jh250044

環境循環型社会の実現に向けた ポリマーインフォマティクスのデータ基盤構築

佐藤正寛（東京大学）

概要

本研究では、環境循環型社会の実現に向け、計算科学に基づくポリマーインフォマティクスのデータ基盤構築を進めた。PoLyInfo 由来の約 1.3 万種のポリマー構造に対する量子化学計算と、約 5,500 種規模の分子動力学計算から多階層物理記述子を整備し、mdx 上で利活用可能なデータベース化を推進した。機械学習では、構造記述子よりも物理ベース記述子が密度などの外挿予測で高いサンプル効率と精度を示し、少量データでも有効であることを確認した。さらに、延べ 6,200 種相当の MD 計算と約 900 種・最大 9 試行のベンチマークにより、15 種類の物性について、計算資源が同等でも「対象種数を増やすべきか」「反復計算を増やすべきか」の最適条件が異なることを明らかにした。以上により、外挿可能な AI モデルと高機能ポリマー設計に資するデータ利活用基盤の有効性を示し、材料開発の高速化に向けた指針を得た。

1. 共同研究に関する情報

(1) 共同利用・共同研究を実施している拠点名

東京大学 情報基盤センター

大阪大学 D3 センター

mdx I

(2) 課題分野

大規模計算科学課題分野

データ科学・データ利活用課題分野

(3) 参加研究者一覧と役割分担

佐藤　正寛	研究統括
熊田　亜紀子	研究副統括
鈴村　豊太郎	データ基盤構築支援
田浦　健次朗	データ基盤構築支援
華井　雅俊	データ基盤構築支援
梅本　貴弘	コード開発・データ基盤構築
WANG　WEIHAO	コード開発・データ基盤構築
RUAN　HAOOU	コード開発・データ基盤構築
横山　尋斗	コード開発・データ基盤構築
井上　勢大	コード開発・データ基盤構築

2. 研究の目的と意義

〈研究計画全体の目的〉

本研究では，計算科学データをもとにしたポリマーインフォマティクスのデータ基盤を構築する。データベースの利活用により，環境循環型社会の実現に必要な高機能ポリマー材料の創出を促すことを目的とする。

〈今年度の目的〉

昨年度は、量子化学・MD 計算の基礎データ収集および外挿可能な AI モデルの初期構築を行い、ポリマー設計における物理記述子の有効性を示した。

今年度は以下を重点的に実施している：

- 高スループット計算によるポリマー物性データの拡張（現時点で約 5,500 種）
- モデルの検証や精度向上に必要となるデータのバラエティとボリュームの評

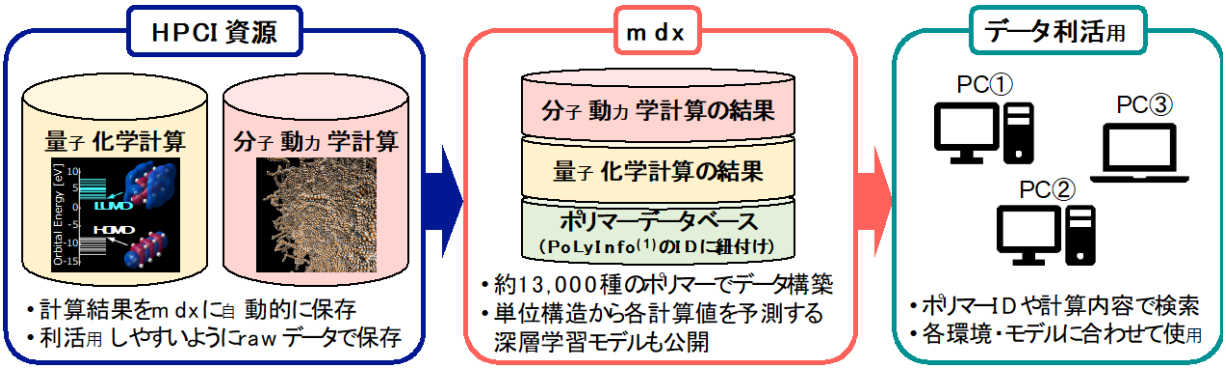


図 1 基盤システムの概略図

価

- 多階層物理記述子を統合した外挿予測モデルの構築
- 得られた物理量のデータベース化とデータ基盤 API 整備
- 機械学習による物理量相関の可視化と設計指針抽出

これらにより、多階層インフォマティクス基盤を実現する。

〈研究の学術的意義・社会的意義〉

本研究で開発される外挿可能な AI モデルは，ポリマーの分子構造からミクロ・メソスケールを経由してマクロ物性に至るまでの過程を定量的に評価できるものであり，ポリマー物性発現の学理構築にも貢献する(図 1)。また，基盤データベースを活用した自身の研究事例を発信すれば，多分野のポリマー材料設計に活かされる。

ポリマー材料は未来の環境循環型社会の実現に必要不可欠である。開発データ・AI 基盤はデータが少ないポリマー分野において、世界に先駆けてデータ利活用基盤を構築し、ポリマーの設計開発の高速化、既存ポリマーの延長上にない外挿的なポリマー創成を可能にする。

