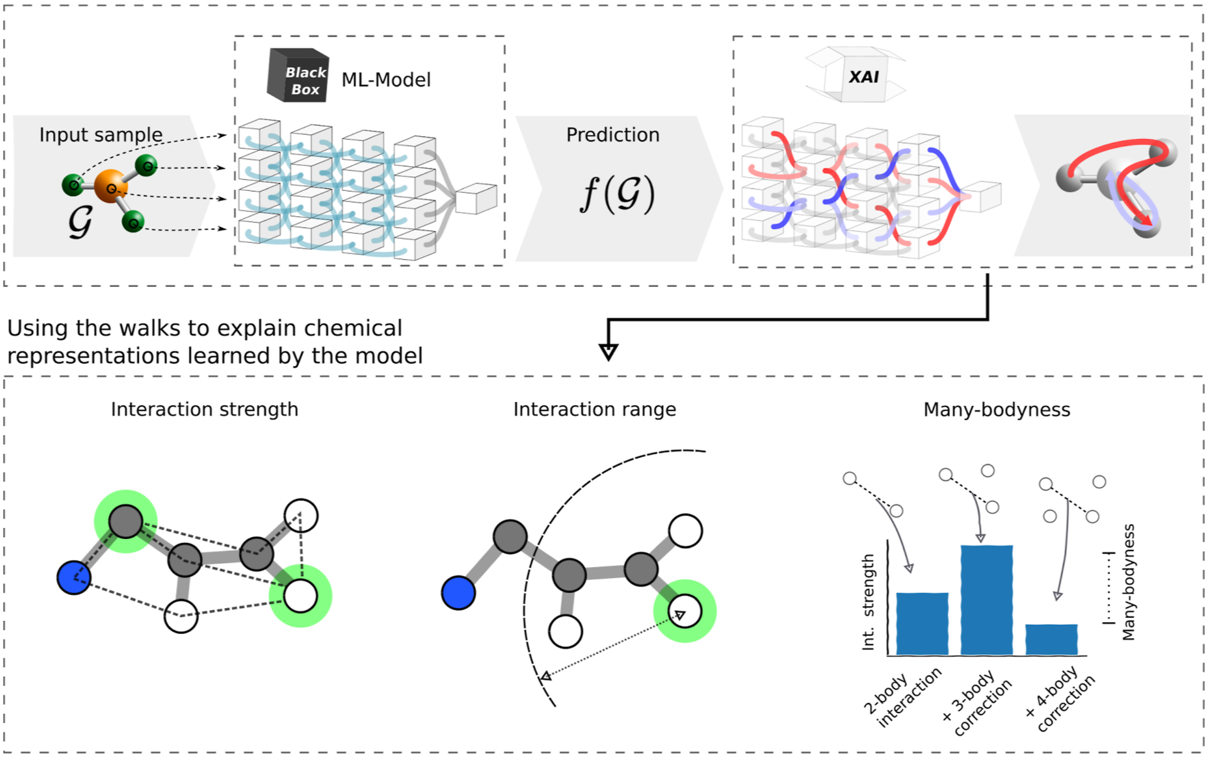


Figure 2. Overview of the explanation framework introduced in this study. A molecular input graph is processed by a black-box ML-model, specifically a GNN. The prediction is related to the input graph in the form of relevant walks on the graph, which are obtained from GNN-LRP.[18] We extend this analysis to quantum chemistry-specific settings: We provide a measure of the interaction strength between two atoms in a molecule (eq 10); we define the range up to which the network considers significant interactions (eq 8); and we specify the many-bodyness, which is a measure for how much the chemical neighborhood influences the interaction strength between two atoms (eq 12).

interaction strength is atom-type dependent. Subfigure 2 shows that the extensive property of energy has a smaller interaction range than the intensive property of HOMO energy. Subfigure 3 shows that the median interaction strength at different atomic distances does not follow a power law, particularly it does not follow r^{-6}. Subfigure 4 shows that interaction strengths between atom pairs at the same distance differ, which is due to the effect of other atoms in the neighborhood, a phenomenon we call "many-bodyness" (see also Figure 2 bottom for an illustration).

While some of these chemical and physical properties might seem to be textbook knowledge, only qualitative guidelines can be formulated with our limited understanding of many-body quantum mechanics. On the other hand, ML models learn a quantitative mapping between structures and QM properties within the chemical space defined by a given data set. Hence, a natural and so far unanswered question is whether these *quantitative* predictions also obey the known *qualitative* chemical and physical principles. This is the main challenge addressed in the current work.

None of the discussed properties is given to the ML models as an inductive bias, i.e., as an explicit part of their architecture or loss function; therefore, it is merely a hope that such principles will be learned from the data. In the current study, we test each of these properties on different MLFFs. Specifically, we show that the closer an MLFF agrees with the above principles, the more stable its MD trajectories are.

Trying to analyze the prediction strategy of graph neural networks (GNNs) applied to molecular data started soon after using GNNs became popular in quantum chemistry. Early approaches analyzed the atom-wise energy contributions or introduced a test charge to measure the model's reaction.[31−33] This approach is still in use today, for instance in assessing the robustness of the prediction.[34]

Using first-order explanation methods like layer-wise relevance propagation[35] can uncover which individual nodes are relevant to the prediction.[36−38] Other approaches yield relevant clusters of atoms.[39−41] Such first-order explanation approaches are useful for a variety of chemical applications,[42,43] but can not go beyond atomic or cluster relevances.

In contrast, higher-order relevance attributions[18] can be associated with many-body interactions[44] and have helped to corroborate the importance of such interactions in coarse-grained protein systems.[45]

1.1. Overview of This Study and Its Contributions.

We focus our analysis on GNNs as a popular implementation of MLFFs. We make use of a recently proposed XAI method, called GNN-LRP,[18] that allows to assign a relevance to sequences of nodes in the graph. In a first step, we review GNNs for quantum chemistry and XAI, specifically the GNN-LRP method, and outline how this method can be used in quantum-chemical applications.

We then extend GNN-LRP specifically for MLFFs. Making use of the fact that molecular "graphs" are embedded in Euclidean space, we propose a distance measure for sequences of nodes and use it to compute the interaction range that a GNN uses to form its prediction. Additionally, we develop a measure for the interaction strength between atoms as seen by the GNN. Lastly, we propose a measure for the many-bodyness of the interaction strength.

We then apply these methods to the popular SchNet[46] and PaiNN[47] architectures in various atomistic settings. SchNet

https://doi.org/10.1021/acs.jctc.4c01424
J. Chem. Theory Comput. 2025, 21, 714−729

and PaiNN use rotationally invariant and equivariant message passing, respectively, which allows us to compare the prediction behavior of both types of architectures. We provide a detailed analysis of each of the four chemical principles stated above and whether they are expressed in the models. Additionally, we provide a way to go beyond the classic "generalization error" as a performance metric, and use our proposed analysis to predict the stability of MD-trajectories.

2. METHODOLOGY

2.1. Graph Neural Networks for Quantum Chemistry.

Most state-of-the-art MLFFs[3] are from the family of GNNs.[48] GNNs for quantum chemistry work in two phases. In the first phase, each atom, indexed by k, gets represented as a point in a high-dimensional "feature space". This is achieved by initializing the atoms to element-specific embeddings and then iterating T-times a "message passing" step between atoms within a certain cutoff distance, resulting for atom k in a vector representation $\boldsymbol{H}_{T,k}$ after the T-th message passing step. After the feature representations are updated by several message-passing steps, they encode the local chemical environment of each atom and thus contain the relevant information about molecular geometry and composition. Then, in the second phase, a feed-forward neural network predicts molecular properties from the atomic feature representations.

SchNet[46,49] and PaiNN[47] are variants of GNNs applied to 3D geometries. They derive a connectivity graph where the graph nodes represent the atoms and the graph edges describe to what extent neighboring atoms are directly interacting. The connectivity of the graph is determined by a cutoff distance, beyond which all direct connections between nodes (atoms) are cut. A "cutoff-function", usually a cosine,[50] is applied to the interactions to ensure that there is a smooth transition toward the cutoff. A single message passing step is represented by a so-called interaction block. For the considered architectures, several interaction blocks are stacked to ensure that also distant nodes can exchange information, as well as to allow the nodes to build a more fine-grained embedding of their atomic neighborhood. While SchNet solely learns scalar feature representations in the first phase, PaiNN in addition learns vectorial features.[47] The rotational equivariant nature of those vectorial feature representations makes PaiNN more data efficient[51] and, as a result, provides more stable MD trajectories.[52]

The first phase of the GNN, the message passing step, can be further divided into two individual steps, the aggregation step and the combination step. In the *aggregation* step the incoming "messages" from an atom's neighboring atoms are aggregated, and in the *combination* step the aggregated messages are combined nonlinearly with the respective atomic feature representation of the node. Hence, the GNN is of the form

$$\boldsymbol{H}_{t+1,k} = C\Big(\sum_{j \in \mathrm{neigh}(k)} \mu(\boldsymbol{H}_{t,k}, \boldsymbol{H}_{t,j}, r_{kj})\Big) \tag{1}$$

where μ and C are message and combine functions, respectively, and r_{kj} is the distance between the atoms indexed by j and k. The set neigh(k) specifies the neighbors of atom k that are within the cutoff distance. The sum over the messages of all neighboring atoms yields the aggregated message.

There is a large variety of models that follow the above message-passing structure. One way to characterize these models is by the rotation order they use for their features (for an overview, see ref 53). For instance, SchNet is a representative example of GNNs that are based on features of rotation order $l = 0$, i.e. features that are invariant to rotation. PaiNN is representative of models that use equivariant message passing and uses features of both rotation order $l = 0$ and $l = 1$ (the "vectorial features", which are equivariant under rotation). For more details about SchNet and PaiNN, see Section S2. Other recent state-of-the-art models are typically also including features with $l \geq 1$. They include NequIP,[51] which can build features of arbitrary rotation order, MACE,[54] which uses an expansion in a spherical basis and relies on many-body messages, and SO3krates,[55] which adds an equivariant attention mechanism.

In the following, we denote for each atom, indexed by i, $\boldsymbol{r}_i \in \mathbb{R}^3$ to be its position. In addition, we consider $f: \mathcal{X} \to \mathbb{R}$ to be the ML model with a scalar prediction. The domain $\mathcal{X}$ of the model in our case is the set of all possible geometric configurations of atoms. Each molecule is represented by the positions $(\boldsymbol{r}_i)_i$ of its atoms, indexed by i, and their respective nuclear charges.

