2025 年 6 月 5 日

情報・システム研究機構 統計数理研究所
三菱ケミカル株式会社

材料研究の新展開：大規模計算データベースと実験データをつなぐスケーリング則を発見

統計数理研究所（以下「統数研」）と三菱ケミカル株式会社（本社：東京都千代田区、社長：筑本 学）の共同研究部門「ISM-MCC フロンティア材料設計拠点」の研究グループは、物質・材料研究機構の研究グループと協力し、材料研究における大規模計算物性データベースと実験データの統合解析において、「Sim2Real 転移学習※1 のスケーリング則※2」と呼ばれる現象を発見しました。本成果（以下「本研究」）をまとめた論文が国際学術誌「npj Computational Materials」に掲載されたことをお知らせいたします（掲載論文の URL=https://doi.org/10.1038/s41524-025-01606-5）。

データ駆動型材料研究では、実験データの不足が原因で AI の予測性能を十分に発揮できないことが大きな課題となっています。この課題を克服するために、物理シミュレーションによって生成された大規模な計算物性データベースの開発が進められています。例えば、計算物性データベースで事前学習されたモデルを、限られた実験データを用いて追加学習することで、直接学習では到達不可能な予測性能を実現できることが知られています。このような統合解析を Sim2Real 転移学習といいます。

本研究では、データ駆動型材料研究の Sim2Real 転移学習において、計算物性データベースの規模が拡大するにつれて、転移モデルの実験物性に対する性能がべき乗則に従い単調に改善していくことを実証しました。材料系の Sim2Real 転移学習において、スケーリング則が存在することを系統的に示したのは、これが初めてです。

スケーリング強度は、データベースの将来価値を評価する定量的な指標となります。また、スケーリング挙動を解析することで AI のモデルが目標性能に到達するために必要なデータ数や到達可能な限界性能を見積もることができます。さらに、スケーリング則の解析は、材料開発プロジェクトにおけるデータプラットフォーム開発の戦略立案やデータ生産プロトコルの効率化につながることが期待されます。

以上

お問合せ先
大学共同利用機関法人
情報・システム研究機構 統計数理研究所
運営企画本部企画室 URA ステーション
TEL: 050-5533-8580、E-mail: ask-ura@ml1.ism.ac.jp

ご参考

研究の背景

データ駆動型研究において最も重要な資源は、言うまでもなくデータです。しかしながら、自然言語処理やコンピュータビジョン、生物、医療などの AI 先進分野に比べると、材料研究のデータ資源は極めて乏しいというのが現状です。この壁を乗り越えるために、材料研究者らは第一原理計算※3 や分子動力学シミュレーション※4 などの物理シミュレーションを駆使し、大規模な計算物性データベースの構築に取り組んできました。無機材料分野では、この分野の先駆けである Materials Project[1] を皮切りに、AFLOW[2]、OQMD[3]、GNoME[4]、OMat24 データセット [5] など、周期表全体を網羅する計算物性データベースが次々と開発されてきました。高分子材料分野では、統数研の研究グループが、高分子材料の計算機実験を全自動化するソフトウェア RadonPy を開発しながら、2 つの国立研究所、 8 大学、37 企業に属する約 260 名が参画する産学連携コンソーシアムを形成し、世界最大級の高分子物性データベースの共同開発を進めています [6]。また、「ISM-MCC フロンティア材料設計拠点」の研究グループ（以下「同グループ」）は、量子化学計算を全自動化するシステムを構築し、高分子材料と溶媒分子の相溶性を網羅的に評価した大規模データベースを開発しています [7]。

材料研究では、転移学習という手法を活用し、大規模な計算機実験のデータと限られた実験データを統合的に活用することで、モデルの予測性能を向上させます。例えば、計算物性データベースを用いてモデルを事前学習し、限られた実験データを用いて現実世界の予測タスク向けに微調整（ファインチューニング）します。このような Sim2Real 転移で得られたモデルは、実験データだけで学習されたモデルでは到達できない高度な予測能力を発揮することが知られています。同グループは、材料開発における実践を通じて、転移学習が限られた実験データの壁を克服するための効果的なアプローチであることを実証してきました [8,9]。

