FORK: First-Order Relational Knowledge Distillation for Machine Learning Interatomic Potentials

Hyukjun Lim, Seokhyun Choung, Jeong Woo Han[*]
Department of Materials Science and Engineering
Seoul National University
{hyukjunlim, schoung9967, jwhan98}@snu.ac.kr

Abstract

State-of-the-art equivariant Graph Neural Networks (GNNs) achieve quantum-level accuracy for molecular simulations but remain computationally prohibitive for large-scale applications. Knowledge distillation (KD) presents a solution by compressing these GNN-based Machine Learning Interatomic Potentials (MLIPs) into efficient models, yet existing distillation methods fail to capture the physics. Current KD approaches rely on simplistic atom-wise feature matching, overlooking the core physical principle of interatomic interactions that define the potential energy surface (PES). We introduce FORK, **F**irst-**O**rder **R**elational **K**nowledge Distillation, a framework that distills relational knowledge from pretrained GNNs by modeling each interatomic interaction as a relational vector. Through a contrastive objective, FORK guides compact student models to preserve the geometric structure of the teacher's learned PES. On the OC20 and SPICE benchmarks, our FORK-trained student outperforms baselines in energy and force prediction, achieving faithful physical knowledge transfer at a fraction of the computational cost. In a practical high-throughput catalyst screening application, the distilled model achieves a 11.9× acceleration while preserving chemical coherency, validating its efficacy for accelerating large-scale materials discovery.

1 Introduction

The advent of Machine Learning Interatomic Potentials (MLIPs), particularly those based on equivariant Graph Neural Networks (GNNs), has enabled molecular simulations with quantum-level accuracy at a fraction of the cost of methods like Density Functional Theory (DFT) [Behler and Parrinello, 2007, Schütt et al., 2017, Unke et al., 2021]. However, a fundamental trade-off persists: state-of-the-art MLIPs rely on large parameter counts and complex operations, creating a computational bottleneck that restricts their use in large-scale applications like high-throughput materials screening and drug discovery [Liao et al., 2023, Unke et al., 2021].

Knowledge Distillation (KD) is a promising strategy to compress these large *teacher* models into computationally efficient *student* models [Hinton et al., 2015, Gou et al., 2021]. However, the direct application of conventional KD to MLIPs is fundamentally limited. Prevailing methods, which minimize the discrepancy between atom-wise hidden representations, overlook a core tenet of chemistry: a system's properties are not defined by isolated atoms but emerge from a complex network of interatomic interactions [Ekström Kelvinius et al., 2023]. Consequently, these approaches fail to capture the relational structure of the Potential Energy Surface (PES), the primary object MLIPs are designed to model.

[*] Corresponding Author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Accelerated Materials Discovery.

This work posits that effective distillation for MLIPs must explicitly transfer the learned physics governing interatomic potentials. To address this, we introduce FORK, **F**irst-**O**rder **R**elational **K**nowledge Distillation, a framework designed to distill this relational physical knowledge (see Figure 2). FORK operates by first representing interatomic interactions as *relational vectors* derived from atomic embeddings. It then employs a **contrastive geometric alignment** objective, based on the InfoNCE loss [Oord et al., 2018], to align the student's relational vector space with the teacher's, thereby preserving the structure of the learned interaction space.

The primary contributions of this paper are threefold: (1) We propose a novel, physics-informed distillation framework that directly transfers first-order interatomic relational knowledge. (2) We formulate a contrastive objective that explicitly preserves the geometric structure of the teacher's learned PES. (3) We provide empirical validation on standard benchmarks and in a high-throughput catalyst screening application, where FORK yields a significant computational speedup while preserving predictive accuracy.

2 Methods

2.1 Relational Vector Representation

To capture the directional nature of interatomic interactions, we define a normalized *relational vector* $\mathbf{r}_k$ for each edge $e_k = (\mathrm{src}, \mathrm{dst})$. This vector is the difference between the L2-normalized final atomic embeddings ($\hat{\mathbf{z}}$) of the constituent atoms for both the teacher (T) and student (S) models. For equivariant models, these embeddings ($\mathbf{z}_i$) are taken from the rotationally invariant $l = 0$ channel. For each atom i, we first compute normalized embedding $\hat{\mathbf{z}}_i = \mathbf{z}_i / \|\mathbf{z}_i\|_2$. The relational vectors for the teacher and student models are then defined as:

$$\mathbf{r}_{T,k} = \hat{\mathbf{z}}_{T,\mathrm{src}} - \hat{\mathbf{z}}_{T,\mathrm{dst}}, \quad \mathbf{r}_{S,k} = \hat{\mathbf{z}}_{S,\mathrm{src}} - \hat{\mathbf{z}}_{S,\mathrm{dst}}. \tag{1}$$

2.2 Contrastive Geometric Alignment

FORK aligns the student's relational space with the teacher's using a contrastive objective based on the InfoNCE loss [Oord et al., 2018]. For a given student relational vector $\mathbf{r}_{S,k}$, its corresponding teacher vector $\mathbf{r}_{T,k}$ is the positive sample, while all other teacher vectors in the batch $\{\mathbf{r}_{T,m}\}_{m \neq k}$ are negative samples. The FORK loss, $\mathcal{L}_{\text{FORK}}$, trains the student to identify the correct positive pair:

$$\mathcal{L}_{\text{FORK}} = -\frac{1}{E_b} \sum_{k=1}^{E_b} \log \frac{\exp(\mathbf{r}_{S,k}^{\top} \mathbf{r}_{T,k} / \tau)}{\sum_{m=1}^{E_b} \exp(\mathbf{r}_{S,k}^{\top} \mathbf{r}_{T,m} / \tau)} \tag{2}$$

where τ is a temperature hyperparameter. This objective enforces geometric alignment between the teacher's and student's learned interaction spaces.

2.3 Training Objective

The student model is trained end-to-end by minimizing a composite loss function that balances task performance with relational knowledge transfer:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda_1 \mathcal{L}_{\text{FORK}} + \lambda_2 L_{\text{KD}} \tag{3}$$

The total loss combines the primary $\mathcal{L}_{\text{task}}$ (e.g., MSE on energy and forces), our relational $\mathcal{L}_{\text{FORK}}$ loss, and an optional conventional distillation loss L_{KD} (e.g., node-to-node feature matching). The hyperparameters λ_1 and λ_2 weight the distillation terms.

