Physics-Informed Weakly Supervised Learning for Interatomic Potentials

Anonymous Authors[1]

Abstract

Machine learning plays an increasingly important role in computational chemistry and materials science, complementing computationally intensive ab initio and first-principles methods. Despite their utility, machine-learning models often lack generalization capability and robustness during atomistic simulations, yielding unphysical energy and force predictions that hinder their real-world applications. We address this challenge by introducing a physics-informed, weakly supervised approach for training machine-learned interatomic potentials (MLIPs). We introduce two novel loss functions, using the concept of conservative forces and extrapolating the total energy via a Taylor expansion. Our approach enhances the accuracy of MLIPs applied to learning tasks with sparse training data set sizes and reduces the need for pre-training computationally demanding models. Particularly, we perform extensive experiments demonstrating reduced energy and force errors—often lower by a factor of two—for various baseline models and benchmark data sets. We also enhance the accuracy of energy and force predictions compared to previous methods that employ data augmentation via a Taylor expansion. Finally, we show that our approach facilitates MLIPs' training in a setting where the computation of forces is infeasible at the reference level, such as those employing complete-basis-set extrapolation. An implementation of our method and scripts for executing experiments are available at `https://anonymous.4open.science/r/PICPS-ML4Sci-1E8F`.

1. Introduction

Ab initio and first-principles methods are inevitable for the computer-aided exploration of molecular and material properties used in the chemical sciences and engineering (Parrinello, 1997; Carloni et al., 2002; Iftimie et al., 2005). However, commonly employed ab initio and first-principles approaches—such as coupled cluster (CC) (Purvis & Bartlett, 1982; Bartlett & Musiał, 2007) and density functional theory (DFT) (Hohenberg & Kohn, 1964; Kohn & Sham, 1965), respectively—require substantial computing resources. Thus, they typically allow only for atomistic simulations of small- to medium-sized atomic systems and restrict the accessible simulation times, which affects the accuracy of estimated molecular and material properties. Classical force fields can extend these length and time scales, providing a computationally efficient alternative to first-principles approaches, but often lack accuracy. Machine-learning-based models hold promise to bridge the gap between first-principles and classical approaches, yielding computationally efficient and accurate machine-learned interatomic potentials (MLIPs) (Smith et al., 2017; Chanussot* et al., 2021; Unke et al., 2021; Merchant et al., 2023; Kovács et al., 2023; Batatia et al., 2023). These MLIPs, however, face several challenges. They require the generation of training data sets that sufficiently cover configurational (atom positions) and compositional (atom types) spaces using, e.g., molecular dynamics simulations based on ab initio or first-principles approaches. Given the high computational cost of the commonly used data generation approaches, the resulting training data sets are often sparse and restrict the application of MLIPs to new molecular and material systems. Active learning (AL) is often used to address this challenge (Li et al., 2015; Vandermause et al., 2020; Zaverkin et al., 2024), while still requiring non-negligible computer resources. Furthermore, MLIPs often lack sufficient robustness and extrapolation capability during atomistic simulations, i.e., they are sensitive to outliers and local perturbations of atomic structures. This sensitivity of ML-based models is caused by existing data sets and data generation techniques not providing sufficient coverage of configurational and compositional spaces (Foret et al., 2020; Andriushchenko & Flammarion, 2022).

Contributions. This paper addresses these challenges using a physics-informed weakly supervised learning (PIWSL)

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review at ICML2024 AI for Science workshop. Do not distribute.

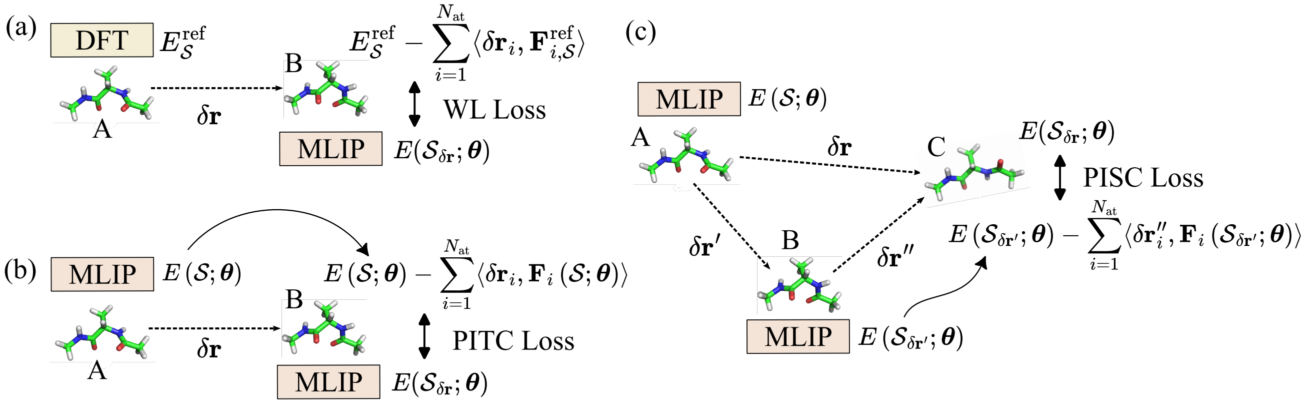
(a)
DFT
$E_{\mathcal{S}}^{\text{ref}}$
A
$\delta\mathbf{r}$
B
$E_{\mathcal{S}}^{\text{ref}} - \sum_{i=1}^{N_{\text{at}}} \langle \delta\mathbf{r}_i, \mathbf{F}_{i,\mathcal{S}}^{\text{ref}} \rangle$
WL Loss
MLIP
$E(\mathcal{S}_{\delta\mathbf{r}};\boldsymbol{\theta})$
(b)
MLIP
$E(\mathcal{S};\boldsymbol{\theta})$
A
$\delta\mathbf{r}$
B
$E(\mathcal{S};\boldsymbol{\theta}) - \sum_{i=1}^{N_{\text{at}}} \langle \delta\mathbf{r}_i, \mathbf{F}_i(\mathcal{S};\boldsymbol{\theta}) \rangle$
PITC Loss
MLIP
$E(\mathcal{S}_{\delta\mathbf{r}};\boldsymbol{\theta})$
(c)
MLIP
$E(\mathcal{S};\boldsymbol{\theta})$
A
$\delta\mathbf{r}$
C
$E(\mathcal{S}_{\delta\mathbf{r}};\boldsymbol{\theta})$
PISC Loss
$\delta\mathbf{r}'$
B
$\delta\mathbf{r}''$
$E(\mathcal{S}_{\delta\mathbf{r}'};\boldsymbol{\theta}) - \sum_{i=1}^{N_{\text{at}}} \langle \delta\mathbf{r}''_i, \mathbf{F}_i(\mathcal{S}_{\delta\mathbf{r}'};\boldsymbol{\theta}) \rangle$
MLIP
$E(\mathcal{S}_{\delta\mathbf{r}'};\boldsymbol{\theta})$

Figure 1: **Schematic illustration of physics-informed weakly supervised losses used in this work.** (a) Taylor expansion-based loss with approximate labels obtained from reference energies and atomic forces (Cooper et al., 2020). (b) Physics-inspired Taylor-expansion-consistency (PITC) loss with approximate labels obtained from energies and atomic forces predicted by an MLIP. (c) Physics-inspired spatial consistency (PISC) loss with approximate labels obtained from energies and atomic forces predicted by an MLIP. Here, $E(\mathcal{S};\boldsymbol{\theta})$ and $\mathbf{F}_i(\mathcal{S};\boldsymbol{\theta})$ denote the total energy and atomic forces predicted by an MLIP parametrized by $\boldsymbol{\theta}$, $\mathcal{S}$ and $\mathcal{S}_{\delta\mathbf{r}}$ define the original atomic structure and the one perturbed by $\delta\mathbf{r}$.

approach. Our method is designed to learn an MLIP, which can accurately predict total energy and atomistic forces for an atomic system exposed to local perturbations. In particular, our contributions are as follows: (i) We introduce PIWSL based on basic physical principles, such as the concept of conservative forces. We combine it with extrapolating the total energy via a Taylor expansion and derive two novel physics-informed loss functions, schematically illustrated in Fig. 1. Particularly, we obtain physics-informed Taylor-expansion-based consistency (PITC) and physics-informed spatial consistency (PISC) losses, which build the basis for the PIWSL approach. (ii) By conducting extensive experiments, we demonstrate that PIWLS allows for training MLIPs without access to large training data sets. (iii) We also observe that PIWSL improves accuracy in predicted total energies and atomic forces, even without access to force labels. This scenario is expected when training MLIPs with reference methods for which calculating atomic forces is infeasible (Smith et al., 2019; 2020). Thus, our results open new possibilities for training MLIPs using highly accurate energy labels, such as those obtained by extrapolating CCSD(T) energies to the complete basis set (CBS) limit (Hobza & Šponer, 2002; Feller et al., 2006). (iv) Finally, introducing local perturbations for atomic structures in the training data sets mitigates sensitivity issues associated with limited sizes of available data sets.

2. Related Work

Machine-Learned Interatomic Potentials. There is a growing interest in using ML-based models for investigating molecular and material systems as they allow performing atomistic simulations with an accuracy on par with first-principles methods but at a fraction of the computational cost. The field of machine-learned interatomic potentials (MLIPs) emerged over two decades ago (Blank et al., 1995) and has been one of the most active research directions since then (Behler & Parrinello, 2007; Artrith et al., 2011; Artrith & Urban, 2016; Smith et al., 2017; Shapeev, 2016; Schütt et al., 2017; Thomas et al., 2018; Unke & Meuwly, 2019; Drautz, 2019; Zaverkin & Kästner, 2020; Zaverkin et al., 2021; Thomas et al., 2018; Schütt et al., 2021; Shuaibi et al., 2021a; Passaro & Zitnick, 2023; Liao et al., 2023; Batzner et al., 2022; Musaelian et al., 2023; Batatia et al., 2022). Though, the development of local higher-body-order representations (Shapeev, 2016; Drautz, 2019; Zaverkin & Kästner, 2020; Zaverkin et al., 2021) and the emergence of equivariant message-passing neural networks (MPNNs) (Thomas et al., 2018; Schütt et al., 2021; Shuaibi et al., 2021a; Passaro & Zitnick, 2023; Liao et al., 2023; Batzner et al., 2022; Musaelian et al., 2023; Batatia et al., 2022) significantly advanced the field. These methods enable the cost-efficient generation of accurate MLIPs for modeling interactions in many-body atomic systems and account for crucial inductive biases as the invariance of the total energy under rotation.

Physics-Informed Machine Learning. Physics-informed ML aims to model physical systems using data-driven techniques and incorporates physics principles into ML-based models. For example, MLIPs based on equivariant MPNNs enforce the invariance of the total energy under rotation and use equivariant features to enrich the building of many-body contributions to it (Thomas et al., 2018; Batzner et al., 2022; Batatia et al., 2022; Musaelian et al., 2023; Liao et al., 2023). Furthermore, physics constraints can be integrated via auxiliary loss functions, prompting ML models to learn important physical relationships, as demonstrated for physics-informed neural networks (PINNs) (Raissi et al., 2019; Cai et al., 2022), which learn to model solutions of partial differential equations by minimizing residuals during training. Applying physics-informed ML to molecular modeling has gained attraction in both ML and computational chemistry communities (Godwin et al., 2021; Ni et al., 2023). As such, prior work (Cooper et al., 2020) has motivated our current research and is discussed in more detail in subsequent sections.

3. Background and Problem Definition

Machine-learned Interatomic Potentials. An atomic configuration, denoted as $\mathcal{S} = \{\mathbf{r}_i, Z_i\}_{i=1}^{N_\mathrm{at}}$, contains N_at atoms and is defined by atom positions $\mathbf{r}_i \in \mathbb{R}^3$ and atom types $Z_i \in \mathbb{N}$. We consider mapping atomic configurations to scalar energies, i.e., $f_{\boldsymbol{\theta}} : \mathcal{S} \mapsto E \in \mathbb{R}$ with $\boldsymbol{\theta}$ denoting trainable parameters. We define $E(\mathcal{S}; \boldsymbol{\theta})$ as the energy predicted by an MLIP for an atomic configuration $\mathcal{S}$. For most MLIPs, atomic forces are computed as the negative gradients of the total energy with respect to atom positions, i.e., $\mathbf{F}_i\left(\mathcal{S}; \boldsymbol{\theta}\right) = -\nabla_{\mathbf{r}_i} E\left(\mathcal{S}; \boldsymbol{\theta}\right)$. In this way, these MLIPs ensure that the resulting forces are conservative (are curl-free) and the total energy is conserved during a dynamic simulation. However, some models are designed to predict atomic forces directly (Hu et al., 2021; Passaro & Zitnick, 2023; Liao et al., 2023; Chanussot* et al., 2021). While this approach avoids expensive gradient computations, it violates energy conservation.

