Platonic Representation of Foundation Machine Learning Interatomic Potentials

Zhenzhu Li[1,2], Aron Walsh[1]
[1]Department of Materials, Imperial College London, London SW7 2AZ, UK
[2]Imperial Global Singapore, CREATE Tower, 138602, Singapore

zhenzhu.li@imperial.ac.uk

Abstract

Foundation machine learning interatomic potentials (MLIPs) are trained on overlapping chemical spaces, yet their latent representations remain model-specific. Here, we show that independently developed MLIPs exhibit statistically consistent geometric organisation of atomic environments, which we term the Platonic representation. By projecting embeddings relative to a set of atomic anchors, we unify the latent spaces of seven MLIPs (spanning equivariant, non-equivariant, conservative, and non-conservative architectures) into a common metric space that preserves chemical periodicity and structural invariants. This unified framework enables direct cross-model optimal transport, interpretable embedding arithmetic, and the detection of representational biases. Furthermore, we demonstrate that geometric distortions in this space can indicate physical prediction failures, including symmetry breaking and incorrect phonon dispersions. Our results show that the latent spaces of diverse MLIPs present consistent statistical geometry shaped by shared physical and chemical constraints, suggesting that the Platonic representation offers a practical route toward interoperable, comparable, and interpretable foundation models for materials science.

References

[1] Huh, M.; Cheung, B.; Wang, T.; Isola, P. Position: The Platonic Representation Hypothesis. Proceedings of the 41st International Conference on Machine Learning. (2024), 20617–20642.
[2] Cubuk, E. D.; Malone, B. D.; Onat, B.; Waterland, A.; Kaxiras, E. Representations in neural network based empirical potentials. The Journal of Chemical Physics 147 (2017), 024104.
[3] Li, Z.; Walsh, A. Platonic Representation of Foundation Machine Learning Interatomic Potentials. (2025) arxiv.org/abs/2512.05349.

Figures

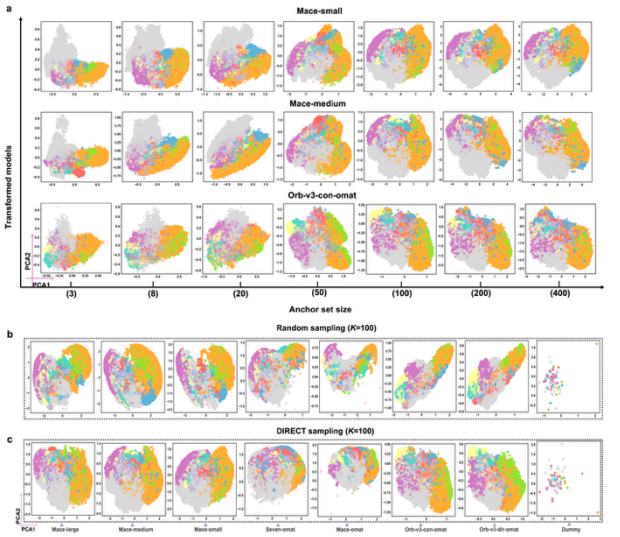

Figure 1. Platonic representations converging with anchor set size and sampling strategy. (a) Transformed representations as a function of anchor set size (K = 3 to 400). (b) 2D PCA projections of converged representations using 100 randomly sampled anchors. (c) Projections using 100 DIRECT-sampled anchors. Despite architectural diversity, all models transformed with DIRECT sampling show substantial alignment. Non-equivariant models (Orb-v3) exhibit systematic skewness. The Dummy model (untrained, random weights) displays no chemical structure, confirming that alignment reflects learned physical knowledge.