CRACK: Contrastive Relational-Aware Compression of Knowledge for Machine Learning Force Fields

Hyukjun Lim, Seokhyun Choung, Jeong Woo Han†

Department of Materials Science and Engineering, College of Engineering, Seoul National University

Computational Catalysis and Emerging Materials Laboratory

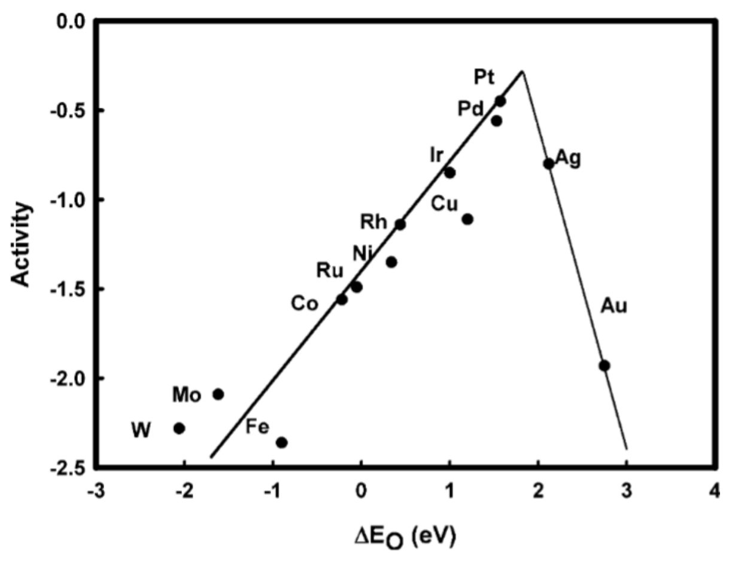
Introduction

Application Domain:

- Accelerating the **discovery of novel catalyst materials** for the **oxygen reduction reaction (ORR)** which is critical in fuel cells.

Governing Problem:

- Developing a **high-performance, low-cost catalyst** for the ORR is a critical challenge to overcome the limitations of platinum (Pt).
  (Nørskov, J. K., et al. JPCB 108.46 (2004): 17886-17892.)



Barrier to Discovery:

- Screening **vast materials search space** to find optimal catalyst.
- Density Functional Theory (DFT) are accurate, but have **infeasible computational cost** for large-scale screening.
- Fundamental **trade-off**: simulation accuracy & scope of the search.
- Research on Machine Learning Force Fields (MLFFs) to break this trade-off by providing both **high accuracy & speed.**
  (Unke, O. T., et al. Chemical Reviews 121.16 (2021): 10142-10186.)

Motivation

Problem Domain:

- Knowledge distillation for Machine Learning Force Fields (MLFFs) to enable efficient molecular simulations.

Existing Challenges:

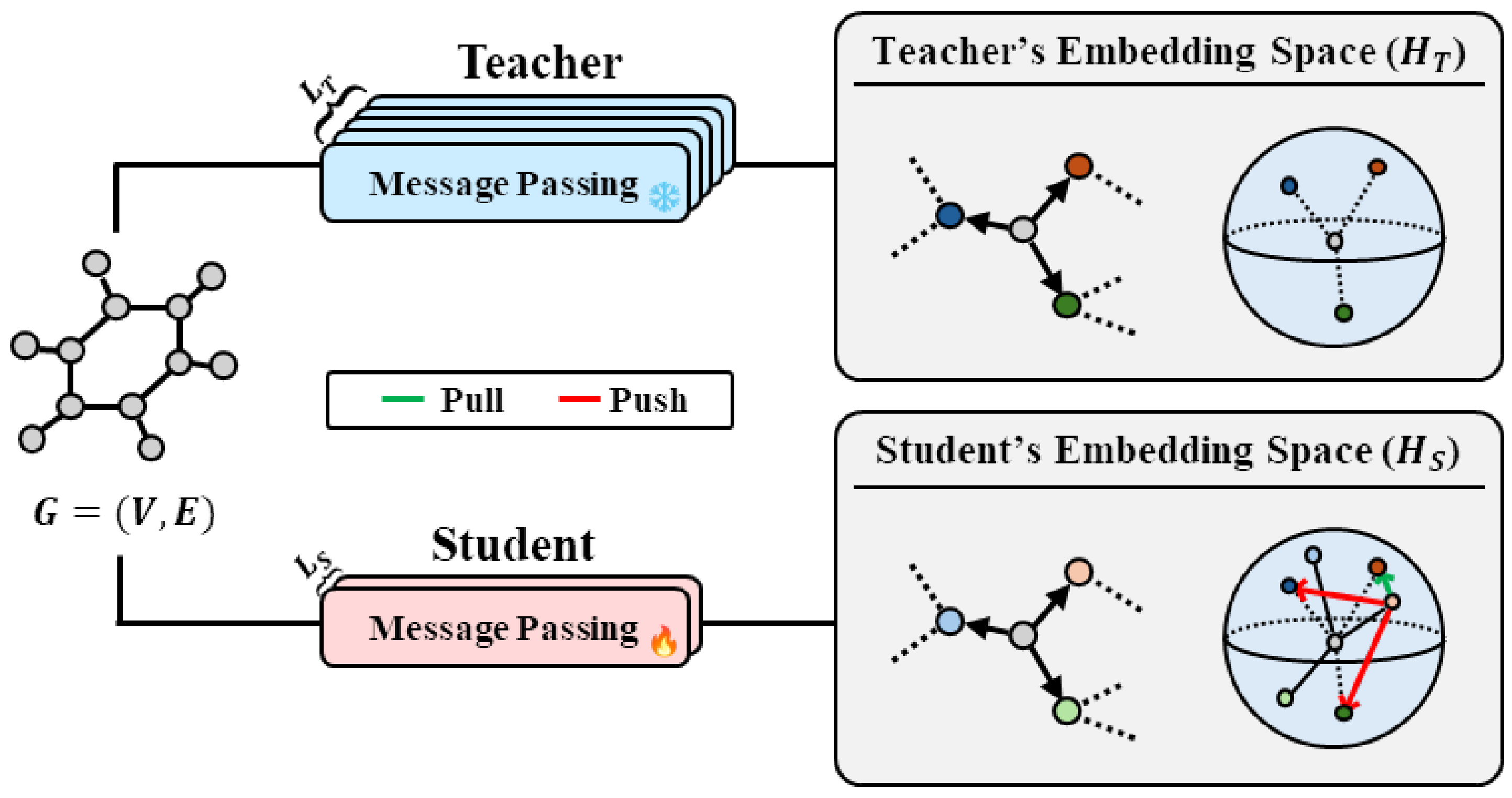
- Trade-off between accuracy and computational efficiency in molecular dynamics simulations.
- State-of-the-art equivariant GNNs (like EquiformerV2) achieve high accuracy but have **substantial computational cost.**
  (Liao, Y., et al. The Twelfth International Conference on Learning Representations (2024))
- This limits adoption in large-scale molecular dynamics, **high-throughput materials screening**, and drug discovery.