3. 当拠点の公募型共同研究として実施した意義

ポリマーインフォマティクスの基盤構築にあたっては、以下の課題を解決する必要がある。まず、計算科学に基づく新たなデータベースの構築と、それを利活用できる基盤の整備が求められる。加えて、ポリマーの多階層構造を考慮した網羅的な計算を実施するためには、大規模な計算資源が不可欠である。

本提案では、図 1 に示す概略図に従い、公募型共同研究を通じてシステム構築を進める。計算負荷の大きい量子化学計算および分子動力学計算については、HPCI 資源を活用した大規模並列計算により実施する。さらに、データ利活用基盤として mdx を用い、検索可能な計算結果のデータベース化を行う。

本提案は、申請者が所属する電気電子分野にとどまらず、材料科学や情報科学分野にも波及効果を持つ学際的な取り組みである。従って、本提案では、工学系研究科・情報理工学系研究科および情報基盤センターと共同でデータベース構築を行うことができる。また、将来的には材料学と情報学の融合による、AI モデル開発から材料設計に至る学際的な共同研究へと発展させることを目指している。

4. 前年度までに得られた研究成果の概要

昨年度は、大規模データベースとマルチスケール計算（第一原理計算・分子動力学計算）を統合した高精度なポリマー物性予測 AI モデルの構築、およびその外挿予測性能の検証に着手した。以下にその具体的な実施内容と成果を記述する。

〈ポリマー構造の取得と計算基盤の整備〉

大規模ポリマーデータベース「PoLy-Info」より、約 13,000 種類のポリマー構造（SMILES）を抽出した。量子化学計算および分子動力学計算における操作性を高めるため、モノマー構造の SMILES から計算に適した CH3 終端のオリゴマー構造を自動生成し、大量の構造に対して一貫した条件下でのシミュレーションを可能にした。

〈量子化学計算（QM）によるミクロ記述子の抽出〉

取得した全 13,000 種類のポリマーに対し、大阪大学サイバーメディアセンターのスーパーコンピュータ「SQUID」上の Gaussian16 を用い、密度汎関数法（DFT）に基づく電子状態計算を実施した。計算レベルは B3LYP/6-31G(d)とし、構造最適化を経て、HOMO/LUMO エネルギー、分極率、溶媒和エネルギー、ESP 電荷など、多岐にわたるミクロ物理量を「QMex 記述子」として網羅的に取得した。大規模計算を遂行するため、並列処理と後処理を効率化するワークフローを構築し、1 万点を超えるデータセットの作成を完了させた。

〈分子動力学計算（MD）によるメゾスケール物性の算出〉

全構造のうち約 5,500 種類に対し、LAMMPS および自動化ライブラリ「RadonPy」を用いて分子動力学計算を実行した。GAFF2 力場を採用し、21 段階の精密な平衡化プロトコルを経て、密度、比熱、熱伝導率、さらにポリマーマトリックス中における Na^+および Cl^-のイオン移動度を含む 12 種類以上の物性値を算出した。これにより、実験値に対応するメゾスケールの物理量データを蓄積し、QM 計算結果と組み合わせたマルチスケールなデータ基盤を構築した。

〈機械学習モデルの開発と外挿予測性能の評価〉

得られた計算物性値を学習データとし、既存手法では困難であった「学習データ範囲外（外挿）」の物性予測手法の開発に注力した。QM 由来のミクロ物理量や MD の力場パラメータ、メゾスケール物理量を入力記述子としたマルチスケールモデルを構築した結果、分子構造ベースの記述子である Extended-Conectivity Fingerprints （ECFP） と比較して実験値に対する外挿予測精度が顕著に向上することを確認した。

5. 今年度の研究成果の詳細

〈外挿予測精度のベンチマークとスケーリング〉

まず、MD シミュレーションから得られる物性を対象にミクロ物理量や MD の力場を用いた予測モデルを構築し、その性能を評価した。具体的には、半径 2、2048 ビット ECFP に加え、グラフ畳み込みニューラルネットワーク （Graph_CN） を比較として評価した。一般的な機械学習のベースラインとしては、部分最小二乗回帰（PLS）とカーネル・リッジ回帰（KRR）を用いた。PLS は線形回帰（LR）、KRR は非線形回帰（NLR）の一種である。

予測性能を評価するために、アンサンブル学習を 3 回実施した。最終予測値は 3 つのアンサンブルの平均で与え、さらにデータ分割方法の違いに起因する予測不確かさも定量化した。ただし、この不確かさ推定は分割に由来するばらつきのみを反映しており、MD で計算された目的変数の不確かさや、選択した力場に由来する系統バイアスは含んでいない。各アンサンブルには二重交差検証を適用した。外側ループではデータセットを 5 分割し、そのうち 1 分割(20%)をテスト、残り 4 分割(80%)を学習に用いた。ハイパーパラメータは、学習セット内で 4 分割交差検証によって選定した。予測精度は、RMSE と決定係数 R^2 で評価した。また、データセットサイズによる精度変化（スケーリング）を調べるため、約 200、500、1000、2000、3000、5000 件のデータをランダムに抽出して予測性能を評価した。