3 Experiments

3.1 Comparison with Baseline KD Methods

We evaluated the performance of FORK on the large-scale Open Catalyst 2020 (OC20) dataset for catalysis [Chanussot et al., 2021] and the SPICE dataset [Eastman et al., 2023] for small molecules. The result for OC20 200K subset, SPICE dataset and experimental details are presented in Appendix B and E.

On the challenging O* subset of the OC20 dataset, we evaluated FORK's ability to distill knowledge from a large, state-of-the-art teacher model into a compact student with over a 7-fold reduction in parameters. The student model trained with a combination of FORK and traditional node-to-node (n2n) distillation consistently achieved the best performance. This combined approach reduced the energy MAE to **231.7 meV**, a substantial improvement over the 252.9 meV from the n2n baseline alone (see Table 1). We found that the methods have a synergetic effect, as FORK alone achieves second-best energy error by capturing the global potential energy landscape, whereas combining it with n2n yields the best energy and force predictions by grounding local atomic features. This highlights the clear benefit of incorporating first-order relational knowledge to effectively close the performance gap to the teacher model.

Table 1: Performance of FORK on O* subset of OC20 dataset. The best results are highlighted in **bold**. Second best results are <u>underlined</u>.

Method	**Params**	**Embedding**		**Energy**	**Force**
		MAE	Cosine Similarity	MAE (meV) ↓	MAE (meV/Å) ↓
Teacher*	153M	-	-	39.8	5.8
vanilla	22M	0.217	0.205	294.5	5.9
pretrained	22M	0.311	0.271	263.6	6.1
n2n	22M	0.078	0.839	252.9	**5.8**
Hessian	22M	1.062	0.073	363.5	26.1
Ours	22M	0.282	0.230	<u>234.1</u>	6.1
Ours (w/ n2n)	22M	0.082	0.820	**231.7**	**5.8**

* The teacher model of EquiformerV2 used for knowledge distillation. Loaded from provided checkpoint.

3.2 Real-World Application: High-Throughput Catalyst Screening

To validate the practical utility of FORK in a real-world scientific application, we conducted a case study on high-throughput catalyst screening. This task is emblematic of the challenges in materials discovery, where millions of potential candidates must be evaluated, making the computational efficiency of MLIPs a critical bottleneck [Abed et al., 2024, Broderick et al., 2023]. Detailed experimental setups are provided in Appendix F.

Computational Efficiency and Performance. The FORK-distilled student exemplifies the core trade-off of knowledge distillation: balancing predictive accuracy with computational efficiency. As detailed in Table 6, the student achieves a **11.9**× increase in inference speed, reducing the average batch inference time from 166.7 ms to **14.0 ms**. Despite a modest trade-off in accuracy, the model's predictive power is sufficient for high-throughput screening, which prioritizes the efficient ranking of candidates from large chemical spaces.

Chemical Coherency. The student model's performance trends mirror the teacher's, correlating directly with the chemical and structural complexity of the catalyst systems (Table 2). For instance, the performance gap (ΔMAE) between models increases from 0.098 eV for simpler binary intermetallics to 0.164 eV for more intricate L_{10}/L_{12} structures. Beyond surface complexity, the student inherits the teacher's nuanced chemical intuition. Despite a >30-fold parameter reduction, the student replicates the teacher's error trends, and the ranking of adsorbates by MAE is identical for both models. This demonstrates that FORK transfers a sophisticated understanding of the underlying physics. Remarkably, the student's predictions are closer to the DFT ground truth than its 39M-parameter teacher in 35% of cases, suggesting that distillation can act as a powerful regularizer.

Implications for Accelerated Materials Discovery. The efficiency of the FORK-distilled model directly accelerates materials discovery. The >10-fold speedup is critical for modern screening campaigns involving millions of calculations, potentially reducing multi-year projects to months Abed et al. [2024]. This trade-off is ideal for a hierarchical workflow: the fast student screens millions of candidates to find a promising subset, which the high-fidelity teacher then re-evaluates for final validation. This methodology significantly lowers the computational barrier for comprehensive screening and paves the way for discovering novel materials.

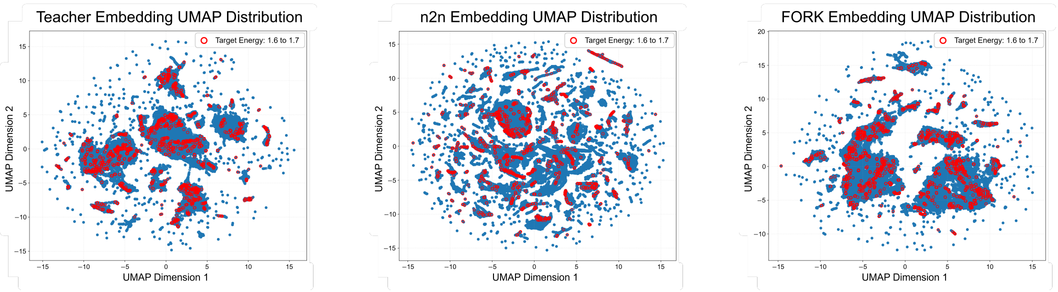
Table 2: Detailed per-adsorbate MAE analysis for FORK Student and GemNet-OC Teacher models across different chemical species and datasets.

Dataset	**Surface Type**	**Adsorbate**	**FORK Student**	**GemNet-OC Teacher**	**ΔMAE**	**FORK Better**
			MAE (eV)	**MAE (eV)**	**(eV)**	**(%)**
Alonso et al.	Binary intermetallics	H	0.214	0.146	0.068	32.8
		O	0.995	0.898	0.097	38.3
		OH	0.491	0.358	0.133	32.3
		Dataset Avg.	*0.565*	*0.467*	*0.098*	*34.5*
Saini et al.	Transition metal alloys	O	0.356	0.243	0.113	39.5
		N	0.333	0.202	0.131	30.1
		CH	0.447	0.335	0.112	33.8
		Dataset Avg.	*0.378*	*0.260*	*0.118*	*34.4*
Li et al.	Binary alloys (L_{10}/L_{12})	C	0.707	0.527	0.180	36.7
		CO	0.582	0.435	0.147	39.9
		Dataset Avg.	*0.645*	*0.481*	*0.164*	*38.3*

ΔMAE = FORK Student MAE - GemNet-OC Teacher MAE (positive values indicate GemNet-OC performs better).
FORK Better (%) indicates percentage of individual reactions where FORK Student outperforms GemNet-OC Teacher.