2.2. Explainable AI.

ML models, in particular deep neural networks, have demonstrated high predictive capabilities for a broad range of tasks, including accurate inferences of molecular electronic properties in the field of quantum chemistry.[49] These models, while achieving high accuracy, are fundamentally black boxes. In other words, they do not achieve the objective of shedding light on the structure of the inferred input−output relation, which is a more fundamental scientific objective.[56] Furthermore, the measured accuracy may conceal whether the learned relation is physically meaningful, or whether it arises from exploiting a confounder in the data, the so-called Clever Hans effect.[17,57,58]

XAI (see e.g. ref 17) is a recent trend in ML, which aims to gain transparency into these highly complex and powerful ML models. Through specific algorithms operating on the structure of the learned ML model, XAI helps clarify the strategy an ML model uses to generate its predictions. XAI has multiple applications: It enables, together with a human expert, to validate an ML model, in particular, detecting features that an ML model uses as part of a Clever Hans strategy (aka. shortcut learning[59]). Another application of XAI is in serving as scientific assistants,[60,61] where, alongside a well-trained ML model, it helps to identify candidate input−output relationships for further testing by human observers in subsequent targeted experiments.

The field and the set of proposed XAI methods is highly heterogeneous. This is partly due to the broad range of meanings of the terms such as "explainability" and "interpretability," as well as the diversity of practical use cases. However, research has coalesced around specific problem formulations, one of which is the problem of attribution.

Attribution assumes an input domain $\mathcal{X}$, an output domain $\mathbb{R}$, typically real-valued, and a prediction function $f: \mathcal{X} \to \mathbb{R}$ linking instances in the input domain to values in the output domain. In a quantum chemistry context, the input can be a set of features describing the molecular geometry, and the output the electronic property (e.g., atomization energy). Focusing on a single prediction $\boldsymbol{x} \to y$ with $\boldsymbol{x} = (x_1, ..., x_d) \in \mathcal{X}$ the collection of input features and $y = f(\boldsymbol{x}) \in \mathbb{R}$ the real-valued output, we would like to compute for each input feature x_i a score $R_i \in \mathbb{R}$ measuring the extent by which this feature has

https://doi.org/10.1021/acs.jctc.4c01424
J. Chem. Theory Comput. 2025, 21, 714−729

contributed to the output y. Many methods have been proposed to compute these scores, e.g., refs 35, 62, and 63 with different properties in terms of robustness, computational efficiency, and applicability. One such method, *Integrated Gradients*,[63] assumes that the function f is differentiable and that the point $\boldsymbol{x}$ of interest is connected to a root point $\tilde{\boldsymbol{x}}$ through a path $\boldsymbol{x}^\alpha$ parametrized by α, particularly $\boldsymbol{x}^0 = \tilde{\boldsymbol{x}}$ and $\boldsymbol{x}^1 = \boldsymbol{x}$, and decomposes the prediction y in terms of input features via the equations

$$y = \int \frac{\partial y(\boldsymbol{x}^\alpha)}{\partial \boldsymbol{x}^\alpha} \frac{\partial \boldsymbol{x}^\alpha}{\partial \alpha} \mathrm{d}\alpha = \sum_{i=1}^{d} \underbrace{\int \frac{\partial y(\boldsymbol{x}^\alpha)}{\partial x_i^\alpha} \frac{\partial x_i^\alpha}{\partial \alpha} \mathrm{d}\alpha}_{R_i}$$

where for practical purposes the integral is discretized, typically into 10−100 steps. Variants of the equation above, involving multiple potentially nonlinear paths, are possible. For those path-based methods to work well, one should assume that the path remains on the data manifold, so that the model's behavior is evaluated on regions of the input space that are physically meaningful. In a quantum chemical scenario, where atomic coordinates or interatomic distances form the input representation, one may be required to define an appropriate path between the current molecule and some reference molecule (e.g., a relaxation path). Such path may however be unknown, or there may be multiple ones.

Alternative approaches to determine the scores R_i, which do not require defining a root point or an integration path, are gradient-based and propagation-based techniques. Both of these techniques are related and only require one forward/backward pass into the network. Propagation techniques, unlike gradient-based techniques, yield explanations that are conservative and continuous (see Section 2.3 and compare ref 64), and we will choose one such propagation-based technique to further develop the explanation framework in this paper. This technique is *Layer-wise Relevance Propagation* (LRP).[35] LRP leverages the structure of the ML model that has produced the prediction. In particular, it assumes the mapping from input to output is given sequentially by the multiple layers of a deep neural network, i.e., $x \to ... (H_j)_j \to (H_k)_k \to ... \to y$, where $(H_j)_j$ and $(H_k)_k$ denote the collection of activation in two consecutive layers. LRP starts at the output of the network, and decomposes the prediction y to neurons in the layer below. These scores are then backpropagated from layer to layer, using purposely defined propagation rules, until the input features are reached. Let $(R_k)_k$ be the scores resulting from propagating from the top layer until the layer with neurons indexed by k. Propagation to neurons of the layer below can be achieved using a rule of the type

$$R_j = \sum_k \frac{z_{jk}}{\sum_{j'} z_{j'k}} R_k \tag{2}$$

where z_{jk} quantifies the contribution of neuron j to the activation of neuron k. The multiple ways the scores z_{jk} are defined give rise to different LRP propagation rules (cf. ref 65 for an overview). Numerous instantiations of LRP have been proposed, covering models as diverse as convolutional neural networks,[35,65] LSTMs,[66] transformers,[67] classical unsupervised learning models,[68,69] and GNNs.[18] Unlike methods based on integrated gradients, LRP benefits from the internal abstractions of the neural network. In the context of a quantum-chemical application, this allows to attribute the prediction in terms of atoms and their relative distances without having to define meaningful paths for the molecule in the input space.

We note that all explanation techniques we have described so far produce an attribution of the prediction onto individual features, which in our quantum chemical scenario could be atoms and interatomic distances. Note that for GNNs, as we treat here, a decomposition onto individual atoms is readily available from the GNN itself, because it predicts atomic contributions to the final predicted quantity. These explanations may provide useful insights into the model, but are strongly limited in their expressive power and their ability to generate useful hypotheses. For example, they do not say anything about whether the property of interest is the result of individual feature contributions (e.g., localized atom-wise contributions), or whether it arises from the interactions of many of these features (e.g., long chains of atoms spanning the whole molecule). To tackle this question, it is essential to move beyond classical attribution techniques and toward higher-order explanations, that are able to capture those more complex interactions.

2.3. Higher-Order Explanations for GNNs.

Classical first-order attribution methods, as specified above, are limited to single feature attributions when predicting molecular properties. Even in simple scenarios, this approach is insufficient to understand the prediction strategy of the model. We believe that it is important to understand not only the relevance of each individual atom but also the nature and strength of the interactions between atoms from the model's perspective.

A seemingly straightforward approach to obtain interaction strengths would be to slightly perturb a given atom A in various directions and record the change in atomic energy contribution at a target atom B. One could then interpret the change in B's energy contribution as an indicator of the interaction strength from A to B, and vice versa. We caution that this approach is not as straightforward as it seems: A change in interatomic distance necessarily induces changes in other interatomic distances with respect to other atoms. Thus, we are left with the original problem of determining the true contributor of the observed energy change, not to mention the risk of moving outside the manifold of the data distribution. Changing the input to the model always carries this risk, which is a known problem for explainability methods.[70,71]

Instead, we can proceed by extracting the contribution of interacting atoms directly from the structure of the MLFF model. This can be achieved in the context of GNN models by the GNN-LRP method,[18] which we present below. We recall from Section 2.1, that a GNN associates at each layer t and for each node (atom) j a representation $\boldsymbol{H}_{t,j}$, which we abbreviate in the following as $\boldsymbol{H}_j$.

A naive application of LRP to this architecture would start at the output and redistribute the predicted value backward, traversing the multiple atom representations at each layer. The procedure would stop when the first layer is reached, where relevance scores can be mapped to atoms according to their representation in the first layer. This procedure, however, does not account for the way the different atoms have exchanged messages in the higher layers. GNN-LRP addresses this shortcoming by recording the path that relevance propagation messages have taken, and this is achieved by applying a slight modification to eq 2

Table 1. Interaction Range and Many-Bodyness Statistics[a]

model	property	data	interaction range measures			many-bodyness $\bar{\gamma}$ (eq 13)
			$\lambda_{0.001}^{\text{thresh}}$ (eq 8)	λ_1^{pow} (eq 9)	λ_4^{pow} (eq 9)	
3L SchNet	energy	QM9	4.14	1.62	2.75	0.85
3L SchNet	dipole	QM9	6.34	2.64	3.37	1.40
3L SchNet	HOMO	QM9	7.04	3.10	3.93	0.87
3L SchNet	LUMO	QM9	7.01	3.03	3.76	0.95
3L PaiNN	energy	QM9	3.88	1.64	2.44	0.92
3L PaiNN	dipole	QM9	4.18	1.68	2.70	1.10
3L PaiNN	HOMO	QM9	6.97	2.56	3.43	0.93
3L PaiNN	LUMO	QM9	6.96	2.63	3.56	1.10
1L PaiNN	energy	Ac-Ala3-NHMe	8.55	2.63	4.41	0.70
2L PaiNN	energy	Ac-Ala3-NHMe	5.57	2.18	3.17	0.54
3L PaiNN	energy	Ac-Ala3-NHMe	4.03	2.26	2.98	0.80
4L PaiNN	energy	Ac-Ala3-NHMe	2.92	1.79	2.50	1.00
5L PaiNN	energy	Ac-Ala3-NHMe	2.57	1.71	2.29	1.70

[a]For the experiments with QM9 data, networks were trained and evaluated on the indicated property. For the experiments with Ac-Ala3-NHMe, networks were trained on energies and forces, and evaluated on the energies. For the interaction range, the thresholded range (eq 8) with $p_{\min} = 0.001$ is displayed, and additionally the first and fourth generalized expectation of the walk-length distribution (eq 9 with $a = 1$ and $a = 4$). The many-bodyness has been evaluated with eq 13. For an extended table, see Table S1.

$$R_{jkl\ldots} = \frac{z_{jk}}{\sum_{j'} z_{j'k}} R_{kl\ldots} \tag{3}$$

In other words, we strip the pooling operation $\sum_k$ and retain the index k in the propagated relevance score. Propagating through all layers of the GNN, we end up accumulating more and more indices, resulting in relevance scores over sequences of nodes (referred to as "walks" $\mathcal{W}$). These walks are of length $T + 1$, where T is the depth of the GNN. So far, for the simplicity of the presentation, we have assumed that each atom is represented by one neuron. However, in real GNNs, it is represented by m neurons, meaning $\boldsymbol{H}_j = (H_j^b)_{b=1}^m$. Taking this into account, we need to extend eq 3 as

$$R_{jkl\ldots}^b = \sum_c \frac{z_{jk}^{bc}}{\sum_{b',j'} z_{j'k}^{b'c}} R_{kl\ldots}^c \tag{4}$$

where b denotes a neuron associated with node j, c denotes a neuron associated with node k, and z_{jk}^{bc} quantifies the contribution of the neuron b in node j to the activation of neuron c in node k.