さらに、化学・記述子・物性からなる結合空間に

おける分布外外挿性能（ここでは、化学・記述子のシフトと、目的物性値レンジのシフトの両方を含めている）に対するスケーリングも評価した。内挿側データを約 50、100、200、500、1000、2000、3000 件とした場合の予測性能を調べた。分割のわずかな違いがあるため、これらの値は概数である。各サイズについて 3 回サンプリングし、各サンプルに対して予測モデルを構築したのち、平均予測精度を算出した。

図 3 は、MD で計算された密度について、内挿予測性能がデータセットサイズにどう依存するかを示している。この場合、データ量が増えるほど予測精度は向上し、とくに ECFP を記述子に用いた場合にその傾向が顕著だった。ECFP は高次元であるため、その予測は利用可能なデータ数の影響を強く受けると考えられる。これらの結果から、ポリマー物性予測においても、データセットの拡張が精度向上に寄与することが示唆された。

図 4 は、内挿予測精度が最も高かった密度について、外挿予測精度が内挿データサイズにどう依存するかを示している。ここでは、分布バイアスを考慮するために、外挿データセットの標準偏差に対する相対 RMSE で外挿精度を評価している。図 5 は、内挿データがおよそ 100 点の場合における、予測密度と MD 計算密度のパリティプロットである。図 5(a)は Graph_CN、図 5(b)は QMex と FF 記述子に基づく線形回帰モデルの結果である。

MD 計算密度の外挿予測では、物理記述子の利用が構造記述子に対して明確な優位性を示した。Graph_CN の外挿性能は、物理記述子を用いた非線形回帰と同程度であったが、物理記述子を用いた線形回帰は、特に少量データ領域でそれらを大きく上回った。内挿データが 100 点の場合、Graph_CN の R^2 は 0.50 であったのに対し、QMex と FF による線形回帰では $R^2 = 0.82$ を達成した。これは、

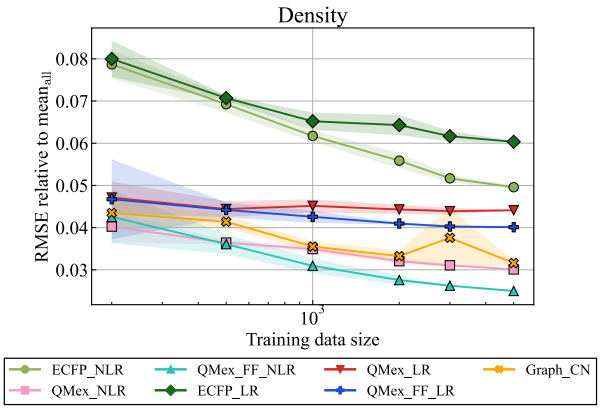


図 3 MD で計算された密度の内挿予測精度のスケーリング。誤差バンドはアンサンブル間の標準偏差を示す。

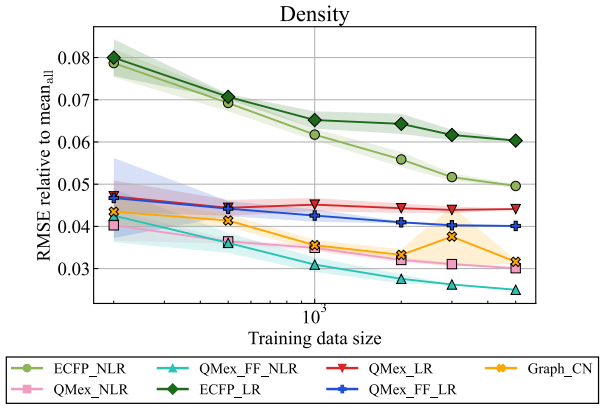
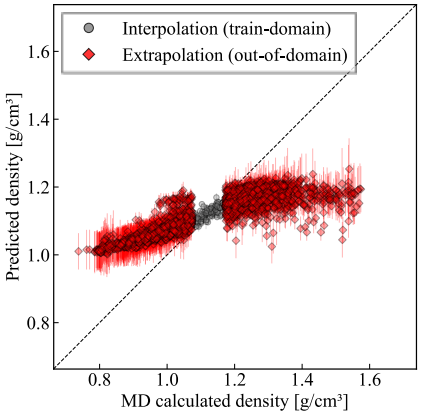
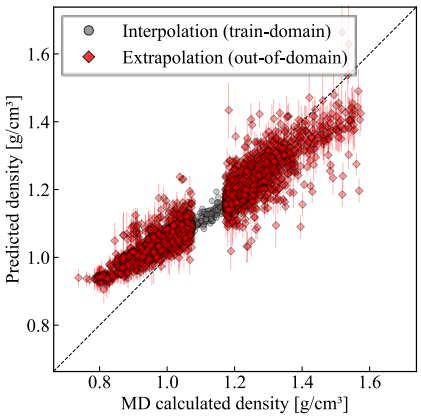