研究内容と成果

本研究において、同グループは、材料研究の多様なタスクにおいて、Sim2Real 転移学習のスケーリング則が成り立つことを明らかにしました（図 1）。統数研の福水健次教授と株式会社 Preferred Networks の共同研究グループは、先行研究においてスケーリング則の存在を理論的に予想し、コンピュータビジョン分野の Sim2Real 転移学習において、この法則が成り立つことを実証しました [10]。この理論によれば、ファインチューニングされたモデルの実験物性に対する予測性能は、計算データベースのサイズ n の増加に伴い、べき乗則 $\text{prediction error} = Dn^{-\alpha} + C$ に従って単調に改善します。データ数 n の増加に対する改善率 α が大きく、転移ギャップ C が小さいデータベースが望ましいと考えられます。転移ギャップは、データベース拡大によって到達可能な限界性能であり、計算物性データベースの将来価値を表す指標となります。

本研究では、RadonPy 高分子物性データベースや高分子相溶性データベースから導かれた転移モデルが、さまざまな実験物性に対して強いスケーリングを示すことが確認されました。実験データの一部は、物質・材料研究機構の PoLyInfo データベース開発チームから提供されました [11]。計算物性データベースは、現実世界の広範な予測タスクに対する転移可能性と強いスケーラビリティを持つことが望ましいと考えられます。これまでにさまざまな計算物性データベースが開発されてきましたが、スケーリング則の観点からその有用性を定量的に示した例は報告されていません。本研究は、多様な

現実系に対して転移学習の強いスケーラビリティを持つことが、計算物性データベースの有用性を表す指標になることを示しました。

スケーリング挙動の解析は、目標精度に到達するために必要なデータ数や、到達可能な性能限界を見積もることに役立ちます。また、スケーリング挙動が収束した場合には、それ以上のデータ生産を停止し、計算資源を他のプロジェクトに再分配するという意思決定が可能になります。さらに本研究では、スケーリング挙動の解析に基づく実験計画の策定や、物理実験と計算機実験の最適な資源配分を決定することが可能であることを示しました。

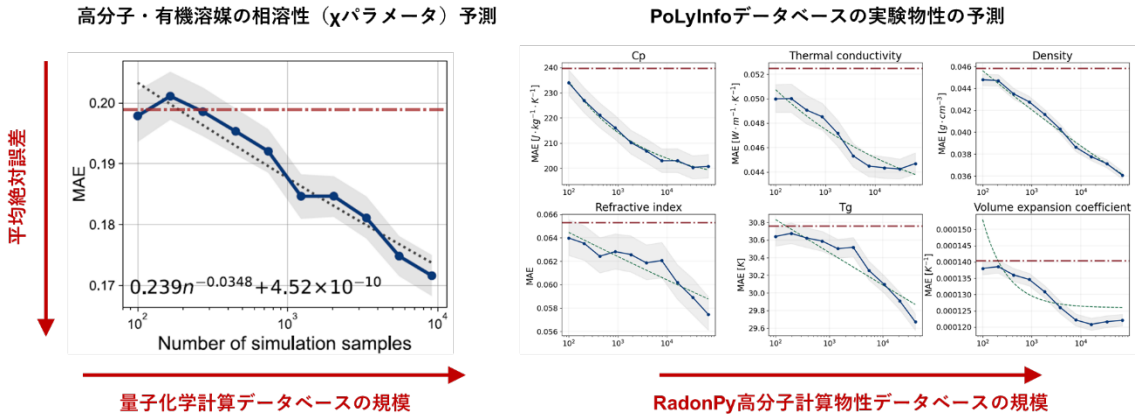

図 1: Sim2Real 転移学習スケーリング則を観測することに成功.

今後の展開

データ駆動型材料研究における重要なマイルストーンの一つは、このようなスケーラブルに転移可能なデータ生産プロトコルと解析ワークフローを確立することです（図 2）。材料開発の多くのドメインでは、データ駆動型研究に必要な十分な量のデータを蓄積できません。この傾向は先端的な研究領域に近づくにつれて顕著になります。そこで、計算機実験のような大量データを生産可能な元ドメインを選定し、機械学習で元ドメインと目標ドメインの間のギャップを埋めるというアプローチが重要になってきます。この際、元ドメインのデータが増加することで、目標ドメインでの予測性能がスケールするようにワークフローを設計することが重要です。あるいは逆に、元ドメインのデータベースから転移可能な目標ドメインを探索していくということも重要になります。

Sim2Real 転移学習やスケーリング則の概念は、計算物性データベースに限らず、あらゆるデータベースの開発に適用できます。ハイスループットなデータ生産プロセスによって基盤的データを構築し、データ生産効率の低い先端研究領域との間の乖離を機械学習でスケーラブルにつなぐことが、データ駆動型材料研究の有効な戦略になります。