A visual description of the method, along with an explanation of how it is used to quantify the model's physical properties, can be seen in Figure 2. The GNN-LRP method is theoretically founded in the higher-order Taylor decomposition of the model's prediction and can be seen as a generalization of LRP[35] and deep Taylor decomposition.[72] Furthermore, as shown in ref 18, it satisfies the axiom of conservation, namely

$$\sum_{\mathcal{W}} R_{\mathcal{W}} = y \tag{5}$$

where y is the predicted value at the output of the GNN. The latter allows us to view the GNN-LRP explanation as a decomposition of the GNN output (e.g., predicted molecular energy) in terms of all the walks $\mathcal{W}$ on the molecular graph. The complexity of the explanation method increases exponentially with the number of layers, however, there are ways to lessen the computational complexity from exponential to polynomial.[73,74]

2.4. Walk-Importance and Walk-Distance.

In the following, we describe how we use the walk-relevances obtained from GNN-LRP[18] to evaluate different properties of the model and its prediction strategy (for an algorithm, see Section S1). One quantity we will use throughout is the measure of *importance* for a walk $\mathcal{W}$ which we define by

$$\mathbb{P}(\mathcal{W}) = \frac{1}{Z}|R_{\mathcal{W}}| \tag{6}$$

where $\mathbb{Z} = \sum_{\mathcal{W}} |R_{\mathcal{W}}|$. Note that $\mathbb{P}(\mathcal{W})$ is a probability distribution of $\mathcal{W}$, i.e., $\mathbb{P}(\mathcal{W})$ has values between 0 and 1, and $\sum_{\mathcal{W}\in\Omega} \mathbb{P}(\mathcal{W}) = 1$, where Ω is the set of all walks for a given atomistic system.

One of the questions we are interested in is how long the range of interactions between atoms, as seen by the model, are. In particular, for any higher-order message $\mathcal{W}$ we can consider some distance $d(\mathcal{W})$ that a walk $\mathcal{W}$ traverses on the molecule. One natural option for such a distance measure is the diameter of the smallest sphere that encloses all atoms in the walk $\mathcal{W}$. This is given by

$$d(\mathcal{W}) := \max_{i,j\in\mathcal{W}} \|\boldsymbol{r}_i - \boldsymbol{r}_j\| \tag{7}$$

where $\|\cdot\|$ denotes the Euclidean norm. In the remainder of this text, we use this distance measure, for example, when we develop more advanced concepts like the interaction range of an MLFF.

2.5. Interaction Range.

An important factor to evaluate is the distance at which atoms still have a significant influence on one another. Although short-range interactions, particularly those between directly bonded atoms, dominate the total energy of a molecule, it is the long-range interactions that, despite their small magnitude, are responsible for interesting macroscopic behavior like protein folding.[6,75,76]

However, modeling long-range interactions in MLFFs also brings a significant computational cost, as the number of interacting atoms scales roughly cubically with distance (due to the increasing volume of the cutoff sphere). For these reasons, it is crucial to get a sense of the range of interaction which the model still takes into account.

We propose to measure interaction range by looking at the maximum distance among walks that are important, i.e., not assigned a non-negligible probability, as measured by eq 6. To this end, we set a probability threshold $p_{\min} = 0.001\max_{\mathcal{W}\in\Omega}\mathbb{P}(\mathcal{W})$ based on which we can search for a walk with maximum distance

$$\lambda_{0.001}^{\text{thresh}} = \max_{\{\mathcal{W}|\mathbb{P}(\mathcal{W})\geq p_{\min}\}} d(\mathcal{W}) \tag{8}$$

Note that not including an importance threshold, or setting it to zero, would be akin to always return the theoretically maximum walk length, which is independent of the solution learned by the GNN model.

As an alternative measure of interaction range, we consider a high-order statistic of the distribution of walk lengths. A simple such statistic, that retains a distance-based interpretation, is the "generalized expectation"

$$\lambda_a^{\text{pow}} = (\mathbb{E}_{\mathcal{W}\sim\mathbb{P}}[d(\mathcal{W})^a])^{1/a} \tag{9}$$

where a is a parameter. Setting $a = 1$ corresponds to measuring the expected distance, and $a = \infty$ the maximum distance. With the same aim of focusing on large distances, but discarding negligibly probable ones, we opt for the value $a = 4$ in our experiments, which is closely related to the kurtosis commonly used to model peaks in a data distribution.

Unless otherwise noted, in all figures in this article, the threshold-based measure defined in eq 8 is used. We consider both measures valuable, and to show that the conclusions drawn in this paper are not dependent on the choice of range measure, both measures are reported for all experiments in Table 1.

2.6. Attributing Atom Interaction Strength. We now want to consider the strength of interaction between two atoms i and j. Chemically, the interaction strength between two atoms in a molecule is not well-defined. The presence of other atoms in the neighborhood and the resulting many-body behavior makes it impossible to measure the 2-body interaction strength in isolation. Nevertheless, multiple approaches to measure the interaction strength exist. For example, the Laplacian of the electron density at a critical point along the bond path can be seen as correlating with the interaction strength.[77]

In this study, we develop a new measure for the interaction strength as seen by a GNN. We focus on two different approaches. In the first approach we want to consider all possible walks $\mathcal{W}$ that traverse the atoms i and j, but can also traverse other atoms in the molecule. We call this the *inclusive* interaction strength, because it is incorporating the context of the interacting atoms as well. We define this interaction strength by

$$s_{ij}^{\text{incl}} := \sum_{\{\mathcal{W}\,|\,i\in\mathcal{W}\wedge j\in\mathcal{W}\}} \mathbb{P}(\mathcal{W}) \tag{10}$$

Another approach to measure the interaction strength would be to consider all walks $\mathcal{W}$ that contain only the atoms i and j. In other words, it consists *exclusively* of walk contributions corresponding to interactions between i and j, without the incorporation of the surrounding atoms. Formally, this can be given by

$$s_{ij}^{\text{excl}} := \sum_{\{\mathcal{W}\,|\,\text{set}(\mathcal{W})=\{i,j\}\}} \mathbb{P}(\mathcal{W})$$

where set($\mathcal{W}$) is the set of atom indices in $\mathcal{W}$.

We decided that it is generally more important to measure the interaction strength of two atoms in the context of their surrounding, therefore we use the inclusive measure s_{ij}^{incl} in the remainder of this text.

2.7. Measuring Many-Bodyness. We refer to many-bodyness as the property where the interaction strength between two atoms is influenced by other atoms in the neighborhood. In other words, we mean by many-bodyness the influence of interactions that are of degree higher than 2-body. In the context of MLFFs, measuring many-bodyness is of particular interest because it highlights a fundamental difference from mechanistic force fields. To illustrate this, assume a simplified force field based on a two-body expansion. In this case, the atom−atom interaction energy is fully isotropic: no matter where in the molecule the two atoms are positioned, the energy contribution will always be the same. Even real-world force-fields that do use higher-order terms usually do not go above 4-body terms. And these 4-body terms are only among chains of covalently bonded atoms. Atom pairs at higher (nonbonded) distances are modeled with 2-body terms only in mechanistic force fields.[29,30]

This is incompatible with physical reality: the atomic neighborhood that atoms are embedded in plays a fundamental role in their interaction. The promise of MLFFs is that they learn to capture this many-body nature better, but it has yet to be shown to which amount this is actually the case.

We propose a definition for the many-bodyness of atom−atom interactions within a molecule. We would like to express by how many orders of magnitude the interaction strength differs for equally distant pairs of atoms. We define

$$\{s\}_R := \{s_{ij} \,\big|\, \|\boldsymbol{r}_i - \boldsymbol{r}_j\| = R\} \tag{11}$$

to be the set of atom−atom interaction strengths for which the distance is R. This formulation is based on the continuous distribution of distances. In practice, where we have limited amounts of data, the condition of equality has to be relaxed to approximate equality: $\|\boldsymbol{r}_i - \boldsymbol{r}_j\| \approx R$. In other words, we are distributing the atom pairs into bins along the interatomic distance. Then, we define the many-bodyness at a distance R as

$$\gamma(R) := \log_{10}\left(\frac{P_{100}(\{s\}_R)}{P_{10}(\{s\}_R)}\right) \tag{12}$$

and, for easier comparison, also the average many-bodyness as a scalar

$$\bar{\gamma} := \frac{1}{R_{\max}} \int_0^{R_{\max}} \gamma(R)\, \mathrm{d}R \tag{13}$$

where $P_{100}(.)$ and $P_{10}(.)$ are percentile functions that return the 100th and 10th percentile. We use the 10th instead of the zeroth percentile to be less sensitive to outliers. Using base 10 for the logarithm instead of e is chosen such that the resulting quantity can be interpreted as orders of magnitude.

Note that this measure of many-bodyness would be equal to 0 for 2-body classical force fields (beyond the bonded cutoff distance), because all atom pairs of same elements at equal distance lead to the same energy contribution.

3. RESULTS

To analyze the first two chemical principles (i) strength of interactions is atom-type and property dependent, and (ii)

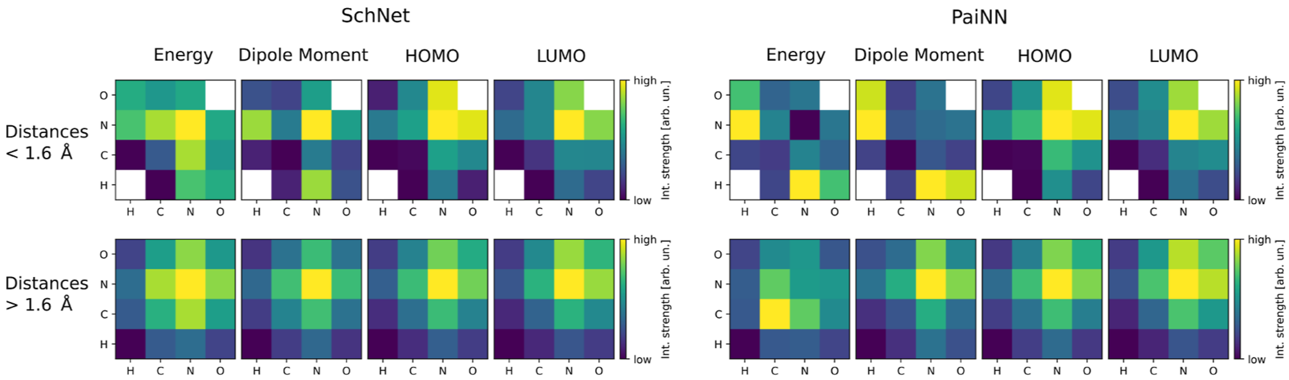


Figure 3. Mean log interaction strength for pairs of elements separated into "bonded" (<1.6 Å) and "nonbonded" (>1.6 Å), and different quantum chemical properties. The networks are 3-layer SchNet and PaiNN trained on the QM9 data set.