図 4 MD で計算された密度の外挿予測精度のスケーリング。誤差バンドはアンサンブル間の標準偏差を示す。



(a) Graph_CN



(b) QM_FF_LR

図 5 約 100 個の内挿データ点を含むデータセットについて、予測された密度の値と MD 計算による密度の値を比較したパリティプロット。内挿データ点の数は 103、外挿データ点の数は 3636 である。エラーバーはアンサンブル間の標準偏差を示している。

量子化学計算と力場パラメータの導出によって、密度予測に本質的な特徴量をうまく設計できたためと考えられる。言い換えれば、物理記述子は著しく高いサンプル効率をもち、1 桁少ない学習点数で高精度に到達できる。実際、図 4 に示されるように、QMex+FF の線形回帰モデルは、学習サンプル 100 点で外挿テストセットに対する RMSE/標準偏差 = 0.43 ± 0.03 を達成したのに対し、Graph_CN が同程度の精度（0.43 ± 0.04）に達するには 1000 点の学習サンプルを必要とした。この比較から、この設定では物理記述子ベースの単純な回帰モデルが、より少ないサンプルで同程度の誤差に到達できることが分かる。

データが少ない状況での予測の安定化・改善は、ポリマー言語モデルを含む最近の事前学習—微調整フレームワークでも示されている。これらはデータ効率を高める既知の手法であり、本研究の物理ベース記述子が与える帰納バイアスを補完するものである。今後は、物理記述子と事前学習表現、さらにマルチタスク学習を組み合わせることで、外挿的な汎化性能をさらに強化できると期待される。

図 6 は、熱伝導率予測の外挿精度が内挿データサイズにどう依存するかを示している。熱伝導率を含め、密度以外の物性については、物理記述子と構造記述子の間で顕著な差は見られなかった。これは、対象物性に対して現行の特徴量が十分な情報を含んでいないことを示唆している。とくに、熱伝導率やイオン移動度のような輸送物性は、純粋に原子レベルの記述子では十分に捉えられない、メゾスケールやトポロジカルな効果の影響を受ける可能性がある。本研究の特徴量セットは DFT と MD 力場に基づいており、主として短距離の構造情報と動的情報を符号化している。

外挿予測が成功した密度については、平均原子質量、原子あたり電子数、平均結合角力定数の 3 つの記述子が、それぞれ密度と強い相関（ピアソンの $r > 0.8$）を示した。これに対し、外挿がうまくいかなかった熱伝導率では、相関はかなり弱く、最も高い HOMO エネルギーでも $r \approx 0.27$ にとどまった。この限界を克服するには、長距離相関や集団モードを捉えられる粗視化記述子や、グラフベースのトポロジカル記述子を表現に追加する必要があることが、本研究の結果から支持される。

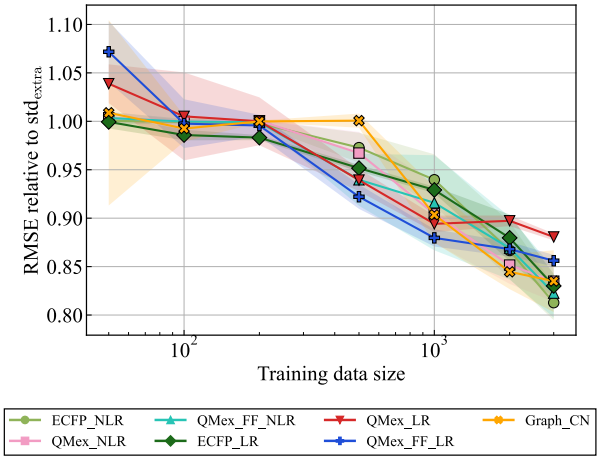
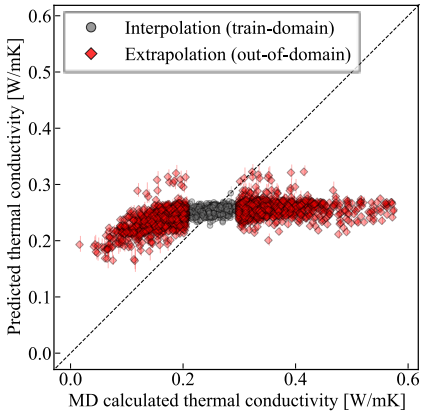
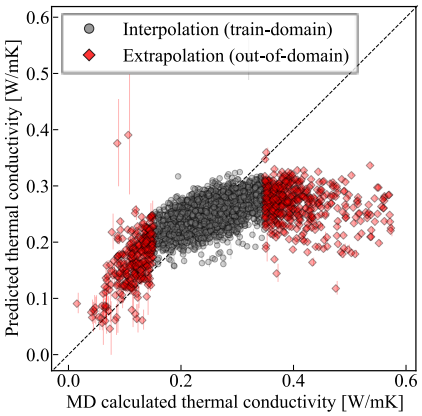