Trainable parameters $\boldsymbol{\theta}$ are optimized by minimizing loss functions on training data $\mathcal{D}$ comprising a total of N_train atomic configurations $\{\mathcal{S}^{(k)}\}_{k=1}^{N_\mathrm{train}}$ as well as their energies $\{E_\mathcal{S}^\mathrm{ref}\}_{\mathcal{S}\in\mathcal{D}}$ and atomic forces $\{\{\mathbf{F}_{i,\mathcal{S}}^\mathrm{ref}\}_{i=1}^{N_\mathrm{at}}\}_{\mathcal{S}\in\mathcal{D}}$

$$
\begin{aligned}
\mathcal{L}\left(\mathcal{D}; \boldsymbol{\theta}\right) &= \sum_{\mathcal{S}\in\mathcal{D}} L\left(S; \boldsymbol{\theta}\right) \\
&= \sum_{\mathcal{S}\in\mathcal{D}} \left[C_\mathrm{e} \ell\left(E\left(\mathcal{S}; \boldsymbol{\theta}\right), E_\mathcal{S}^\mathrm{ref}\right) + C_\mathrm{f} \sum_{i=1}^{N_\mathrm{at}} \ell\left(\mathbf{F}_i(\mathcal{S}; \boldsymbol{\theta}), \mathbf{F}_{i,\mathcal{S}}^\mathrm{ref}\right) \right] .
\end{aligned} \tag{1}
$$

Here, ℓ denotes a point-wise loss function such as the absolute and squared error between the predicted and reference total energies and atomic forces. Typically, reference energies $E_\mathcal{S}^\mathrm{ref}$ and atomic forces $\mathbf{F}_{i,\mathcal{S}}^\mathrm{ref}$ are provided by ab initio or first principles methods such as CC or DFT, respectively. The relative contributions of energies and forces in Eq. (1) are balanced with the coefficients C_e and C_f.

Weakly Supervised Learning. Generating many reference labels with a first-principles approach is challenging due to the high computational cost. Furthermore, the calculation of atomic forces can be infeasible for some high-accuracy ab initio methods, e.g., for CCSD(T)/CBS (Hobza & Šponer, 2002; Feller et al., 2006). We focus on weakly supervised learning methods to improve the performance of MLIPs in scenarios when only limited data set sizes are available. These involve the generation of approximate but physically motivated total energies for atomic structures generated by small displacements of their atomic positions, i.e., $\mathcal{S}_{\delta\mathbf{r}} = \{\mathbf{r}_i + \delta\mathbf{r}_i, Z_i\}_{i=1}^{N_\mathrm{at}}$ with an atomic displacement vector $\delta\mathbf{r}$ where $\delta\mathbf{r}_i$ is the displacement vector for atom i. Approximate labels are computed with MLIPs trained using reference total energies and atomic forces.

4. Physics-informed Weakly Supervised Learning

For MLIPs, the generation of approximate labels employed in weakly supervised losses is highly non-trivial. Small displacements in atomic structures can lead to significant changes in energies and atomic forces. Thus, standard approaches, effective for many other ML tasks, are typically inefficient (Yang et al., 2022). To address this problem, we propose physics-informed weakly supervised learning approaches that involve (i) a Taylor expansion of the total energy for computing the response to atomic displacements and (ii) spatial consistency to estimate the displaced potential energy, based on the concept of conservative forces. We finally introduce the PIWSL loss term, combining both classes of weakly supervised loss functions with the supervised loss.

4.1. Physics-Informed Taylor-Expansion Based Consistency Loss

This section introduces the physics-informed Taylor-expansion-based consistency (PITC) loss. Particularly, we relate the energy predicted directly for a displaced atomic configuration with the energy obtained by the Taylor expansion from the original configuration; see Fig. 1 (b). We estimate the energy for an atomic structure $\mathcal{S}$ drawn from the training data set with atomic positions displaced by a vector $\delta\mathbf{r}$: $\mathcal{S}_{\delta\mathbf{r}} = \{\mathbf{r}_i + \delta\mathbf{r}_i, Z_i\}_{i=1}^{N_\mathrm{at}}$. For this atomic configuration, we expand the energy predicted by an MLIP in its first-order Taylor series[1] around the atomic displacement vector $\delta\mathbf{r}_i$ and obtain

$$
\begin{aligned}
E\left(\mathcal{S}_{\delta\mathbf{r}}; \boldsymbol{\theta}\right) &\approx E\left(\mathcal{S}; \boldsymbol{\theta}\right) \\
&\quad - \sum_{i=1}^{N_\mathrm{at}} \langle \delta\mathbf{r}_i, \mathbf{F}_i\left(\mathcal{S}; \boldsymbol{\theta}\right) \rangle + \mathcal{O}\left(\|\delta\mathbf{r}\|^2\right) ,
\end{aligned} \tag{2}
$$

[1] In general, employing a more sophisticated higher-order ordinary differential equation solver is a viable option. However, this approach does not consistently improve performance due to the MLIP prediction error. As a result, the increased computational expense associated with higher-order methods may not be justified.

where $\langle\cdot\rangle$ denotes the inner product. Here, we used that atomic forces are defined as the negative gradients of the total energy. For small magnitudes of $\delta\mathbf{r}_i$, the second order term $\mathcal{O}\left(\|\delta\mathbf{r}\|^2\right)$ in Eq. (2) can be neglected. Using approximate labels $E\left(\mathcal{S}_{\delta\mathbf{r}};\boldsymbol{\theta}\right)$, we define the PITC loss as

$$L_{\mathrm{PITC}}\left(\mathcal{S};\boldsymbol{\theta}\right) = \ell\left(E\left(\mathcal{S}_{\delta\mathbf{r}};\boldsymbol{\theta}\right), E\left(\mathcal{S};\boldsymbol{\theta}\right) - \sum_{i=1}^{N_{\mathrm{at}}}\langle\delta\mathbf{r}_i, \mathbf{F}_i\left(\mathcal{S};\boldsymbol{\theta}\right)\rangle\right), \quad (3)$$

where ℓ denotes a point-wise loss for regression problems and $\delta\boldsymbol{r}$ is sampled or determined adversarially; see Section 4.4 for more details. Hence, whenever we encounter a structure $\mathcal{S}$ in a batch during gradient descent, a new $\delta\boldsymbol{r}$ is computed for $\mathcal{S}$.

4.2. Physics-Informed Spatial-Consistency Loss

This section introduces a physics-informed approach for generating weak labels based on the concept of conservative forces. Thus, we leverage the fact that the energy difference between two points on the potential energy surface is independent of the path taken between them. We consider two paths from the reference data points to some target points, composed of three points and displacement vectors around an atom. We estimate the potential energy at the final point via Eq. (3). An example of two paths is demonstrated in Fig. 1 (c). The figure relates the energy obtained when displacing atomic positions of the original configuration $\mathcal{S}$ (denoted by A in the figure) by $\delta\mathbf{r}$ (from configuration A to C) with the energy obtained through consecutive displacements $\delta\mathbf{r}'$ (from configuration A to B) and $\delta\mathbf{r}''$ (from configuration B to C).

For the first path, we directly predict the energy with an MLIP, i.e., $E\left(\mathcal{S}_{\delta\mathbf{r}};\boldsymbol{\theta}\right)$, which is related to the approximated energy at $\mathbf{r}+\delta\mathbf{r}$ using Eq. (3) through PITC loss. For the second path, we directly compute the synthetic energy $E\left(\mathcal{S}_{\delta\mathbf{r}'};\boldsymbol{\theta}\right)$ for atomic positions displaced by $\delta\mathbf{r}'$ and use it to approximate $E\left(\mathcal{S}_{\delta\mathbf{r}};\boldsymbol{\theta}\right)$ after applying the second displacement vector $\delta\mathbf{r}'' \equiv \delta\mathbf{r} - \delta\mathbf{r}'$. The physics-informed spatial consistency (PISC) loss can be defined as

$$L_{\mathrm{PISC}}\left(\mathcal{S};\boldsymbol{\theta}\right) = \ell\left(E\left(\mathcal{S}_{\delta\mathbf{r}};\boldsymbol{\theta}\right), E\left(\mathcal{S}_{\delta\mathbf{r}'};\boldsymbol{\theta}\right) - \sum_{i=1}^{N_{\mathrm{at}}}\langle\delta\mathbf{r}_i'', \mathbf{F}_i\left(\mathcal{S}_{\delta\mathbf{r}'};\boldsymbol{\theta}\right)\rangle\right), \quad (4)$$

where $\delta\mathbf{r}$ is sampled or determined adversarially; see Section 4.4. After joint training of PITC and PISC losses, the three different estimations at $\mathcal{S}_{\delta\mathbf{r}}$ become spatially consistent. Note that our conservative forces-based approach is not limited to relations between two displacement paths or three displacement vectors. We discuss several other possible configurations in Appendix E.

4.3. Combined Physics-Informed Weakly Supervised Loss (PIWSL)

Together with the usual MLIP loss function given in Eq. (1), the overall objective, which we refer to as the PIWSL loss, can be written as

$$\min_{\boldsymbol{\theta}} \tilde{\mathcal{L}}\left(\mathcal{D};\boldsymbol{\theta}\right) = \min_{\boldsymbol{\theta}} \sum_{\mathcal{S}\in\mathcal{D}} \left(L\left(\mathcal{S};\boldsymbol{\theta}\right) + C_{\mathrm{PITC}} L_{\mathrm{PITC}}\left(\mathcal{S};\boldsymbol{\theta}\right) + C_{\mathrm{PISC}} L_{\mathrm{PISC}}\left(\mathcal{S};\boldsymbol{\theta}\right)\right), \quad (5)$$

where C_{PITC} and C_{PISC} are the weights of the weakly supervised PITC and PISC losses.

4.4. Determining Perturbation Directions and Magnitudes

The effectiveness of the proposed approach hinges on appropriate choices of the perturbation vectors $\delta\mathbf{r}$. Therefore, we introduce and justify various generation strategies for the perturbations $\delta\mathbf{r}$ used in Eq. (3) and Eq. (4). We can express any vector $\delta\mathbf{r}$ as $\delta\mathbf{r} \equiv \epsilon\mathbf{g}/\|\mathbf{g}\|_2$, where ϵ is the magnitude of $\delta\mathbf{r}$ and $\mathbf{g}/\|\mathbf{g}\|_2$ represents the direction of $\delta\mathbf{r}$. Physical constraints are considered when determining ϵ. Specifically, the maximum length of the displacement can be obtained from the validity of the Taylor expansion in Eq. (2), which is typically given as at most 30% of the original bond length whose shortest example is the carbon and hydrogen bond, approximately 1.09 Å; see also Fig. 2 (c), (d). The specific values of ϵ chosen for our experiments are provided in Appendix B.

To determine $\mathbf{g}/\|\mathbf{g}\|_2$ we explore two strategies. First, we compute it as the unit vector of a perturbation vector sampled from a multivariate Gaussian distribution, with each independent Gaussian having zero mean and unit variance

$$\delta\mathbf{r}_{\mathrm{rnd}} \equiv \epsilon\mathbf{g}/\|\mathbf{g}\|_2. \quad (6)$$

Second, we compute an adversarial direction, as proposed in (Goodfellow et al., 2014; Miyato et al., 2018), which involves defining it as the direction (the gradients) in which the loss error increases the most at the current atom coordinates $\mathbf{r}$ and for the current energy. Assuming the norm of adversarial perturbation as L_2, the adversarial direction can be approximated by (Miyato et al., 2018)

$$\delta\mathbf{r}_{\mathrm{adv}} \equiv \epsilon\mathbf{g}/\|\mathbf{g}\|_2, \text{ where } \mathbf{g} = \nabla_{\mathbf{r}} L_{\mathrm{dist}}(\mathbf{y}_{\mathrm{pred}}, \mathbf{y}_{\mathrm{label}}), \quad (7)$$

where L_{dist} is a distance measure function to be maximized by adding $\delta\mathbf{r}_{\mathrm{adv}}$, and $\mathbf{y}_{\mathrm{pred}}$ and $\mathbf{y}_{\mathrm{label}}$ are the ML model prediction and the label values. Due to their computational efficiency, we mainly use Eq. (6) in our experiments. A quantitative comparison between the random and adversarial directions is provided in Section 5.5.