3.3 Ablation Studies

We conduct a series of ablation studies to validate our key design choices. Further quantitative ablations, including analysis on the importance of relational contrastive learning and the impact of the temperature parameter, are in Appendix G.



(a) Teacher.

(b) n2n.

(c) FORK.

Figure 1: UMAP visualization of final-layer embeddings on the OC20 O* subset, comparing the **(a)** Teacher model, **(b)** a student trained with n2n distillation, and **(c)** a student trained with our FORK framework.

Geometric Fidelity of the Distilled Representation. To qualitatively assess the transfer of the teacher's learned potential energy surface, we projected the final-layer atom embeddings onto a 2D manifold using UMAP [McInnes et al., 2018], as shown in Figure 1. The visualization demonstrates that the FORK model's embedding space more closely replicates the geometric structure and cluster separation of the teacher's space than the disordered embeddings from a student trained with node-to-node (n2n) distillation. This visual evidence corroborates our quantitative results, highlighting FORK's ability to transfer the essential geometric features of the teacher's learned representation.

4 Conclusion

We present FORK, a **F**irst-**O**rder **R**elational **K**nowledge Distillation framework that transfers the physical understanding of a large-scale teacher model to a compact Machine Learning Interatomic Potential (MLIP). By preserving the geometric relationships between atoms rather than matching isolated features, FORK enables a student model to faithfully capture the teacher's learned potential energy surface. This relational approach outperforms conventional distillation methods in terms of energy and force prediction accuracy on challenging benchmarks like OC20 and SPICE, and achieves a 11.9× acceleration in a practical catalyst screening application while preserving chemical coherency. Future work can extend this framework by incorporating higher-order physical interactions, paving the way for efficient large-scale molecular simulations.

Acknowledgments and Disclosure of Funding

This paper was supported by Korea Institute for Advancement of Technology(KIAT) grant funded by the Korea Government(Ministry of Education)(P0025681-G02P22450002201-10054408, Semiconductor-Specialized University)

References

Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical review letters*, 98(14):146401, 2007.

Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in neural information processing systems*, 30, 2017.

Oliver T Unke, Stefan Chmiela, Huziel E Sauceda, Michael Gastegger, Igor Poltavsky, Kristof T Schutt, Alexandre Tkatchenko, and Klaus-Robert Müller. Machine learning force fields. *Chemical Reviews*, 121(16):10142–10186, 2021.

Yi-Lun Liao, Brandon Wood, Abhishek Das, and Tess Smidt. Equiformerv2: Improved equivariant transformer for scaling to higher-degree representations. *arXiv preprint arXiv:2306.12059*, 2023.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

Filip Ekström Kelvinius, Dimitar Georgiev, Artur Toshev, and Johannes Gasteiger. Accelerating molecular graph neural networks via knowledge distillation. *Advances in Neural Information Processing Systems*, 36:25761–25792, 2023.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, et al. Open catalyst 2020 (oc20) dataset and community challenges. *Acs Catalysis*, 11(10):6059–6072, 2021.

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, et al. Spice, a dataset of drug-like molecules and peptides for training machine learning potentials. *Scientific Data*, 10(1):11, 2023.

Jehad Abed, Jiheon Kim, Muhammed Shuaibi, Brook Wander, Boris Duijf, Suhas Mahesh, Hyeonseok Lee, Vahe Gharakhanyan, Sjoerd Hoogland, Erdem Irtem, et al. Open catalyst experiments 2024 (ocx24): Bridging experiments and computational models. *arXiv preprint arXiv:2411.11783*, 2024.

Kirby Broderick, Eric Lopato, Brook Wander, Stefan Bernhard, John Kitchin, and Zachary Ulissi. Identifying limitations in screening high-throughput photocatalytic bimetallic nanoparticles with machine-learned hydrogen adsorptions. *Applied Catalysis B: Environmental*, 320:121959, 2023.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Dávid Péter Kovács, J Harry Moore, Nicholas J Browning, Ilyes Batatia, Joshua T Horton, Yixuan Pu, Venkat Kapil, William C Witt, Ioan-Bogdan Magdau, Daniel J Cole, et al. Mace-off: Short-range transferable machine learning force fields for organic molecules. *Journal of the American Chemical Society*, 147(21):17598–17611, 2025.

Johannes Gasteiger, Florian Becker, and Stephan Günnemann. Gemnet: Universal directional graph neural networks for molecules. *Advances in Neural Information Processing Systems*, 34: 6790–6802, 2021.

Ishan Amin, Sanjeev Raja, and Aditi Krishnapriyan. Towards fast, specialized machine learning force fields: Distilling foundation models via energy hessians. *arXiv preprint arXiv:2501.09009*, 2025.

Johannes Gasteiger, Muhammed Shuaibi, Anuroop Sriram, Stephan Günnemann, Zachary Ulissi, C Lawrence Zitnick, and Abhishek Das. Gemnet-oc: developing graph neural networks for large and diverse molecular simulation datasets. *arXiv preprint arXiv:2204.02782*, 2022.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3967–3976, 2019.

Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. Deep graph contrastive representation learning. *arXiv preprint arXiv:2006.04131*, 2020.

Petar Veličković, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. Deep graph infomax. *arXiv preprint arXiv:1809.10341*, 2018.

Kirsten T Winther, Max J Hoffmann, Jacob R Boes, Osman Mamun, Michal Bajdich, and Thomas Bligaard. Catalysis-hub. org, an open electronic structure database for surface reactions. *Scientific data*, 6(1):75, 2019.

Carmen Martínez-Alonso, Valentin Vassilev-Galindo, Benjamin M Comer, Frank Abild-Pedersen, Kirsten T Winther, and Javier LLorca. Application of machine learning to discover new intermetallic catalysts for the hydrogen evolution and the oxygen reduction reactions. *Catalysis Science & Technology*, 14(13):3784–3799, 2024.