図 6 MD で計算された熱伝導率の外挿予測精度のスケーリング。誤差バンドはアンサンブル間の標準偏差を示す。



(a) 内挿データ数が 500 の場合



(b) 内挿データ数が 3000 の場合

図 7 QMex および FF 記述子に基づく線形回帰モデルを用いて、予測された熱伝導率の値と MD 計算による熱伝導率の値を比較したパリティプロット。データセットの内挿データ点は、(a) 500 点および (b) 3000 点である。エラーバーは、アンサンブル間の標準偏差を示す。

また、どのポリマー系統やどの物性で予測失敗が生じやすいのかを系統的に明らかにすることも、依然として重要な未解決課題である。これに取り組むことで、より具体的な特徴量設計やベースラインモデルの改善につながる可能性がある。

図 4 と図 6 は、内挿データ数が増えるほど外挿性能も改善する傾向を示している。これは、データベース拡張が外挿予測精度の向上にも有効であることを意味する。図 7 は、内挿データ数が (a) 500 点、(b) 3000 点の場合における、熱伝導率の MD 計算値と予測値のパリティプロットである。熱伝導率のように外挿性能がまだ十分高くない物性については、利用可能データ量をさらに増やすことで改善できる可能性がある。

物理ベース記述子に基づく本研究のモデルも、既存手法と同様に、学習データセットが大きくなるほど信頼性の高い予測を与える傾向がある。とくに、記述子が目的物性を支配する重要な物理・化学要因とうまく整合している場合には、外挿予測すら可能になる。他方で、長距離相関や集団的モードが支配的な量は、現在の記述子設計では高精度に予測することが難しい。こうした場合には、マルチスケール情報や長距離記述子を取り込んで特徴表現を豊かにすることが有効と考えられる。

次に、量子化学計算値、力場に加えて MD 計算値を記述子に加えたポリマー物性実験値予測モデルのベンチマーク結果を示す。高分子材料が持つ階層構造に着目し、マルチスケール計算から得られる物理量を記述子として利用することで、外挿予測性能を向上させることを目的とする。

使用したデータセットは、PoLyInfo データベースから取得した高分子物性データである。対象物性として、密度およびガス拡散係数 (Cgd) を選択した。密度データは 890 点、Cgd データは 203 点である。分子構造は SMILES 形式で取得し、物性値は複数の測定値が存在する場合には中央値を採用した。

予測精度の比較のため、分子構造ベースの記述子として、ECFP および 2D 記述子フィンガープリント（2DFP）を計算した。ECFP は半径 2、ビット長 2048 の設定で生成した。2DFP は原子種、結合種類、部分電荷などに基づく化学情報を数値ベクトルとして表現する。機械学習モデルとして、線形回帰モデルである PLS 回帰および非線形回帰モデルである KRR を使用した。モデル評価には RMSE および決定係数 R^2 を用い、5 分割交差検証を 5 回繰り返すことで評価を行った。また、分子グラフを用いる深層学習モデルとして Graph Isomorphism Network（GIN）を導入し、比較を行った先行研究での有効性に倣い、各ポリマーを，繰り返し単位の周期性を捉える環状分子グラフとして表現した。環状構造に含める繰り返し単位数は 5 とした。GIN を用いるモデル（以下 Mol_GIN)ではまずハイパーパラメータ最適化を行い、続いて最適化パラメータで 5 分割交差検証を実施した。

学習データを超えた汎化性能を評価するため、外挿課題も実施した。外挿は二種類、すなわち目的物性の範囲に関する外挿と、分子構造の非類似性に基づく外挿を考えた。内挿データは全データセットの約 30% に設定した。分子構造類似度は平均タニモト係数により評価した。外挿予測は、内挿データ部分集合のみで学習したモデルを用いて実施した。

まず、QM 記述子、FF 記述子、MD 記述子の 3 種について，非ゼロの全組合せに対し，LR および NLR モデルを用いて外挿性能を評価した。物性範囲外挿では LR モデルの方が良好な傾向がある一方で、構造類似度外挿では NLR モデルが概して優れた性能を示した。この傾向は先行研究で報告された結果と整合する。記述子別には、FF 記述子と MD 記述子を用いたモデルが相対的に高い予測精度を達成した。これは、目的物性である密度と Cgd がいずれも原子および分子の熱運動の影響を受ける物理量であることに起因すると考えられる。したがって、FF 記述子や MD 記述子のように分子間相互作用を効果的に捉える記述子は、これらの予測課題に特に適している。本結果はまた、この種の熱機械的物性に対しては、QM 記述子そのもの

