June 5, 2025

The Institute of Statistical Mathematics
MITSUBISHI CHEMICAL CORPORATION

A New Horizon in Data-Driven Materials Research

Unveiling Scaling Laws Bridging Extensive Computational Databases and Limited Experimental Data

The research group at the ISM-MCC Frontier Materials Design Laboratory (a joint research division of Mitsubishi Chemical Corporation (MCC) and the Institute of Statistical Mathematics (ISM)) has discovered a phenomenon known as the "scaling law[*a] of Sim2Real transfer learning[*b]" in the integrated analysis of large-scale computational materials property databases and experimental data, in collaboration with research groups from the National Institute for Materials Science (NIMS).

In materials research, the development of extensive computational materials property databases generated through physical simulations is progressing to address the challenge of limited experimental data. Notably, it has been shown that models pre-trained on extensive computational databases can achieve predictive performance unattainable through direct learning, when fine-tuned with limited experimental data. Such integrated analysis is referred to as Sim2Real transfer learning.

This study demonstrated that in Sim2Real transfer learning of data-driven materials research, the performance of predictive models fine-tuned using experimental data improves monotonically according to a power law scaling as the size of the computational database increases. The existence of scaling laws in transfer learning using materials computational databases has been empirically and systematically validated for the first time. Furthermore, it was confirmed that the computational database for polymer materials developed by the same group exhibits strong scaling for various downstream tasks in real-world applications.

The scaling strength serves as a quantitative measure to assess the future value of the database. Analyzing scaling behavior enables us to estimate the amount of data required for a model to reach a desired performance, as well as the potential upper limit of that performance. Furthermore, the analysis of scaling laws leads to strategic planning for data platform development and the optimization of data production protocols in materials development projects.

These research findings were published in *npj Computational Materials* on May 24, 2025(URL=https://doi.org/10.1038/s41524-025-01606-5).

End of document

Contact:

The Institute of Statistical Mathematics, Research Organization of Information and Systems
URA Station, Planning Unit, Administration Planning and Coordination Section
TEL: +81-50-5533-8580
E-mail: ask-ura@ism.ac.jp

Mitsubishi Chemical Corporation
General Affairs & Communication Office
Media Relations Department
TEL:+81-3-6748-7140

Reference

Research Content

In data-driven research, the most crucial resource is data. However, compared to AI-advanced fields such as natural language processing, computer vision, biology, and medicine, the data resources in materials research are extremely limited. To overcome this barrier, materials researchers have utilized physical simulations, such as first-principles calculations[*c] and molecular dynamics simulations[*d], to construct extensive computational materials databases. In the field of inorganic materials, pioneering efforts like Materials Project[1] have led to the development of computational materials databases that span the entire periodic table, including AFLOW[2], OQMD[3], GNoME[4], and OMat24 dataset[5]. In the field of polymer materials, the research group at ISM has developed RadonPy, a software platform that fully automates computational experiments on polymer materials. They have formed an industry-academia consortium involving two national institutes, eight universities, and 37 companies, collaborating on the joint development of one of the world's largest polymer properties databases[6]. Furthermore, in collaboration with MCC, ISM has established the "ISM-MCC Frontier Materials Design Laboratory," focusing on automating quantum chemistry calculations and jointly developing a large-scale database that comprehensively evaluates the miscibility between polymer materials and solvent molecules[7].

In materials research, utilizing techniques like transfer learning integrates vast computational data with limited experimental data to enhance model predictive performance. For instance, models pretrained using extensive computational materials databases are fine-tuned for real-world prediction tasks using limited experimental data. Models derived from such Sim2Real transfer learning are known to exhibit superior predictive capabilities compared to those trained solely on experimental data. Through practical applications in materials development, the group has demonstrated that transfer learning is a powerful approach to overcoming the limitations posed by scarce experimental data[8,9].

In this study, the group demonstrated that scaling laws for Sim2Real transfer learning hold across various tasks in materials research (Figure 1). A joint research team led by Professor Kenji Fukumizu of ISM and Preferred Networks, Inc. had previously shown the existence of scaling laws in their theoretical work, and validated their applicability in Sim2Real transfer learning for computer vision[10]. According to this theory, the predictive performance of fine-tuned models on experimental properties improves monotonically with the size n of the computational database, following a power law relationship: $\text{prediction error} = Dn^{-\alpha} + C$. A database with a larger decay rate α and a smaller transfer gap (C) is considered ideal. The transfer gap represents the performance improvement limit attainable through database expansion and serves as a key indicator for the future potential of computational property databases.