Jiang Li, Joakim Halldin Stenlid, Michael T Tang, Hong-Jie Peng, and Frank Abild-Pedersen. Screening binary alloys for electrochemical co 2 reduction towards multi-carbon products. *Journal of Materials Chemistry A*, 10(30):16171–16181, 2022.

Shikha Saini, Joakim Halldin Stenlid, and Frank Abild-Pedersen. Electronic structure factors and the importance of adsorbate effects in chemisorption on surface alloys. *npj Computational Materials*, 8(1):163, 2022.

Richard Tran, Janice Lan, Muhammed Shuaibi, Brandon M Wood, Siddharth Goyal, Abhishek Das, Javier Heras-Domingo, Adeesh Kolluru, Ammar Rizvi, Nima Shoghi, et al. The open catalyst 2022 (oc22) dataset and challenges for oxide electrocatalysts. *ACS Catalysis*, 13(5):3066–3084, 2023.

A Overview Figure of FORK

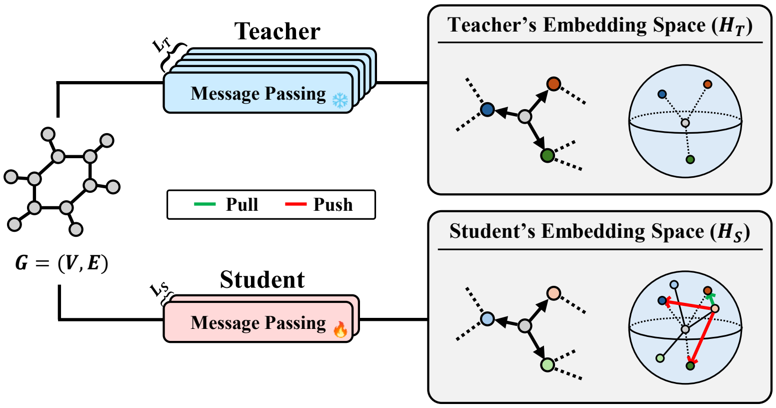


Figure 2: An overview of the FORK, First-Order Relational Knowledge distillation framework. A student model is trained to replicate the geometry of the teacher's embedding space (H_T) using a contrastive objective. For each interatomic interaction, the student's corresponding relational vector is "pulled" toward the teacher's (positive sample) and "pushed" away from others (negative samples).

B Experimental Details

B.1 Experimental Design for Baseline Comparison

Catalytic Systems. We leverage Open Catalyst 2020 (OC20), which is a massive dataset designed to accelerate the discovery of catalysts for renewable energy applications [Chanussot et al., 2021]. We use two standard subsets for our experiments: the 200K subset, which is a diverse chemical dataset of 200,000 datapoints, and the O* subset, which is specifically designed to test the model to 12,182 O* adsorption reactions. The teacher is a 153M-parameter pretrained EquiformerV2 [Liao et al., 2023] with 20 message passing layers, while the student employs the same architecture reduced to just 2 layers (22M parameters). This architectural consistency enables a 'pretrained' baseline, where the student is initialized directly from the teacher's first two layers. We show hyperparamters used in Table 3.

Table 3: Key Hyperparameters for Training on OC20 Dataset.

Hyperparameter	Value
Optimizer	AdamW
Learning Rate Schedule	Cosine Decay w/ Warmup
Peak Learning Rate	5×10^{-4}
Warmup Steps	30,000
Batch Size	4
$\mathcal{L}_{\text{FORK}}$ weight (λ_1)	10.0
L_{KD} weight (λ_2)	1.0
Temperature (τ)	0.15

Small Molecules. We use SPICE dataset, which consists of a diverse set of small, drug-like molecules [Eastman et al., 2023]. Its focus on smaller, conformationally flexible molecules provides a complementary challenge to the rigid surfaces of OC20, allowing us to test FORK's robustness and applicability in a different chemical domain. We use subset of monomers, solvated amino acids, and systems with iodine. We distill MACE-OFF Large model as a teacher [Kovács et al., 2025], and GemNet-dT as a student [Gasteiger et al., 2021]. We use idential hyperparmeters to train models provided in Amin et al. [2025], with $\mathcal{L}_{\text{FORK}} = 100$.

Implementation Details. All models were trained using the AdamW optimizer. The learning rate was warmed up to a peak value of 5×10^{-4} over 30,000 steps and then decayed using a cosine schedule. The batch size was set to 4. The loss balancing hyperparameters were empirically set to $\lambda_1 = 10.0$ for the FORK loss and $\lambda_2 = 1.0$ for the n2n feature matching loss. Based on our ablation studies, the temperature for the contrastive loss was set to $\tau = 0.15$. All experiments were conducted on one NVIDIA A6000 or L40S GPU, separately.

C Related Works

C.1 Machine Learning Interatomic Potentials

The development of Machine Learning Interatomic Potentials (MLIPs) represents a pivotal advancement in computational science, aiming to bridge the gap between the accuracy of quantum mechanical methods like Density Functional Theory (DFT) and the efficiency of classical force fields [Behler and Parrinello, 2007]. Early models such as SchNet introduced deep learning architectures to predict chemical properties directly from atomic coordinates [Schütt et al., 2017]. More recent equivariant Graph Neural Networks (GNNs), including GemNet-OC and EquiformerV2, have set new standards in accuracy by incorporating geometric symmetries directly into the model architecture, enabling highly precise predictions of energy and interatomic forces at a fraction of DFT's computational cost [Gasteiger et al., 2022, Liao et al., 2023]. However, the computational demands of these state-of-the-art models motivate the need for effective model compression techniques like knowledge distillation.

C.2 Knowledge Distillation for Molecular GNNs

Knowledge distillation (KD) has evolved from a general model compression technique into a promising strategy for deploying deep learning models in resource-intensive scientific applications [Hinton et al., 2015, Gou et al., 2021]. For molecular GNNs, KD has shown significant promise in accelerating simulations. A common approach, known as feature-based distillation, involves training a student model to mimic the atom-wise hidden representations of a teacher [Ekström Kelvinius et al., 2023]. This is typically achieved by minimizing a regression loss (e.g., L1 or L2 norm) between the student's and teacher's final node embeddings.