5. Experiments

We evaluate our method with an extensive set of experiments with the following objectives: (1) a comparison of the PIWSL method with existing baselines, (2) a fine-grained analysis of the impact of the PIWSL method using the aspirin molecule, (3) an exploration of the ability of PIWSL to improve energy and force predictions when force labels are inaccessible, (4) a comparison to prior work that used a simple weakly supervised approach, (5) an ablation study, and (6) a comparison between random and adversarial generation of the displacement vector. In general, we focus on the data-scarce setting where the number of training samples is between 100 and 1000 since the computational cost of ab initio and first-principles approaches to generate large datasets is prohibitive.

5.1. MLIPs and Data Sets

We trained the following representative models that are provided in the OpenCatalyst code base (Chanussot* et al., 2021): SchNet (Schütt et al., 2017), PaiNN (Schütt et al., 2021), SpinConv (Shuaibi et al., 2021a), eSCN (Passaro & Zitnick, 2023), and Equiformer v2 (Liao et al., 2023), covering MLIPs with a smaller (SchNet, SpinConv) and larger number of parameters (eSCN, Equiformer v2). Moreover, we also considered the highly parameter-efficient model PaiNN. Unless otherwise mentioned, except for SchNet, forces are directly predicted and not computed through the negative gradient of the energy. The results where forces are computed as negative energy gradients are analyzed in Section 5.4 and Appendix K. To evaluate the effect and dependency of the physics-informed weakly supervised approach in detail, we performed the training on various datasets: ANI-1x (Smith et al., 2020) as a comprehensive molecular data set, TiO2 (Artrith & Urban, 2016) as a data set for inorganic materials, the revised MD17 (rMD17) data set (Chmiela et al., 2017; 2018) representing a smaller and single-molecule data set, and LMNTO as another material data set (Cooper et al., 2020) (The results for rMD17 and LMNTO are provided in Appendix M). The detailed description of each data set is provided in Appendix C.

5.2. Benchmark Results

We compare models trained with the PIWSL loss (see Eq. (5)) with baseline models trained with the standard supervised loss only (see Eq. (1)). We also compare to a recently proposed data augmentation method that incorporates the task of denoising random perturbations of the atomic coordinates into the learning objective (NoisyNode) (Godwin et al., 2021). More details are provided in Appendix B.

ANI-1x: Large Molecular Data Set. The results provided in Table 1 show that our approach improves the baseline models' performance in almost all cases. In particular, the error reduction for the predicted energies is often between 10% and more than 50%. Interestingly, we do not only observe an increase in accuracy for the total energies but also the atomic forces because of the inclusion of force prediction in PITC and PISC losses, different from the previous work (Cooper et al., 2020). In most cases, the data augmentation method (NoisyNode) reduces the accuracy of the MLIPs. This is the case, as this method does not incorporate the concept of conservative forces in its loss function. SchNet is one notable exception here, which we hypothesize to be related to SchNet being an invariant MLIP, which can be less expressible than equivalent MLPIs.

Dependence on Number of Samples for ANI-1x. We train the MLIPs with training set sizes in $[50, 10^2, 10^3, 10^4]$. The results are plotted in Fig. 2 (a), (b). Although the observed error reduction depends strongly on the type of MLIP used, the benefit of the weakly supervised losses often decreases slightly with the number of training samples. This is an expected result as the area covered by the weakly supervised losses is also gradually covered by the reference data as the number of training samples increases. Moreover, the gain in accuracy of energy predictions is larger than that for forces that are only indirectly trained through the consistency constraint in PITC (Eq. (3)). Finally, it is shown that the improvement is more significant for highly parameterized MLIPs, which benefit the most from increasing the training data size.

TiO2: Data Set for Metal Oxides. Titanium dioxide (TiO_2) is a highly relevant metal oxide for industrial applications, featuring several high-pressure phases. Thus, ML models should be able to predict total energies and atomic forces for various high-pressure phases of TiO_2, considering periodic boundaries (relevant when aggregating over the local atomic neighborhood). The results for trained models are provided in Table 2. Similar to the ANI-1x data set, our approach improves the accuracy of predicted energies and atomic forces. Interestingly, although the error in the total energy for $N_{\mathrm{train}} = 1000$ training configurations reaches small RMSE values, from 2 to 4 kcal/mol, the PIWSL still provides a further error reduction. This observation indicates strong evidence of the effectiveness of PIWSL applied to material data sets.

5.3. Qualitative Impact of Using PIWSL

We evaluate the atom-level impact and robustness of the PIWSL method using the aspirin molecule, focusing on the potential energy's dependence on the carbon-hydrogen (C-H) bond length. We train PaiNN models on the rMD17 aspirin dataset with sample sizes 100 and 200, both with and without the PIWSL loss. The detailed training setup and errors of the used MLIPs are summarized in Appendix F.

Table 1: **Energy and atomic force root-mean-square errors (RMSEs) for the ANI-1x data set.** The results are obtained by averaging over three independent runs. Energy RMSE is given in kcal/mol, while force RMSE is in kcal/mol/Å.

		$N_{\text{train}} = 100$			$N_{\text{train}} = 1000$		
		Baseline	Noisy Nodes	PIWSL	Baseline	Noisy Nodes	PIWSL
Schnet	energy	65.09 ± 2.42	**57.39 ± 0.05**	60.30 ± 1.77	31.49 ± 0.01	**31.10 ± 0.00**	31.50 ± 0.00
	force	29.06 ± 0.19	**25.62 ± 0.01**	28.20 ± 0.60	18.94 ± 0.01	**18.10 ± 0.00**	18.93 ± 0.00
PaiNN	energy	168.01 ± 1.22	464.55 ± 6.91	**109.89 ± 11.46**	56.62 ± 2.80	305.76 ± 33.93	**24.53 ± 0.48**
	force	21.33 ± 0.10	20.82 ± 0.03	**18.76 ± 0.30**	12.96 ± 0.06	14.25 ± 0.18	**11.43 ± 0.05**
SpinConv	energy	162.14 ± 7.55	147.73 ± 2.23	**130.97 ± 8.58**	43.59 ± 1.71	299.33 ± 419.10	**39.44 ± 1.31**
	force	21.22 ± 0.43	**21.08 ± 0.43**	21.61 ± 0.44	14.51 ± 1.07	15.83 ± 0.75	**13.59 ± 0.20**
eSCN	energy	214.52 ± 7.55	521.92 ± 12.05	**183.70 ± 9.79**	59.59 ± 8.92	241.34 ± 20.16	**21.03 ± 0.56**
	force	20.07 ± 0.27	23.68 ± 0.11	**19.69 ± 0.05**	12.50 ± 0.78	14.42 ± 0.84	**11.83 ± 0.12**
Equiformer	energy	398.71 ± 13.69	632.38 ± 0.11	**154.98 ± 8.83**	54.52 ± 4.52	854.33 ± 317.7	**20.89 ± 0.50**
	force	20.71 ± 0.05	21.82 ± 0.01	**20.55 ± 0.05**	10.10 ± 0.00	24.79 ± 2.05	**9.68 ± 0.03**

Table 2: **Energy and atomic force root-mean-square errors (RMSEs) for the TiO_2 data set.** The results are obtained by averaging over three independent runs. Energy RMSE is given in kcal/mol, while force RMSE is in kcal/mol/Å.

		$N_{\text{train}} = 100$			$N_{\text{train}} = 1000$		
		Baseline	Noisy Nodes	PIWSL	Baseline	Noisy Nodes	PIWSL
SchNet	energy	18.85 ± 0.00	**17.48 ± 0.00**	17.58 ± 0.00	35.58 ± 0.00[a]	58.08 ± 18.44	**15.28 ± 0.12**
	force	2.74 ± 0.00	**2.51 ± 0.00**	2.74 ± 0.00	6.54 ± 0.00[a]	18.40 ± 0.00	**3.61 ± 0.27**
PaiNN	energy	13.93 ± 0.16	n/a[b]	n/a[b]	2.42 ± 0.05	n/a[b]	n/a[b]
	force	1.52 ± 0.00	n/a[b]	n/a[b]	0.29 ± 0.03	n/a[b]	n/a[b]
SpinConv	energy	20.00 ± 0.42	18.76 ± 0.74	**16.98 ± 0.99**	4.17 ± 0.76	4.09 ± 0.65	**2.50 ± 0.40**
	force	1.58 ± 0.03	**1.53 ± 0.03**	1.59 ± 0.03	0.65 ± 0.02	0.71 ± 0.16	**0.58 ± 0.05**
eSCN	energy	16.41 ± 1.10	20.92 ± 0.00	**12.63 ± 0.78**	3.31 ± 1.18	20.90 ± 0.01	**1.40 ± 0.10**
	force	1.57 ± 0.04	1.66 ± 0.00	**1.44 ± 0.03**	0.46 ± 0.23	1.66 ± 0.00	**0.21 ± 0.00**
Equiformer	energy	18.21 ± 0.02	19.06 ± 0.02	**13.93 ± 0.09**	3.67 ± 0.03	18.75 ± 0.05	**1.82 ± 0.34**
	force	1.56 ± 0.01	1.64 ± 0.00	**1.51 ± 0.19**	**0.17 ± 0.01**	1.58 ± 0.00	**0.17 ± 0.01**

[a] We observed an increase in energy and force RMSEs for SchNet for larger training set sizes, which may be explained by the limited expressive power of SchNet features.
[b] Predicted values become NaN when adding displacements to atomic configurations.

We examine the potential energy across a range of aspirin atom-coordinates, varying the C-H bond length from 0.9 Å to 1.4 Å[2]. The results in Fig. 2 (c), (d) show that the PIWSL method enhances the learned potential energy function, indicating improved robustness under atomic coordinate perturbations because of the improvement of the potential curve prediction without spurious oscillations. Although the estimated potential energies do not always match actual values, the direction of the line between the reference and estimated points with a deviation length of $||\delta\mathbf{r}|| = 0.01$Å, illustrated by arrows in Fig. 2, consistently reproduces the correct gradient of the potential energy, i.e., the correct force direction. This suggests the PIWSL method's effectiveness comes from the consistency condition, ensuring model predictions of potential energy and force alignment, resulting in accurate potential energy estimation and consistent force prediction. As discussed in Section 4.2, this addresses the issue of recent MLIPs with separate force branches not guaranteeing predicting conservative forces. The proposed method shows a reduction of the rotation of the predicted forces, as explored in Appendix L, though the reduction is still not enough to realize the rotation free conservative force prediction. In conclusion, the weakly supervised losses reduce individual energy and force errors, making the MLIPs more physically accurate.

[2] The equilibrium C-H bond length is about 1.09 Å

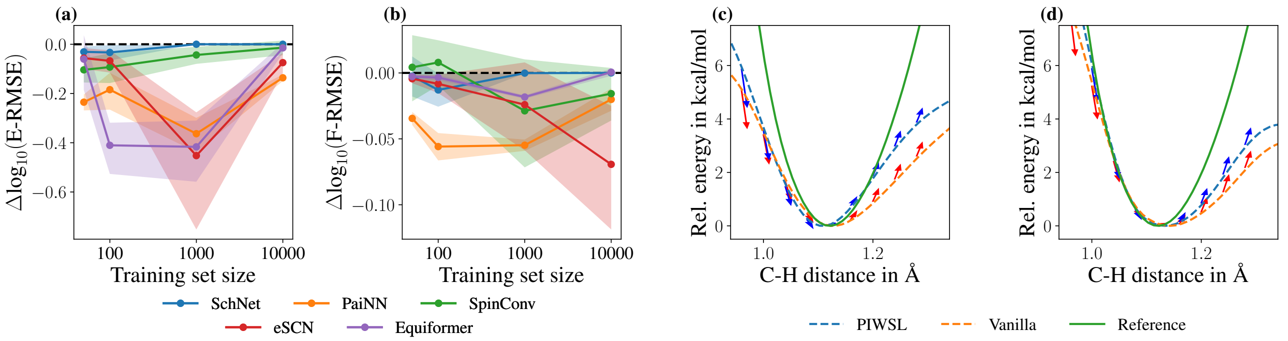

Figure 2: **Relative performance gains for MLIPs trained with physics-informed weakly supervised losses compared to the same MLIPs trained with a standard loss function.** Performance gains are evaluated for (a) energy (E-) and (b) force (F-) RMSEs. The results are for ANI-1x. (b) Plots of the aspirin molecule's potential energy vs. C-H bond length (distance). The red and blue arrows indicate the reference, $E(\mathcal{S};\boldsymbol{\theta})$, to perturbed potential energy, $E(\mathcal{S}_{\delta\mathbf{r}};\boldsymbol{\theta})$, predicted by Eq. (8), on the baseline and PIWSL model predictions, respectively.