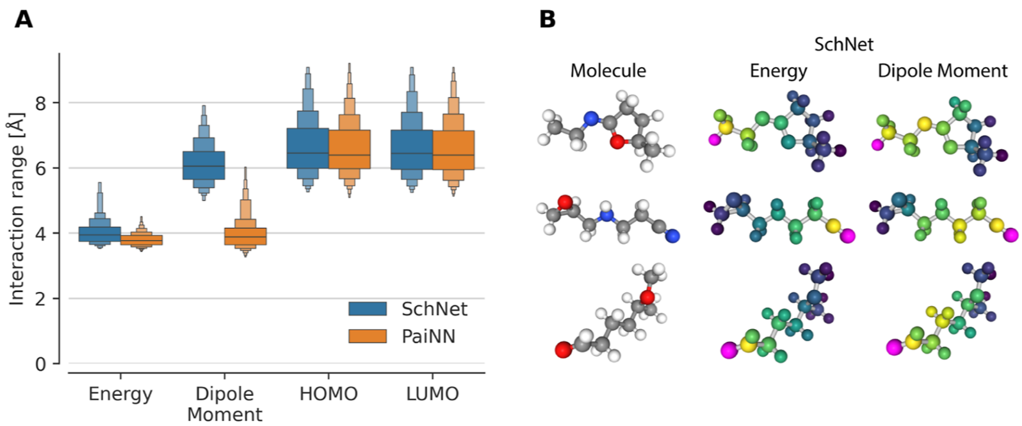


Figure 4. Interaction range. (A) SchNet and PaiNN interaction ranges (eq 8) calculated for models trained on various properties of the QM9 data set. Energy is an extensive property, while dipole moment, HOMO and LUMO are intensive properties. (B) Examples of interaction strength for several molecules from QM9. The chosen atom is highlighted in purple, and the color scale ranges from dark blue (indicating weak interactions) to yellow (indicating strong interactions).

intensive properties require larger interaction range than extensive properties, we train SchNet and PaiNN models on four properties of the QM9 data set. The properties are atomization energy, dipole moment, and highest occupied molecular orbital (HOMO) energies and lowest unoccupied molecular orbital (LUMO) energies. For the other two chemical principles (iii)−(iv), we trained SchNet and PaiNN on the molecule Ac-Ala3-NHMe from the MD22 data set. The models were trained using SchNetPack.[78,79] For details of how the networks in this study were trained, cf. Section S6.

3.1. Chemical Principle 1: The Strength of Interactions Is Atom-Type and Property Dependent. The nature of chemical interactions is fundamentally tied to the electronic configurations and corresponding atomic numbers of the elements involved. Within MLFFs, atomic numbers are encoded per-atom during training, resulting in learned interaction strengths that differ for each atom type, as expected. Figure 3 illustrates the averaged interaction strength between four pairs of elements in the QM9 data set. We removed molecules containing fluorine from our evaluation set due to its low occurrence, with only 3 molecules in the evaluation set containing it. The atom pairs for four models each of the SchNet and PaiNN architectures are categorized into two length-scales: bonded (<1.6 Å) and nonbonded (>1.6 Å) interactions. Examination of these matrices reveals a clear atom-type and property dependence that the models have captured during training.

Interestingly, the interaction patterns differ significantly when comparing models within one architecture trained on different properties, and, for some properties, are drastically different when comparing the relevance between two architectures. This disparity is particularly evident when analyzing the "bonded" interactions occurring within the 0.5−1.6 Å range (first row in Figure 3). For instance, while nitrogen−nitrogen interactions are deemed to be the strongest for energy prediction in SchNet, they are the weakest in the PaiNN architecture. Although we cannot definitively assert which representation is more accurate, it is reasonable to assume that a correct quantum projection exists for an atom-centered molecular basis representation.[80] In particular, a recent study introduced a second quantization framework that partitions long-range many-body dispersion interaction energy into atom−atom (or fragment−fragment) contributions.[81] While this approach has not yet been extended to the total interaction energy, it can, in principle, be generalized to accomplish this. This in turn would enable a similar interaction strength analysis as was done in the current study and serve as a sort of "ground truth" for correct interaction strength distributions. The discrepancy of interaction patterns between architectures underscores the importance of employing explainable artificial intelligence (XAI) techniques to analyze and interpret these complex relationships.

As we extend our examination to nonbonded interactions, we observe that the interaction patterns become more

https://doi.org/10.1021/acs.jctc.4c01424
J. Chem. Theory Comput. 2025, 21, 714−729

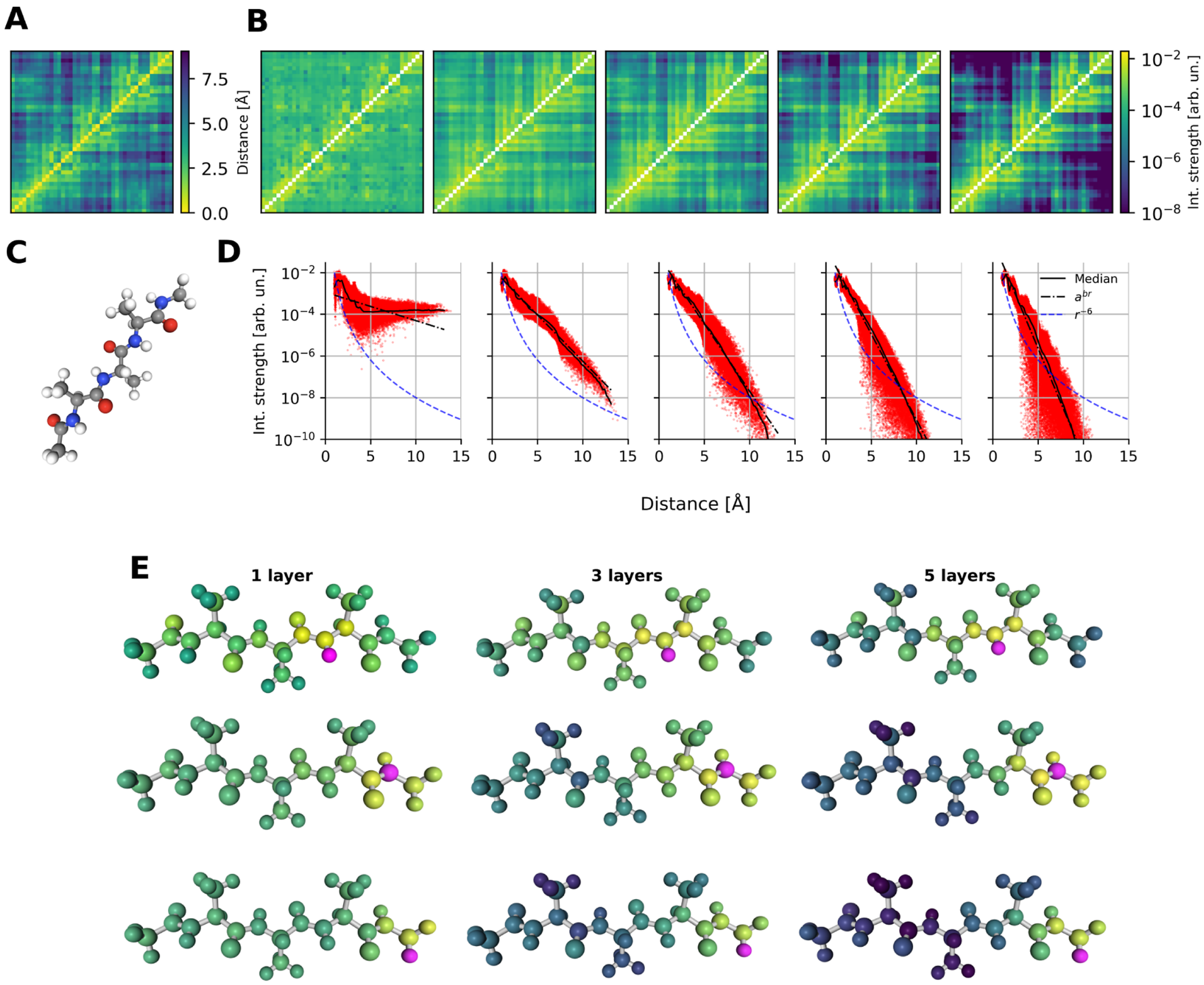


Figure 5. Interaction strength of atom pairs in the tetrapeptide Ac-Ala3-NHMe. (A) Distance matrix of one conformation of Ac-Ala3-NHMe. (B) Atomic interaction strengths for 1 to 5-layer PaiNNs. All matrices were computed for the same randomly chosen conformer as in A. (C) Structure of Ac-Ala3-NHMe. (D) Atomic interaction strengths as a function of distances for PaiNNs with 1 to 5 interaction layers, evaluated on Ac-Ala3-NHMe from MD22. The black lines indicate the median and the best fit of an exponential function, while the blue dashed line represents decay with r^{-6}, a common model for London dispersion decay. (E) Examples of interaction strength for chosen (purple) atoms averaged over 100 conformations. The colorscale is consistent with B.

consistent across all properties and architectures. The atom-type dependence is visually preserved; however, it is now more closely related to the average distance between atom pairs. There is one interesting exception: As already seen at the "bonded" distance, once again for the PaiNN energy model, the nitrogen−nitrogen interaction strength is weak, whereas it is the strongest in all the other settings. The observed decay of interaction strength with distance will be analyzed in more detail within Section 3.4.

3.2. Chemical Principle 2: Intensive Properties Require Larger Interaction Range than Extensive Properties. The QM9 data set provides several quantum chemical properties, of which some are extensive and some are intensive. Extensive properties can be thought of as "the whole is the sum of the parts", i.e., additive local contributions sum toward the final quantity. Intensive properties are the opposite, where the quantity can only be determined by taking into consideration the entire molecule.

The extensive property we considered is the atomization energy. The intensive properties were the dipole moment, HOMO and LUMO energies. As described above, separate models of SchNet and PaiNN were trained for each property. The interaction range was determined with Formula 8. Figure 4A shows that training on the intensive properties causes the models to learn to use a longer interaction range than the extensive properties, which is what we expected. Figure 4B illustrates a difference of 2 Å in the interaction range of SchNet models trained on energies and dipole moments for three molecules from the QM9 data set.

The interaction ranges learned by SchNet and PaiNN are remarkably similar for three properties. SchNet and PaiNN have different message-passing schemes, with SchNet being rotation invariant, and PaiNN rotation equivariant. The only property in which they diverge is the dipole moment. This discrepancy is interesting, because although the dipole moment is considered an intensive property, it can potentially also be

https://doi.org/10.1021/acs.jctc.4c01424
J. Chem. Theory Comput. 2025, 21, 714−729

seen as an extensive property: Since the overall dipole is computed as a charge-weighted sum of (centered) position vectors, each atom only needs to predict its local charge density. This local charge density prediction could potentially be treated as similar to localized energy-contributions, so from the perspective of the model, the dipole moment could be treated like an extensive property. PaiNN's interaction range for the dipole moment is similar to that of the energy, which would indicate that PaiNN indeed treats the dipole moment like an extensive property.