Recently, more physically-informed methods have emerged. For instance, Amin et al. [2025] proposed distilling knowledge by matching the Hessians of the energy predictions. This second-order approach allows the student to learn the curvature of the Potential Energy Surface (PES), which is vital for modeling vibrational properties. While powerful, these higher-order methods can be computationally intensive and may not be universally applicable. Our work, FORK, complements these approaches by focusing on efficiently capturing fundamental first-order relational information.

C.3 Relational Knowledge Distillation

Relational Knowledge Distillation (RKD), pioneered by Park et al. [2019], shifted the distillation paradigm from matching individual data points to preserving the relationships between them. Traditional RKD computes pairwise relations either between different samples in a batch or between features within a single instance. While innovative, applying standard RKD to molecular systems reveals critical limitations. Treating atoms across a batch as samples leads to physically meaningless comparisons between non-interacting atoms in different molecules. Furthermore, instance-level RKD, which compares all atom pairs within a molecule, suffers from quadratic complexity and incorrectly treats all pairs equally, ignoring the fundamental distinction between bonded and non-bonded interactions. These limitations highlight the need for a domain-specific RKD that respects the physical hierarchies of molecular structures.

C.4 Contrastive Learning on Graphs

Contrastive learning has become a dominant paradigm for self-supervised representation learning on graphs [Oord et al., 2018]. Methods like GRACE [Zhu et al., 2020] and Deep Graph Infomax [Veličković et al., 2018] learn powerful node embeddings by maximizing the similarity between different augmented views of the same node or graph (positive pairs) while minimizing similarity with other nodes or graphs (negative pairs).

C.5 Preliminaries

We represent a molecular system as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of N atoms (nodes) and $\mathcal{E}$ is the set of E interatomic interactions (edges) within a given cutoff radius. Each atom $i \in \mathcal{V}$ is described by an initial feature vector $\mathbf{h}_i^{(0)}$, typically encoding its atomic number. The primary task of an MLIP is to learn a mapping from the atomic positions $\mathbf{x}$ to the total potential energy $U(\mathbf{x})$ and the per-atom forces $\mathbf{F}_i = -\nabla_{\mathbf{x}_i} U(\mathbf{x})$.

Our knowledge distillation framework comprises the teacher and student model. The teacher Model, denoted as f_T with parameters θ_T, is a pretrained, high-capacity equivariant GNN. For a given graph $\mathcal{G}$, it computes a set of final node embeddings $\mathbf{Z}_T = \{\mathbf{z}_{T,1}, \ldots, \mathbf{z}_{T,N}\}$, where each $\mathbf{z}_{T,i} \in \mathbb{R}^{D_T}$. The student model, f_S with parameters θ_S, is a more compact GNN architecture. It produces lower-dimensional embeddings $\mathbf{Z}'_S \in \mathbb{R}^{N \times D_S}$. To facilitate knowledge transfer, a Multi-Layer Perceptron (MLP) projection head $P : \mathbb{R}^{D_S} \to \mathbb{R}^{D_T}$ maps the student's embeddings into the teacher's latent space, yielding the projected embeddings $\mathbf{Z}_S = P(\mathbf{Z}'_S)$.

D Training Algorithm

This section provides the detailed pseudocode for the end-to-end training procedure of FORK, as described in the main paper.

Algorithm 1 First-Order Relational Knowledge Distillation (FORK) Training

```
Input: Training data loader $\mathcal{D}$, pretrained teacher model $f_T$, student model $f_S$, projection head $P$.
Input: Hyperparameters: learning rate $\eta$, loss weights $\lambda_1, \lambda_2$, temperature $\tau$.
Initialize parameters $\theta_S$ of $f_S$ and $\theta_P$ of $P$.
Freeze parameters of the teacher model $f_T$.
for each training epoch do
    for each batch of molecular graphs $\{\mathcal{G}\}$ in $\mathcal{D}$ do
        // Generate embeddings from teacher and student models
        With no gradient tracking for $f_T$:
        $\mathbf{Z}_T \leftarrow f_T(\{\mathcal{G}\})$ {Teacher atom embeddings, size $N_{batch} \times D_T$}
        $\mathbf{Z}'_S \leftarrow f_S(\{\mathcal{G}\})$ {Student atom embeddings, size $N_{batch} \times D_S$}
        $\mathbf{Z}_S \leftarrow P(\mathbf{Z}'_S)$ {Projected student embeddings, size $N_{batch} \times D_T$}
        // Compute standard task loss (Energy and Forces)
        $U_S, \mathbf{F}_S \leftarrow$ Predictions from $f_S$
        $\mathcal{L}_{\text{task}} \leftarrow \text{Loss}((U_S, \mathbf{F}_S), (U_{true}, \mathbf{F}_{true}))$
        // Compute optional node-to-node KD loss
        $L_{\text{KD}} \leftarrow \frac{1}{N_{batch}} \sum_{i=1}^{N_{batch}} ||\mathbf{z}_{S,i} - \mathbf{z}_{T,i}||_2^2$
        // Construct Relational Vectors for all E edges in the batch
        For each edge $e_k = (\text{src}, \text{dst})$:
        Normalize atom embeddings: $\hat{\mathbf{z}} = \mathbf{z}/||\mathbf{z}||_2$
        Compute normalized difference vectors: $\mathbf{r}_k = \hat{\mathbf{z}}_{\text{src}} - \hat{\mathbf{z}}_{\text{dst}}$
        This yields teacher set $\{\mathbf{r}_{T,k}\}$ and student set $\{\mathbf{r}_{S,k}\}$.
        // Compute FORK contrastive loss
        $\mathcal{L}_{\text{FORK}} \leftarrow -\frac{1}{E_{batch}} \sum_{k=1}^{E_{batch}} \log \frac{\exp(\mathbf{r}_{S,k} \cdot \mathbf{r}_{T,k}/\tau)}{\sum_{m=1}^{E_{batch}} \exp(\mathbf{r}_{S,k} \cdot \mathbf{r}_{T,m}/\tau)}$
        // Compute total loss and update student model
        $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{task}} + \lambda_1 \mathcal{L}_{\text{FORK}} + \lambda_2 L_{\text{KD}}$
        Update parameters $(\theta_S, \theta_P)$ using gradient descent on $\mathcal{L}_{\text{total}}$.
    end for
end for
Output: Trained student model $f_S$. The projection head $P$ is discarded after training.
```

E Additional Results on Benchmarks

E.1 Performance on the OC20 200K Subset

To further validate our method, we evaluated its performance on the large and chemically diverse 200K subset of the OC20 benchmark. The results, summarized in Table 4, reinforce the conclusions drawn from the O* subset.

