Beyond Scaling: Chemical Intuition as Emergent Ability of Universal Machine Learning Interatomic Potentials

Shinnosuke Hattori
Advanced Research Laboratory
Sony Group Corporation
Atsugi, Kanagawa, Japan
shinnosuke.hattori@sony.com

Kohei Shimamura
Department of Physics
Kumamoto University
Kumamoto, Japan

Ken-ichi Nomura
Collaboratory for Advanced
Computing and Simulation
University of Southern California
Los Angeles, CA, USA
knomura@usc.edu

Aiichiro Nakano
Collaboratory for Advanced
Computing and Simulation
University of Southern California
Los Angeles, CA, USA

Rajiv K. Kalia
Collaboratory for Advanced
Computing and Simulation
University of Southern California
Los Angeles, CA, USA

Priya Vashishta
Collaboratory for Advanced
Computing and Simulation
University of Southern California
Los Angeles, CA, USA

Abstract

Machine Learning Interatomic Potentials (MLIPs) have successfully demonstrated power-low scaling in their training performance, however, the emergence of novel capabilities at scale remains unexplored. We have developed Edge-wise Emergent Decomposition (E3D) framework to investigate how an MLIP develops the ability to derive physically meaningful local representations of chemical bonds without explicit supervision. Employing an E(3)-equivariant network (Allegro) trained on molecular data (SPICE 2), we found that the model by itself has acquired the knowledge of bond dissociation energy (BDE) for archetypal bond types. The emergent BDE values quantitatively agree with literature and are found to be robust across distinct organic and inorganic training sets. E3D employs a set of internal representations, probability distribution, and associated information entropy to enable visual inspection and quantitative assessment of various model training scenarios. We apply E3D framework and discuss the synergetic effect of hybrid training set along with its potential to overcome the scaling wall for transition state energy prediction problem.

1 Introduction

Machine Learning Interatomic Potentials (MLIPs) have revolutionized computational simulation by delivering near-quantum accuracy at substantially reduced computational cost [1, 2, 3]. Recent progress has been driven primarily by scaling—increasing model size and dataset diversity—leading to predictable power-law improvements in accuracy [4, 5]. This scaling success has enabled novel

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

software frameworks [6, 7, 8, 9, 10, 11, 12, 13], also large and diverse datasets including SPICE [14, 15], MPTrj [13], Alexandria [16], TM23 [17], OMat24 [18], OMol25 [19], and MatPES [20]. These advances have produced increasingly generalizable MLIPs for applications ranging from battery materials [21], catalysts [22, 23], drug discovery [24, 25], and nanodevices [12, 26, 27, 28].

While current trends of MLIP development follow "bigger and more diverse" strategy [4, 5, 29], it has become increasingly clear that the strategy faces challenges in the prediction of complex chemical reactions [22].

Accurate prediction of chemical reactions is one of long-sought capabilities of MLIPs. Figures 1**a** and 1**b** schematically present the potential energy surface (PES) and reaction pathway learned by an MLIP. The mechanistic understanding of reaction pathways and energy barriers may substantially accelerate the development and synthesis of novel materials. Therefore, great efforts have been made to date, such as large-scale data generation initiatives, e.g. the Open Catalyst Project [22, 30, 23], and active learning for reaction modeling including ANI potential [29]. A recent development in "foundation model" has also demonstrated surprising predictability of chemical reactions. These models have shown enhanced performance in chemical reaction prediction through training on a hybrid dataset that combines organic and inorganic materials data [31, 32].

Despite these advancements, the lack of our understanding of how a model acquires chemical intuition during training remains a great barrier for scalable MLIPs to address the wide aspects of chemical reactions. For example, Figure 1c illustrates a disparity in the scaling behavior of reaction energy (ΔE) and activation energy (E_a) using Allegro [33] models trained on SPICE 2 dataset [14] and evaluated on Transition1x (T1x) dataset [34]. The accuracy of ΔE improves consistently with increasing training data in all tensor sizes, thus demonstrating the scaling behavior previously reported. In contrast, the predicted E_a plateaus around the data size of 10^5, hitting a "scaling wall" where additional data provides little to no benefit. This disparity is observed consistently across larger model sizes, suggesting limitations beyond the model capacity.

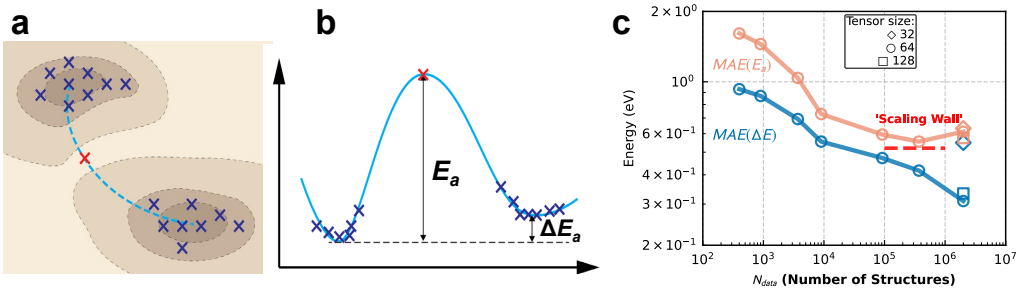

Figure 1: **Schematic of potential energy surface (PES) learning a**. A contour plot of PES learned by MLIP. Markers indicate sampled data points near the ground states (blue) and at the transition state (red). **b**. An energy profile along reaction coordinate. ΔE and E_a indicate the reaction energy (the energy difference between initial and final states) and the activation barrier (the energy difference between the initial state and the saddle point), respectively. **c**. Scaling behavior of prediction errors for ΔE and E_a on the Transition1x dataset [34]. Mean Absolute Error (MAE) versus number of SPICE 2 [14] training structures for Allegro models with tensor sizes 32 (diamonds), 64 (circles), and 128 (squares). While ΔE accuracy improves consistently, E_a predictions hit a "scaling wall."