Table 3: **Experimental results of no-force label training on the ANI-1x dataset with 1000 training samples.** "FB" means the force is estimated by the force branch, and "GB" means the force is estimated by the gradient of the potential energy wrt the atomic coordinates.

Model	Case		Baseline	PIWSL
PaiNN	FB	energy	42.36 ±0.30	**25.42** ±0.72
		force	24.25 ±0.00	**20.54** ±0.08
	GF	energy	41.83 ±1.81	**29.71** ±0.55
		force	83.36 ±2.85	**24.02** ±0.95
Equiformer	FB	energy	43.14 ±0.86	**29.48** ±0.51
		force	24.25 ±0.00	**21.99** ±0.49
	GF	energy	42.55 ±0.99	**32.66** ±1.11
		force	35.70 ±0.78	**21.83** ±0.27

5.4. Training MLIPs when Reference Forces are Missing

In the following, we explore scenarios where only potential energy labels are available. This situation commonly arises when calculating energy labels with chemically accurate approaches, such as CCSD(T)/CBS (Hobza & Šponer, 2002; Feller et al., 2006), for which force calculation is infeasible. To consider practical applications, we examine two cases: 1) predicting force by a force branch (FB) and 2) predicting force via the spatial gradient of the predicted potential energy (GF). The former enables fast force prediction and is popular in the machine learning community, while the latter requires additional gradient calculation but yields curl-free force prediction. It is popular in computational chemistry as it ensures the conservation of the total energy during MD simulations. The results are provided in Table 3[3]. Notably, our PIWSL method consistently demonstrates superior performance across all the cases. Interestingly, a larger improvement in the force prediction performance is observed in the GF case. We attribute this phenomenon to the inherent nature of PIWSL, which requires consistency between potential energy variation and the forces, as discussed in Section 5.3. This result aligns with our expectations, confirming the capability of our PIWSL method to enable ML models to estimate the force label. Thus, it opens a new possibility for training MLIP models using highly accurate reference methods, such as CCSD(T)/CBS.

5.5. Detailed Analysis

We provide several detailed analyses of our approach. As base MLIPs, we use the Equiformer v2 and PaiNN since they use equivariant features and have obtained a high accuracy on the ANI-1x dataset with $N = 1000$ training samples.

Comparison to Taylor Expansion Based Weak Label Approach. We compare our proposed method with the Taylor expansion-based weak label (WL) (Cooper et al., 2020) method. For the WL method, we utilize Equation (9) as the loss function. We only consider the PITC loss, Equation (3) for simplicity. For a fair comparison, we consider the following two cases. First, we train with reference forces and energies (w. RF). Second, we train the methods without reference forces and only the reference energies. For the training with reference forces, we set the numeric coefficient of the PITC loss to 1.0; for the training without reference forces, the coefficient is set to 0.1. The results are provided in Table 4. First, our PITC loss shows the best accuracy in

[3] To simulate no force label training, we simply set the coefficient of the force loss as 0

Table 4: **A comparison with the conventional weak-label case.** "WL" means the Taylor expansion-based method using reference energies and forces Equation (9). The listed values are the RMS errors for energy [kcal/mol] and force [kcal/mol/Å] on the ANI-1x dataset with 1000 samples, with and without using reference forces (RF) for training.

Model	Case		Baseline	PITC	WL
PaiNN	(w RF)	energy	56.62 ±2.80	**30.94**±0.56	81.86±9.39
		force	12.96±0.06	**12.04**±0.04	14.54±0.12
	(w/o RF)	energy	42.36±0.30	**25.42**±0.72	35.79±0.70
		force	24.25±0.00	**20.54**±0.08	24.25±0.01
Equiformer	(w RF)	energy	54.52±4.52	**23.16**±0.19	31.02±3.99
		force	10.10±0.00	**10.03**±0.05	13.43±0.92
	(w/o RF)	energy	43.14±0.86	**29.48**±0.51	30.08±1.97
		force	24.25±0.00	**21.99**±0.49	24.25±0.00

Table 5: **Experimental results on ANI-1x data set with 1000 training samples when ablating proposed weakly supervised losses.** The results are obtained by averaging over three independent runs. Energy RMSE is given in kcal/mol, while force RMSE in kcal/mol/Å.

Model	PITC	PISC	energy	force
PaiNN	✗	✗	56.62 ± 2.80	12.96 ± 0.06
	✓	✗	24.60 ± 0.18	11.51 ± 0.03
	✗	✓	58.30 ± 2.10	13.18 ± 0.29
	✓	✓	**24.53 ± 0.48**	**11.43 ± 0.05**
Equiformer	✗	✗	54.52 ± 4.52	10.10 ± 0.00
	✓	✗	32.64 ± 26.48	**9.64 ± 0.03**
	✗	✓	48.96 ± 4.96	10.30 ± 0.06
	✓	✓	**20.89 ± 0.50**	9.68 ± 0.03

all settings. Interestingly, PaiNN failed to learn the potential energy with the WL loss and reference forces. We hypothesize this to be due to the imbalance of the training between the energies and forces, i.e., the WL loss only trains the potential energy. This hypothesis is supported by the results for the training without reference forces, where the error in energy is reduced compared to the baseline. However, the proposed PITC loss still performs better here. In summary, the PITC loss enables MLIPs to learn both the energies and forces that are consistent with each other and does it better than the previously proposed WL method.

Ablation Study. Using an ablation experiment, we also investigate the effect of the PITC and PISC losses. The results are provided in Table 5 and show that the gain in accuracy is mainly attributable to the PITC loss, particularly for PaiNN. Training with the PISC loss only does not always provide gains in accuracy. However, we can see that the PISC loss stabilizes the training with the PITC loss and improves accuracy. The combined efficiency of the losses is particularly pronounced for the Equiformer v2[4].

Table 6: **A result of the spatial-deviation vector selection dependence.** The listed numerical values are the root mean square errors for the energy [kcal/mol] and force [kcal/mol/Å] on the ANI-1x dataset with 1000 samples.

		Baseline	Random (Eq. (6))	Adversarial (Eq. (7))
PaiNN	energy	56.62±2.80	**24.53**±0.48	33.67 ±1.12
	force	12.96±0.18	**11.43**±0.05	12.74±0.14
Equiformer	energy	54.52±4.52	23.16±0.50	**20.54**±0.21
	force	10.10±0.00	10.03±0.03	**9.93**±0.04

Adversarial Direction as Spatial Deviation Vector. In the following, we discuss the performance dependence on the selection of the spatial-deviation vector $\delta\mathbf{r}$ in Equation (5). The detailed implementation and setups are provided in Appendix B. Table 6 shows the experimental result of the spatial-deviation vector selection dependence. The result shows that both selections improve the performance compared to the baselines without weak supervision, though the improvement depends on the model selection.

6. Discussion and Limitations

This work introduces the PIWSL method, encompassing two distinct physics-informed weakly supervised loss functions, for learning MLIPs. These losses operate explicitly (PITC loss) and implicitly (PISC loss), which enables any MLIP models to improve their accuracy, particularly in scenarios characterized by small training data set sizes, which is common when investigating a new molecular or material system. Our extensive experiments demonstrated notable efficacy and efficiency of our method from various aspects: (i) dependence on the training dataset size, (ii) dependence on the variability in molecular structures within the dataset, and (iii) selection of the deviation vector. In particular, it is shown that our PIWSL method enables ML models to improve the force prediction even without force labels, thereby opening a new possibility for training MLIPs using highly accurate reference methods, such as CCSD(T)/CBS.

Limitations. The proposed PIWSL method is tailored to ML models for predicting forces and energies of atomic systems. It cannot be applied to other ML problems unrelated to application in computational chemistry.

[4] To reduce the effect from an outlier for the PITC loss only for the Equiformer, we repeated the experiment 5 times to obtain the result in Table 5.

7. Potential Broader Impact and Ethical Aspects

This paper presents work whose goal is to advance the field of machine learning, in particular, machine learning for science. Due to the generic nature of pure science, there are many potential societal consequences of our work in the far future, none of which we feel must be specifically highlighted here.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

Andriushchenko, M. and Flammarion, N. Towards understanding sharpness-aware minimization. In *International Conference on Machine Learning*, pp. 639–668. PMLR, 2022.

Artrith, N. and Urban, A. An implementation of artificial neural-network potentials for atomistic materials simulations: Performance for tio2. *Computational Materials Science*, 114:135–150, 2016. ISSN 0927-0256. doi: https://doi.org/10.1016/j.commatsci.2015.11.047. URL https://www.sciencedirect.com/science/article/pii/S0927025615007806.

Artrith, N., Morawietz, T., and Behler, J. High-dimensional neural-network potentials for multicomponent systems: Applications to zinc oxide. *Physical Review B*, 83(15): 153101, 2011.

Bartlett, R. J. and Musiał, M. Coupled-cluster theory in quantum chemistry. *Rev. Mod. Phys.*, 79: 291–352, Feb 2007. doi: 10.1103/RevModPhys.79.291. URL https://link.aps.org/doi/10.1103/RevModPhys.79.291.

Batatia, I., Kovacs, D. P., Simm, G., Ortner, C., and Csányi, G. Mace: Higher order equivariant message passing neural networks for fast and accurate force fields. *Advances in Neural Information Processing Systems*, 35: 11423–11436, 2022.

Batatia, I., Benner, P., Chiang, Y., Elena, A. M., Kovács, D. P., Riebesell, J., Advincula, X. R., Asta, M., Baldwin, W. J., Bernstein, N., Bhowmik, A., Blau, S. M., Cărare, V., Darby, J. P., De, S., Pia, F. D., Deringer, V. L., Elijošius, R., El-Machachi, Z., Fako, E., Ferrari, A. C., Genreith-Schriever, A., George, J., Goodall, R. E. A., Grey, C. P., Han, S., Handley, W., Heenen, H. H., Hermansson, K., Holm, C., Jaafar, J., Hofmann, S., Jakob, K. S., Jung, H., Kapil, V., Kaplan, A. D., Karimitari, N., Kroupa, N., Kullgren, J., Kuner, M. C., Kuryla, D., Liepuoniute, G., Margraf, J. T., Magdău, I.-B., Michaelides, A., Moore, J. H., Naik, A. A., Niblett, S. P., Norwood, S. W., O'Neill, N., Ortner, C., Persson, K. A., Reuter, K., Rosen, A. S., Schaaf, L. L., Schran, C., Sivonxay, E., Stenczel, T. K., Svahn, V., Sutton, C., van der Oord, C., Varga-Umbrich, E., Vegge, T., Vondrák, M., Wang, Y., Witt, W. C., Zills, F., and Csányi, G. A foundation model for atomistic materials chemistry, 2023.

Batzner, S., Musaelian, A., Sun, L., Geiger, M., Mailoa, J. P., Kornbluth, M., Molinari, N., Smidt, T. E., and Kozinsky, B. E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature communications*, 13(1):2453, 2022.

Behler, J. and Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.*, 98(14):146401, 2007.

Blank, T. B., Brown, S. D., Calhoun, A. W., and Doren, D. J. Neural network models of potential energy surfaces. *The Journal of chemical physics*, 103(10):4129–4137, 1995.

Briganti, V. and Lunghi, A. Efficient generation of stable linear machine-learning force fields with uncertainty-aware active learning. *Mach. Learn.: Sci. Technol.*, 4(3):035005, jul 2023. doi: 10.1088/2632-2153/ace418. URL https://dx.doi.org/10.1088/2632-2153/ace418.

Cai, S., Mao, Z., Wang, Z., Yin, M., and Karniadakis, G. E. Physics-informed neural networks (pinns) for fluid mechanics: A review. *Acta Mechanica Sinica*, pp. 1–12, 2022.

Carloni, P., Rothlisberger, U., and Parrinello, M. The role and perspective of ab initio molecular dynamics in the study of biological systems. *Acc. Chem. Res.*, 35(6): 455–464, 2002. doi: 10.1021/ar010018u.

Chanussot*, L., Das*, A., Goyal*, S., Lavril*, T., Shuaibi*, M., Riviere, M., Tran, K., Heras-Domingo, J., Ho, C., Hu, W., Palizhati, A., Sriram, A., Wood, B., Yoon, J., Parikh, D., Zitnick, C. L., and Ulissi, Z. Open catalyst 2020 (oc20) dataset and community challenges. *ACS Catalysis*, 2021. doi: 10.1021/acscatal.0c04525.

Chmiela, S., Tkatchenko, A., Sauceda, H. E., Poltavsky, I., Schütt, K. T., and Müller, K.-R. Machine learning of accurate energy-conserving molecular force fields. *Science advances*, 3(5):e1603015, 2017.