3.3. MD Stability as an Additional Performance Measure. Data sets used for benchmarking MLFFs often come from single trajectory MD simulations.[82,83] In such data sets, the vast majority of samples are drawn from a small set of metastable states (local minima of the potential energy surface). As a result, state-of-the-art MLFFs achieve very low test errors, but this accuracy stems from the fact that the network has a rather "easy" interpolation task where many conformations are close to the energy minima. As an additional performance measure, it has been proposed to perform MD simulations with the MLFF and count how many simulations are unstable.[21,52]

All following experiments were computed from networks trained on the Ac-Ala3-NHMe molecule (Figure 5C) from the MD22 data set.[83] Ac-Ala3-NHMe is a tetrapeptide containing 42 atoms and can exist in a folded and unfolded state, which makes it particularly interesting to analyze.

To show how the explainability framework in this study can be used to identify models which use chemically implausible prediction strategies, we trained 5 versions of PaiNN with different hyperparameters. The goal was to create a spectrum of models which range between common hyperparameters[46,84] to rather extreme hyperparameters which we anticipated to lead to models with chemically implausible representations. We varied the amount of interaction layers L of PaiNN between 1 and 5 and adjusted the cutoff c such that $L{\cdot}c = 15$ Å, which is more than enough to cover the entire length of Ac-Ala3-NHMe even in its unfolded state. The range of networks, with "unreasonable" parametrizations at both ends, was chosen purely for didactic purposes: All networks were able to achieve a very low test error (see paragraph below) and interesting behavior can be observed at both ends of the spectrum.

For instance, it was known from previous studies[46,84] that cutoff lengths around 5 Å are well suited for MLFFs, and going significantly below 5 Å, as we did here, impedes performance, but it is not entirely clear *why* such short cutoff lengths do not work well (see Section 3.5).

Additionally to the number of interaction layers and the cutoff length, the number of radial basis functions and the embedding sizes were adjusted to keep the five networks comparable (see Table 2): The number of radial basis functions were varied such that their spacing along the distance between atoms is the same for all models. The embedding size was varied such that the total number of parameters is roughly equal. The 1-layer network is an exception, we set its embedding size to a larger value to keep its generalization error similar to the other networks. While it was not possible to keep the final generalization errors exactly equal, we note that all errors were far below what is generally considered "chemically accurate" (1 kcal/mol).

We conducted 30 MD simulations of Ac-Ala3-NHMe, with three simulations for each of 10 different starting configurations. The time-step of the integrator was 0.5 fs and the MD trajectories were run for 1 ns, totaling 2 million time-steps. The simulations were performed in the canonical (*NVT*) ensemble at 500 K with a time constant of 5 fs. An MD trajectory was considered unstable if the potential energy of the molecule went outside the range −200 to 200 kcal/mol. In practice, this typically means that an atom of the molecule dissociated, which leads to an abrupt change in the energy. Table 2 shows how many MD trajectories per network were unstable.

Table 2. Test Accuracies and MD Instability for Versions of PaiNN with Various Amounts of Interaction Layers[a]

N_L	cutoff	N_{rbf}	$N_{emb.}$	property	RMSE	MAE	MD failures
1	15	60	512	energy	0.13	0.10	22/30
				forces	0.25	0.19	
2	7.5	30	157	energy	0.17	0.13	0/30
				forces	0.25	0.18	
3	5	20	128	energy	0.11	0.09	0/30
				forces	0.14	0.10	
4	3.75	15	111	energy	0.28	0.22	4/30
				forces	0.27	0.19	
5	3	12	100	energy	0.41	0.33	11/30
				forces	0.33	0.24	

[a]The cutoffs are measured in Å, energy in kcal/mol, and forces in kcal/mol/Å. RMSE: root mean squared error, MAE: mean absolute error. Data splits (n_{total} = 85k): $n_{train} = 0.85 \times n_{total}$, $n_{val} = 0.1 \times n_{total}$, $n_{test} = 0.05 \times n_{total}$.

It can be seen that the 1- and 5-layer networks were unstable, whereas the 2-, 3- and 4-layer networks were mostly stable. Note that the differences in force RMSE (root mean squared error) between the stable and unstable networks were negligible, with the most unstable network (1 layer) having one of the lower force errors. This indicates that the test error is not a good measure of how well a network will actually generalize, a finding that we replicate from other studies.[20,21,52] In the following sections, we will relate the interaction range and many-bodyness obtained from the adapted explainability framework introduced in this paper to the MD stability of these networks.

3.4. Chemical Principle 3: Interaction Strength Decreases Polynomially with Distance. It is generally expected that the interaction strength between atoms beyond covalent bonds decreases with distance. However, there is no universal functional form to express this decrease. For example, the Coulomb force decreases with the inverse of the squared distance r^{-2}. Due to electric field screening effects, the effective decrease is typically much more rapid. In the Lennard-Jones potential, London dispersion forces decrease with r^{-7} (due to the r^{-6} term in the potential). What most decay laws have in common is that the decrease is proportional to a polynomial of the distance. We therefore expected to find that the interaction strength as seen by MLFFs would also decrease polynomially, which would directly imply that the interaction energy and the forces similarly decrease polynomially.

We tested this hypothesis using the 1−5-layer networks introduced above. Figure 5B presents exemplary interaction matrices for a single conformation of the tetrapeptide, alongside a matrix of pairwise atomic distances. These interaction matrices display the interaction strengths between atom pairs for the five different models, enabling qualitative comparison. In the 1-layer model, the interaction strengths are nearly uniformly distributed across all distances. As the number of layers increases, the interaction matrices become more diverse, indicating stronger interactions between atoms in close proximity and diminishing interaction strengths as the distance grows. This qualitative observation is quantitatively supported by statistical analyses across multiple samples, as shown in Figure 5D. It shows the relationship between the interatomic distance and the interaction strength. We first note that all networks show some decay with distance, but the degree with which the strength decays differs considerably. The 1-layer network plateaus after around 5 Å, which is chemically implausible and an indication for the poor MD stability of this network.

While the interaction strength decays in all of the four other networks, none of them exhibit a power law decay. In all cases, a decay modeled with an exponential is a better fit (see the helper lines in the plots indicating the best fit of an exponential curve). With each added interaction layer, the decay is faster. However, it is not immediately clear from this why the 5-layer network is unstable in MD trajectories, whereas the 4L network is mostly stable. We return to this question in Section 3.5.

In GNNs, the number of walks between two atoms decreases exponentially with distance. For an approximate formula for this decrease, see Section S3. The number of walks between two atoms is directly related to their interaction strength. Recall from eq 10 that the interaction strength is formed as a sum of the relevances of each walk between two atoms. Therefore, what the fact that the number of walks decreases exponentially means is that an exponential decrease of the interaction strength is "baked in" to GNNs, as long as they have a cutoff which is shorter than the length of the atomistic system that they operate on. Such an architectural constraint is also called an inductive bias in ML literature. An obvious question is what would happen in the absence of such an inductive bias, i.e. what representation would the model learn if a decay of the interaction strength is not architecturally forced. We turn to this question next.

3.4.1. Interaction Strength without a Cutoff. For the 1-layer network discussed above, we observed that it does not learn a consistent decay of the interaction strength and plateaus after about 5 Å. However, GNNs are usually trained with at least three interaction layers to form many-body representations, so the fact that the 1-layer network did not learn a decay does not imply that GNNs will fundamentally fail to learn a decay of the interaction strength.

In order to test whether the failure to learn a decay after 5 Å is an isolated issue of having only one interaction layer, we trained PaiNNs with 3 interaction layers, with a cutoff length of 15 Å, which is longer than the maximum length of the molecule. To further "free" the network from range constraints, we also removed the cosine cutoff, which is applied in many GNN-MLFF architectures and forces a cosine-shaped decay of the message features toward the end of the cutoff distance.

The results show that in the absence of a cutoff and even without a cosine cutoff function applied, the model does not learn a chemically appropriate decay of the interaction strength (Figure 6, right). While the interaction strength does decay initially, it increases again at higher distances. The tendency of the model training to increase the interaction strengths at higher distances is an effect which we observed throughout this study, and is explored in more depth in Section 3.6. For the PaiNN model with a cosine-cutoff function applied, the interaction strength does decay with distance and does not increase again. However, due to the results of the model

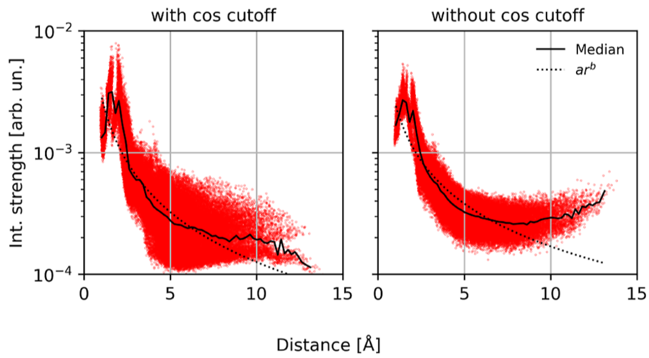


Figure 6. 3-layer PaiNNs on Ac-Ala3-NHMe, with a cutoff length of 15 Å. This length is more than the maximum length of the molecule, so each atom in each layer of the GNN "sees" all other atoms directly. Left: with a cosine cutoff function; Right: without any cutoff function. The forces RMSE was 0.163 and 0.159 kcal/mol/Å respectively, so both networks had an almost equal validation error. The MD instability was 1/30 with cosine cutoff function, and 24/30 without. The dotted line is the best fit of a monomial function to the data.

without cosine cutoff, we know that this continuous decrease of the interaction strength is the effect of the cosine cutoff and not a learned behavior. This indicates that the models indeed will not learn the correct quantum chemical representation without architectural constraints.

To investigate whether this chemical implausibility of the interaction strength reveals weaknesses in the learned representations, we performed the same MD-stability tests as described in Section 3.3. Note that the forces RMSE of both networks were similar, at 0.163 and 0.159 kcal/mol/Å, respectively. Despite this excellent validation error, the MD stability differed drastically. Only 1 out of 30 trajectories of the model with the cosine cutoff was unstable, compared to 24 out of 30 trajectories of the model without cosine cutoff. This indicates that networks that appear to be chemically implausible based on our analysis, do in fact extrapolate badly to new data, even if they seem indistinguishable from well functioning networks based solely on validation error.

3.5. Chemical Principle 4: Many-Bodyness. We defined the many-bodyness of atomic interactions as the base-10 log ratio of the strongest to the weakest interaction strength for atom pairs at the same distance (Section 2.7). The expectation is that atomic interactions are influenced by other atoms in the neighborhood, modulating the interaction.