This observation of the scaling disparity has given rise to several scientific questions; Why does the model successfully learn to predict ΔE while struggling with E_a? Any underlying data representations that could explain the learning behaviors? Is there a robust metric to capture the mechanisms behind these scaling disparities and help quantify chemical information that is internally learned by an MLIP?

To address these questions, we propose Edge-wise Emergent Energy Decomposition (E3D) framework that leverages the chemical bond and its associated Bond Dissociation Energy (BDE). Unlike the conventional approach, where the total energy is defined as the sum of atomic energies, we consider the total energy as the sum of the bond energies, which provides multiple advantages. Because the bond energy does not explicitly appear in the loss function, it serves a rigorous check of whether a

model has learned the fundamental chemistry. The bond energy is also an experimentally measurable quantity, therefore existing chemistry databases may be used for quantitative validation. Not only E3D framework is capable to examine the bond energy by pair atom type (e.g., C-H, C-C etc) but also their order (e.g. single, double, and triple bonds). This "bond type decomposition" approach substantially improves the interpretability and explainability of model prediction validated by literature.

E3D further leverages a set of internal representations (IRs), i.e. the symmetric (D_{ij}) and asymmetric (A_{ij}) terms given in Equation 7 in Method section. The two-dimensional probability distribution of D_{ij} vs. A_{ij} and associated Shannon entropy H_{2D} enable quantitative assessment of training progress and a direct access to when and how an MLIP acquires the chemical intuition during training.

In this study, we employ the E(3)-equivariant Allegro architecture [33] and demonstrate the capability of E3D framework that uncovers rich chemistry learned by the model. We also apply E3D to "the unreasonable effectiveness of hybrid data" [35], in which the energy prediction of transition state (TS) may be improved by hybridizing organic and inorganic material datasets [31, 32]. We examine the benefit by hybridizing seemingly incoherent datasets and discuss its potential to tackle the scaling wall problem.

The key contributions and findings of this study are as follows.

- Discovery of emergent BDE: A scalable model trained on large materials dataset automatically acquires chemical intuitions for various bond types, which is robust and observed across material datasets.
- Probability distribution of the IRs and its information entropy as a novel analysis tool for qualitative and quantitative evaluation in a variety of training scenarios.
- Synergetic effect of hybrid dataset to improve the accuracy of TS energy prediction. The result suggests a novel approach to overcome the scaling wall during training.

2 Methods

N-body Interaction

The potential energy E_{system} of a system composed of N atoms can be described using a body-order expansion that includes terms up to the N-body interaction:

$$E_{\mathrm{system}} = \sum_i E_i^{(1)} + \sum_{ij} E_{ij}^{(2)} + \sum_{ijk} E_{ijk}^{(3)} + \cdots + \ (N\mathrm{-body\ term}). \tag{1}$$

Since the computational cost of N-body interactions scales as $O(N^N)$, directly incorporating such terms into MLIPs is not practical. Many of the current MLIPs allow the E_{system} to be expressed as a sum of single-atom energies $\tilde{E}_i^{(1)}$ that include many-body effects:

$$E_{\mathrm{system}} = \sum_i \tilde{E}_i^{(1)}. \tag{2}$$

To accomplish this, the existing MLIPs, such as Behler-Parrinello neural network potentials [1], describe many-body interactions by combining two-body and three-body descriptors through non-linear functions, such as activation functions used in neural networks. In particular, Atomic Cluster Expansion (ACE) [9], which enables the construction of many-body descriptors in a computationally efficient manner, has significantly improved the accuracy of the representation in Eq. 2.

However, it is not necessary to represent the total energy with the single-atom energy $\tilde{E}_i^{(1)}$. Here, the cohesive energy is defined as $E_{\mathrm{coh}} \equiv E_{\mathrm{system}} - \sum_i E_i^{(1)}$, and then Eq. 1 becomes,

$$E_{\mathrm{coh}} = \sum_{ij} E_{ij}^{(2)} + \sum_{ijk} E_{ijk}^{(3)} + \cdots + \ (N\mathrm{-body\ term}). \tag{3}$$

E_{coh} is preferred to E_{system} for the training target of MLIPs, because it eliminates dependencies on computational settings such as pseudopotentials in first-principles calculations, thereby providing an unbiased energy reference.

Similar to that the representation from Eq. 1 to Eq. 2 was achieved, MLIPs should be able to expand E_{coh} in Eq. 3 using two-body interactions $\tilde{E}_{ij}^{(2)}$ that include many-body effects, as follows:

$$E_{\mathrm{coh}} = \sum_{ij} \tilde{E}_{ij}^{(2)}. \tag{4}$$

Since $\tilde{E}_{ij}^{(2)}$ can be naturally interpreted as a quantitative measure of the attractive or repulsive interatomic interactions, Eq. 4 serves as a fundamental concept for evaluating interatomic bond strengths using MLIPs themselves.

Model and Energy Decomposition Formulation

We employ the E(3)-equivariant Allegro architecture [33] for our analysis due to its inherent energy decomposability shown in Eq. 4. We trained Allegro models with internal tensor representation sizes of 32, 64, and 128, spherical harmonics expansion $l_{\mathrm{max}} = 2$, and radial cutoff $r_{\mathrm{cut}} = 5.2$ Å. For most analyses, we employed a tensor size of 64, prioritizing computational efficiency without compromising accuracy. This specific size proved optimal: a smaller tensor (size 32) resulted in lower accuracy due to its limited parameter count and thus reduced representational capacity, whereas a larger tensor (size 128) demonstrated comparable performance to size 64.