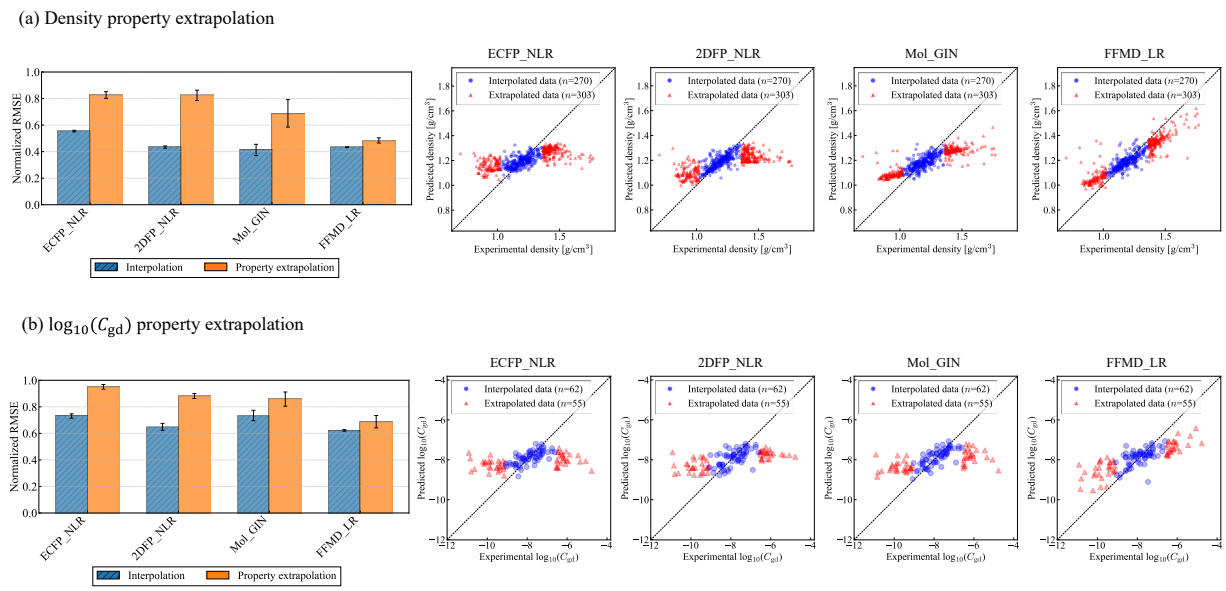


図 8　(a) 密度および (b) Cgd における物性範囲の外挿予測精度の比較。補間および外挿の両方の予測精度を示している。実験値と予測値のパリティプロットも含まれている。モデル選択は正規化 RMSE に基づいて行われている。

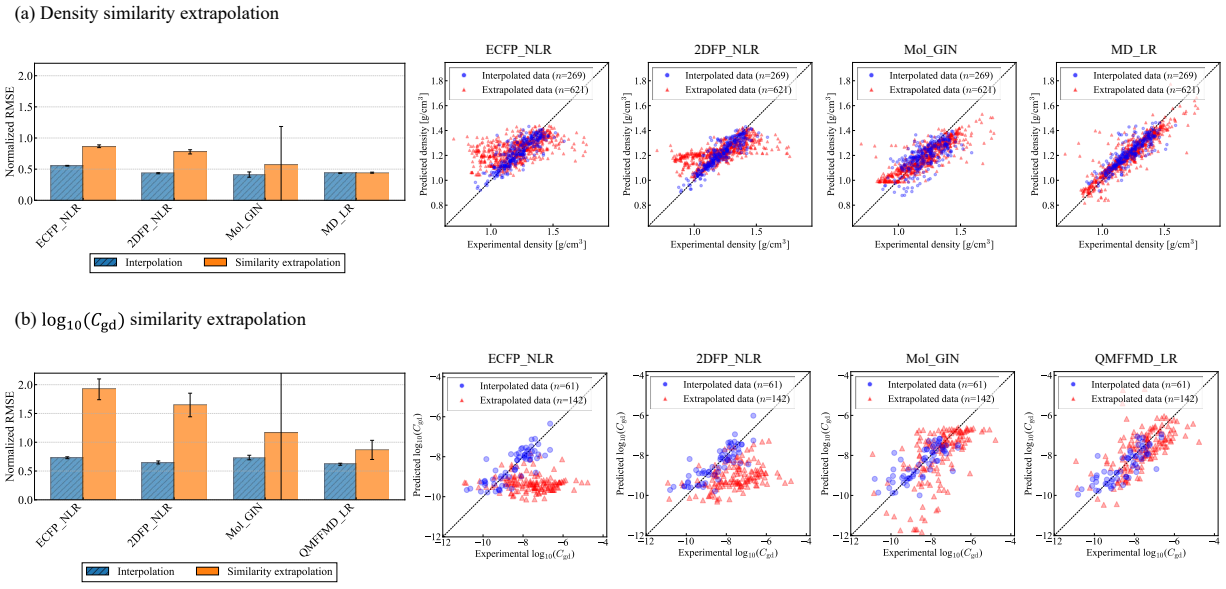


図 9　(a) 密度および (b) Cgd における分子構造の類似性に関する外挿予測精度の比較。補間および外挿の両方の予測精度を示している。実験値と予測値のパリティプロットも含まれている。モデル選択は正規化 RMSE に基づいて行われている。

よりも、QM 計算から蒸留された FF 記述子の方が情報量が高いことを示唆する。これは、FF 記述子が関連する分子間相互作用をより直接的に符号化しているためである。