This study also confirmed that Sim2Real transferred models derived from the RadonPy database and the polymer miscibility database, both developed by the group, exhibit strong scaling across various experimental properties. Some of the experimental data were provided by the PoLyInfo database development team at NIMS[11]. Computational property databases with broad transferability and strong scalability are desirable for addressing extensive real-world prediction tasks. While various computational property databases have been developed, no prior studies have quantitatively demonstrated their utility from the perspective of scaling laws. This study highlighted that strong scalability in transfer learning for diverse real-world systems can serve as a key indicator of the utility of computational property databases.

Analyzing scaling behavior offers several practical benefits. It enables the estimation of the amount of data required to achieve a target accuracy and the attainable performance limits.

Additionally, when scaling behavior converges, it allows for informed decisions to halt further data production and reallocate computational resources to other projects. Furthermore, this study demonstrated that it is possible to formulate experimental plans and determine the optimal allocation of resources between real-world experiments and computer simulations based on observed scaling behaviors.

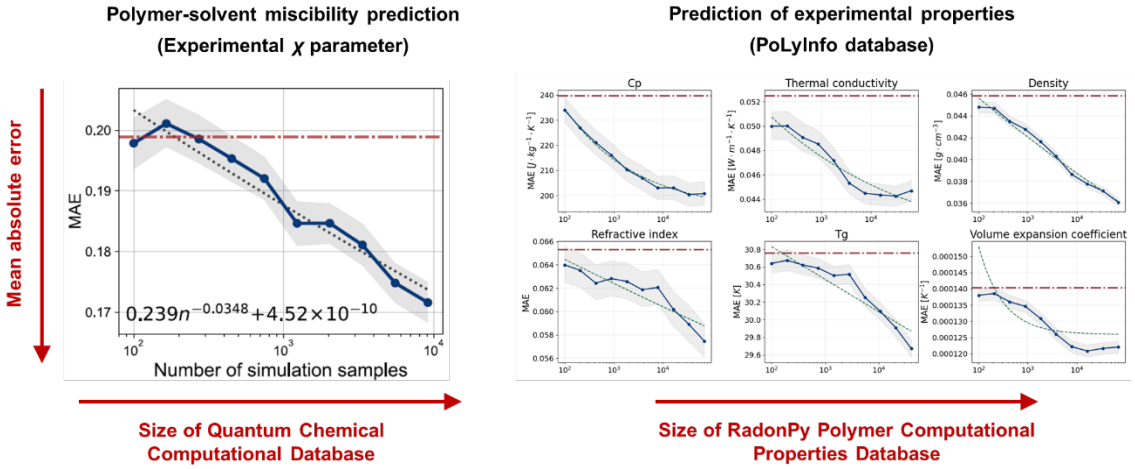

Figure 1: Successfully observed the scaling law for Sim2Real transfer learning.

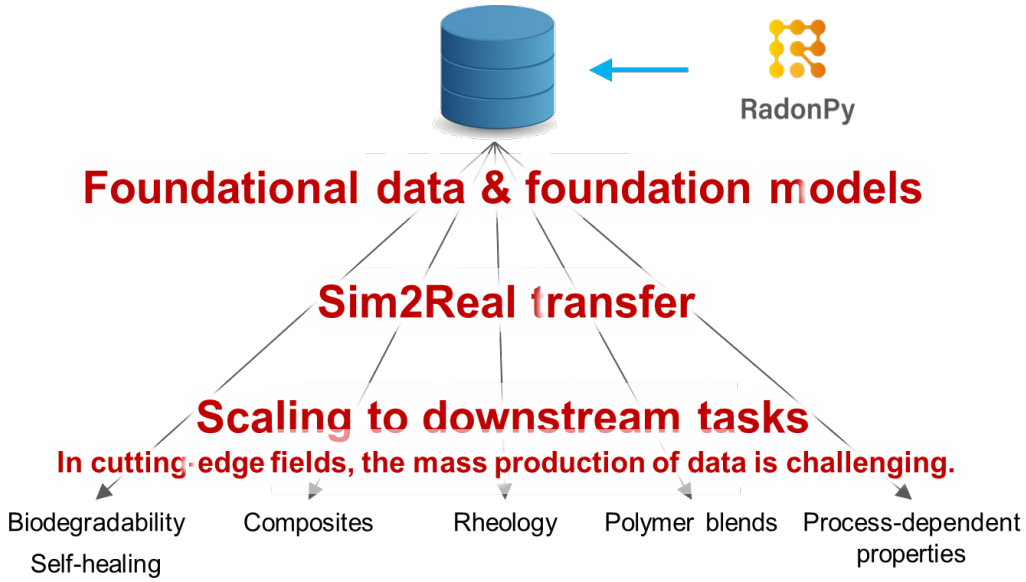

Figure 2: Data platform development strategy based on the scaling law of Sim2Real transfer learning.