Allegro decomposes total system energy E_{system} into per-node (atom) energies ε_i and per-edge energies ε_{ij}[33]:

$$\begin{aligned} E_{\mathrm{system}} &= \sum_{i=1}^{N} (\sigma_{Z_i} \varepsilon_i + \mu_{Z_i}), \\ \varepsilon_i &= \sum_{j \in N(i)} \sigma_{Z_i Z_j} \varepsilon_{ij}, \end{aligned} \tag{5}$$

where σ_{Z_i}, $\sigma_{Z_i Z_j}$ are learnable per-species scale parameters, μ_{Z_i} are learnable shift parameters for atoms of species Z_i and Z_j, N is the total number of atoms, and $N(i)$ represents atom i's local neighborhood. The Allegro model, which is based on the ACE formalism [9] and combined with their connection via nonlinear functions, enables the expression of E_{system} in terms of the per-edge energies ε_{ij}[33].

Therefore, in order to implement an architecture that satisfies Eq. 4 within the Allegro model, we standardize the normalization parameters by setting the shift parameters $\mu_{Z_i} = 0$ and scale factors $\sigma_{Z_i} = \sigma_{Z_i Z_j} = 1.0$ in Eq. 5. Furthermore, instead of learning E_{system}, the model is trained to predict E_{coh}. With these settings, Eq. 5 becomes a direct sum of edge energies ε_{ij}:

$$E_{\mathrm{coh}} = \sum_{i=1}^{N} \sum_{j \in N(i)} \varepsilon_{ij}, \tag{6}$$

which is equivalent to Eq. 4.

Importantly, ε_{ij} and ε_{ji} are generally not the same (i.e., $\varepsilon_{ij} \neq \varepsilon_{ji}$) reflecting the local environments of i and j atoms independently. Thus, these energy values differ based on which atom index to be taken as the central atom. We define a set of metrics, i.e. the symmetric component D_{ij} and asymmetric component A_{ij}, to capture the degree of the bond energy strength and its asymmetry:

$$\begin{aligned} D_{ij} &= \varepsilon_{ij} + \varepsilon_{ji}, \\ A_{ij} &= \varepsilon_{ij} - \varepsilon_{ji}. \end{aligned} \tag{7}$$

Here, D_{ij} represents total bond energy between ith and jth atoms. The summation of D_{ij} for all atomic pairs in the system, i.e., $\sum_i^N \sum_{j \in N(i), j>i} D_{ij}$ is equal to E_{coh}. The asymmetric component A_{ij} quantifies energy imbalance arising from the difference in the local environment of the two atoms. For example if ith and jth atoms are of the same element, A_{ij} reflects the difference between the two local environments.

We trained the Allegro models modified with unscaled ε_{ij} outputs using NequIP framework[10]. Detailed training procedure, hyperparameters and datasets are provided in Supplemental Material.

E3D Protocol

We assigned bond multiplicities (single, double, triple, aromatic) using General Amber Force Field (GAFF) atom types from `antechamber` in AmberTools 23[36] with manual validation for ambiguous cases. Reference BDE values from experimental measurements [37] enable quantitative comparison with learned D_{ij} distributions, providing a direct assessment of the physical meaningfulness of the learned representation.

In addition we define two key metrics, the BDE distribution shift Δ_{BDE} and the BDE distribution width σ_{BDE}, to enable a unified evaluation metric across different model training settings. Δ_{BDE} and σ_{BDE} are computed from a single distribution combining the BDE distributions of all bond types, each of which is shifted by their reference BDE value. See the inset in Figure 2 in main text.

To analyze correlations between the symmetric (D_{ij}) and asymmetric (A_{ij}) components of learned bond energies, we generated $2D$ histograms of D_{ij} versus A_{ij}. This analysis focused on C–C, C–H, C–O, and C–N from T1x dataset structures, considering covalent bonds shorter than 2.2 Å to ensure chemical relevance. We constructed these histograms using uniform 0.1 eV energy bin widths for both axes over their observed ranges.

Shannon entropy quantifies organization structure of representational spaces:

$$H_{2D} = -\sum_{k,l} p_{kl} \log p_{kl} \tag{8}$$

where p_{kl} is the normalized frequency in bin (k, l) of the $2D$ histogram. Lower entropy values indicate more organized, well-defined internal representations, suggesting that the model has developed structured chemical understanding. This metric provides a quantitative measure of how clearly the model distinguishes different chemical environments and structure.

Table 1 summarizes the metrics employed in the E3D framework to characterize learned representations and their relationship to physical quantities.

Table 1: Key metrics in E3D framework.

Metric	Symbol	Description
Symmetric Bond Energy	D_{ij}	Total interaction energy: $\varepsilon_{ij} + \varepsilon_{ji}$
Asymmetric Bond Energy	A_{ij}	Energy imbalance due to asymmetry: $\varepsilon_{ij} - \varepsilon_{ji}$
BDE Distribution Shift	Δ_{BDE}	Mean deviation from reference BDEs
BDE Distribution Width	σ_{BDE}	Standard deviation of D_{ij} distributions
$2D$ Map Information Entropy	H_{2D}	Shannon entropy of the D_{ij} vs. A_{ij} map

3 Results

Emergent Bond Dissociation Energy

First, we demonstrate that Allegro model can acquire the knowledge of BDE without explicit supervision. The constraints for Allegro were implemented to achieve output completion as described in Eq. 6. We trained the Allegro model using SPICE 2 [14] and evaluated the generalization performance using T1x [34] dataset. Subsequently, we analyzed changes in IRs with increasing amounts of training data and verified whether generalization monitoring functions appropriately.

Figure 2**a** compares D_{ij} against BDEs from literature. The model generates a distribution that aligns well with established chemical trends. Not only the order of bond energy of C–H, N–H, and O–H are correctly predicted, but also the average of the bond energies shows a quantitative agreement with the reference BDE. Surprisingly, the model appears to have learned the concept of bond order correctly, for example single, double, and triple bonds for C–C, C–N, C–O, and N–N bonds, despite the fact that the model was trained solely on the total system energy. We call the novel model capability as 'emergent BDE.'