The student model trained with the combined FORK and node-to-node (n2n) distillation strategy again demonstrated superior performance. This approach decreased the energy MAE to **371.1 meV**, significantly improving upon the 412.8 meV achieved by the n2n baseline. The consistent improvement across both the O* and 200K subsets confirms that our relational distillation approach robustly enhances the accuracy of the student model across different chemical distributions.

Table 4: Performance of FORK on 200K subset of OC20 dataset. The best results are highlighted in **bold**. Second best results are underlined.

Method	**Params**	**Embedding**		**Energy**	**Force**
		MAE	Cosine Similarity	MAE (meV) ↓	MAE (meV/Å) ↓
Teacher*	153M	-	-	171.5	12.4
vanilla	22M	0.309	0.233	474.9	51.8
pretrained	22M	0.181	0.460	410.8	37.6
n2n	22M	0.096	0.816	412.8	34.8
Hessian	22M	0.351	0.180	419.3	48.6
Ours	22M	0.190	0.424	373.8	35.8
Ours (w/ n2n)	22M	0.097	0.811	**371.1**	**34.1**

* The teacher model of EquiformerV2 used for knowledge distillation. Loaded from provided checkpoint.

E.2 Applicability to Small Molecules

Beyond the large-scale catalytic systems of OC20, we further evaluated FORK's applicability to the distinct chemical domain of small, drug-like molecules using the SPICE dataset [Eastman et al., 2023]. The results, detailed in Table 5, highlight FORK's ability to act synergistically with other advanced distillation methods.

On this benchmark, second-order methods that distill Hessian information prove to be particularly effective, consistently outperforming other student models. This underscores the importance of capturing PES curvature for small, flexible molecules. However, our key finding is that combining FORK with Hessian distillation yields the most robust and accurate student. While this combined model performs on par with the Hessian baseline on simpler subsets, it achieves state-of-the-art performance on the most challenging "Systems with Iodine" subset, surpassing all other methods in both energy and force prediction. This suggests that while Hessian distillation effectively transfers second-order knowledge, supplementing it with FORK's first-order relational knowledge provides a critical performance advantage in more complex chemical environments.

Table 5: Performance of FORK on SPICE dataset. The best results are highlighted in **bold**. Second best results are underlined.

Subset	Method	Params	Energy	Force
			MAE (meV/atom) ↓	MAE (meV/Å) ↓
Monomers	Teacher*	4.7M	0.65	6.6
	vanilla	0.67M	2.2	13.4
	n2n	0.67M	2.3	14.5
	Hessian	0.67M	**1.2**	**8.5**
	Ours (w/ n2n)	0.67M	2.1	13.8
	Ours (w/ Hessian)	0.67M	1.4	8.9
Solvated Amino Acids	Teacher*	4.7M	1.3	19.4
	vanilla	0.67M	1.7	22.9
	n2n	0.67M	1.5	21.4
	Hessian	0.67M	**0.4**	**11.4**
	Ours (w/ n2n)	0.67M	1.2	21.3
	Ours (w/ Hessian)	0.67M	**0.4**	11.9
Systems with Iodine	Teacher*	4.7M	1.3	15.3
	vanilla	0.67M	3.2	25.4
	n2n	0.67M	3.0	25.9
	Hessian	0.67M	2.4	19.6
	Ours (w/ n2n)	0.67M	3.2	25.8
	Ours (w/ Hessian)	0.67M	**2.2**	**19.4**

* The teacher model of MACE-OFF Large used for knowledge distillation. Loaded from provided checkpoint.

F Details on Catalyst Screening Experiments

This section provides detailed methodology for the high-throughput catalyst screening experiments described in the main paper.

F.1 Dataset Preparation and Processing

Data Sources. We obtained three benchmark datasets from the CatalysisHub database [Winther et al., 2019], each representing different catalyst screening applications. The Alonso dataset [Martínez-Alonso et al., 2024] contains 2,628 DFT-calculated adsorption energies for H*, O*, and OH* on binary intermetallic compounds with AB, A_2B, and A_3B stoichiometries, where surface structures were prepared using the lowest energy facets with biaxial strain considerations. The Li dataset [Li et al., 2022] comprises 337 binding energy calculations for C* and CO* adsorbates on binary alloys with L_{10} and L_{12} crystal structures, specifically curated for CO_2 reduction reaction screening. The Saini dataset [Saini et al., 2022] includes 441 chemisorption energy calculations for O*, N*, and CH* on transition metal alloy surfaces with systematic variations in d-band properties.

Data Preprocessing. Each dataset underwent standardized preprocessing to ensure consistency across evaluations. We first extracted atomic structures from CatalysisHub JSON format and converted them to ASE Atoms objects with proper periodic boundary conditions. All structures were verified to meet convergence criteria with force tolerance below 0.05 eV/Å. Duplicate structures were identified and removed based on chemical formula and energy values. Finally, adsorbate positions were standardized relative to the surface normal to maintain consistent geometric representations across different surface types.

Distillation Setup. We distilled knowledge from a 39M-parameter GemNet-OC [Gasteiger et al., 2022], trained on the OC20 and OC22 datasets [Chanussot et al., 2021, Tran et al., 2023]. The student was a compact, 1M-parameter GemNet-OC architecture, trained using the FORK framework. The distillation process was conducted on a 2M subset of the OC20 dataset, chosen to cover a wide range of adsorbate-catalyst systems.

F.2 Reaction Energy Calculation Methodology

Energy Computation Protocol. For each reaction in the datasets, we computed reaction energies by evaluating individual species energies and applying stoichiometric coefficients. Gas-phase molecules were identified based on system size (fewer than 10 atoms) or unit cell volume (exceeding 8000 Å^3). For gas molecules, we utilized GemNet-OC energies exclusively for both models to maintain consistency, as these species typically require different computational treatment than surface-bound structures. For slab and adslab configurations, both FORK and GemNet-OC models computed energies independently. Reaction energies were then calculated as $\Delta E = \sum_{\text{products}} \nu_i E_i - \sum_{\text{reactants}} \nu_j E_j$, where ν represents stoichiometric coefficients and E represents species energies.