We contrast the expectation of many-bodyness of the interaction as seen by MLFFs with classical force fields. The 2-body terms in classical force fields are fully isotropic, as the effect of other atoms can not be taken into consideration by definition. 3- and 4-body terms do take other atoms into consideration and would lead to the possibility of at least some many-bodyness even in classical force fields. However, 3-, 4- and higher order terms are typically only applied to (chains of) covalently bonded atoms. This means that atom pairs at higher distances will experience strictly isotropic interactions in classical force fields.

As seen already in Section 3.4 (Figure 5DE), the interaction strengths as seen by MLFFs differ significantly at the same distance. At first thought, one may assume that the 1-layer network has no many-bodyness, because it considers only 2-body terms. However, this is not true: In the 1-layer network, each atom receives input from all other atoms in the molecule, and then integrates all of these "messages" into its final

https://doi.org/10.1021/acs.jctc.4c01424
J. Chem. Theory Comput. 2025, 21, 714−729

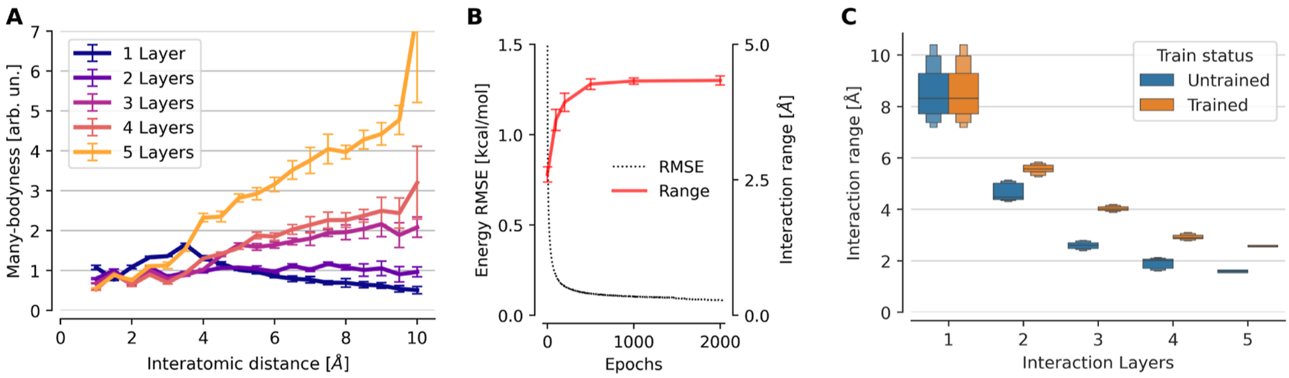


Figure 7. (A) The many-bodyness γ measured on several bins along the atom-pair distance. Note that the measure of the many-bodyness (eq 12) is logarithmic (base 10), i.e., a value of 1 unit of many-bodyness indicates that the lowest to the highest interaction strength in a bin differs by a factor of 10. Networks: 1−5-layer PaiNNs trained and evaluated on Ac-Ala3-NHMe. (B) Evolution of the energy RMSE and the mean interaction range (eq 8) during training. Error bars represent standard deviation. The 3-layer PaiNN architecture was trained on Ac-Ala3-NHMe. (C) Interaction range (eq 8) for untrained and trained (on Ac-Ala3-NHMe) variants of the PaiNNs with 1−5 interaction layers.

prediction. A 1-layer GNN is therefore not equivalent to 2-body terms.

We also compared the many-bodyness across interaction distances (Figure 7A). It is known that local interactions should be influenced by the other atoms in the neighborhood, which is therefore what we expected to find in GNNs. We found that the many-bodyness grows with increasing distance in the 3-, 4- and 5-layer networks. For the 2-layer network, the many-bodyness stays roughly constant and increases only slightly for larger interatomic distances, and for the 1-layer network it decreases with distance.

The many-bodyness of the 5-layer network seems excessively high: It reaches a value above 4 even at relatively short distances (below 8 Å), which means that the interaction strengths differ by a factor of more than 10,000. We hypothesize that this high many-bodyness is not physical and is an indicator as to why the MD-trajectories that were run with this network are often unstable.

3.6. How Does Training a GNN Change the Interaction Range?

Using our measure of the interaction range can not only be used on the fully trained network. Instead, the evolution of these measures can be tracked throughout the training of the GNN. Doing this analysis uncovers that the interaction range is increased significantly during training, but only after the error on the test set already almost converged (Figure 7B). As the validation error approaches a plateau, the interaction range keeps increasing. We hypothesize that this is because the error can initially be reduced by taking into consideration only the immediate surrounding of each atom, whereas to remove the last remaining bits of error, a wider context needs to be considered.

A noteworthy finding is that training of the model increases the interaction range in all cases, even when the range in the untrained model (i.e., a model with randomly initialized weights) starts out higher than what is likely physically appropriate, as is the case in the 1-layer network. Figure 7C shows this effect: For each of the 1- up to 5-layer PaiNNs, the trained variant has a longer interaction range than the untrained one. For the 1-layer variant, the intuitive interaction range measure which is shown in this figure does not distinguish between trained and untrained, because in both cases, the threshold for the range cutoff is higher than the length of the molecule. The interaction range measure based on the fourth moment of the walk distribution (Table 1) however shows that the range of the trained 1-layer network is indeed significantly higher.

4. DISCUSSION

MLFFs have recently become highly popular, because they are considered a useful compromise between classical force fields (quick, less accurate) and first-principles electronic structure calculations (slow, more accurate). A zoo of different kernel and neural network models (e.g., refs 13, 51, 52, 54, 85−87) have emerged, offering many possible modeling approaches.

Throughout the use of such MLFFs, the community has been striving to gain a deeper understanding of their potential limitations. For general applications in the sciences, XAI methods have proven to be invaluable.[61,88−91] Moreover, XAI methods have been used to debug models, to gain novel insights and to capture whether or not suspected/expected structures or knowledge are embodied in the respective ML architectures.[17,92] However, the use of XAI methods in theoretical chemistry has so far been rather limited (some approaches are e.g. refs 42, 80, and 93), which may be partly due to the fact that first-order explanation techniques are insufficient to capture the complexities of atomistic systems.

In this work we have developed an explanation framework based on higher-order explanations[18] and applied it to two popular MLFFs (SchNet and PaiNN). This framework was then used to examine to what extent these models reflect known chemical principles after training. We found that the models were able to extract physical relationships from data just by learning to predict a set of energies and forces.

At the same time, one important property, namely that the interaction strength between atom pairs should decrease with a power law, was violated. Indeed, we showed theoretically and experimentally that a fundamental limitation of current GNN architectures is that the interaction strength decreases exponentially. Especially when imposing a cutoff distance of 4−5 Å, as is common in state-of-the-art MLFFs, this exponential decay leads to distances above 10 Å being barely reachable. This finding can be taken as guidance to design improved GNNs that fulfill power-law properties (or interaction distributions as proposed in ref 81) and can in this manner closer reflect chemistry and physics.

A somewhat troubling finding was that several different instantiations of the GNNs we used (e.g., the variants using too few or too many layers, or unsuitable cutoffs) differed

https://doi.org/10.1021/acs.jctc.4c01424
J. Chem. Theory Comput. 2025, 21, 714−729

significantly in their learned prediction strategy, despite them all having a very low test-set error. This means that a model does not necessarily have to reflect the known chemical principles in order to yield a good test-set error. It had been shown previously[20,21,52] that the test-set error is not necessarily indicative of MD stability—a finding clearly replicated in this study. However, with the XAI-based analysis we propose, we can obtain deeper insights. We can show that models which deviate too far from the principles we proposed will produce unstable MD trajectories, despite these models' low test-set error.

Our findings suggest that ML models applied to chemical systems can still benefit from several improvements. This could lead to enhanced transferability in compositional and structural chemical spaces as well as scalability in terms of system size.

These results show a tangible benefit of analyzing MLFFs with explainability methods. Specifically, they confirm that MLFFs can indeed learn the fundamental physical and chemical principles as expected, which allows a more confident transition of MLFFs from exploratory research to real-world applications.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.4c01424.

Algorithm for the computation of interaction strength, interaction range and many-bodyness, details on the considered GNNs and corresponding relevance attribution, theoretical formula for the exponential decrease of number of walks, additional results on the influence of training data on interaction range, computational details regarding GNN training and MD, analysis of a model overfitting (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Malte Esders − *BIFOLD—Berlin Institute for the Foundations of Learning and Data, 10587 Berlin, Germany; Machine Learning Group, Berlin Institute of Technology, 10587 Berlin, Germany;* orcid.org/0000-0002-9136-914X; Email: esders@tu-berlin.de

Alexandre Tkatchenko − *Department of Physics and Materials Science, University of Luxembourg, L-1511 Luxembourg City, Luxembourg;* orcid.org/0000-0002-1012-4854; Email: alexandre.tkatchenko@uni.lu

Klaus-Robert Müller − *BIFOLD—Berlin Institute for the Foundations of Learning and Data, 10587 Berlin, Germany; Machine Learning Group, Berlin Institute of Technology, 10587 Berlin, Germany; Google Deepmind, 10963 Berlin, Germany; Department of Artificial Intelligence, Korea University, 136-713 Seoul, Korea; Max Planck Institute for Informatics, 66123 Saarbrücken, Germany;* Email: klaus-robert.mueller@tu-berlin.de

Authors

Thomas Schnake − *BIFOLD—Berlin Institute for the Foundations of Learning and Data, 10587 Berlin, Germany; Machine Learning Group, Berlin Institute of Technology, 10587 Berlin, Germany*

Jonas Lederer − *BIFOLD—Berlin Institute for the Foundations of Learning and Data, 10587 Berlin, Germany; Machine Learning Group, Berlin Institute of Technology, 10587 Berlin, Germany*

Adil Kabylda − *Department of Physics and Materials Science, University of Luxembourg, L-1511 Luxembourg City, Luxembourg*

Grégoire Montavon − *Department of Mathematics and Computer Science, Free University of Berlin, 14195 Berlin, Germany; BIFOLD—Berlin Institute for the Foundations of Learning and Data, 10587 Berlin, Germany; Machine Learning Group, Berlin Institute of Technology, 10587 Berlin, Germany*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.4c01424

Author Contributions

$^{\nabla}$T.S. and J.L. contributed equally to this work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was in part supported by the German Ministry for Education and Research (BMBF) under grants 01IS14013A-E, 01GQ1115, 01GQ0850, 01IS18025A, 031L0207D, 01IS18037A, and BIFOLD24B and by BASLEARN—TU Berlin/BASF Joint Laboratory, cofinanced by TU Berlin and BASF SE. K.-R.M. was partly supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (No. RS-2019-II190079, Artificial Intelligence Graduate School Program, Korea University and No. RS-2024-00457882, AI Research Hub Project). A.K. acknowledges financial support from the Luxembourg National Research Fund (FNR AFR Ph.D. Grant 15720828). A.T. was funded by the European Research Council (ERC Advanced Grant FITMOL). We would like to thank Stefan Chmiela, Stefan Gugler, J. Thorben Frank and Stefan Blücher for valuable discussions, and Oliver T. Unke for helpful feedback on the manuscript.