Our E3D framework enables analysis of how the quality of IR changes with dataset size, providing a means to analyze the scaling wall. As shown in Figures 2**b** and 2**c**, the mean difference Δ_{BDE} and

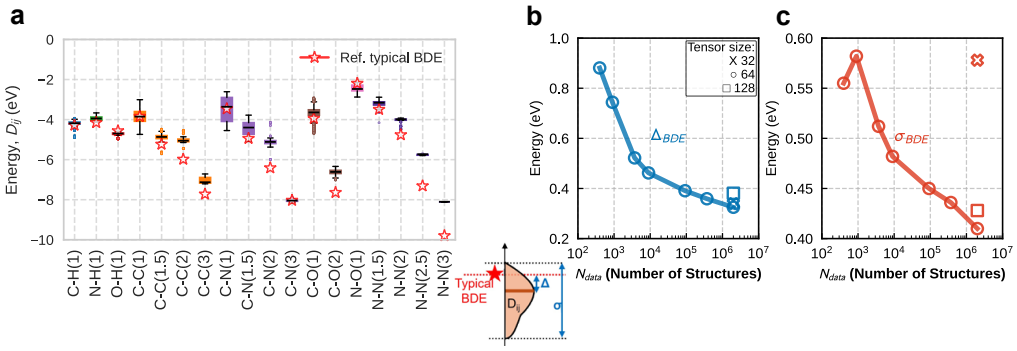

Figure 2: **Emergent BDE. a**. Distributions of learned symmetric bond energies (D_{ij}) from Allegro model (tensor size 64) trained on SPICE 2 dataset [14], compared with reference BDE values (star markers) [37] for the bond types commonly found in SPICE 2 and T1x datasets. The model successfully predicts the value and order of chemical bonds. The numbers (1), (2), and (3) represent single, double, and triplet bonds, respectively. The numbers (1.5) and (2.5) represent single and double bonds, respectively, in aromatic rings. **b.** BDE distribution shift with training data size (Δ_{BDE}): MAE between learned D_{ij} means and reference BDEs. **c.** BDE distribution width (σ_{BDE}): three standard deviations of learned D_{ij} distributions. While Δ_{BDE} exceeded 10^4, the slope became smaller and showed a tendency toward saturation. σ_{BDE} continued to show a monotonic decreasing trend.

variance σ_{BDE} relative to reference BDEs show signs of convergence in energy error reduction trends when training structure counts exceed $10^{4\sim5}$. This corresponds to the position where the scaling wall in Fig. 1, suggesting that BDE emergence may be related to predictive capability for reaction tasks.

Figures 2**b** and 2**c** also illustrates the effect from tensor size to the scaling wall. While increasing the tensor size to 128 produces negligible changes in Δ_{BDE} and σ_{BDE}, reducing it to 32 significantly degrades σ_{BDE} performance, suggesting an optimal tensor size of 64.

To further assess the dataset dependency of the emergent BDE, we have also applied E3D to an Allegro model trained on MatPES dataset, majority of which consists of inorganic crystalline data. Figure 3 shows the obtained D_{ij} distribution, in which the model also acquires the chemical intuition, i.e. bond type, multiplicity, BDE value, using MatPES.

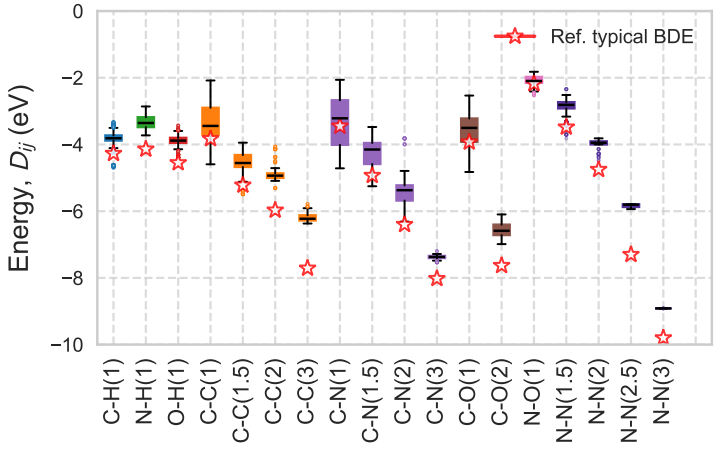

Figure 3: Emergent BDEs. Distributions of learned symmetric bond energies (D_{ij}) from an Allegro model (tensor size 128) trained on the MatPES dataset[38], compared with reference BDE values (stars markers). Distributions are qualitatively consistent with those from the SPICE 2-trained model (Figure 2).

