Transferability and Scalability of Growing Computational Database in Sim2Real Materials Informatics

Ryo Yoshida
The Institute of Statistical Mathematics, Tokyo, Japan
Email: yoshidar@ism.ac.jp

ID: INV04
Invited Talk
Aug. 23, 13.30 - 14.00

Keywords: Polymer Database, Molecular Dynamics, Transfer Learning, Small Data.

In the past decade, machine learning has shown the potential to greatly accelerate the of discovery of new materials in various systems. However, the major obstacle in data-driven materials research, the lack of data resources, remains unresolved. High experimental costs and the cultural barrier of researchers to disclose their laboratory data contribute to this issue, making it difficult to solve in the short term. To address data scarcity, the development of open databases using computational methods such as first-principles calculations and molecular dynamics (MD) simulations is underway. However, for polymeric materials, no comprehensive database exists due to technical barriers in automating computer experiments. To generate data resources for machine learning, we developed RadonPy, an open-source software that fully automates polymer property calculations based on all-atom classical MD simulations [1]. Using RadonPy, we create the world's largest polymer property databases, spanning a broad chemical space of more than $10^5 - 10^7$ polymer species.

The methodology of transfer learning, particularly the simulation-to-real (Sim2Real) transfer learning, enables the integration of extensive simulation data with limited experimental data [2,3,4]. Transfer learning leverages data or pretrained models from a source domain to enhance tasks in a target domain. Here, we present a measure for quantitatively evaluating the transferability and scalability of growing computational materials database. Our work revealed a scaling law in Sim2Real transfer learning using the RadonPy database. By observing the scaling behavior of transferred predictors, we estimate their expected generalization performance achievable by further increasing simulation data, serving as an indicator of the database's potential value. Additionally, multidimensional scaling, considering both physical and computer experiments, provides a statistical estimate for the equivalent sample size of experimental and computational data. Furthermore, observing the scaling strength of individual materials, we can gain insights into material groups sharing physical mechanisms across different material systems.

References

1. Y. Hayashi, J. Shiomi, J. Morikawa, and R. Yoshida, "RadonPy: automated physical property calculation using all-atom classical molecular dynamics simulations for polymer informatics", npj Comput. Mater. **8**, 222 (2022).
2. Wu *et al.*, "Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm", npj Comput. Mater. **5**, 66 (2019).
3. Yamada *et al.*, "Predicting materials properties with little data using shotgun transfer learning", ACS Central Science **5**, 1717 (2019).
4. Aoki *et al.*, "Multitask machine learning to predict polymer-solvent miscibility using Flory-Huggins interaction parameters", Macromolecules **56**, 5446 (2023).