図 8 および図 9 では、ECFP および 2DFP 記述子を用いたモデル (NLR および Mol_GIN) と、物理ベース記述子の最適組合せを用いたモデルの外挿性能を比較する。図 8 は物性範囲外挿の結果を、図 9 は構造類似度外挿の結果を示す。図 8(a) と図 9(a) は密度予測に、図 8(b) と図 9(b) は Cgd 予測に対応する。参考として、全データセットを用いて得た内挿予測精度も併記し、内挿性能は全データの標準偏差で正規化した。誤差棒は、5 つのアンサンブルモデル間の予測性能の標準偏差を表す。2DFP を用いた LR モデルは、著しい過学習により予測精度が大きく低下したため除外した。

図 8 および図 9 から、物理ベース記述子を用いるモデルは、外挿課題において構造ベースモデル（ECFP NLR、2DFP NLR、Mol_GIN）を一貫して上回ることが分かる。全ての外挿シナリオで内挿に比べ予測精度は低下するものの、その低下は物理ベース記述子に基づくモデルの方が小さい。これらの結果は、物理ベースモデルが学習データ外領域に適用された場合でも、より信頼性の高い予測を与えることを示唆する。構造ベースモデルの中では、環状グラフを用いる Mol_GIN が最も高い外挿性能を示したが、これは ECFP や 2DFP のようなモノマー基準の記述子よりも、大きなスケールの構造パターンを捉えられるためと考えられる。すなわち、ポリマー物性予測ではモノマー水準を超えた特徴抽出が高い予測性能の達成に重要である。一方で、分子構造類似度に基づく外挿課題では、Mol_GIN は予測精度のばらつきが大きく、予測安定性に課題があることも示された。これは、本研究で扱った中で Mol_GIN のみが深層学習モデルであり、本研究規模の小さなデータセットではモデル複雑性により過学習が生じやすいことに起因すると考えられる。

〈データのバラエティとボリュームの評価〉

モデルの検証や精度向上に必要となるデータのバラエティとボリュームの評価に関して、MD 計算で得られる物理量は初期構造や平衡化過程，乱数シード等に依存し，系統誤差に加えて統計的なばらつきが不可避である。ハイスループットスクリーニングに耐える AI モデルを構築するには，学習データである MD 計算値の不確かさを可能な限り減らす必要がある。複数試行の平均化はサンプリング由来の統計的不確かさを低減しうるが，限られた計算資源の下で「試行回数を増やすべきか」「対象ポリマー種を広げるべきか」というトレードオフの最適解は未確立である。

そこで，対象ポリマー900 種類を選定し，各ポリマーについて独立 MD 試行を合計 9 回揃えるベンチマークを構築する。これによって、試行間の分散・標準誤差を評価することで，平均化がもたらす統計的不確かさ低減効果を定量化する。既存データには 1〜5 回の試行が含まれており，とくに約 750 種は 5 回，約 160 種は 4 回の試行結果を有する。資源配分の最適化は，同一総計算資源の下で条件を対比する設計により行う。すなわち，広い化学空間を浅く探索する「900 種類×1 回試行」と，狭い化学空間を深く反復する「100 種類×9 回試行」などを代表ケースとして並列に評価し，モデルの精度・外挿性・校正の観点からデータベース構築のための指針を示す。

本研究では、RadonPy ライブラリを用いて、平衡化計算から熱伝導率算出のための非平衡計算に至るまで、一連の計算を実施した。シミュレーションセル内のポリマー鎖数や平衡化手順を含む各種計算条件は、既存データセットに準拠させた。既存データに含まれる全 1,077 種のポリマーに対して DFT 計算を行い、原子電荷を取得できた 972 種を MD 計算の対象とした。これらの MD 計算は 16 コア並列で実行した。MD 計算では、初期構造によっては本プロトコルにおいて平衡化に到達しない場合があるほか、異常終了などによりデータ取得に失敗する場合も想定される。このため、全データに対して 6 回ずつ計算を実施した。さらに、これらの計算において熱伝導率の合計計算サンプル数が 9 回に達しなかった 74 種については、追加で 5 回の計算を実施した。最終的に、9 回の計算サンプルを 900 種類以上確保できなかった物性は熱伝導率のみであり、その対象数は 879 種であったため、これをデータの最大数として扱った。

次に、得られた MD 計算データを用いて物性予測モデルを構築した。具体的には、学習に用いるデータ数を 100、300、900、計算サンプル数を 1、3、9 とした 9 通りの条件について予測精度を評価し、各条件におけるモデル性能を比較した。予測精度の指標には、学習に用いた MD データのラベルの標準偏差で規格化した RMSE（Normalized RMSE）を用いた。予測モデルには、既報の QM 記述子および FF 記述子を用いた KRR を採用した。