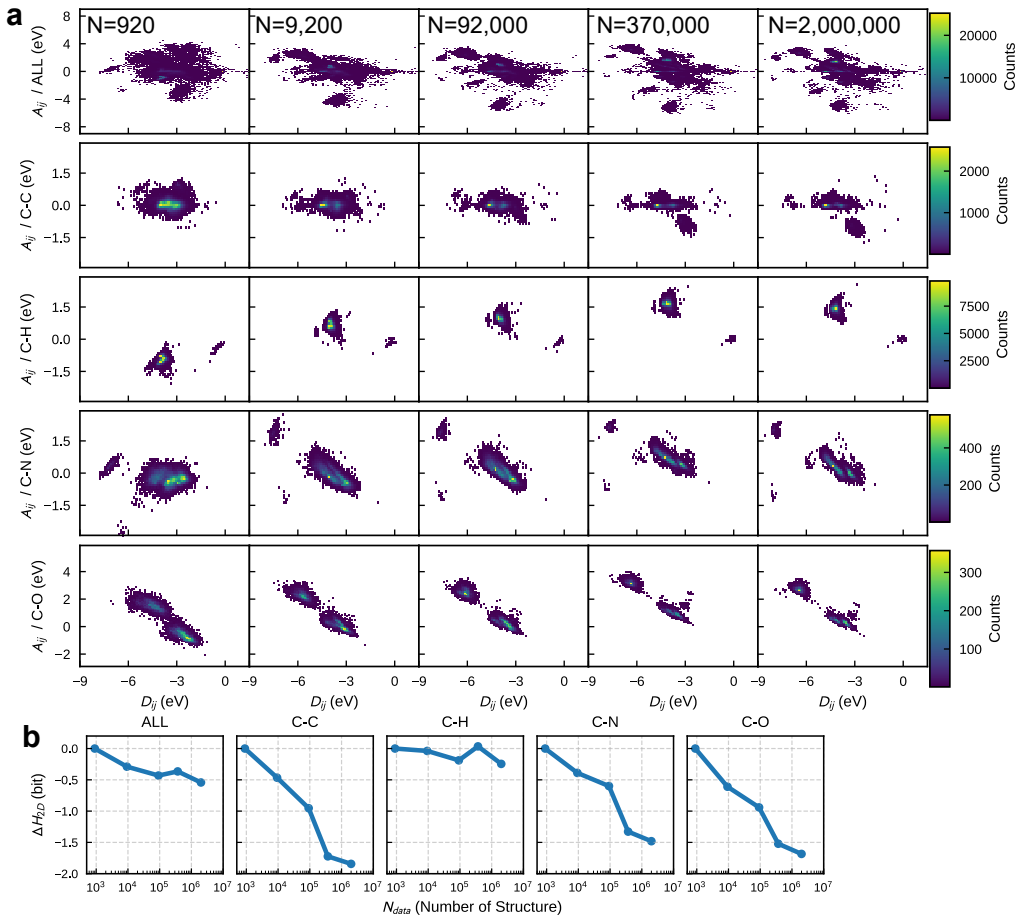

Figure 4: **Evolution of bond energy representations a.** IR distributions for C−C, C−H, C−O, and C−N bonds at varying training data volumes N_{data}. Here, we employ the structure of the test set, which has been separated from the training set in SPICE 2. **b.** H_{2D} plot as a function of N_{data}, in which the entropy value at N_{data} = 920 as reference. As N_{data} increases more refined structures and boundaries has developed that result in the consistent reduction of H_{2D}.

Distribution of Internal Representation and Shannon Entropy

Fig. 4**a** shows the distribution of the IRs, the symmetric D_{ij} and asymmetric A_{ij} terms given by Eq. 7. Using the Shannon entropy (H_{2D}), we also quantify the changes in the distribution during training as a function of the training set size. We have examined the distributions over archetypal chemical bond types (C−H, C−C, C−O, C−N) across varying training conditions. Here each training set with different size is randomly sampled from the entire SPICE 2 dataset. The distribution function shows progressive development and structural refinement that evolve into isolated peaks as training size increase. This result indicates an increasing development of complex and multi-modal features inside the model. Each island of the distribution corresponds to distinctive bond characteristics such as bond order, therefore, the obtained IR distribution may be seen as a fingerprint of given atomic pair.

Figure 4**b** shows consistent decreasing treads in H_{2D} with respect to the size of training set. We found that the BDE representations emerge even with relatively small size training set and converges by the training data size reaches approximately $10^{4\sim5}$.

Effect of Hybrid Dataset on Transition State Energy Prediction

Lastly, we examine the effect of hybrid dataset on TS energy prediction. We trained the models on either SPICE 2 alone, or a hybrid dataset consists of SPICE 2 + MatPES (labeled as Hybrid).

Table 2 summarizes the model performance using MAE of E_a and ΔE. As previously reported, we have also confirmed that the performance of TS energy prediction improves consistently using the hybrid dataset regardless of the tensor size, for example, achieving a E_a MAE of 0.44 eV (Hybrid) compared to 0.58 eV (SPICE 2).

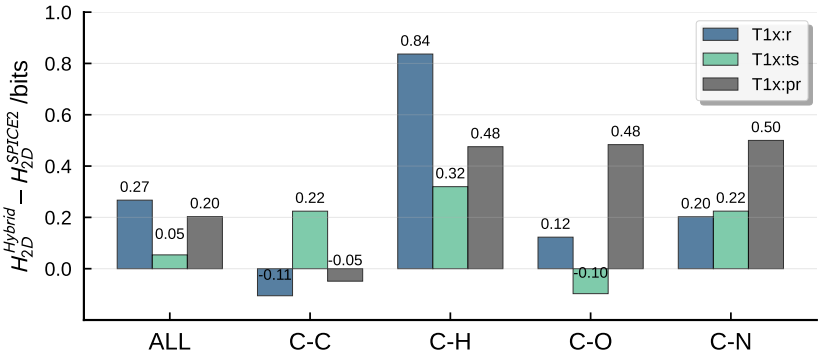

Figure 5: Entropy change in IR distributions using models trained on SPICE 2 and Hybrid (SPICE 2 + MatPES). The entropy is evaluated using T1x dataset, which is divided into two mutually exclusive sets of atomic configurations; the reactant structure (T1x:r), transition state structures (T1x:ts), and the product structures (T1x:pr). Four types of chemical bond (C−H, C−C, C−O, and C−N) are examined here. Entropy increases in various chemical bonds. C−H bonds show this across all T1x datasets. C−O and C−N bonds show especially high entropy increases in T1x:pr.

Figure 5 shows the breakdown of H_{2D} by bond types tested on the T1x dataset. Here we divide T1x into three groups, that is, the reactant structures (labeled T1x:r), transition state structures (labeled T1x:ts), and the product structures (labeled T1x:pr). We found that entropy tended to increase across various bond types, particularly in C−H bonds across all T1x groups.

Table 2: Performance improvements with hybrid dataset training on T1x reactive properties.

Dataset	Tensor size	E_a MAE (eV)	ΔE MAE (eV)
SPICE 2	64	0.61	0.31
	128	0.58	0.33
Hybrid	64	0.49	0.27
	128	**0.44**	**0.25**

4 Discussion

E3D framework has revealed for the first time the internal development of BDE within MLIP. The emergent BDE may explain the unprecedented generalizability of MLIP because it would be possible to construct and stably simulate an extremely large protein molecule [5] if an MLIP understood the basic building blocks of chemistry. Currently, most of MLIPs are trained on system energy, atomic force, and system stress; therefore, other MLIPs that exhibit remarkable scalability [11, 12, 10] may have acquired a similar IR.