Chmiela, S., Sauceda, H. E., Müller, K.-R., and Tkatchenko, A. Towards exact molecular dynamics simulations with machine-learned force fields. *Nature communications*, 9 (1):3887, 2018.

Christensen, A. S. and von Lilienfeld, O. A. On the role of gradients for machine learning of molecular energies and forces. *Mach. Learn.: Sci. Technol.*, 1(4):045018, oct 2020. doi: 10.1088/2632-2153/abba6f.

Cooper, A. M., Kästner, J., Urban, A., and Artrith, N. Efficient training of ann potentials by including atomic forces via taylor expansion and application to water and a transition-metal oxide. *npj Computational Materials*, 6 (1):54, 2020.

Drautz, R. Atomic cluster expansion for accurate and transferable interatomic potentials. *Physical Review B*, 99(1): 014104, 2019.

Feller, D., Peterson, K. A., and Crawford, T. D. Sources of error in electronic structure calculations on small chemical systems. *J. Chem. Phys.*, 124(5):054107, 2006. doi: 10.1063/1.2137323. URL https://doi.org/10.1063/1.2137323.

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

Fu, X., Wu, Z., Wang, W., Xie, T., Keten, S., Gomez-Bombarelli, R., and Jaakkola, T. S. Forces are not enough: Benchmark and critical evaluation for machine learning force fields with molecular simulations. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=A8pqQipwkt. Survey Certification.

Godwin, J., Schaarschmidt, M., Gaunt, A., Sanchez-Gonzalez, A., Rubanova, Y., Veličković, P., Kirkpatrick, J., and Battaglia, P. Simple gnn regularisation for 3d molecular property prediction & beyond. *arXiv preprint arXiv:2106.07971*, 2021.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Hobza, P. and Šponer, J. Toward true dna base-stacking energies: Mp2, ccsd(t), and complete basis set calculations. *J. Am. Chem. Soc.*, 124(39):11802–11808, 2002. doi: 10.1021/ja026759n. URL https://doi.org/10.1021/ja026759n. PMID: 12296748.

Hohenberg, P. and Kohn, W. Inhomogeneous electron gas. *Phys. Rev.*, 136(3B):B864–B871, Nov 1964. doi: 10.1103/PhysRev.136.B864. URL https://link.aps.org/doi/10.1103/PhysRev.136.B864.

Hu, W., Shuaibi, M., Das, A., Goyal, S., Sriram, A., Leskovec, J., Parikh, D., and Zitnick, C. L. Forcenet: A graph neural network for large-scale quantum calculations. *arXiv preprint arXiv:2103.01436*, 2021.

Iftimie, R., Minary, P., and Tuckerman, M. E. Ab initio molecular dynamics: Concepts, recent developments, and future trends. *Proc. Natl. Acad. Sci.*, 102(19):6654–6659, 2005. doi: 10.1073/pnas.0500193102.

Kohn, W. and Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.*, 140 (4A):A1133–A1138, Nov 1965. doi: 10.1103/PhysRev.140.A1133. URL https://link.aps.org/doi/10.1103/PhysRev.140.A1133.

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. Mace-off23: Transferable machine learning force fields for organic molecules, 2023.

Li, Z., Kermode, J. R., and De Vita, A. Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces. *Phys. Rev. Lett.*, 114(9):096405, 2015.

Liao, Y.-L., Wood, B., Das*, A., and Smidt*, T. Equiformerv2: Improved equivariant transformer for scaling to higher-degree representations. *arxiv preprint arxiv:2306.12059*, 2023.

Merchant, A., Batzner, S., Schoenholz, S. S., Aykol, M., Cheon, G., and Cubuk, E. D. Scaling deep learning for materials discovery. *Nature*, 624:80–85, 2023. doi: 10.1038/s41586-023-06735-9.

Miyato, T., Maeda, S.-i., Koyama, M., and Ishii, S. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018.

Musaelian, A., Batzner, S., Johansson, A., Sun, L., Owen, C. J., Kornbluth, M., and Kozinsky, B. Learning local equivariant representations for large-scale atomistic dynamics. *Nature Communications*, 14(1):579, 2023.

Ni, Y., Feng, S., Ma, W.-Y., Ma, Z.-M., and Lan, Y. Sliced denoising: A physics-informed molecular pre-training method. *arXiv preprint arXiv:2311.02124*, 2023.

Parrinello, M. From silicon to rna: The coming of age of ab initio molecular dynamics. *Solid State Commun.*, 102(2):107–120, 1997. ISSN 0038-1098. doi: https://doi.org/10.1016/S0038-1098(96)00723-5. URL https://www.sciencedirect.com/science/article/pii/S0038109896007235. Highlights in Condensed Matter Physics and Materials Science.

Passaro, S. and Zitnick, C. L. Reducing SO(3) Convolutions to SO(2) for Efficient Equivariant GNNs. In *International Conference on Machine Learning (ICML)*, 2023.

Podryabinkin, E. V. and Shapeev, A. V. Active learning of linearly parametrized interatomic potentials. *Comput. Mater. Sci.*, 140:171–180, 2017. doi: https://doi.org/10.1016/j.commatsci.2017.08.031.

Purvis, G. D. and Bartlett, R. J. A full coupled-cluster singles and doubles model: The inclusion of disconnected triples. *J. Chem. Phys.*, 76(4):1910–1918, 1982. doi: 10.1063/1.443164. URL https://doi.org/10.1063/1.443164.

Raissi, M., Perdikaris, P., and Karniadakis, G. E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, February 2019. doi: 10.1016/j.jcp.2018.10.045.

Schütt, K., Kindermans, P.-J., Sauceda Felix, H. E., Chmiela, S., Tkatchenko, A., and Müller, K.-R. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in neural information processing systems*, 30, 2017.

Schütt, K., Unke, O., and Gastegger, M. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International Conference on Machine Learning*, pp. 9377–9388. PMLR, 2021.

Shapeev, A. V. Moment tensor potentials: A class of systematically improvable interatomic potentials. *Multiscale Model. Simul.*, 14(3):1153–1173, 2016. doi: 10.1137/15M1054183.

Shuaibi, M., Kolluru, A., Das, A., Grover, A., Sriram, A., Ulissi, Z., and Zitnick, C. L. Rotation invariant graph neural networks using spin convolutions. *arXiv preprint arXiv:2106.09575*, 2021a.

Shuaibi, M., Sivakumar, S., Chen, R. Q., and Ulissi, Z. W. Enabling robust offline active learning for machine learning potentials using simple physics-based priors. *Mach. Learn.: Sci. Technol.*, 2(2):025007, dec 2021b. doi: 10.1088/2632-2153/abcc44. URL https://dx.doi.org/10.1088/2632-2153/abcc44.

Smith, J. S., Isayev, O., and Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.*, 8(4):3192–3203, 2017. doi: 10.1039/C6SC05720A. URL http://dx.doi.org/10.1039/C6SC05720A.

Smith, J. S., Nebgen, B. T., Zubatyuk, R., Lubbers, N., Devereux, C., Barros, K., Tretiak, S., Isayev, O., and Roitberg, A. E. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat. Commun.*, 10(1):2903, 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-10827-4.

Smith, J. S., Zubatyuk, R., Nebgen, B., Lubbers, N., Barros, K., Roitberg, A. E., Isayev, O., and Tretiak, S. The ani-1ccx and ani-1x data sets, coupled-cluster and density functional theory properties for molecules. *Scientific data*, 7(1):134, 2020.

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

Unke, O. T. and Meuwly, M. Physnet: A neural network for predicting energies, forces, dipole moments, and partial charges. *J. Chem. Theory Comput.*, 15(6):3678–3693, 2019.

Unke, O. T., Chmiela, S., Sauceda, H. E., Gastegger, M., Poltavsky, I., Schütt, K. T., Tkatchenko, A., and Müller, K.-R. Machine Learning Force Fields. *Chem. Rev.*, 121:10142–10186, 2021. doi: 10.1021/acs.chemrev.0c01111.

Vandermause, J., Torrisi, S. B., Batzner, S., Xie, Y., Sun, L., Kolpak, A. M., and Kozinsky, B. On-the-fly active learning of interpretable bayesian force fields for atomistic rare events. *npj Comput. Mater.*, 6(20), 2020.

Yang, X., Song, Z., King, I., and Xu, Z. A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

Zaverkin, V. and Kästner, J. Gaussian Moments as Physically Inspired Molecular Descriptors for Accurate and Scalable Machine Learning Potentials. *J. Chem. Theory Comput.*, 16(8):5410–5421, 2020. ISSN 1549-9618. doi: 10.1021/acs.jctc.0c00347.

Zaverkin, V., Holzmüller, D., Steinwart, I., and Kästner, J. Fast and Sample-Efficient Interatomic Neural Network Potentials for Molecules and Materials Based on Gaussian Moments. *J. Chem. Theory Comput.*, 17(10):6658–6670, 2021. ISSN 1549-9618. doi: 10.1021/acs.jctc.1c00527.

Zaverkin, V., Holzmüller, D., Christiansen, H., Errica, F., Alesiani, F., Takamoto, M., Niepert, M., and Kästner, J. Uncertainty-biased molecular dynamics for learning uniformly accurate interatomic potentials. *npj Computational Materials*, 10(1):83, 2024.

A. Additional Related Work

Mitigation of data-sparsity problem in ML models for chemistry. Obtaining sufficient training datasets for ML models for material science is challenging, especially when considering unexplored materials, as this requires a substantial number of expensive ab initio or first principle approaches. To alleviate this challenge, one prevalent strategy is the application of active learning (AL) methods by applying physics and chemistry insights to explore the target molecules' phase space (Li et al., 2015; Podryabinkin & Shapeev, 2017; Vandermause et al., 2020; Shuaibi et al., 2021b; Briganti & Lunghi, 2023; Zaverkin et al., 2024). On the other hand, it is frequently reported that using equivariant ML models reduces the required amount of training data because equivariant ML models do not require additional data to understand translation and rotation symmetry, which is often required to train a high-performance image recognition model with the usual convolution and attention layer.

B. Experimental Setup and Datasets

Experiment Scripts. We developed our code based on the scripts provided in (Fu et al., 2023) with extension by taking into account the latest Open Catalyst Project code base (Chanussot* et al., 2021). We follow all the hyperparameters introduced in the Open Catalyst Project, which are highly tuned to their large dataset for the potential energy and force prediction [5]. We have only adjusted the learning rate scheduler, which can be found in our repository. The loss functions for the potential energy and forces are the mean-absolute error (MAE) and L_2-norm (L2MAE) losses with coefficients typically $1, 100$, respectively. The detailed model hyper-parameters are provided in our repository. For the PITC and PISC loss functions, we use the mean square error (MSE) loss.

Datasets. The datasets are split into train, validation, and test datasets for our experiments. For this purpose, we first randomly shuffled the dataset and sampled the training dataset with the assumed number. For the validation dataset, we sampled the same number as the training dataset when the number was more than 100; We sampled 100 samples when the training sample number was less than 100. Note that in the case of rMD17 dataset, we followed the previous work convention used in (Fu et al., 2023), which prepared a validation dataset with 9000 and a test dataset with 10000 samples. For the test dataset, we used the same data set for all the training sample number cases. For ANI-1x and rMD17 datasets, we prepared 10000 samples as the test datasets; For TiO2 and LMNTO datasets, we prepared 1000 samples as the test datasets because of their smaller dataset size.

Table 7: The mini-batch size for each dataset for all the models.

	ANI-1x	TiO2	rMD17	LMNTO
mini-Batch Size	6	4	16	4

Experiment Setup. As mentioned, we followed the setup in the Open Catalyst Project code for the hyper-parameters for each model, such as the learning rate, model structural parameters, and so on. For simplicity, we used the same mini-batch size for all the models, which is provided in Table 7. Note that the mini-batch size is determined as the maximum value for 1 A100 GPU memory. All training runs were performed on a single NVIDIA A100 GPU.

To prevent over-fitting and to simulate a real-world application with an active learning approach, we set the maximum iteration number of the training as the typical number when the training loss reaches a plateau. The concrete total iteration number is provided in Table 8.

As discussed in Section 4.4, we used a uniform random number to determine the spatial-deviation vector $\delta \mathrm{r}$. More concretely, the expression of the vector is given as $\delta \mathbf{r} \equiv \delta r_0 \, \mathbf{g}$, where $\mathbf{g}$ is determined by the uniform random number $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ with $\mathbf{0} \in \mathbb{R}^{N_{\mathrm{at}}}$. In our experiment, we also determined the length δr_0 as a uniform random number with $\delta r_0 < \delta r_{\max}$[6].

