

Fine-tuning a universal potential for accelerating surface property prediction.

Materials design and discovery has high potential to transform technology, though traditional trial-and-error approaches limit the turnaround time of novel materials. Computational tools, such as density functional theory (DFT), are becoming increasingly powerful high-throughput methods of predicting hypothetical-bulk material properties. The accurate prediction of surface properties in a high throughput framework is a vastly less explored and equally important task. This proposal therefore seeks to begin addressing a related gap in the computational design of materials framework: the development of universal surface property prediction by integrating machine learning with first-principles thermodynamics.

Predictive synthesis methods have made rapid advances in the space of bulk synthesis, however, methods for predictive nanoscale synthesis continue to lag due to the number of surface energy calculations necessary for making accurate predictions. Computing the bulk phase diagram for a simple ternary chemical space requires ~10-100 DFT calculations, feasible in the span of days to weeks. In comparison, (exhaustively) computing the surface phase diagram of the same ternary space may require upwards of 1,000-10,000 DFT calculations because of the complexities related to surface faceting, reconstruction, and terminating compositions.

Recent developments in machine learning technologies present a unique solution to reduce the computational load and subsequent calculation time. Machine learned interatomic potentials (MLIPs) have been shown to accurately predict bulk material properties with computational speed-ups from days or weeks to seconds, when compared to DFT. Due to the availability of bulk relaxation trajectories, from sources such as Materials Project or Open Quantum Materials Database, and subsequent lack of large-scale surface property databases, MLIPs are often solely trained on bulk data. Having not been exposed to surface complexities, it is expected that such models will perform poorly for predicting surface properties without further tuning.

The specific aim of this proposal is to explore fine-tuning a pre-trained universal interatomic potential, CHGNet,[1] with the expectation of achieving greater predictive accuracy for surface properties. In recent years, the Open Catalyst Project (OCP) has publicized several datasets including first-principles calculations of surface properties, for the sake of accelerating the understanding of catalytic mechanisms. (Add more here related to specifics of the datasets? “OC20” and “OC22”. “OC22” is the actual one we want I believe) Our goal is to re-purpose this data for fine-tuning the prediction of surface energy and related properties.

Several important milestones are:

1. Testing the “zero-shot” performance of CHGNet for surface energy prediction.

“Zero-shot” implies the use of the pre-trained CHGNet model and therefore this task requires no additional training. The purpose of this task is to familiarize the group to the CHGNet code (available open source via GitHub). While becoming familiar with the model’s functionality, we will be appropriately isolating a testing set from the OCP dataset. Assuming the anticipated poor performance, the following milestones will be pursued.

2. Fine-tune the final “X” layers of CHGNet using OCP data and re-test/compare to the “zero-shot” performance.

Fine-tuning does not have a well-defined framework associated with it, therefore the depth (“X”) of layers that we re-train will be treated as a hyperparameter and explored accordingly. Methods for improving this selection, such as AutoFreeze,[2] exist but are beyond the scope of our proof of concept.

3. Explore the effects of emerging data selection approaches.

If time allows, the use of more efficient sampling approaches may be explored and discussed in terms of improved accuracy. One interesting approach would be the inclusion of DIRECT-sampling,[3] which has been shown to improve predictive accuracy of the universal potential M3Gnet,[4] for under-represented portions of the original sample space. The approach uses a combination of principal component analysis and stratified sampling to evenly select training points across a lower dimensional representation of the sample space.

[1] B. Deng *et al.*, “CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling,” *Nat. Mach. Intell.*, vol. 5, no. 9, Art. no. 9, Sep. 2023, doi: 10.1038/s42256-023-00716-3.

[2] Y. Liu, S. Agarwal, and S. Venkataraman, “AutoFreeze: Automatically Freezing Model Blocks to Accelerate Fine-tuning.” arXiv, Apr. 03, 2021. Accessed: Aug. 09, 2023. [Online]. Available: <http://arxiv.org/abs/2102.01386>

[3] J. Qi, T. W. Ko, B. C. Wood, T. A. Pham, and S. P. Ong, “Robust Training of Machine Learning Interatomic Potentials with Dimensionality Reduction and Stratified Sampling.” arXiv, Jul. 24, 2023. Accessed: Jul. 28, 2023. [Online]. Available: <http://arxiv.org/abs/2307.13710>

[4] C. Chen and S. P. Ong, “A universal graph deep learning interatomic potential for the periodic table,” *Nat. Comput. Sci.*, vol. 2, no. 11, pp. 718–728, Nov. 2022, doi: 10.1038/s43588-022-00349-3.