■ REFERENCES

(1) Pfau, D.; Axelrod, S.; Sutterud, H.; von Glehn, I.; Spencer, J. S. Accurate computation of quantum excited states with neural networks. *Science* **2024**, *385*, No. eadn0137.

(2) Szabó, P. B.; Schätzle, Z.; Entwistle, M. T.; Noé, F. An improved penalty-based excited-state variational Monte Carlo approach with deep-learning ansatzes. *J. Chem. Theory Comput.* **2024**, *20*, 7922−7935.

(3) Unke, O. T.; Chmiela, S.; Sauceda, H. E.; Gastegger, M.; Poltavsky, I.; Schütt, K. T.; Tkatchenko, A.; Müller, K.-R. Machine learning force fields. *Chem. Rev.* **2021**, *121*, 10142−10186.

(4) Szabo, A.; Ostlund, N. S. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*; Courier Corporation, 2012.

(5) Piela, L. *Ideas of Quantum Chemistry*; Elsevier, 2006.

(6) Sherrill, C. D.; Sumpter, B. G.; Sinnokrot, M. O.; Marshall, M. S.; Hohenstein, E. G.; Walker, R. C.; Gould, I. R. Assessment of standard force field models against high-quality ab initio potential curves for prototypes of π−π, CH/π, and SH/π interactions. *J. Comput. Chem.* **2009**, *30*, 2187−2193.

(7) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. How robust are protein folding simulations with respect to force field parameterization? *Biophys. J.* **2011**, *100*, L47−L49.

(8) Bengio, Y.; Goodfellow, I.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2017; Vol. *1*.

(9) Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. E. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat. Commun.* **2019**, *10*, 2903.

(10) Zubatyuk, R.; Smith, J. S.; Leszczynski, J.; Isayev, O. Accurate and transferable multitask prediction of chemical properties with an atoms-in-molecules neural network. *Sci. Adv.* **2019**, *5*, No. eaav6490.

(11) Devereux, C.; Smith, J. S.; Huddleston, K. K.; Barros, K.; Zubatyuk, R.; Isayev, O.; Roitberg, A. E. Extending the applicability of the ANI deep learning molecular potential to sulfur and halogens. *J. Chem. Theory Comput.* **2020**, *16*, 4192−4202.

(12) Illarionov, A.; Sakipov, S.; Pereyaslavets, L.; Kurnikov, I. V.; Kamath, G.; Butin, O.; Voronina, E.; Ivahnenko, I.; Leontyev, I.; Nawrocki, G.; Darkhovskiy, M.; Olevanov, M.; Cherniavskyi, Y. K.; Lock, C.; Greenslade, S.; Sankaranarayanan, S. K.; Kurnikova, M. G.; Potoff, J.; Kornberg, R. D.; Levitt, M.; Fain, B. Combining Force Fields and Neural Networks for an Accurate Representation of Chemically Diverse Molecular Interactions. *J. Am. Chem. Soc.* **2023**, *145*, 23620−23629.

(13) Kovács, D. P.; Moore, J. H.; Browning, N. J.; Batatia, I.; Horton, J. T.; Kapil, V.; Magdău, I.-B.; Cole, D. J.; Csányi, G. MACE-OFF23: Transferable machine learning force fields for organic molecules. *arXiv* **2023**, arXiv:2312.15211.

(14) Anstine, D.; Zubatyuk, R.; Isayev, O. AIMNet2: a neural network potential to meet your neutral, charged, organic, and elemental-organic needs. *ChemRxiv* **2023**.

(15) Kabylda, A.; Frank, J. T.; Dou, S. S.; Khabibrakhmanov, A.; Sandonas, L. M.; Unke, O. T.; Chmiela, S.; Müller, K.; Tkatchenko, A. Molecular Simulations with a Pretrained Neural Network and Universal Pairwise Force Fields. *ChemRxiv* **2024**.

(16) Montavon, G.; Samek, W.; Müller, K.-R. Methods for interpreting and understanding deep neural networks. *Digit. Signal Process.* **2018**, *73*, 1−15.

(17) Samek, W.; Montavon, G.; Lapuschkin, S.; Anders, C. J.; Müller, K.-R. Explaining deep neural networks and beyond: A review of methods and applications. *Proc. IEEE* **2021**, *109*, 247−278.

(18) Schnake, T.; Eberle, O.; Lederer, J.; Nakajima, S.; Schütt, K. T.; Müller, K.-R.; Montavon, G. Higher-Order Explanations of Graph Neural Networks via Relevant Walks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 7581−7596.

(19) Miksch, A. M.; Morawietz, T.; Kästner, J.; Urban, A.; Artrith, N. Strategies for the construction of machine-learning potentials for accurate and efficient atomic-scale simulations. *Mach. Learn.: Sci. Technol.* **2021**, *2*, 031001.

(20) Stocker, S.; Gasteiger, J.; Becker, F.; Günnemann, S.; Margraf, J. T. How robust are modern graph neural network potentials in long and hot molecular dynamics simulations? *Mach. Learn.: Sci. Technol.* **2022**, *3*, 045010.

(21) Fu, X.; Wu, Z.; Wang, W.; Xie, T.; Keten, S.; Gomez-Bombarelli, R.; Jaakkola, T. Forces are not enough: Benchmark and critical evaluation for machine learning force fields with molecular simulations. *arXiv* **2022**, arXiv:2210.07237.

(22) Wang, Z.; Wu, H.; Sun, L.; He, X.; Liu, Z.; Shao, B.; Wang, T.; Liu, T.-Y. Improving machine learning force fields for molecular dynamics simulations with fine-grained force metrics. *J. Chem. Phys.* **2023**, *159*, 035101.

(23) Gong, S.; Yan, K.; Xie, T.; Shao-Horn, Y.; Gomez-Bombarelli, R.; Ji, S.; Grossman, J. C. Examining graph neural networks for crystal structures: Limitations and opportunities for capturing periodicity. *Sci. Adv.* **2023**, *9*, No. eadi3245.

(24) Fonseca, G.; Poltavsky, I.; Tkatchenko, A. Force Field Analysis Software and Tools (FFAST): Assessing Machine Learning Force Fields under the Microscope. *J. Chem. Theory Comput.* **2023**, *19*, 8706−8717.

(25) Atkins, P. W.; De Paula, J.; Keeler, J. *Atkins' Physical Chemistry*; Oxford University Press, 2023.

(26) Stone, A. *The Theory of Intermolecular Forces*; Oxford University Press, 2013.

(27) Stone, A.; Price, S. Some new ideas in the theory of intermolecular forces: anisotropic atom-atom potentials. *J. Phys. Chem.* **1988**, *92*, 3325−3335.

(28) Eramian, H.; Tian, Y.-H.; Fox, Z.; Beneberu, H. Z.; Kertesz, M. On the anisotropy of van der Waals atomic radii of O, S, Se, F, Cl, Br, and I. *J. Phys. Chem. A* **2013**, *117*, 14184−14190.

(29) Case, D. A.; Cheatham III, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz Jr, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber biomolecular simulation programs. *J. Comput. Chem.* **2005**, *26*, 1668−1688.

(30) Brooks, B. R.; Brooks, C. L., III; Mackerell, A. D., Jr; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; et al. CHARMM: The biomolecular simulation program. *J. Comput. Chem.* **2009**, *30*, 1545−1614.

(31) Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* **2017**, *8*, 13890.

(32) Xie, T.; Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **2018**, *120*, 145301.

(33) Schütt, K. T.; Gastegger, M.; Tkatchenko, A.; Müller, K.-R. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*; Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., Müller, K.-R., Eds.; Springer International Publishing: Cham, 2019; pp 311−330.

(34) Chong, S.; Grasselli, F.; Ben Mahmoud, C.; Morrow, J. D.; Deringer, V. L.; Ceriotti, M. Robustness of local predictions in atomistic machine learning models. *J. Chem. Theory Comput.* **2023**, *19*, 8020−8031.

(35) Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One* **2015**, *10*, No. e0130140.

(36) Ying, Z.; Bourgeois, D.; You, J.; Zitnik, M.; Leskovec, J. GNNExplainer: Generating Explanations for Graph Neural Networks. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 9244−9255.

(37) Luo, D.; Cheng, W.; Xu, D.; Yu, W.; Zong, B.; Chen, H.; Zhang, X. Parameterized Explainer for Graph Neural Network. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 19620−19631.

(38) Cho, H.; Lee, E. K.; Choi, I. S. Layer-wise relevance propagation of InteractionNet explains protein−ligand interactions at the atom level. *Sci. Rep.* **2020**, *10*, 21155.

(39) Collins, E. M.; Raghavachari, K. Interpretable Graph-Network-Based Machine Learning Models via Molecular Fragmentation. *J. Chem. Theory Comput.* **2023**, *19*, 2804−2810.

(40) El-Samman, A. M.; Husain, I. A.; Huynh, M.; De Castro, S.; Morton, B.; De Baerdemacker, S. Global geometry of chemical graph neural network representations in terms of chemical moieties. *Digit. Discov.* **2024**, *3*, 544−557.

(41) Lederer, J.; Gastegger, M.; Schütt, K. T.; Kampffmeyer, M.; Müller, K.-R.; Unke, O. T. Automatic identification of chemical moieties. *Phys. Chem. Chem. Phys.* **2023**, *25*, 26370−26379.

(42) Jiménez-Luna, J.; Grisoni, F.; Schneider, G. Drug discovery with explainable artificial intelligence. *Nat. Mach. Intell.* **2020**, *2*, 573−584.

(43) McCloskey, K.; Taly, A.; Monti, F.; Brenner, M. P.; Colwell, L. J. Using attribution to decode binding mechanism in neural network models for chemistry. *Proc. Natl. Acad. Sci. U.S.A.* **2019**, *116*, 11624−11629.

(44) Schnake, T.; Jafaria, F. R.; Lederer, J.; Xiong, P.; Nakajima, S.; Gugler, S.; Montavon, G.; Müller, K.-R. Towards Symbolic XAI − Explanation Through Human Understandable Logical Relationships Between Features. *Information Fusion* **2025**, 102923.

(45) Bonneau, K.; Lederer, J.; Templeton, C.; Rosenberger, D.; Müller, K.-R.; Clementi, C. Peering inside the black box: Learning the relevance of many-body functions in Neural Network potentials. *arXiv* **2024**, arXiv:2407.04526.