Evaluation Metrics. We assessed model performance using multiple complementary metrics to capture different aspects of prediction quality. Mean Absolute Error (MAE) was computed as the primary metric: $\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |E^i_{\text{pred}} - E^i_{\text{DFT}}|$, providing an overall measure of prediction accuracy. Additionally, we tracked per-reaction accuracy through binary indicators showing when FORK predictions were closer to DFT than GemNet-OC predictions. Spearman's rank correlation coefficient ρ evaluated whether models preserved the correct energy ordering crucial for catalyst screening.

F.3 Computational Efficiency and Performance Trade-off

A primary motivation for knowledge distillation is to accelerate inference without catastrophically compromising accuracy. As summarized in Table 6, the FORK-distilled student model achieves a transformative leap in computational performance. The student model is **11.9× faster** than its teacher, reducing the average inference time per batch from 166.7 ms to just **14.0 ms** on identical hardware. While this acceleration comes with a trade-off in accuracy, the student's overall Mean Absolute Error (MAE) is higher than the teacher's, its performance remains well within a reasonable range for high-throughput screening, where the primary goal is to rank and identify promising candidates from vast chemical spaces.

Table 6: High-throughput catalyst screening performance comparison between GemNet-OC teacher and FORK-distilled student models.

Model	**Params**	**Energy**	**Force**	**Inference Time (ms/batch)**	**Speedup**
		MAE (meV/adsorption) ↓	MAE (meV/Å) ↓		
Teacher	38.9M	107.8	18.4	166.7	1.0×
Student	1.2M	436.1	50.3	14.0	**11.9×**

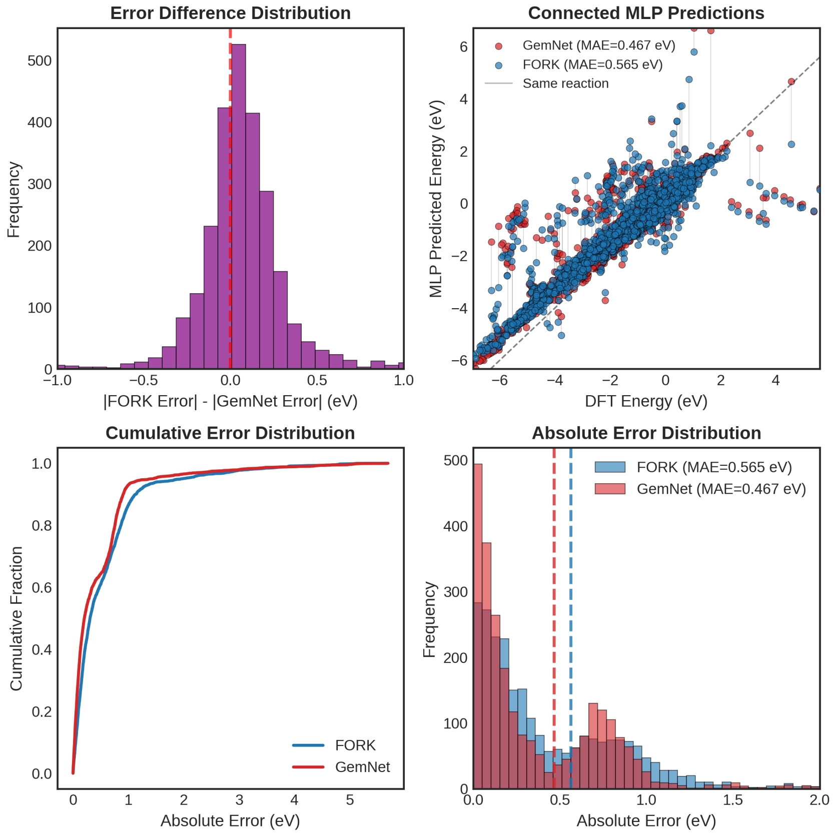


Figure 3: Performance comparison on the Alonso et al. binary intermetallics dataset (2,628 reactions). Error difference distribution showing FORK outperforms GemNet-OC in 34.5% of reactions despite higher average MAE. Connected parity plot visualizing individual reaction predictions, with gray lines connecting FORK and GemNet-OC predictions for the same reaction. Cumulative error distribution demonstrating the models' relative error profiles across all reactions. Absolute error distribution with MAE values of 0.565 eV (FORK) and 0.467 eV (GemNet-OC).

