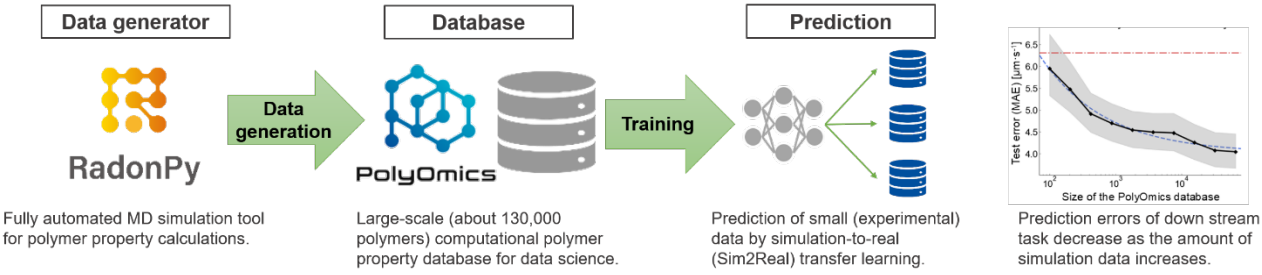
G1-09 データ不足の壁を乗り越える Sim2Real ポリマーインフォマティクス

林　慶浩（統数研・総研大・理研）

データ駆動型材料科学の進展には、大規模なデータベースが不可欠である。自然言語処理や画像分野とは異なり、高分子科学においては大規模オープンデータベースの構築が困難であった。その背景には、高分子合成や物性測定に要する高いコスト、化学空間の広大さ、高分子材料の構造的複雑性がある。そのため、既存の高分子データベースは、規模、スコープ、物性多様性の点で限定的である。

この課題を克服するため、我々は全原子分子動力学（MD）シミュレーションを完全自動化するパイプライニングライブラリ RadonPy[1] を用いて、約 13 万種の多様なポリマーに対して最大 43 種類の物性を収録する大規模高分子データベース PolyOmics[2,3] を開発した。PolyOmics は、RadonPy とスーパーコンピューター「富岳」の膨大な計算資源約 1 億ノード時間積を使用して構築された。また、3 国所、8 大学、37 社の企業から 260 名以上の研究者が参加する大規模産学コンソーシアムによりプロジェクトを推進している。PolyOmics は Hugging Face サイト上で全データが公開されており[3]、合成高分子や、セルロース誘導体、ラダーポリマー、有機低分子などの多様なサブセットを含んでいる。

PolyOmics のデータは自動化された MD シミュレーションによって大量生成されたため、データには偏りやばらつきが含まれる。また、実験データにも同様の偏りやばらつきが存在する。このような計算と実験間のギャップは、計算データで事前学習したモデルから実験値へと転移する機械学習手法である simulation-to-real (Sim2Real) 転移学習により埋めることに成功している。さらに、Sim2Real 転移学習の重要な性質として、事前学習に用いた計算データサイズの増加に伴い、実験値予測の汎化誤差がべき乗則により単調に減少することが理論的に予想されている。PolyOmics の大規模データで事前学習された機械学習モデルは、幅広い高分子特性予測タスクにおいて、このべき乗則スケーリングを示した[2,4]。様々な実験値の予測タスクに対して、転移学習により早期適応可能な基盤モデルを開発すべく、PolyOmics に収録されている 43 物性全てを 1 つのモデルで事前学習した基盤モデルの構築を進めている。この基盤モデルの詳細は当日発表する。



1）Y. Hayashi, J. Shiomi, J. Morikawa, R. Yoshida, *npj Computational Materials*, **8**, 222 (2022).
2）R. Yoshida and Y. Hayashi et al., *arXiv preprint*, arXiv:2511.11626 (2025).
3）https://huggingface.co/datasets/yhayashi1986/PolyOmics
4) S. Minami, Y. Hayashi, S. Wu, K. Fukumizu et al., npj Computational Materials, 11, 146 (2025).

PROFILE

林　慶浩（統計数理研究所 先端データサイエンス研究系 マテリアルズインフォマティクス研究推進センター　准教授・総合研究大学院大学　准教授・理化学研究所 TRIP-AGIS　客員研究員）

2014 年に東京工業大学大学院理工学研究科有機・高分子物質専攻博士課程を修了し博士（工学）取得、2014 年～2019 年まで東京工業大学の研究員・特任助教を経て、2020 年より統計数理研究所に特任助教として着任。2022 年より同助教、2026 年から同准教授。専門は主に有機物・高分子を対象とした計算化学とマテリアルズインフォマティクス。近年の研究では、高分子物性計算を完全自動化する RadonPy の開発や、約 13 万種のポリマーを包含する大規模高分子計算物性データベース PolyOmics の開発を行っている。

Copyright(c); 2026, The Chemical Society of Japan