[5] The Open Catalyst Project (OC) provides three kinds of tasks and predicting force and potential energy is one of them. However, we do not use it in the present work. This is because our main focus is on general-purpose MLIPs, which should be able to run proper molecular dynamics simulations but also energy relaxation. OC is designed to investigate only the latter, making it less suitable for the current study.

[6] Note that the definition of $\delta \mathbf{r}$ is slightly different from Eq. (6) for computational efficiency.

Table 8: The total iteration number for each dataset and each sample number. The number in the parentheses approximates the corresponding total epoch number.

sample number	ANI-1x	TiO2	rMD17	LMNTO
50	7.5×10^3 (900)	–	–	–
100	10^4 (600)	10^4 (400)	7.5×10^3 (1200)	10^4 (400)
1K	4×10^4 (240)	10^4 (100)	10^4 (160)	10^4 (100)
10K	10^5 (60)	–	–	–

Table 9: A table for the hyper-parameter used for the training in the paper. Each case means as: $(C_{\mathrm{PITC}}, C_{\mathrm{PISC}}, \delta r_{\max})$= Case A: (1.2, 0.8, 0.05), Case B: (1.0, 0, 0.02), Case C: (0.1, 0.01, 0.02), Case D: (1.2, 0.01, 0.05), Case E: (1.2, 0.01, 0.02), Case F: (1.2, 0.01, 0.03), and Case G: (0.01, 0.001, 0.05)

Dataset	Size	Equiformer	eSCN	PaiNN	Spinconv	SchNet
ANI-1x	50	A	C	B	A	A
	100	A	C	A	D	A
	1K	D	D	D	B	B
	10K	G	C	B	C	C
TiO2	100	A	A	–	A	A
	1K	G	A	–	A	A
rMD17 (Aspirin)	100	E	B	F	D	A
	1K	B	B	B	B	B
rMD17 (Benzene)	100	B	B	B	B	B
	1K	G	B	B	B	B
rMD17 (Naphthalene)	100	G	B	B	B	B
	1K	G	B	B	B	B
LMNTO	100	B	B	B	A	B
	1K	B	A	B	B	B

The remaining hyper-parameters are the coefficient of the PITC and PISC losses $(C_{\mathrm{PITC}}, C_{\mathrm{PISC}})$ and the absolute value of the spatial-deviation vector $\|\delta\mathbf{r}\|$. Those hyper-parameters are tuned with the Optuna (Akiba et al., 2019) for PaiNN and Equiformer v2 using the ANI-1x dataset with a 1K sample that is differently sampled from the one used for the actual training. Because of the nature of the hyper-parameter search, Optuna found several best hyper-parameters for each time. So, we selected the following representative combinations $(C_{\mathrm{PITC}}, C_{\mathrm{PISC}}, \delta r_{max})$ = Case A: (1.2, 0.8, 0.05), Case B: (1.0, 0, 0.02), Case C: (0.1, 0.01, 0.02), Case D: (1.2, 0.01, 0.05), Case E: (1.2, 0.01, 0.02), Case F: (1.2, 0.01, 0.03), and Case G: (0.01, 0.001, 0.05).

NoisyNode Implementation. In our experiment, we used a NoisyNode (Godwin et al., 2021) as one of the baseline methods. This method aims to improve the performance of ML models by adding a perturbation to the node features, such as the atomic coordinate, and requiring ML models to recover the original label. This enables ML models to be more robust to noise in the data. Although the original method recommends adding a decoder network to learn the denoising process effectively, we do not utilize a decoder network following previous work (Liao et al., 2023).

Loss Function of The Weak Label by Cooper et al., 2020. Cooper et al., 2020 proposed a similar Taylor-based weak label method. Nonetheless, the loss in Eq. (3) is different from the one proposed by Cooper et al., 2020, where reference energy and atomic force labels are used to estimate weak label of energies $E^{\mathrm{ref}}_{\mathcal{S}_{\delta\mathbf{r}}}$ for displaced atomic configurations $\mathcal{S}_{\delta\mathbf{r}}$

$$E^{\mathrm{ref}}_{\mathcal{S}_{\delta\mathbf{r}}} \approx E^{\mathrm{ref}}_{\mathcal{S}} - \sum_{i=1}^{N_{\mathrm{at}}} \left\langle \delta\mathbf{r}_i, \mathbf{F}^{\mathrm{ref}}_{i,\mathcal{S}} \right\rangle + \mathcal{O}\left(||\delta\mathbf{r}||^2\right). \tag{8}$$

Thus, trainable parameters of MLIPs are optimized by minimizing the weak label (WL) loss

$$L_{\mathrm{WL}}\left(\mathcal{S};\boldsymbol{\theta}\right)=\ell\left(E\left(\mathcal{S}_{\delta\mathbf{r}};\boldsymbol{\theta}\right),E_{\mathcal{S}}^{\mathrm{ref}}-\sum_{i=1}^{N_{\mathrm{at}}}\left\langle\delta\mathbf{r}_i,\mathbf{F}_{i,\mathcal{S}}^{\mathrm{ref}}\right\rangle\right). \tag{9}$$

Fig. 1 (a) illustrates the approach pioneered by Cooper et al., 2020, which computes the energy of a displaced atomic configuration with a Taylor expansion based on reference energy and atomic force labels. This approach was applied to train MLIPs without explicit force labels.

Training with Adversarial Direction Setup. In our experiments with adversarial direction (Section 5.5), we determined the adversarial direction using Eq. (7). More concretely, we only considered the potential energy as the $\mathbf{y}_{\mathrm{pred}}$ and $\mathbf{y}_{\mathrm{label}}$ to avoid Hessian calculation involved in the gradient calculation of forces. In addition, we considered the loss function for the potential energy as L_{dist} Then, the expression of $\mathbf{g}=\nabla_{\mathbf{r}}L_{\mathrm{dist}}$ reduces to:

$$\mathbf{g}_{\mathcal{S}}=\nabla_{\mathbf{r}_i}\sqrt{(E\left(\mathcal{S};\boldsymbol{\theta}\right)-E_{\mathcal{S}}^{\mathrm{ref}})^2}=-\frac{1}{2L_{\mathrm{dist}}}(\mathbf{F}_i\left(\mathcal{S};\boldsymbol{\theta}\right)-\mathbf{F}_{i,\mathcal{S}}^{\mathrm{ref}})(E\left(\mathcal{S};\boldsymbol{\theta}\right)-E_{\mathcal{S}}^{\mathrm{ref}}). \tag{10}$$

In our experiment, we also randomly flip the signature of the direction to avoid an overfitting to the adversarial direction.

C. Description and References for the Datasets

In this section, we provide a short description of each dataset.

ANI-1x ANI-1x dataset (Smith et al., 2020) includes 63865 molecules whose size ranges from 4 to 64 and the ML model requires to learn quantum mechanical feature (potential energy and force) on various molecules from small number of samples for each molecule.

TiO2 Titanium dioxide (TiO_2) is an industrially relevant and well-studied material. TiO2 dataset (Artrith & Urban, 2016) includes 7815 structures of several titanium dioxide phases whose reference energy and forces were obtained from DFT calculations. The number of atoms in one sample is typically 95 with the periodic boundary condition.

rMD17 The rMD17 dataset (Christensen & von Lilienfeld, 2020) includes ten relatively small-size molecules each of which has 10^5 samples collected by performing MD simulations. ML models require to learn quantum mechanical feature (potential energy and force) on one-molecule in a steady-state. In this "revised" version, the molecules are taken from the original MD17 dataset but the energies and forces are recalculated at the PBE/def2-SVP level of theory using very tight SCF convergence and very dense DFT integration grid. Consequently, the dataset is practically free from numerical noise.

LMNTO The Li-Mo-Ni-Ti oxide (LMNTO), is of technological significance as a potential high-capacity positive electrode material for lithium-ion batteries. This compound exhibits substitutional disorder, with Li, Mo, Ni, and Ti all sharing the same sublattice. This dataset include LMNTO with the composition $Li_8Mo_2Ni_7Ti_7O_{32}$ using MD simulation, resulting in approximately 2600 structures.

D. Differences in Gradients for Physics-informed Losses

In this section, we consider gradients of the proposed two losses. First, considering squared errors for L we obtain the following gradients of the loss in Eq. (8) with respect to trainable parameters

$$\frac{\mathrm{d}\mathcal{L}_{\mathrm{WL}}}{\mathrm{d}\boldsymbol{\theta}}=2\left(E\left(\mathcal{S}_{\delta\mathbf{r}};\boldsymbol{\theta}\right)-E_{\mathcal{S}}^{\mathrm{ref}}+\sum_{i=1}^{N_{\mathrm{at}}}\langle\delta\mathbf{r}_i,\mathbf{F}_{i,\mathcal{S}}^{\mathrm{ref}}\rangle\right)\frac{\mathrm{d}E\left(\mathcal{S}_{\delta\mathbf{r}};\boldsymbol{\theta}\right)}{\mathrm{d}\boldsymbol{\theta}}. \tag{11}$$

In contrast, for the PITC loss in Eq. (3) we obtain

$$\begin{aligned}\frac{\mathrm{d}\mathcal{L}_{\mathrm{PITC}}}{\mathrm{d}\boldsymbol{\theta}} =&2\left(E\left(\mathcal{S}_{\delta\mathbf{r}};\boldsymbol{\theta}\right)-E\left(\mathcal{S};\boldsymbol{\theta}\right)+\sum_{i=1}^{N_{\mathrm{at}}}\langle\delta\mathbf{r}_i,\mathbf{F}_i\left(\mathcal{S};\boldsymbol{\theta}\right)\rangle\right)\times\\ &\left(\frac{\mathrm{d}E\left(\mathcal{S}_{\delta\mathbf{r}};\boldsymbol{\theta}\right)}{\mathrm{d}\boldsymbol{\theta}}-\frac{\mathrm{d}E\left(\mathcal{S};\boldsymbol{\theta}\right)}{\mathrm{d}\boldsymbol{\theta}}+\sum_{i=1}^{N_{\mathrm{at}}}\frac{\mathrm{d}\langle\delta\mathbf{r}_i,\mathbf{F}_i\left(\mathcal{S};\boldsymbol{\theta}\right)\rangle}{\mathrm{d}\boldsymbol{\theta}}\right).\end{aligned}\tag{12}$$

The above equations indicate that the direction of the derivative of the PITC loss Eq. (12) is different from that of the weak label loss because of the incorporation of potential energy and force at the reference point, which may potentially lead to the avoidance of local minima, not only the improvement of their accuracy.

Secondly, the gradient of PISC loss in Eq. (4) is given as:

$$\begin{aligned}\frac{\mathrm{d}\mathcal{L}_{\mathrm{PISC}}}{\mathrm{d}\boldsymbol{\theta}}=&2\left(E\left(\mathcal{S};\boldsymbol{\theta}\right)-\sum_{i=1}^{N_{\mathrm{at}}}\langle\delta\mathbf{r}_i,\mathbf{F}_i\left(\mathcal{S};\boldsymbol{\theta}\right)\rangle-E\left(\mathcal{S}_{\delta\mathbf{r}'};\boldsymbol{\theta}\right)+\sum_{i=1}^{N_{\mathrm{at}}}\langle\delta\mathbf{r}''_i,\mathbf{F}_i\left(\mathcal{S}_{\delta\mathbf{r}'};\boldsymbol{\theta}\right)\rangle\right)\\ &\times\left(\frac{\mathrm{d}E\left(\mathcal{S};\boldsymbol{\theta}\right)}{\mathrm{d}\boldsymbol{\theta}}-\sum_{i=1}^{N_{\mathrm{at}}}\frac{\mathrm{d}\langle\delta\mathbf{r}_i,\mathbf{F}_i\left(\mathcal{S};\boldsymbol{\theta}\right)\rangle}{\mathrm{d}\boldsymbol{\theta}}-\frac{\mathrm{d}E\left(\mathcal{S}_{\delta\mathbf{r}'};\boldsymbol{\theta}\right)}{\mathrm{d}\boldsymbol{\theta}}+\sum_{i=1}^{N_{\mathrm{at}}}\frac{\mathrm{d}\langle\delta\mathbf{r}''_i,\mathbf{F}_i\left(\mathcal{S}_{\delta\mathbf{r}'};\boldsymbol{\theta}\right)\rangle}{\mathrm{d}\boldsymbol{\theta}}\right).\end{aligned}\tag{13}$$

E. Variations of the Configuration for the Physics-informed Spatial-consistency Loss

Table 10: A result for the spatial configuration dependence study of PISC loss. The listed numerica values are the root mean square errors on the ANI-1x dataset (Artrith & Urban, 2016). Energy [kcal/mol] and force [kcal/mol/Å] errors of different models on trained 1000 configurations with three different initial weight parameters to suppress statistical fluctuations. The case 1, 2, 3 corresponds with Eq. (4), Eq. (15) and Eq. (16).