The emergent BDE is achieved neither with loss function nor inductive bias, simply learned from data, a great advantage of MLIPs that learn highly nonlinear interactions automatically without tweaking fitting parameters. At the same time, it also leaves room for further improvement; one may come up with a novel network design with an inductive bias suitable for learning BDE.

We found that the emergent BDE is observed with the two widely-used large-scale materials datasets, SPICE 2 and MatPES. The result suggests that the emergent model capability is not specific to SPICE 2 dataset that predominantly consists of the bond types we examined, but rather a generic capability of MLIPs that are equipped with sound scalability.

The IR distribution and information entropy provide novel insights of the MLIP development and enable us to inspect the process from multiple aspects. By changing the data size, the distributions exhibit the formation of domains that go through the refinement of their boundaries and the shrinkage in size. In addition, such domains can split into further smaller clusters as the training set increases. See Fig. 4**a**. The result indicates that a better understanding of the local atomic environment and bond energies has been developed inside the model.

The change in entropy with respect to the size of the training set provides great insight about model training and highly correlates with the model scalability. It is particularly useful for comparing several training scenarios, for example, SPICE 2 vs. MatPES or the size of training set.

A declining trend shown in Fig. 4**b** suggests that the IR distribution converges to narrower peaks with a smaller variance. How many peaks (possible BDE values) exist depends on the training set, for example, the C-C bond with different multiplicity. The value of entropy could remain the same when the distribution barely changes (for example, C-H bond in Fig. 4**a**) or increases if more complex structures emerge as model training proceeds.

We found that the hybrid dataset exceeds the scaling wall with the SPICE 2 dataset on the prediction of TS energy, achieving a further reduction of the MAE loss by approximately 25% for E_a and ΔE. The increased entropy suggests the formation of more complex multi-modal IR structures than the one with SPICE 2 alone. Material datasets that consist of molecules or bulk systems usually have very different characteristics due to their symmetries, boundary conditions, and underlying quantum mechanics theory used to generate training set, therefore, a common practice is to train separate MLIPs depending on target applications. E3D framework has shown a synergetic effect of the hybrid dataset on TS energy prediction, i.e. an improved out-of-distribution performance using SPICE 2 and MatPES.

In summary, we have developed the E3D framework that uncovers the underlying mechanisms for the unprecedented generalization of universal MLIPs. Current MLIP advancements largely rely on methodologies designed for generic AI models, therefore, a physically-inspired method such as E3D has great potential to benefit MLIP research as well as to create a novel guideline to overcome the scaling wall.

Acknowledgements

This research was supported by the U.S. Department of Energy, Office of Basic Energy Sciences, Chemical Sciences, Geosciences, and Bioscience Division, Geosciences Program under Award DE-SC0025222. This research used resources from the Argonne Leadership Computing Facility, a U.S. DOE Office of Science user facility at Argonne National Laboratory, which is supported by the Office of Science of the U.S. DOE under Contract No. DE-AC02-06CH11357. Model development and evaluation were performed on the Sophia supercomputers at Argonne Leadership Computing Facility under the Aurora Early Science Program (ESP) U.S. DOE Innovative and Novel Computational Impact on Theory and Experiment (INCITE) Program. KN was supported by NSF grant OAC-2118061.

Author contributions

S.H., K.S. and K.N. designed the project and wrote the manuscript. K.S. developed the basic Methods of MLIP Energy Decomposition. S.H. developed fixed Allegro model and code. S.H. performed MLIP training with Data Volume Change and data analyses. K.N. supervised the project. All authors contributed to the discussion of the results as well as the writing and editing of the manuscript.

References

[1] Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.*, 98(14):146401, April 2007.

[2] Albert P Bartók, Mike C Payne, Risi Kondor, and Gábor Csányi. Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.*, 104(13):136403, April 2010.

[3] Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M Elena, Dávid P Kovács, Janosh Riebesell, Xavier R Advincula, Mark Asta, William J Baldwin, Noam Bernstein, Arghya Bhowmik, Samuel M Blau, Vlad Cărare, James P Darby, Sandip De, Flaviano Della Pia, Volker L Deringer, Rokas Elijošius, Zakariya El-Machachi, Edvin Fako, Andrea C Ferrari, Annalena Genreith-Schriever, Janine George, Rhys E A Goodall, Clare P Grey, Shuang Han, Will Handley, Hendrik H Heenen, Kersti Hermansson, Christian Holm, Jad Jaafar, Stephan Hofmann, Konstantin S Jakob, Hyunwook Jung, Venkat Kapil, Aaron D Kaplan, Nima Karimitari, Namu Kroupa, Jolla Kullgren, Matthew C Kuner, Domantas Kuryla, Guoda Liepuoniute, Johannes T Margraf, Ioan-Bogdan Magdău, Angelos Michaelides, J Harry Moore, Aakash A Naik, Samuel P Niblett, Sam Walton Norwood, Niamh O'Neill, Christoph Ortner, Kristin A Persson, Karsten Reuter, Andrew S Rosen, Lars L Schaaf, Christoph Schran, Eric Sivonxay, Tamás K Stenczel, Viktor Svahn, Christopher Sutton, Cas van der Oord, Eszter Varga-Umbrich, Tejs Vegge, Martin Vondrák, Yangshuai Wang, William C Witt, Fabian Zills, and Gábor Csányi. A foundation model for atomistic materials chemistry. *arXiv [physics.chem-ph]*, December 2023.

[4] Amil Merchant, Simon Batzner, Samuel S Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, November 2023.