(46) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K. R. Schnet−a deep learning architecture for molecules and materials. *J. Chem. Phys.* **2018**, *148*, 241722.

https://doi.org/10.1021/acs.jctc.4c01424
J. Chem. Theory Comput. 2025, 21, 714−729

(47) Schütt, K. T.; Unke, O.; Gastegger, M. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *Proceedings of the 38th International Conference on Machine Learning*, 2021; pp 9377−9388.
(48) Zhou, J.; Cui, G.; Hu, S.; Zhang, Z.; Yang, C.; Liu, Z.; Wang, L.; Li, C.; Sun, M. Graph neural networks: A review of methods and applications. *AI Open* **2020**, *1*, 57−81.
(49) Schütt, K. T.; Kindermans, P.-J.; Felix, H. E. S.; Chmiela, S.; Tkatchenko, A.; Müller, K.-R. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. *Adv. Neural Inf. Process. Syst.* **2017**, 992−1002.
(50) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.
(51) Batzner, S.; Musaelian, A.; Sun, L.; Geiger, M.; Mailoa, J. P.; Kornbluth, M.; Molinari, N.; Smidt, T. E.; Kozinsky, B. E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.* **2022**, *13*, 2453.
(52) Frank, J. T.; Unke, O. T.; Müller, K. R.; Chmiela, S. A Euclidean transformer for fast and stable machine learned force fields. *Nat. Commun.* **2024**, *15*, 6539.
(53) Duval, A.; Mathis, S. V.; Joshi, C. K.; Schmidt, V.; Miret, S.; Malliaros, F. D.; Cohen, T.; Lio, P.; Bengio, Y.; Bronstein, M. A Hitchhiker's Guide to Geometric GNNs for 3D Atomic Systems. *arXiv* **2023**, arXiv:2312.07511.
(54) Batatia, I.; Kovacs, D. P.; Simm, G.; Ortner, C.; Csanyi, G. MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 11423−11436.
(55) Frank, J. T.; Unke, O.; Müller, K.-R. So3krates: Equivariant attention for interactions on arbitrary length-scales in molecular systems. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 29400−29413.
(56) Hoffmann, R.; Malrieu, J.-P. Simulation vs. understanding: a tension, in quantum chemistry and beyond. Part A. Stage setting. *Angew. Chem., Int. Ed.* **2020**, *59*, 12590−12610.
(57) Anders, C. J.; Weber, L.; Neumann, D.; Samek, W.; Müller, K.-R.; Lapuschkin, S. Finding and removing clever hans: Using explanation methods to debug and improve deep models. *Inf. Fusion* **2022**, *77*, 261−295.
(58) Kauffmann, J.; Dippel, J.; Ruff, L.; Samek, W.; Müller, K.-R.; Montavon, G. The Clever Hans Effect in Unsupervised Learning. *Nature Machine Intelligence* **2025**, in press.
(59) Geirhos, R.; Jacobsen, J.; Michaelis, C.; Zemel, R. S.; Brendel, W.; Bethge, M.; Wichmann, F. A. Shortcut learning in deep neural networks. *Nat. Mach. Intell.* **2020**, *2*, 665−673.
(60) Zednik, C.; Boelsen, H. Scientific Exploration and Explainable Artificial Intelligence. *Minds Mach.* **2022**, *32*, 219−239.
(61) Klauschen, F.; Dippel, J.; Keyl, P.; Jurmeister, P.; Bockmayr, M.; Mock, A.; Buchstab, O.; Alber, M.; Ruff, L.; Montavon, G.; Müller, K.-R. Toward Explainable Artificial Intelligence for Precision Pathology. *Annu. Rev. Pathol.: Mech. Dis.* **2024**, *19*, 541−570.
(62) Strumbelj, E.; Kononenko, I. An Efficient Explanation of Individual Classifications using Game Theory. *J. Mach. Learn. Res.* **2010**, *11*, 1−18.
(63) Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, 2017; pp 3319−3328.
(64) Montavon, G. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*; Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., Müller, K.-R., Eds.; Springer-Verlag: Berlin, Heidelberg, 2022; pp 253−265.
(65) Montavon, G.; Binder, A.; Lapuschkin, S.; Samek, W.; Müller, K.-R.; Explainable, A. I. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*; Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., Müller, K.-R., Eds.; Springer, 2019; Vol. *11700*, pp 193−209.
(66) Arras, L.; Arjona-Medina, J. A.; Widrich, M.; Montavon, G.; Gillhofer, M.; Müller, K.-R.; Hochreiter, S.; Samek, W.; Explainable, A. I. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*; Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., Müller, K.-R., Eds.; Springer, 2019; Vol. *11700*, pp 211−238.
(67) Ali, A.; Schnake, T.; Eberle, O.; Montavon, G.; Müller, K.-R.; Wolf, L. XAI for Transformers: Better Explanations through Conservative Propagation. In *Proceedings of the 39th International Conference on Machine Learning*, 2022; pp 435−451.
(68) Kauffmann, J. R.; Müller, K.-R.; Montavon, G. Towards explaining anomalies: A deep Taylor decomposition of one-class models. *Pattern Recogn.* **2020**, *101*, 107198.
(69) Kauffmann, J. R.; Esders, M.; Ruff, L.; Montavon, G.; Samek, W.; Müller, K.-R. From Clustering to Cluster Explanations via Neural Networks. *IEEE Transact. Neural Networks Learn. Syst.* **2024**, *35*, 1926−1940.
(70) Dombrowski, A.-K.; Gerken, J. E.; Müller, K.-R.; Kessel, P. Diffeomorphic Counterfactuals With Generative Models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 3257−3274.
(71) Blücher, S.; Vielhaben, J.; Strodthoff, N. Decoupling pixel flipping and occlusion strategy for consistent XAI benchmarks. *arXiv* **2024**, arXiv:2401.06654.
(72) Montavon, G.; Lapuschkin, S.; Binder, A.; Samek, W.; Müller, K. R. Explaining NonLinear Classification Decisions with Deep Taylor Decomposition. *Pattern Recogn.* **2017**, *65*, 211−222.
(73) Xiong, P.; Schnake, T.; Montavon, G.; Müller, K.-R.; Nakajima, S. Efficient computation of higher-order subgraph attribution via message passing. In *Proceedings of the 39th International Conference on Machine Learning*, 2022; pp 24478−24495.
(74) Xiong, P.; Schnake, T.; Gastegger, M.; Montavon, G.; Müller, K.-R.; Nakajima, S. Relevant walk search for explaining graph neural networks. In *Proceedings of the 40th International Conference on Machine Learning*, 2023; pp 38301−38324.
(75) Tkatchenko, A.; Scheffler, M. Accurate Molecular Van Der Waals Interactions from Ground-State Electron Density and Free-Atom Reference Data. *Phys. Rev. Lett.* **2009**, *102*, 073005.
(76) Ambrosetti, A.; Ferri, N.; DiStasio, R. A.; Tkatchenko, A. Wavelike charge density fluctuations and van der Waals interactions at the nanoscale. *Science* **2016**, *351*, 1171−1176.
(77) Bader, R. F. W. *Atoms in Molecules: A Quantum Theory*; Clarendon Press, 1990.
(78) Schütt, K. T.; Kessel, P.; Gastegger, M.; Nicoli, K. A.; Tkatchenko, A.; Müller, K.-R. SchNetPack: A Deep Learning Toolbox For Atomistic Systems. *J. Chem. Theory Comput.* **2019**, *15*, 448−455.
(79) Schütt, K. T.; Hessmann, S. S. P.; Gebauer, N. W. A.; Lederer, J.; Gastegger, M. SchNetPack 2.0: A neural network toolbox for atomistic machine learning. *J. Chem. Phys.* **2023**, *158*, 144801.
(80) Gallegos, M.; Vassilev-Galindo, V.; Poltavsky, I.; Martín Pendás, Á.; Tkatchenko, A. Explainable chemical artificial intelligence from accurate machine learning of real-space chemical descriptors. *Nat. Commun.* **2024**, *15*, 4345.
(81) Gori, M.; Kurian, P.; Tkatchenko, A. Second quantization of many-body dispersion interactions for chemical and biological systems. *Nat. Commun.* **2023**, *14*, 8218.
(82) Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K. R. Machine learning of accurate energy-conserving molecular force fields. *Sci. Adv.* **2017**, *3*, No. e1603015.
(83) Chmiela, S.; Vassilev-Galindo, V.; Unke, O. T.; Kabylda, A.; Sauceda, H. E.; Tkatchenko, A.; Müller, K. R. Accurate global machine learning force fields for molecules with hundreds of atoms. *Sci. Adv.* **2023**, *9*, No. eadf0873.
(84) Gasteiger, J.; Groß, J.; Günnemann, S. Directional Message Passing for Molecular Graphs. In *International Conference on Learning Representations*, 2020.
(85) Noé, F.; Tkatchenko, A.; Müller, K.-R.; Clementi, C. Machine learning for molecular simulation. *Annu. Rev. Phys. Chem.* **2020**, *71*, 361−390.
(86) Fuchs, F.; Worrall, D.; Fischer, V.; Welling, M. SE(3)-Transformers: 3D Roto-Translation Equivariant Attention Networks. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1970−1981.
(87) Keith, J. A.; Vassilev-Galindo, V.; Cheng, B.; Chmiela, S.; Gastegger, M.; Müller, K.-R.; Tkatchenko, A. Combining machine

learning and computational chemistry for predictive insights into chemical systems. *Chem. Rev.* **2021**, *121*, 9816−9872.
(88) Eberle, O.; Büttner, J.; Kräutli, F.; Müller, K.-R.; Valleriani, M.; Montavon, G. Building and interpreting deep similarity models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 1149−1161.
(89) Sidak, D.; Schwarzerová, J.; Weckwerth, W.; Waldherr, S. Interpretable machine learning methods for predictions in systems biology from omics data. *Front. Mol. Biosci.* **2022**, *9*, 926623.
(90) El-Hajj, H.; Eberle, O.; Merklein, A.; Siebold, A.; Shlomi, N.; Büttner, J.; Martinetz, J.; Müller, K.-R.; Montavon, G.; Valleriani, M. Explainability and transparency in the realm of digital humanities: toward a historian XAI. *Int. J. Digit. Humanities* **2023**, *5*, 299−331.
(91) Flora, M. L.; Potvin, C. K.; McGovern, A.; Handler, S. A Machine Learning Explainability Tutorial for Atmospheric Sciences. *Artif. Intell. Earth Syst.* **2024**, *3*, No. e230018.
(92) Oviedo, F.; Ferres, J. L.; Buonassisi, T.; Butler, K. T. Interpretable and explainable machine learning for materials science and chemistry. *Acc. Mater. Res.* **2022**, *3*, 597−607.
(93) Roy, S.; Dürholt, J. P.; Asche, T. S.; Zipoli, F.; Gómez-Bombarelli, R. Learning a reactive potential for silica-water through uncertainty attribution. *Nat. Commun.* **2024**, *15*, 6030.

https://doi.org/10.1021/acs.jctc.4c01424
J. Chem. Theory Comput. 2025, 21, 714−729