Dataset	Model		Baseline	PISC (Case 1)	PISC (Case 2)	PISC (Case 3)
ANI-1x	PaiNN	energy	60.11	45.24	46.32	57.29
		force	13.10	**12.33**	12.42	13.28

In Section 4.2, we consider the following form of the PISC loss:

$$L_{\mathrm{PISC}}\left(\mathcal{S};\boldsymbol{\theta}\right)=\ell\left(E\left(\mathcal{S}_{\delta\mathbf{r}};\boldsymbol{\theta}\right),E\left(\mathcal{S}_{\delta\mathbf{r}'};\boldsymbol{\theta}\right)-\sum_{i=1}^{N_{\mathrm{at}}}\langle\delta\mathbf{r}''_i,\mathbf{F}_i\left(\mathcal{S}_{\delta\mathbf{r}'};\boldsymbol{\theta}\right)\rangle\right),\tag{14}$$

where $\delta\mathbf{r},\delta\mathbf{r}',\delta\mathbf{r}''$ are related as: $\delta\mathbf{r}'+\delta\mathbf{r}''=\delta\mathbf{r}$. In this section, as a variant of Eq. (4), we also consider the following three PISC losses:

$$L_{\mathrm{PISC2}}\left(\mathcal{S};\boldsymbol{\theta}\right)=\ell\left(E\left(\mathcal{S};\boldsymbol{\theta}\right)-\sum_{i=1}^{N_{\mathrm{at}}}\langle\delta\mathbf{r}_i,\mathbf{F}_i\left(\mathcal{S};\boldsymbol{\theta}\right)\rangle,E\left(\mathcal{S}_{\delta\mathbf{r}'};\boldsymbol{\theta}\right)-\sum_{i=1}^{N_{\mathrm{at}}}\langle\delta\mathbf{r}''_i,\mathbf{F}_i\left(\mathcal{S}_{\delta\mathbf{r}'};\boldsymbol{\theta}\right)\rangle\right),\tag{15}$$

$$L_{\mathrm{PISC3}}\left(\mathcal{S};\boldsymbol{\theta}\right)=\ell\left(E\left(\mathcal{S}_{\delta\mathbf{r}'};\boldsymbol{\theta}\right),E\left(\mathcal{S}_{\delta\mathbf{r}};\boldsymbol{\theta}\right)-\sum_{i=1}^{N_{\mathrm{at}}}\langle-\delta\mathbf{r}''_i,\mathbf{F}_i\left(\mathcal{S}_{\delta\mathbf{r}};\boldsymbol{\theta}\right)\rangle\right),\tag{16}$$

where the point at $\mathbf{r}+\delta\mathbf{r}$ is the point where PIRC loss is imposed. The results are provided in Table 10, which indicates that Eq. (4) (Case 1) shows the best or better performance than the other cases for both the potential energy and the force predictions. In this study, we used the ANI-1x dataset with 1000 training samples that are different from the one used to train the model used in the main body. For the coefficient of the PITC and PISC losses, we used 0.1 and 0.001.

F. Detailed Setup for Aspirin C-H Bond Potential Variation

This section describes the detailed setup and procedures for Section 5.3. First, we trained PaiNN models with and without PIWSL losses using the aspirin data of rMD17 with training samples: 100 and 200. For PIWSL, we used

Table 11: The performance of the models used for plotting Fig. 2B. All the models other than the reference model (train = 1000) are the OC20's implementation. For the reference model, we tuned the PaiNN model following the original paper (Schütt et al., 2021).

		$N_{\text{train}} = 100$		$N_{\text{train}} = 200$		$N_{\text{train}} = 1000$
		Baseline	PIWSL	Baseline	PIWSL	Baseline
PaiNN	energy	6.55	5.64	5.11	4.48	0.68
	force	7.38	7.36	3.95	3.97	1.44

$(C_{\text{PITC}}, C_{\text{PISC}}, \delta r_{\max}) = (1.2, 0.01, 0.03)$. The other experimental setups are the same as the other rMD17 experiments given in Appendix B. We used the PaiNN model with gradient-based force to obtain the reference model and tuned the model hyper-parameter with Optuna (Akiba et al., 2019). The obtained models' performance is provided in Table 11. Then, we prepared the aspirin molecule data with its atomic coordinate and atomic type and perturbed one of the C-H bond lengths from 0.8 Å to 1.8 Å divided by 100-point and estimated the corresponding potential energy with the pre-trained models. The aspirin data is provided in our publicly available source code.

G. Loss Function Dependence of the Physics-informed Taylor-Expansion Based Consistency Loss

Table 12: A result for the loss function dependence study. The listed numerica values are the root mean square errors on the ANI-1x dataset (Artrith & Urban, 2016). Energy [kcal/mol] and force [kcal/mol/Å] errors of different models on trained 1000 configurations with three different initial weight parameters to suppress statistical fluctuations. "MAE" means the mean-absolute-error and "MSE" means the mean-square-error.

Dataset	Model		Baseline	PITC MAE loss	PITC MSE loss	PITC ReLU loss
ANI-1x	PaiNN	energy	60.11	58.84	**47.09**	60.47
		force	13.10	13.18	**12.19**	13.06

Table 12 provides the result of the loss function dependence of the PIWSL. For simplicity, we only consider the PITC loss (the coefficient of the PITC and PISC losses are set as 0.1 and 0). For the ReLU loss, we considered:

$$L_{\text{ReLU}}(\mathcal{S};\boldsymbol{\theta}) = \text{ReLU}\left(\left|E(\mathcal{S};\boldsymbol{\theta}) - \sum_{i=1}^{N_{\text{at}}} \langle \delta \mathbf{r}_i, \mathbf{F}_i(\mathcal{S};\boldsymbol{\theta})\rangle - E(\mathcal{S}_{\delta\mathbf{r}};\boldsymbol{\theta})\right| - E(\mathcal{S}_{\delta\mathbf{r}};\boldsymbol{\theta})\,||\delta\mathbf{r}||^2\right), \tag{17}$$

This loss is zero when the difference between the two terms is less than the second-order term in $\delta\mathbf{r}$. The result indicates that taking the second-order term into account does not improve the performance and that the MSE loss function shows the best performance. In this study, we used the ANI-1x dataset with 1000 training samples that are different from the one used to train the model used in the main body.

H. Deviation Length Dependence

Table 13: A result for the deviation length dependence study. The listed numerical values are the root mean square errors on the ANI-1x dataset (Artrith & Urban, 2016). Energy [kcal/mol] and force [kcal/mol/Å] errors of different models on trained 1000 configurations with three different initial weight parameters to suppress statistical fluctuations.

Dataset	Model		Baseline	$\delta r_{\max} = 0.001$	$\delta r_{\max} = 0.01$	$\delta r_{\max} = 0.1$
ANI-1x	PaiNN	energy	60.11	60.43	**47.09**	109.17
		force	13.10	12.75	12.19	**11.70**

In this section, we provide the result of the deviation length dependence study: $||\delta\mathbf{r}||$. For simplicity, we only consider the PITC loss (the coefficient of the PITC and PISC losses are set as 0.1 and 0). The results are provided in Table 13 which indicates that the longer deviation vector length is fruitful for the force but too large value is harmful for the potential energy prediction performance. In this study, we used the ANI-1x dataset with 1000 training samples that is different from the one used to train the model used in the main body.

I. Perturbed Atom Number Dependence

Table 14: A result for the perturbed atom number dependence study. The listed numerical values are the root mean square errors on the ANI-1x dataset (Artrith & Urban, 2016). Energy [kcal/mol] and force [kcal/mol/Å] errors of different models on trained 1000 configurations with three different initial weight parameters to suppress statistical fluctuations.

Dataset	Model		Baseline	10%	20%	50%	75%	90%	100 %
ANI-1x	PaiNN	energy	60.11	46.68	52.37	54.51	46.94	**45.92**	46.32
		force	13.10	13.03	12.62	12.16	**12.14**	12.24	12.42

This section provides the result of the perturbed atom number dependence study. For simplicity, we only consider the PIRC loss (the coefficient of the PIRC and PISC losses are set as 0.1 and 0). In this study, we randomly selected the atoms in a training sample following the ratio of $0\%, 10\%, 20\%, 50\%, 75\%, 90\%, 100\%$. The results are provided in Table 14, which indicates that around 75% to 100% ratio cases result in the best performance for the force and the potential energy prediction. However, the number dependence is not monotonic but rather complicated. In the main body, we perturbed all the atoms (100 %) as a conservative choice. In this study, we used the ANI-1x dataset with 1000 training samples that are different from the one used to train the model used in the main body.

J. Dependence on the Number of Training Iterations

To show the effectiveness of our approach even in the case of longer training, we provide the result of the training iteration number dependence study. In this study, we performed training twice as long as in the main part, that is, 80K iterations for ANI-1x with 1K training samples. The results are provided in Table 15 and indicate that our approach performs better in the longer training case. On the other hand, the training without PIWSL losses shows an overfitting to the validation dataset, reducing its performance compared to the shorter training iteration case. In this study, we used the ANI-1x dataset with 1000 training samples that are different from the one used to train the model used in the main body. For the coefficient of the PITC and PISC losses, we used 1.2 and 0.01.

Table 15: The training iteration number dependence result. The listed numerical values are the root mean square errors on the ANI-1x dataset (Artrith & Urban, 2016). Energy (E, kcal/mol) and force (F, kcal/mol/Å) errors of different models on trained 1000 configurations with three different initial weight parameters to suppress statistical fluctuations.

Model	Iteration Number		Baseline	PIWSL
PaiNN	40K	energy	56.62	**24.53**
		force	12.96	**11.43**
	80K	energy	59.92	**23.78**
		force	13.10	**11.50**

K. Additional Experiments with Gradient-Based Force Prediction

In this section, we provide the result of the training with the gradient-based force prediction. The results are provided in Table 16, which shows that our PIWSL loss function also enables a better force prediction, even in the case of the gradient-based force prediction task. This also indicates that our PIWSL method can improve the ML model performance in the case of MLIP and other generic property prediction tasks by calculating their first derivatives in terms of the atomic coordinate and utilizing the proposed loss functions. In this study, we used the ANI-1x dataset with 1000 training samples that are different from the one used to train the model used in the main body. For the coefficient of the PITC and PISC losses, we used 0.1 and 0.01 with $\|\delta\mathbf{r}\|_{\max} = 0.02$.

Table 16: The results with gradient-based force prediction experiments. The listed numerical values are the root mean square errors on the ANI-1x dataset (Artrith & Urban, 2016). Energy [kcal/mol] and force [kcal/mol/Å] errors of different models on trained 1000 configurations with three different initial weight parameters to suppress statistical fluctuations.

Model		Baseline (GF)	PIWSL (GF)
PaiNN	energy	23.57 ± 0.62	**20.23 ± 0.18**
	force	11.32 ± 0.08	**11.13 ± 0.04**
Equiformer	energy	29.07 ± 2.32	**19.53 ± 0.32**
	force	**11.90 ± 0.13**	11.99 ± 0.03

L. Force-Rotation Improvement in the case of Models with Force-branch

Table 17: A comparison of the rotation of force in the case of the models with force-branch. The listed numerical values are the absolute value of the total rotation of the force on the ANI-1x dataset (Artrith & Urban, 2016). The models are trained 1000 configurations with three different initial weight parameters to suppress statistical fluctuations.

Model	Baseline	PITC
PaiNN	45.18	**39.06**
Equiformer	29.62	**23.42**

In this section, we study the effect of our loss functions on the rotation of force in the case of the model with the force branch. The results are provided in Table 17 which shows that our PI loss function reduces the rotation of the predicted force, allowing a better energy conservation property when used for MD simulations. In this study, we used the ANI-1x dataset with 1000 training samples that are different from the one used to train the model used in the main body. In this experiment, we used $(C_{\rm PITC}, C_{\rm PISC}, \delta r_{\rm max}) = (1.2, 0.01, 0.05)$.

M. Further Experimental Results

In this section, we provide additional results. Table 18 provides the detailed results on the experiments for ANI-1x dataset with either 50 or 10K training samples, depicted in Fig. 2. Table 20 provides the results on the Benzene and Naphthalene data in rMD17; And the other is LMNTO (Cooper et al., 2020) presented in Table 19.