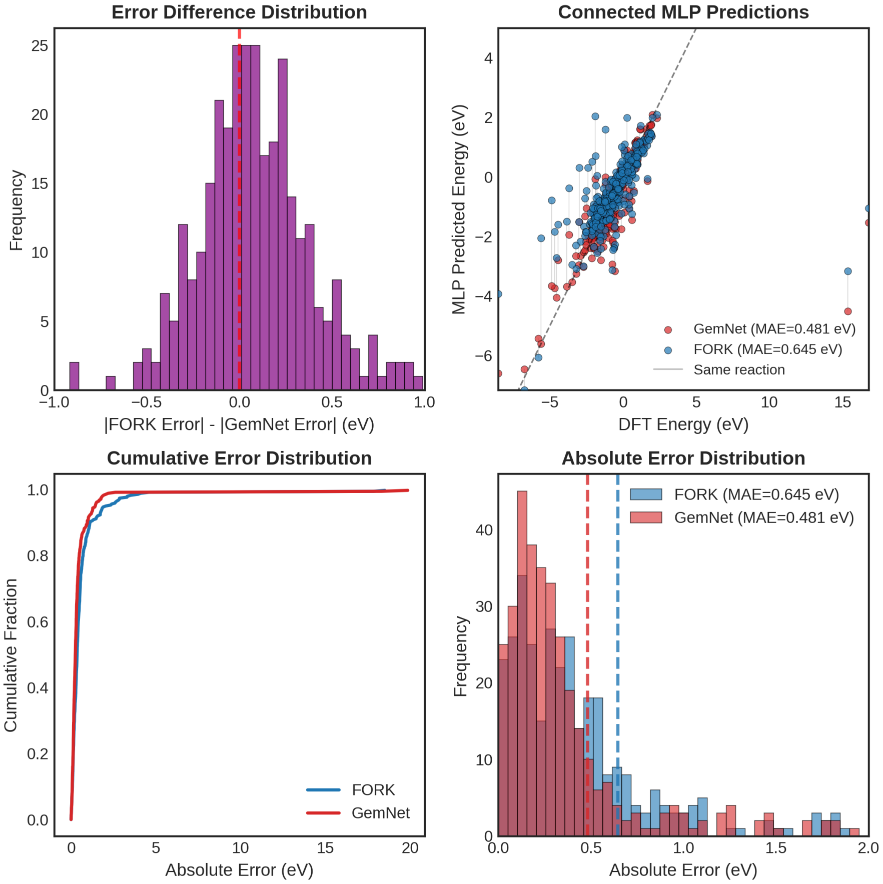


Figure 4: Performance comparison on the Li et al. binary alloys dataset (337 reactions) for CO_2 reduction catalysts. Error difference distribution showing FORK outperforms GemNet-OC in 38.3% of reactions. Connected parity plot for C* and CO* binding energies. Cumulative error distribution. Absolute error distribution with MAE values of 0.645 eV (FORK) and 0.481 eV (GemNet-OC).

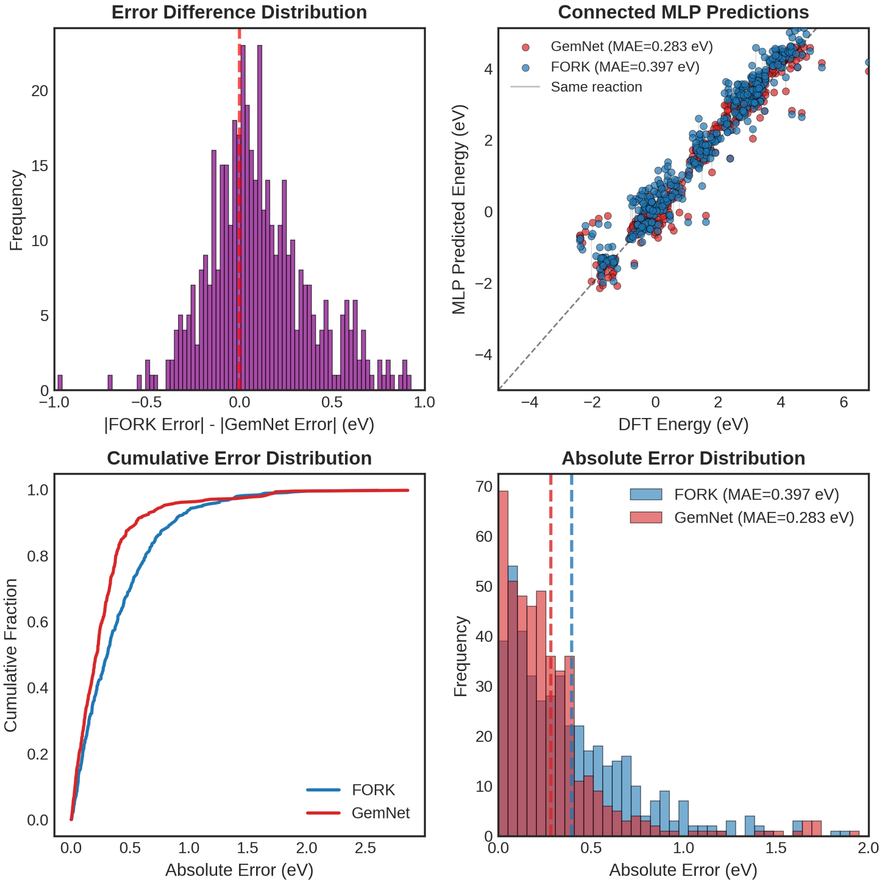


Figure 5: Performance comparison on the Saini et al. transition metal alloys dataset (441 reactions). Error difference distribution showing FORK outperforms GemNet-OC in 34.4% of reactions. Connected parity plot for O*, N*, and CH* chemisorption energies. Cumulative error distribution. Absolute error distribution with MAE values of 0.378 eV (FORK) and 0.260 eV (GemNet-OC).

G Additional Ablation Studies

G.1 Importance of Relational Contrastive Learning

The performance of instance-level contrastive loss applied directly to atom embeddings $(\mathbf{z}_{S,i}, \mathbf{z}_{T,i})$ was compared to FORK's relational approach. Table 7 shows the results of the experiment conducted on the O* subset of OC20 dataset. The consistent superiority of the relational method, especially in energy MAE, validates our central hypothesis: distilling interactions is more effective than distilling isolated atom features.

Table 7: Performance of instance-level contrastive loss compared to relational-level contrastive loss. The best results are highlighted in **bold**. Second best results are underlined.

Method	**Params**	**Embedding**		**Energy**	**Force**
		MAE	Cosine Similarity	MAE (meV) ↓	MAE (meV/Å) ↓
instance-level	22M	0.258	0.210	241.5	6.1
instance-level, w/ n2n	22M	0.081	0.828	235.7	**5.8**
relational-level	22M	0.282	0.230	234.1	6.1
relational-level, w/ n2n	22M	0.082	0.820	**232.0**	**5.8**

G.2 Impact of Temperature in Contrastive Loss

The temperature τ controls the difficulty of the contrastive task. A low τ increases discrimination but risks instability, while a high τ may wash out important details. An optimal $\tau = 0.15$ was found empirically to balance these trade-offs.

Table 8: Performance of FORK with different temperature τ. The top-2 best results are highlighted in **bold**.

τ	**Params**	**Embedding**		**Energy**	**Force**
		MAE	Cosine Similarity	MAE (meV) ↓	MAE (meV/Å) ↓
0.05	22M	0.085	0.806	234.0	5.8
0.07	22M	0.084	0.810	232.9	5.8
0.1	22M	0.083	0.814	**232.0**	5.8
0.15	22M	**0.082**	0.820	**232.0**	5.8
0.2	22M	**0.082**	**0.823**	232.9	5.8