[5] Boris Kozinsky, Albert Musaelian, Anders Johansson, and Simon Batzner. Scaling the leading accuracy of deep equivariant models to biomolecular simulations of realistic size. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, number Article 2 in SC '23, pages 1–12, New York, NY, USA, November 2023. Association for Computing Machinery.

[6] Kristof T Schütt, Stefaan S P Hessmann, Niklas W A Gebauer, Jonas Lederer, and Michael Gastegger. SchNetPack 2.0: A neural network toolbox for atomistic machine learning. *J. Chem. Phys.*, 158(14):144801, April 2023.

[7] Jinzhe Zeng, Duo Zhang, Denghui Lu, Pinghui Mo, Zeyu Li, Yixiao Chen, Marián Rynik, Li'ang Huang, Ziyao Li, Shaochen Shi, Yingze Wang, Haotian Ye, Ping Tuo, Jiabin Yang, Ye Ding, Yifan Li, Davide Tisi, Qiyu Zeng, Han Bao, Yu Xia, Jiameng Huang, Koki Muraoka, Yibo Wang, Junhan Chang, Fengbo Yuan, Sigbjørn Løland Bore, Chun Cai, Yinnian Lin, Bo Wang, Jiayan Xu, Jia-Xin Zhu, Chenxing Luo, Yuzhi Zhang, Rhys E A Goodall, Wenshuo Liang, Anurag Kumar Singh, Sikai Yao, Jingchao Zhang, Renata Wentzcovitch, Jiequn Han, Jie Liu, Weile Jia, Darrin M York, E, Weinan, Roberto Car, Linfeng Zhang, and Han Wang. DeePMD-kit v2: A software package for deep potential models. *J. Chem. Phys.*, 159(5), August 2023.

[8] Yury Lysogorskiy, Cas van der Oord, Anton Bochkarev, Sarath Menon, Matteo Rinaldi, Thomas Hammerschmidt, Matous Mrovec, Aidan Thompson, Gábor Csányi, Christoph Ortner, and Ralf Drautz. Performant implementation of the atomic cluster expansion (PACE) and application to copper and silicon. *npj Computational Materials*, 7(1):1–12, June 2021.

[9] Ralf Drautz. Atomic cluster expansion for accurate and transferable interatomic potentials. *Phys. Rev. B Condens. Matter*, 99(1):014104, January 2019.

[10] Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.*, 13(1):2453, May 2022.

[11] Ilyes Batatia, Dávid Péter Kovács, Gregor N C Simm, Christoph Ortner, and Gábor Csányi. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. *arXiv [stat.ML]*, June 2022.

[12] Dávid Péter Kovács, J Harry Moore, Nicholas J Browning, Ilyes Batatia, Joshua T Horton, Venkat Kapil, William C Witt, Ioan-Bogdan Magdău, Daniel J Cole, and Gábor Csányi. MACE-OFF23: Transferable machine learning force fields for organic molecules. *arXiv [physics.chem-ph]*, December 2023.

[13] Bowen Deng, Peichen Zhong, Kyujung Jun, Janosh Riebesell, Kevin Han, Christopher J Bartel, and Gerbrand Ceder. CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nature Machine Intelligence*, 5(9):1031–1041, September 2023.

[14] Peter Eastman, Benjamin P Pritchard, John D Chodera, and Thomas E Markland. Nutmeg and SPICE: Models and data for biomolecular machine learning. *J. Chem. Theory Comput.*, 20(19):8583–8593, October 2024.

[15] Peter Eastman, Pavan Kumar Behara, David L Dotson, Raimondas Galvelis, John E Herr, Josh T Horton, Yuezhi Mao, John D Chodera, Benjamin P Pritchard, Yuanqing Wang, Gianni De Fabritiis, and Thomas E Markland. SPICE, a dataset of drug-like molecules and peptides for training machine learning potentials. *Sci. Data*, 10(1):11, January 2023.

[16] Jonathan Schmidt, Hai-Chen Wang, Tiago F T Cerqueira, Silvana Botti, and Miguel A L Marques. A dataset of 175k stable and metastable materials calculated with the PBEsol and SCAN functionals. *Sci Data*, 9(1):1–8, March 2022.

[17] Cameron J Owen, Steven B Torrisi, Yu Xie, Simon L Batzner, Kyle Bystrom, J Coulter, Albert Musaelian, Lixin Sun, and B Kozinsky. Complexity of many-body interactions in transition metals via machine-learned force fields from the TM23 data set. *Npj Comput. Mater.*, 10(1):1–16, February 2023.

[18] Luis Barroso-Luque, Muhammed Shuaibi, Xiang Fu, Brandon M Wood, Misko Dzamba, Meng Gao, Ammar Rizvi, C Lawrence Zitnick, and Zachary W Ulissi. Open materials 2024 (OMat24) inorganic materials dataset and models. *arXiv [cond-mat.mtrl-sci]*, October 2024.

[19] Daniel S Levine, Muhammed Shuaibi, Evan Walter Clark Spotte-Smith, Michael G Taylor, Muhammad R Hasyim, Kyle Michel, Ilyes Batatia, Gábor Csányi, Misko Dzamba, Peter Eastman, Nathan C Frey, Xiang Fu, Vahe Gharakhanyan, Aditi S Krishnapriyan, Joshua A Rackers, Sanjeev Raja, Ammar Rizvi, Andrew S Rosen, Zachary Ulissi, Santiago Vargas, C Lawrence Zitnick, Samuel M Blau, and Brandon M Wood. The open molecules 2025 (OMol25) dataset, evaluations, and models. *arXiv [physics.chem-ph]*, May 2025.

[20] Aaron D Kaplan, Runze Liu, Ji Qi, Tsz Wai Ko, Bowen Deng, Janosh Riebesell, Gerbrand Ceder, Kristin A Persson, and Shyue Ping Ong. A foundational potential energy surface dataset for materials. *arXiv [cond-mat.mtrl-sci]*, March 2025.