ANI-1x: 50 and 10K Training Sample Cases Table 18 provides the detailed results on the experiments for ANI-1x dataset with either 50 or 10K training samples. It shows that although the small training sample number case ($N_{\rm train} = 50$) shows a large error reduction by PIWSL, we can still find around 5 to 25 % error reduction in the case with large training sample number case ($N_{\rm train} = 10$K), indicating the effectiveness of the PIWSL method on relatively large training sample regime, in particular, multi-molecule case.

LMNTO The results on the LMNTO dataset is provided in Table 19, where PIWLS shows the error reduction for most cases, in particular, the small training sample number case ($N_{\rm train} = 100$).

rMD17: Small-Molecular Dynamics Trajectory Here, we aim to analyze the effect of the PIWSL on smaller molecules with single-molecule system, different from the multiple-molecule system as ANI-1x dataset. To this purpose, we pick-up the benzene ($n_{\rm atom} = 12$), Naphthalene ($n_{\rm atom} = 18$), and aspirin ($n_{\rm atom} = 21$) as small to middle size molecules. The results are given in Table 20, which indicates that our approach is still effective in the case of single and smaller molecule. In particular, the PIWLS performance in the case of Benzene has nearly no-gain in terms of the Baseline case. We consider that this can be attributed by the too small variation in the Benzene dataset, which can be too easy for ML models to learn. We also note that the obtained results are somewhat worse than the original report (Schütt et al., 2017; 2021). This is because of the modified re-implementation of the OpenCatalyst project code and the use of the force-branch instead of the gradient-based force. The effect of the gradient-based force is given in Appendix K.

Table 18: Energy and atomic force root-mean-square errors (RMSEs) for the ANI-1x data set (Smith et al., 2020). The results are obtained by averaging over three independent runs. Energy RMSE is given in kcal/mol, while force RMSE is in kcal/mol/Å.

		$N_{\text{train}} = 50$			$N_{\text{train}} = 10K$		
		Baseline	Noisy Nodes	PIWSL	Baseline	Noisy Nodes	PIWSL
Schnet	energy	90.08 ± 1.24	**76.83 ± 0.75**	83.90 ± 2.82	24.88 ± 0.01	**24.86 ± 0.00**	24.88 ± 0.00
	force	35.49 ± 0.36	**31.13 ± 0.13**	35.30 ± 0.87	13.36 ± 0.01	13.36 ± 0.00	13.36 ± 0.00
PaiNN	energy	212.64 ± 1.14	440.11 ± 11.68	**121.36 ± 4.13**	19.14 ± 0.38	165.25 ± 4.87	**14.10 ± 0.14**
	force	22.61 ± 0.04	22.50 ± 0.22	**20.83 ± 0.28**	8.24 ± 0.10	9.22 ± 0.09	**7.89 ± 0.02**
SpinConv	energy	222.75 ± 7.12	219.85 ± 6.99	**175.38 ± 9.77**	19.42 ± 0.67	46.31 ± 10.31	**18.81 ± 0.60**
	force	**24.88 ± 0.88**	24.61 ± 0.35	25.12 ± 0.58	10.31 ± 0.33	10.78 ± 0.66	**9.94 ± 0.12**
eSCN	energy	517.17 ± 31.98	583.90 ± 33.04	**454.40 ± 11.10**	12.65 ± 0.63	165.30 ± 33.11	**10.66 ± 0.31**
	force	22.51 ± 0.09	24.04 ± 0.15	**22.28 ± 0.08**	5.11 ± 0.30	11.51 ± 0.23	**4.35 ± 0.15**
Equiformer	energy	498.58 ± 17.44	630.32 ± 0.32	**433.88 ± 79.63**	8.03 ± 0.21	970.95 ± 236.90	**7.77 ± 0.14**
	force	22.86 ± 0.04	22.92 ± 0.00	**22.72 ± 0.04**	**2.97 ± 0.00**	29.28 ± 5.63	2.98 ± 0.00

Table 19: Root mean square errors on the LMNTO dataset (Cooper et al., 2020). Energy [kcal/mol] and force [kcal/mol/Å] errors of different models on trained either 100 or 1000 configurations with three different initial weight parameters to suppress statistical fluctuations.

Dataset	Model		Baseline	NoisyNode	PIWSL	Baseline	NoisyNode	PIWSL
				$N_{\text{train}} = 100$			$N_{\text{train}} = 1000$	
LMNTO ($n_{\text{atom}} = 56$)	SchNet	energy	4.46 ± 0.00	6.10 ± 0.00	**4.45 ± 0.00**	**3.09 ± 0.00**	3.25 ± 0.00	**3.09 ± 0.00**
		force	9.24 ± 0.00	**8.31 ± 0.00**	9.24 ± 0.00	**5.09 ± 0.00**	5.21 ± 0.00	**5.09 ± 0.00**
	PaiNN	energy	6.91 ± 0.02	7.09 ± 0.04	**5.99 ± 0.02**	3.26 ± 0.01	4.61 ± 0.03	**2.98 ± 0.01**
		force	**4.75 ± 0.00**	7.20 ± 0.01	**4.75 ± 0.00**	**2.03 ± 0.00**	2.55 ± 0.00	**2.03 ± 0.00**
	SpinConv	energy	7.90 ± 0.00	7.83 ± 0.04	**7.83 ± 0.01**	4.90 ± 0.33	7.20 ± 0.06	**3.95 ± 0.02**
		force	**4.63 ± 0.01**	5.14 ± 0.04	4.71 ± 0.02	1.81 ± 0.01	2.33 ± 0.00	**1.74 ± 0.00**
	eSCN	energy	7.92 ± 0.00	7.92 ± 0.00	7.92 ± 0.00	7.93 ± 0.00	7.93 ± 0.00	**6.40 ± 0.14**
		force	4.67 ± 0.01	7.59 ± 0.02	**4.64 ± 0.01**	1.54 ± 0.00	1.98 ± 0.06	**1.53 ± 0.00**
	Equiformer v2	energy	7.40 ± 0.03	7.92 ± 0.00	**7.32 ± 0.08**	**3.57 ± 0.05**	7.04 ± 0.03	3.60 ± 0.02
		force	4.26 ± 0.00	7.60 ± 0.02	**4.24 ± 0.02**	**1.34 ± 0.00**	1.99 ± 0.00	**1.34 ± 0.00**

Table 20: Root mean square errors on the Naphthalene in rMD17 dataset (Chmiela et al., 2018). Energy (E, kcal/mol) and force (F, kcal/mol/Å) errors of different models on trained either 100 or 1000 configurations with three different initial weight parameters to suppress statistical fluctuations.

Dataset	Model		Baseline	NoisyNode	PIWSL	Baseline	NoisyNode	PIWSL
				$N_{\text{train}} = 100$			$N_{\text{train}} = 1000$	
Benzene ($n_{\text{atom}} = 12$)	Schnet	energy	**0.23 ± 0.00**	0.58 ±0.00	**0.23 ±0.00**	**0.17 ± 0.00**	0.32 ± 0.00	**0.17 ± 0.00**
		force	**2.32 ± 0.00**	3.61 ± 0.00	**2.32 ± 0.00**	**1.27 ± 0.00**	2.51 ± 0.00	**1.27 ± 0.00**
	PaiNN	energy	**0.90 ± 0.02**	2.29 ± 0.50	**0.89 ±0.03**	**0.47 ± 0.03**	0.75 ± 0.02	**0.49 ± 0.03**
		force	**0.57 ± 0.00**	5.33 ± 0.16	**0.57 ± 0.00**	**0.23 ± 0.00**	2.50 ± 0.00	0.30 ± 0.00
	SpinConv	energy	2.27 ± 0.09	2.32 ± 0.00	**1.61 ±0.28**	**0.90 ± 0.12**	2.35 ± 0.01	1.07 ± 0.00
		force	**0.61 ± 0.01**	3.56 ± 0.00	0.65 ± 0.01	**0.39 ± 0.00**	2.33 ± 0.00	0.43 ± 0.00
	eSCN	energy	**0.59 ± 0.01**	3.47 ± 0.04	**0.58 ± 0.03**	0.20 ± 0.00	1.01 ± 0.01	**0.19 ± 0.00**
		force	**0.74 ± 0.01**	8.43 ± 0.18	**0.75 ± 0.02**	**0.14 ± 0.00**	2.99 ± 0.01	**0.14 ± 0.00**
	Equiformer	energy	1.55 ± 0.01	2.08 ± 0.01	**1.52 ± 0.01**	0.281 ± 0.01	1.68 ± 0.02	**0.276 ± 0.01**
		force	**0.72 ± 0.00**	10.32 ± 0.04	**0.72 ± 0.01**	0.15 ± 0.00	2.89 ± 0.00	**0.13 ± 0.00**
Naphthalene ($n_{\text{atom}} = 18$)	Schnet	energy	**1.41 ± 0.00**	1.92 ± 0.00	**1.41 ± 0.00**	**1.05 ± 0.00**	1.49 ± 0.00	**1.05 ± 0.00**
		force	**5.76 ± 0.00**	5.96 ± 0.00	**5.76 ± 0.00**	**3.80 ± 0.00**	4.08 ± 0.00	**3.80 ± 0.00**
	PaiNN	energy	3.63 ± 0.01	5.13 ± 0.06	**3.54 ± 0.02**	1.37 ± 0.02	2.22 ± 0.04	**1.33 ± 0.01**
		force	**1.98 ± 0.01**	10.99 ± 0.05	**1.99 ± 0.00**	**0.72 ± 0.00**	2.56 ± 0.01	**0.72 ± 0.00**
	SpinConv	energy	2.96 ± 0.22	5.73 ± 0.00	**2.88 ± 0.02**	**1.80 ± 0.02**	3.39 ± 0.00	2.40 ± 0.18
		force	2.04 ± 0.01	3.91 ± 0.00	**1.99 ± 0.01**	0.97 ± 0.00	2.46 ± 0.00	**0.96 ± 0.00**
	eSCN	energy	**2.07 ± 0.03**	7.63 ± 0.05	2.12 ± 0.01	**0.56 ± 0.01**	2.15 ± 0.29	0.58 ± 0.01
		force	**2.28 ± 0.01**	9.68 ± 0.23	2.32 ± 0.23	**0.42 ± 0.01**	2.86 ± 0.03	**0.42 ± 0.01**
	Equiformer	energy	4.37 ± 0.03	5.70 ± 0.05	**4.27 ± 0.01**	**0.71 ± 0.02**	3.70 ± 0.10	**0.72 ± 0.02**
		force	1.93 ± 0.03	12.73 ± 0.06	**1.89 ± 0.00**	0.43 ± 0.02	3.20 ± 0.02	**0.38 ± 0.06**
Aspirin ($n_{\text{atom}} = 21$)	Schnet	energy	3.76 ± 0.00	**3.56 ± 0.00**	3.74 ± 0.00	**2.77 ± 0.00**	3.08 ± 0.00	**2.77± 0.00**
		force	12.32 ± 0.00	**11.59 ± 0.00**	12.20 ± 0.00	**6.63 ± 0.00**	7.03 ± 0.00	**6.63 ± 0.00**
	PaiNN	energy	6.55 ± 0.03	9.36 ± 0.08	**5.64 ± 0.02**	4.07 ± 0.01	4.10 ± 0.01	**3.99± 0.01**
		force	**7.38 ± 0.02**	20.37 ± 0.04	**7.36 ± 0.03**	2.17 ± 0.00	2.17 ± 0.00	**2.16 ± 0.01**
	SpinConv	energy	5.71 ± 0.04	6.11 ± 0.00	**5.03 ± 0.01**	4.04 ± 0.09	4.12 ± 0.06	**3.42 ± 0.20**
		force	**8.68 ± 0.03**	10.17 ± 0.00	8.94 ± 0.02	1.88 ± 0.00	1.89 ± 0.00	**1.83± 0.01**
	eSCN	energy	5.14 ± 0.02	6.44 ± 0.04	**4.82± 0.13**	**1.28 ± 0.03**	1.28 ± 0.02	1.29 ± 0.01
		force	6.14 ± 0.03	13.88± 0.16	**6.10 ± 0.03**	1.30 ± 0.01	**1.29± 0.02**	1.30 ± 0.01
	Equiformer	energy	4.79 ± 0.02	5.75 ± 0.08	**4.66 ± 0.04**	1.83 ± 0.04	1.83 ± 0.06	**1.75 ± 0.01**
		force	**4.86 ± 0.03**	16.80 ± 0.05	**4.86 ± 0.03**	1.00 ± 0.03	1.00 ± 0.00	**0.94 ± 0.08**