Limitation of Previous Approaches:

- Simplistic atom-wise feature matching that treats atoms as independent entities.
- Missing the crucial physics: how atoms **interact with neighbors** to define the potential energy surface (PES).

Methods

CRACK Architecture:



Key Components:

1. **Relational Vectors:**
   - Derived from learned atomic embeddings of bonded atom pairs ($\boldsymbol{z}_{src} - \boldsymbol{z}_{dst}$).
   - Serve as proxies for teacher's learned representation of interatomic potentials.
2. **Contrastive Learning:**
   - InfoNCE loss trains student to generate relational vectors uniquely identifiable with teacher counterparts.
   - Each teacher relational vector forms positive pair with corresponding student vector.
   - All other student vectors in batch serve as negatives.

Advantages:

Physics-Informed

O(E) Scalability

Applicability

Results

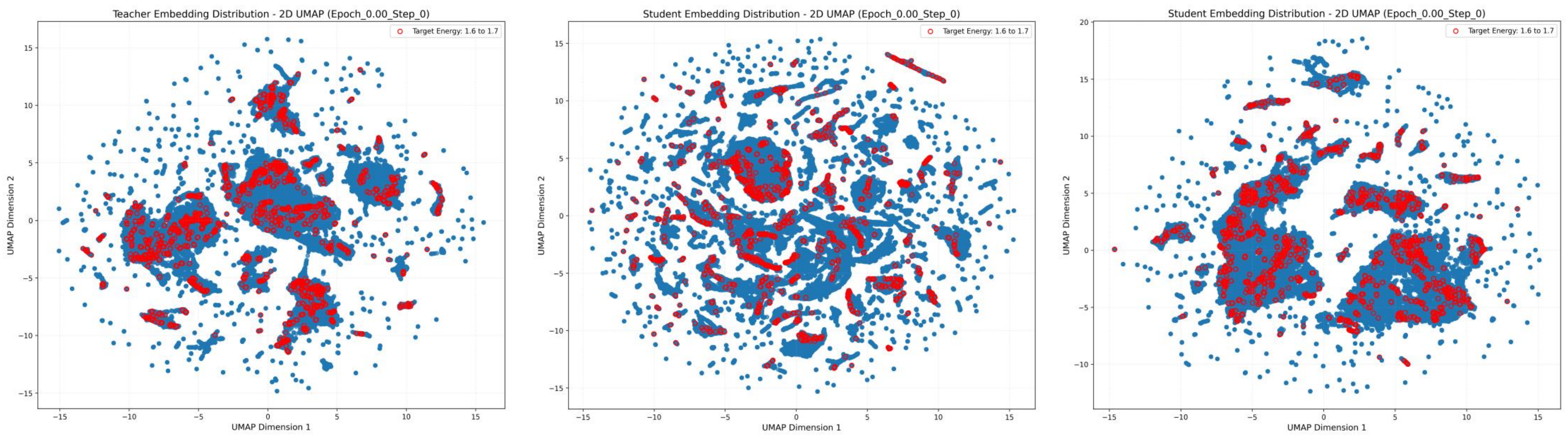
O* Subset of OC20 dataset:

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**	**234.1**	**6.1**
Ours (w/ n2n)	**22M**	**0.082**	**0.820**	**231.7**	**5.8**

200K Subset of OC20 dataset:

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**

Visualization of Embeddings for O* Subset:



Teacher **n2n** **CRACK**

Conclusion & Future Work

Key Contributions:

- First KD framework to directly distill **first-order interatomic relational knowledge** for MLFFs.
- Novel **relational-contrastive loss** that captures geometry of teacher's learned PES.
- **State-of-the-art performance** on OC20, significantly outperforming baselines.

Future Directions:

- More **sophisticated relational descriptors** incorporating higher-order structural features.
- Application to **high-throughput materials discovery** by enabling rapid screening of large material databases with compressed yet accurate models.

Contact us: hyukjunlim@snu.ac.kr

Scan for Project Page