Efficient and Effective Spatial Graph Representation Learning via Canonical Location Transformation

Supplementary Material

SI. Performance on Open Catalyst and Comparison with Physically Informed Deep Neural Networks

Open Catalyst project. The Open Catalyst Project [ZCD+20] is a joint initiative between Meta's Fundamental AI Research (FAIR) team and Carnegie Mellon University's Department of Chemical Engineering. Its goal is to leverage AI for modeling and discovering novel catalysts to advance renewable energy storage and combat climate change.

A key challenge in catalyst research is to identify low-cost materials that can drive reactions at high rates. Quantum mechanical simulations, such as Desnity Functional Theory (DFT), allow for testing and evaluating new catalyst structures. The Open Catalyst Project has released the Open Catalyst 2020 (OC20) [CDG+21] and Open Catalyst 2022 (OC22) [TLS+23] datasets for training machine learning models. Together, these datasets include 1.3 million molecular relaxations and results from over 260 million DFT calculations.

Structure to Total Energy and Forces. We included in our experiments the Structure to Total Energy and Forces (S2EF) task of the Open Catalyst 2022 (OC22) dataset. This task involves predicting the total energy and per-atom forces based on atomic positions as computed by DFT. In this context, energy refers to adsorption energy, which is defined as the difference between the total energy of the combined surface-adsorbate system and the sum total of the relaxed slab and gas-phase adsorbate molecule energies. Forces are determined as the negative gradient of the energy with respect to atomic positions. The S2EF task serves as a general benchmark for catalyst research due to its large-scale dataset and inclusion of inorganic and organic materials. The OC22 dataset specifically consists of $62,331$ DFT relaxations spanning various oxide materials, coverages, and adsorbates. It provides precomputed LMDB files for training, validation, and testing, which incudes input structures from relaxation trajectories along with their energy and atomic forces. The validation and test sets are further divided into in-distribution (ID) and out-of-distribution (OOD) splits based on material similarity to the training set.

While this task has been explored in the context of Physically Informed Neural Networks (PINNs), it has not been a primary benchmark for the other methods we compare against. Therefore, we include it in the appendix as an additional evaluation to assess the generalization of our approach beyond the main experiments.

1) Experimental Results on Open Catalyst.: Testing Metrics To assess model performance and practical applicability, we use evaluation metrics based on DFT calculations. The four key metrics for this task include:

- Energy MAE (lower better): Measures the Mean Absolute Error (MAE) between the computed and ground truth energy values.
- Force MAE (lower better): MAE between the predicted and ground truth per-atom forces, computed only for free catalyst and adsorbate atoms.
- Force Cosine Similarity (higher better: Measures the alignment between predicted and ground truth force vectors by calculating the mean cosine of their angles, considering only free atoms.
- Energy and Forces within Threshold (EFwT) (higher better): A practical benchmark that evaluates whether both energy and forces are close to the ground truth. A prediction is considered "correct" if the energy error is within 0.2 *eV* and the maximum per-atom force error is below 0.3 *eV/Å*. EFwT represents the percentage of structures meeting this criteria.

TABLE SI
Results for the S2EF task of the OC22 dataset for our SGRLvCLT coupled with SIR-GN, compared to state-of-the-art physically informed neural networks. Values for SchNet, DimeNet++, and GemNet, are literature produced [TLS+23]. The best performing values are highlighted in bold.

S2EF	Energy MAE ⇓		Force MAE ⇓		Force Cosine ⇑		EFwT ⇑	
	ID	OOD	ID	OOD	ID	OOD	ID	OOD
SchNet	7.92	7.93	0.06	.082	.363	.22	0%	0%
DimeNet++	2.08	2.48	.043	.059	.61	.44	0%	0%
GemNet	.37	**.83**	.03	**.04**	.69	.55	**.02%**	0%
SGRLvCLT-SIRGN	**.37**	.86	**.03**	.04	**.7**	**.57**	.01%	0%

Comparison to the state of the art – PINN. In our experiment, we compare our approach to three state-of-the-art physically informed neural networks (PINNs) [KGMG20], [GSS+22], [SKF+17], which incorporate specialized loss functions to enforce specific physical constraints. Performance results for these models were obtained from literature values [TLS+23] to provide a direct comparison.

As shown in Table SI, our method, when combined with the SIR-GN model, achieves performance comparable to or exceeding that of specialized PINN-based approaches such as GemNet. Notably, our approach remains general and independent of any explicit physical constraints, demonstrating its effectiveness without requiring domain-specific modifications.