[21] Suyeon Ju, Jinmu You, Gijin Kim, Yutack Park, Hyungmin An, and Seungwu Han. Application of pretrained universal machine-learning interatomic potential for physicochemical simulation of liquid electrolytes in li-ion battery. *Digit. Discov.*, May 2025.

[22] Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, Aini Palizhati, Anuroop Sriram, Brandon Wood, Junwoong Yoon, Devi Parikh, C Lawrence Zitnick, and Zachary Ulissi. The open catalyst 2020 (OC20) dataset and community challenges. *arXiv [cond-mat.mtrl-sci]*, October 2020.

[23] Brook Wander, Muhammed Shuaibi, John R Kitchin, Zachary W Ulissi, and C Lawrence Zitnick. CatTSunami: Accelerating transition state energy calculations with pretrained graph neural networks. *ACS Catal.*, pages 5283–5294, March 2025.

[24] Elena Gelžinytė, Mario Öeren, Matthew D Segall, and Gábor Csányi. Transferable machine learning interatomic potential for bond dissociation energy prediction of drug-like molecules. *J. Chem. Theory Comput.*, 20(1):164–177, January 2024.

[25] Shinnosuke Hattori and Qiang Zhu. Revisiting aspirin polymorphic stability using a machine learning potential. *ACS Omega*, 0(0):null, August 2024.

[26] Bohayra Mortazavi, Xiaoying Zhuang, Timon Rabczuk, and Alexander V Shapeev. Atomistic modeling of the mechanical properties: the rise of machine learning interatomic potentials. *Mater. Horiz.*, 10(6):1956–1968, 2023.

[27] Hugo X Rodrigues, Hudson R Armando, Daniel A da Silva, João Paulo J da Costa, Luiz A Ribeiro, Jr, and Marcelo L Pereira, Jr. Machine learning interatomic potential for modeling the mechanical and thermal properties of naphthyl-based nanotubes. *J. Chem. Theory Comput.*, 21(5):2612–2625, March 2025.

[28] Kohei Shimamura, Shinnosuke Hattori, Ken-Ichi Nomura, Akihide Koura, and Fuyuki Shimojo. Thermal conductivity calculation using homogeneous non-equilibrium molecular dynamics simulation with allegro. *Int. J. Heat Mass Transf.*, 234(126106):126106, December 2024.

[29] Shuhao Zhang, Małgorzata Z Makoś, Ryan B Jadrich, Elfi Kraka, Kipton Barros, Benjamin T Nebgen, Sergei Tretiak, Olexandr Isayev, Nicholas Lubbers, Richard A Messerly, and Justin S Smith. Exploring the frontiers of condensed-phase chemistry with a general reactive machine learning potential. *Nat. Chem.*, 16(5):727–734, May 2024.

[30] Richard Tran, Janice Lan, Muhammed Shuaibi, Brandon M Wood, Siddharth Goyal, Abhishek Das, Javier Heras-Domingo, Adeesh Kolluru, Ammar Rizvi, Nima Shoghi, Anuroop Sriram, Félix Therrien, Jehad Abed, Oleksandr Voznyy, Edward H Sargent, Zachary Ulissi, and C Lawrence Zitnick. The open catalyst 2022 (OC22) dataset and challenges for oxide electrocatalysts. *ACS Catal.*, 13(5):3066–3084, March 2023.

[31] K Nomura, Shinnosuke Hattori, Satoshi Ohmura, Ikumi Kanemasu, Kohei Shimamura, Nabankur Dasgupta, A Nakano, R Kalia, and P Vashishta. Allegro-FM: Towards equivariant foundation model for exascale molecular dynamics simulations. *The Journal of Physical Chemistry Letters*, 0(0):6637–6644, February 2025.

[32] Shiota Tomoya, Ishihara Kenji, Tuan Minh Do, Mori Toshio, and Mizukami Wataru. Taming multi-domain, -fidelity data: Towards foundation models for atomistic scale simulations. *arXiv [physics.chem-ph]*, December 2024.

[33] Albert Musaelian, Simon Batzner, Anders Johansson, Lixin Sun, Cameron J Owen, Mordechai Kornbluth, and Boris Kozinsky. Learning local equivariant representations for large-scale atomistic dynamics. *Nat. Commun.*, 14(1):579, February 2023.

[34] Mathias Schreiner, Arghya Bhowmik, Tejs Vegge, Jonas Busk, and Ole Winther. Transition1x - a dataset for building generalizable reactive machine learning potentials. *Sci. Data*, 9(1):779, December 2022.

[35] Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *IEEE Intell. Syst.*, 24(2):8–12, March 2009.

[36] David A Case, Hasan Metin Aktulga, Kellon Belfon, David S Cerutti, G Andrés Cisneros, Vinícius Wilian D Cruzeiro, Negin Forouzesh, Timothy J Giese, Andreas W Götz, Holger Gohlke, Saeed Izadi, Koushik Kasavajhala, Mehmet C Kaymak, Edward King, Tom Kurtzman, Tai-Sung Lee, Pengfei Li, Jian Liu, Tyler Luchko, Ray Luo, Madushanka Manathunga, Matias R Machado, Hai Minh Nguyen, Kurt A O'Hearn, Alexey V Onufriev, Feng Pan, Sergio Pantano, Ruxi Qi, Ali Rahnamoun, Ali Risheh, Stephan Schott-Verdugo, Akhil Shajan, Jason Swails, Junmei Wang, Haixin Wei, Xiongwu Wu, Yongxian Wu, Shi Zhang, Shiji Zhao, Qiang Zhu, Thomas E Cheatham, 3rd, Daniel R Roe, Adrian Roitberg, Carlos Simmerling, Darrin M York, Maria C Nagan, and Kenneth M Merz, Jr. AmberTools. *J. Chem. Inf. Model.*, 63(20):6183–6191, October 2023.

[37] James Speight. *Lange's handbook of chemistry, seventeenth edition*. McGraw-Hill Education, Columbus, OH, 17 edition, October 2016.

[38] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv [cs.LG]*, January 2020.