今回の研究により、RadonPy プロジェクトにおけるデータベース開発や、ポリマー・溶媒系の相溶性予測モデルを構築するための量子化学計算・深層学習統合解析プラットフォームの設計指針が確立しました。今後もデータ生産を継続しながら、下流タスクにおける転移モデルの予測性能を向上させていく予定です。

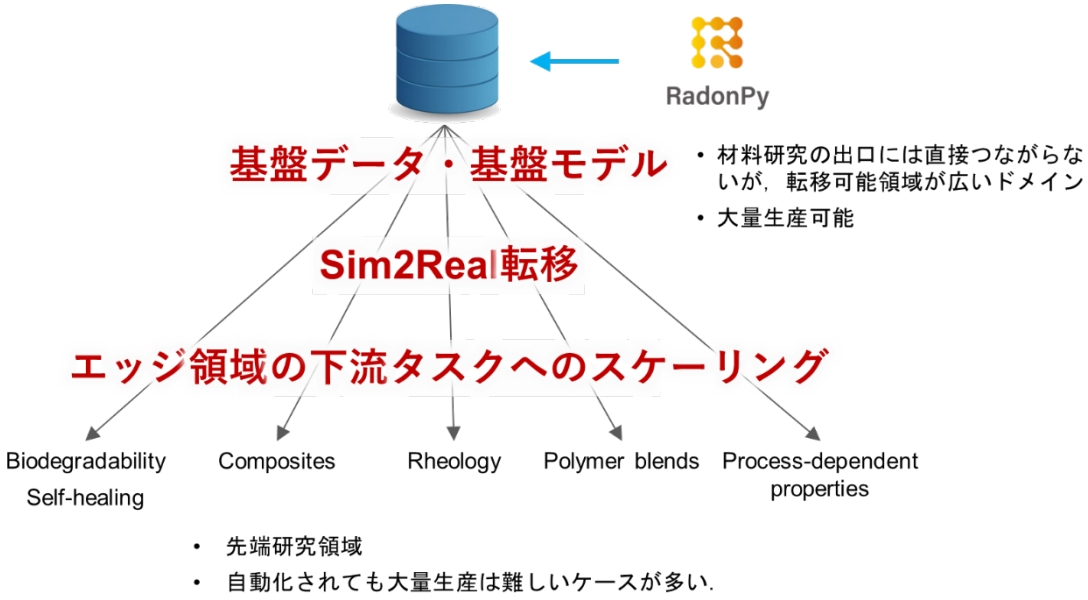

図 2: Sim2Real 転移学習のスケーリング則に基づくデータプラットフォーム開発戦略

掲載論文

論文題目: Scaling law of Sim2Real transfer learning in expanding computational materials databases for real-world predictions

著者: Shunya Minami, Yoshihiro Hayashi, Stephen Wu, Kenji Fukumizu, Hiroki Sugisawa, Masashi Ishii, Isao Kuwajima, Kazuya Shiratori, Ryo Yoshida

雑誌: npj Computational Materials 11, 146

DOI: https://doi.org/10.1038/s41524-025-01606-5

掲載日時: 2025 年 5 月 24 日

謝辞

本研究の一部は、文部科学省「富岳」成果創出加速プログラム「データ駆動型高分子材料研究のデータ基盤」(hp210264) 、科学技術振興機構 CREST（JPMJCR19I3、JPMJCR22O3、JPMJCR2332）の一環として実施されたものです。また、本研究を実施するにあたり、高分子物性データベース PoLyInfo を提供していただいた国立研究開発法人 物質・材料研究機構 技術開発・共用部門の石井真史氏ならびに桑島功氏に感謝の意を表します。

参考文献

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

用語解説

※1　計算物性データベースで事前に学習されたモデルを、実験データを用いて追加学習し実験物性の予測モデルを構築する。

※2　AI のスケーリング則とは、機械学習モデルの性能（例：予測精度）が、学習データの量の増加に伴って、べき乗則に従って改善していくという経験則です。

※3　量子力学の原理に基づいて、材料の電子構造・エネルギー・反応性などを理論的に解析する手法です。

※4　原子や分子の運動をニュートンの運動方程式に基づいて時間発展的に追跡する計算手法です。粒子間の相互作用をポテンシャル関数で表し、物質の構造変化・拡散・熱伝導などの物理特性を原子スケールで解析します。