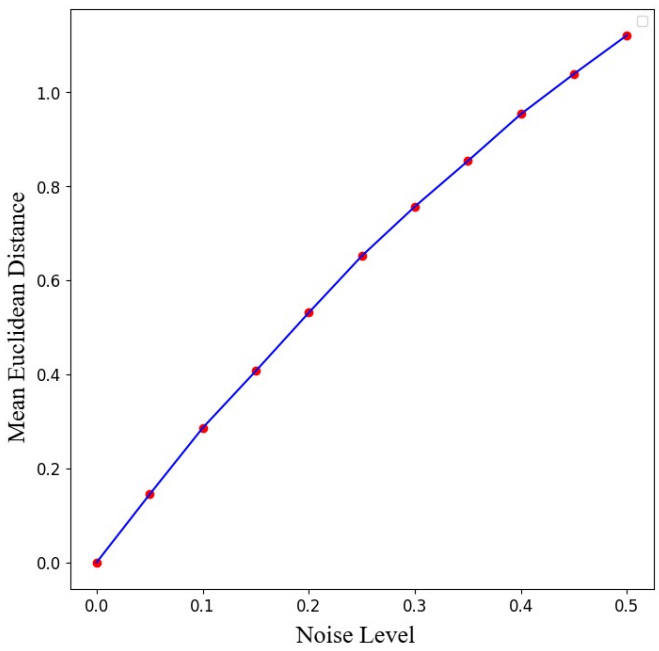

Fig. SI. Mean Euclidean distance between the canonical locations of a perturbed graph, and its unperturbed counterpart. The perturbation applies random rotations and translations, with random noise in the range $\gamma \in (0, 0.5)$ with increments of 0.05.

SII. Robustness to noise of our rotation-translation invariant transformation

To evaluate the robustness of the proposed SGRLvCLT method and confirm its invariance to rotation and translation, Figure SI presents the mean Euclidean distance between the canonical locations of a perturbed graph, $\tilde{G}$, and its unperturbed counterpart, G. The perturbation applies random rotations and translations, with random noise on top, which ranges from $\gamma \in (0, 0.5)$ with increments of 0.05. When $\gamma = 0$, the mean distance is zero, verifying invariance to rotation and translation. As γ increases, the distance increases and follows a generally linear trend. This confirms that our approach successfully preserves spatial relationships under these transformations, validating its effectiveness in ensuring geometric invariance.

References

[CDG+21] 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. Open catalyst 2020 (oc20) dataset and community challenges. *ACS Catalysis*, 11(10):6059–6072, 2021.

[GSS+22] Johannes Gasteiger, Muhammed Shuaibi, Anuroop Sriram, Stephan Günnemann, Zachary W. Ulissi, C. Lawrence Zitnick, and Abhishek Das. GemNet-OC: Developing graph neural networks for large and diverse molecular simulation datasets. *TMLR*, 2022.

[KGMG20] Johannes Klicpera, Shankari Giri, Johannes T. Margraf, and Stephan Günnemann. Fast and uncertainty-aware directional message passing for non-equilibrium molecules. *CoRR*, abs/2011.14115, 2020.

[SKF+17] 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. In *Proc. of NeurIPS Conf.*, pages 991–1001, 2017.

[TLS+23] Richard Tran, Janice Lan, Muhammed Shuaibi, Brandon Wood*, Siddharth Goyal, Abhishek Das, Javier Heras-Domingo, Adeesh Kolluru, Ammar Rizvi, Nima Shoghi, Anuroop Sriram, Zachary Ulissi, and C. Lawrence Zitnick. The open catalyst 2022 (oc22) dataset and challenges for oxide electrocatalysts. *ACS Catalysis*, 2023.

[ZCD+20] C. Lawrence Zitnick, Lowik Chanussot, Abhishek Das, Siddharth Goyal, Javier Heras-Domingo, Caleb Ho, Weihua Hu, Thibaut Lavril, Aini Palizhati, Morgane Riviere, Muhammed Shuaibi, Anuroop Sriram, Kevin Tran, Brandon M. Wood, Junwoong Yoon, Devi Parikh, and Zachary W. Ulissi. An introduction to electrocatalyst design using machine learning for renewable energy storage. *CoRR*, abs/2010.09435, 2020.