Future Outlook

One of the critical milestones in data-driven materials research is establishing scalable and transferable data production protocols and analytical workflows that enable effective transfer learning (Figure 2). In many target domains, it is challenging to accumulate the data required for data-driven research. This tendency becomes more pronounced as we approach advanced research areas. Therefore, selecting source domains capable of producing large volumes of data, such as computational experiments, and bridging the gap between the source and target domains using machine learning is an increasingly important approach. In this context, it is crucial to design workflows such that as the data from the source domain increases, predictive

performance in the target domain scales accordingly. Conversely, exploring target domains that can benefit from transfer learning from source domain databases is equally important.

Note that the concepts of Sim2Real transfer learning and scaling laws are not limited to computational databases; they can be applied to the development of any database. Building foundational data through high-throughput data production processes and leveraging machine learning to bridge the gap between these foundational data and advanced research domains with lower data production efficiency provides a scalable and effective strategy for data-driven materials research.

This study has established design guidelines for the development of databases in the RadonPy project and the integration of quantum chemical calculations and deep learning for building solubility prediction models of polymer-solvent systems. Moving forward, we plan to continue data production while improving the predictive performance of transfer models in downstream tasks.

Publication

Title: Scaling law of Sim2Real transfer learning in expanding computational materials databases for real-world predictions
Authors: Shunya Minami, Yoshihiro Hayashi, Stephen Wu, Kenji Fukumizu, Hiroki Sugisawa, Masashi Ishii, Isao Kuwajima, Kazuya Shiratori, Ryo Yoshida
Journal: npj Computational Materials 11, 146
DOI: https://doi.org/10.1038/s41524-025-01606-5

Acknowledgements

This research was partially supported the Ministry of Education, Culture, Sports, Science and Technology (MEXT) "Fugaku" Program for Promoting Research to Accelerate Scientific Breakthroughs (hp210264), as well as the Japan Science and Technology Agency (JST) CREST projects (JPMJCR19I3, JPMJCR22O3, JPMJCR2332). We also express our gratitude to Dr. Masashi Ishii and Mr. Isao Kuwajima of the Technical Development and Shared Facilities Division at NIMS for providing the polymer property database PoLyInfo.

References

1) Jain et al., The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater* **1**, 011002 (2013). https://doi.org/10.1063/1.4812323
2) Curtarolo et al., AFLOW: An automatic frame-work for high-throughput materials discovery. *Comput Mater Sci* **58**, 218–226 (2012). https://doi.org/10.1016/j.commatsci.2012.02.005
3) Kirklin et al., The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Comput Mater* **1**, 15010 (2015). https://doi.org/10.1038/npjcompumats.2015.10
4) Merchant et al., Scaling deep learning for materials discovery. *Nature* **624**, 80–85 (2023). https://doi.org/10.1038/s41586-023-06735-9
5) Barroso-Luque et al., Open materials 2024 (omat24) inorganic materials dataset and models. *arXiv preprint* arXiv:2410.12771 (2024). https://doi.org/10.48550/arXiv.2410.12771
6) Hayashi et al., RadonPy: automated physical property calculation using all-atom classical molecular dynamics simulations for polymer informatics. *npj Comput Mater* **8**, 222 (2022). https://doi.org/10.1038/s41524-022-00906-4
7) Aoki et al., Multitask machine learning to predict polymer–solvent miscibility using Flory–Huggins interaction parameters. *Macromolecules* 56, 5446-5456 (2023). https://doi.org/10.1021/acs.macromol.2c02600

8) Wu et al., Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. *npj Comput Mater* **5**, 66 (2019). https://doi.org/10.1038/s41524-019-0203-2
9) Yamada et al., Predicting materials properties with little data using shotgun transfer learning. *ACS Cent Sci* 5, 1717-1730 (2019). https://doi.org/10.1021/acscentsci.9b00804
10) Mikami et al., A scaling law for syn2real transfer: How much is your pre-training effective? *Machine Learning and Knowledge Discovery in Databases*, 477–492 (2023). https://doi.org/10.1007/978-3-031-26409-2_29
11) Ishii et al., NIMS polymer database PoLyInfo (I): an overarching view of half a million data points. *STAM-M* **4**, 2354649 (2024). https://doi.org/10.1080/27660400.2024.2354649

Terminology

*a) The scaling law of AI is an empirical law that the performance of the accuracy of a machine learning model, e.g. prediction accuracy, improves according to the power law as the amount of training data increases.

*b) Prediction of experimental properties by the model trained by adding experimental data to a model pre-trained by computational database.

*c) A method to theoretically analyze the electronic structure, energy, reactivity, etc. of materials based on the principles of quantum mechanics.

*d) A method to calculate the trajectory of atoms and molecules based on Newton's equations of motion. The interaction between particles is represented by potential functions. Based on this method, physical properties such as structural change, diffusion and heat conduction of materials are analyzed on an atomic scale.