図 10 に、density について、計算サンプル数を 1、

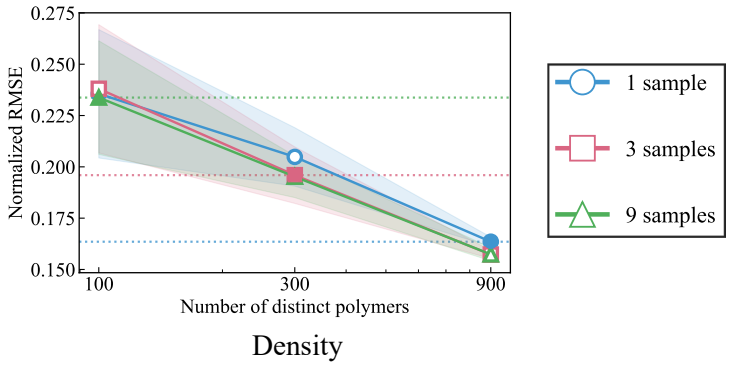


図 10 QM および FF 記述子を用いた KRR モデルにおける MD 計算 density の予測精度スケーリング。ポリマー種類数の増加に対する予測精度の変化を、計算サンプル数が 1、3、9 の場合について比較した。塗りつぶし記号は、計算資源が同等となる条件を表す。

3、9 とし、学習に用いるポリマー種数を 100、300、900 とした場合のスケーリングを示す。最終的な予測精度としては、ポリマー種の抽出方法およびサンプリングの組合せをそれぞれランダムに 5 回変更した場合の予測精度の平均値を示し、エラーバーにはその標準偏差を付した。

density については、いずれのサンプリング数においても予測精度がスケールすることが示された。また、同一の計算条件で比較した場合、1 回計算を 900 種のポリマーに対して実施した場合に最も高い予測精度が得られた。この結果は、サンプル数を増やす効果よりも、データの種類を増やす効果の方が大きいことを示している。density の場合は対象とした物性の中でも特に計算値の不確かさが小さかったため考えられる。

そのほかに，例えば bulk modulus に関しても予測精度がスケールすることが示された。また、同一の計算条件で比較した場合、3 回計算を 300 種のポリマーに対して実施した場合に最も高い予測精度が得られた。この結果は、サンプル数を増やす効果とデータの種類を増やす効果がともに同程度効いていることを示している。本研究で対象とした 15 種の物性のうち、3 回計算を 300 種のポリマーに対して実施した場合がもっとも高精度になるケースが一番多かった。

volume expansion に関しても予測精度がスケールすることが示された。また、同一の計算条件で比較した場合、9 回計算を 100 種のポリマーに対して実施した場合に最も高い予測精度が得られた。ポリマー種を増やすことよりもサンプルを集めて MD 計算値の不確かさを減らすことが効果的であることが示唆された。volume expansion や linear expansion は文献でも指摘されている通り MD 計算値ラベルの不確かさが大きく、したがってサンプル数を増やすことでその不確かさの影響を減らすことができたと考えられる。

Cp に関しては、サンプル数が 3 や 9 の場合で特にスケーリング傾向が見られた。同一の計算条件で比較した場合、9 回計算を 100 種のポリマーに対して実施した場合に最も高い予測精度が得られた。Cp の予測精度は対象とした物性の中でも少ないデータ種類で高精度予測が可能であり、データ種類の増加に対する影響力が小さかったためと考えられる。

さらに，MD 計算ラベルの不確かさと、同じ計算量でのデータのバラエティとボリュームの最適条件の対応関係について検討した。概ね MD 計算ラベルの不確かさが大きいと各ポリマー種で必要な計算回数が増える傾向にある。Cp や Cv のように、不確かさが小さいにも関わらず必要な計算回数が多い場合もあるが、これは先述のとおり、すでに高精度予測するための十分なデータ種類があるなど、データ種類増加による効果が少なくなっているためと考えられる。

以上の結果から、データのバラエティとボリュームの最適条件は MD 計算値の不確かさおよびデータ種類増加に伴うスケール度合によって決まると考えられる。本研究で作成されたデータセットは

MD 計算値データベースを構築するうえで、計算資源の有効活用方針の設定に重要な知見を提供しうる。

6. 進捗状況の自己評価と今後の展望

昨年度の計画に対する進捗は概ね順調である。量子化学計算および分子動力学計算の実行フェーズについては、予定していた初期ターゲットを 100%達成しており、さらなる拡張段階にある。今年度は、今後のさらなるデータベースの拡張に向けて、限られた計算資源の下で「試行回数を増やすべきか」「対象ポリマー種を広げるべきか」という点に着目し、延べ 6,200 種類相当の MD 計算を実施した。15 種類の物性に対して、特に内挿予測を想定した場合における試行回数とポリマー種の最適条件を算出した。今後は外挿タスクにおける最適条件の評価を行う。メゾ物理量を活用した AI モデルの構築に関しては、データ数に対するスケーリングを評価しつつ、数種類の主要物性において予測性能を確認しており（進捗 80%）、現在は対象物性の拡大とデータベース化（進捗 70%）を並行して進めている。
また、物性予測のみならず、原子数 13 以下の低分子を対象とした網羅的な分子生成手法の検討も開始しており、次世代の分子設計手法としての可能性を示した。

※7.研究業績はウェブ入力です