実験計画法による高分子物性自動計算と機械学習の融合

南條　舜　　総合研究大学院大学　統計科学専攻　博士課程 3年

1. 背景

材料設計のパラメータ空間は広大である. マテリアルズインフォマティクス（Materials Informatics: MI）の目的は，材料データとデータ科学の先進技術を活用し，広大な探索空間から革新的特性を持つ新材料を発見することである. MI の基本的なワークフローは，順問題と逆問題からなる[1]. 順問題では，材料の組成や構造 X から特性 Y を予測する入出力のデータ集合に教師あり学習を適用し，順方向の予測モデル Y = f(X) を構築する. 逆問題では，モデルの逆写像 $X \fallingdotseq f^{-1}(Y)$ を求め，所望の特性 Y=Y* を有する材料 X を予測する. データ駆動型材料研究における最も大きな壁は，体系的且つ包括的なデータの不足である. 特に，高分子材料のデータ資源の乏しさは際立っている. 現在の高分子物性データベースはいずれもデータの量が非常に少なく，たとえば高分子データベース（PoLyInfo）に物性値が公開されているホモポリマーの数は 19,000程度である. さらに，データ自動抽出のための API も整備されていない. また，我々が目指す“革新的な材料”の周辺には，そもそもデータが存在しない. したがって，限られたデータの壁を乗り越えるデータ科学の方法論が MI の基本問題の解決につながる. データの不足を補う手段の一つとして，材料研究では機械学習とシミュレーションの融合が重要な役割を担う. 本研究では，高分子材料（ポリマー）に目標を定め，データ駆動型材料研究における機械学習とシミュレーションの融合技術を創出することにより上記問題の解決を目指す.

2. 研究方法

限られたデータを乗り越えるための方法論として適応的実験計画に焦点を定め，ポリマー設計の機械学習[2]と全原子古典分子動力学シミュレーション（MD: molecular dynamics）による高分子物性自動計算の融合を図る. 類似研究として低分子化合物や単結晶のようなシンプルな系では，第一原理計算と機械学習を融合した物質探索の方法論やソフトウェアが既に確立されている. しかしながら，高分子材料系では，シミュレーションによる物性評価の自動化・高速化が技術的な障壁となり，全くもって研究が進んでいない. そこで，統計数理研究所が開発に成功した世界で唯一の高分子物性自動計算ソフトウェア RadonPy[3] と機械学習を融合することで，この壁を打ち破る. 具体的には下記の手順に従い，ベイズ最適化や能動学習等の実験計画法を適用し，計算機実験データを逐次的に生成する. なお，今回の実験においてサロゲートモデルにGaussian Process（GP）を使用し，入力には MDシミュレーションの入力パラメータをカーネル平均埋め込み[4]により固定長化した170次元のベクトルを用いた.

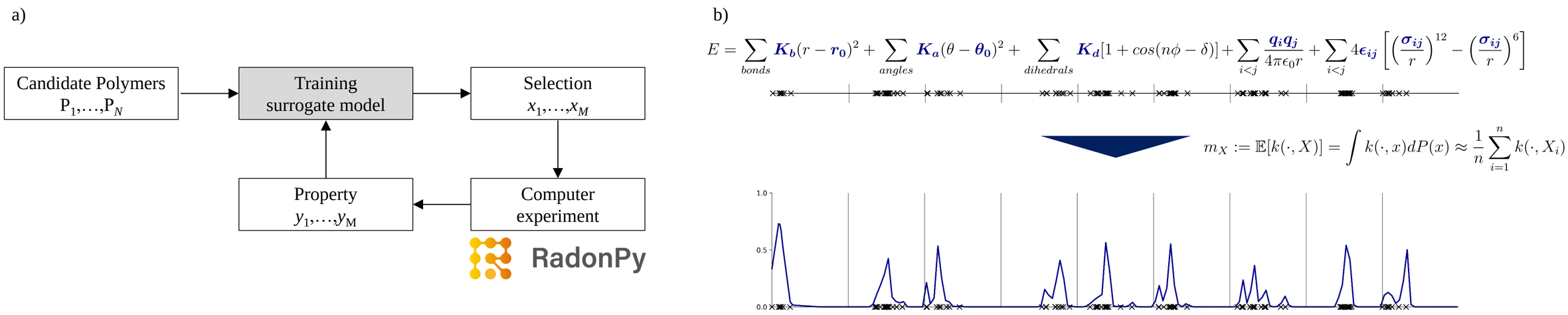

Figure 1. a）高分子物性計算と実験計画法の融合に関するワークフロー.　b）サロゲートモデルの訓練に用いた入力 xの計算手順.

3. 実験

上記ワークフローの動作検証として，二つの物性（定圧モル比熱：C_p，屈折率：Refractive index）が所望の領域に入るポリマーの探索を実施した. まず，両物性値が共に低い領域に存在する初期データ 10点をランダムに選定し，異なる目標領域を設定した独立の三つの実験を実施した. なお，GPモデルの訓練後に次のシミュレーション候補を選定する際の獲得関数 A(x)に関しては Probability of Improvement（PI）を用いた. 本実験において目標物性は二つ存在するため，PIはそれぞれの物性値の予測結果から得られた確率密度関数を所望の範囲で積分することに得られた. そして，獲得関数の大きいポリマーの選定，RadonPyを使用した物性計算，GPモデルの更新のサイクルを繰り返した.

$$A(x) = P(\alpha_1 < y_1 < \beta_1) \cdot P(\alpha_2 < y_2 < \beta_2)$$
$$= \int_{\alpha_1}^{\beta_1} p(y_1|x)\mathrm{d}y_1 \cdot \int_{\alpha_2}^{\beta_2} p(y_2|x)\mathrm{d}y_2$$

4. 結果

探索結果の推移を示す. 異なる目標領域（赤，青，緑）を与えたいずれの場合においても探索サイクル数が増えるにつれて目標領域およびその周辺に物性値が観測される様子を確認することができた. また，それぞれの目標領域に含まれるポリマーは共通の構造的特徴を有することも明らかとなった. さらに，シミュレーション候補ポリマーを獲得関数 PIに基づいて選定する場合とランダムに選定する場合の結果を比較したところ，前者の方が効率よく目標領域内に物性値が観測されることが示唆された.

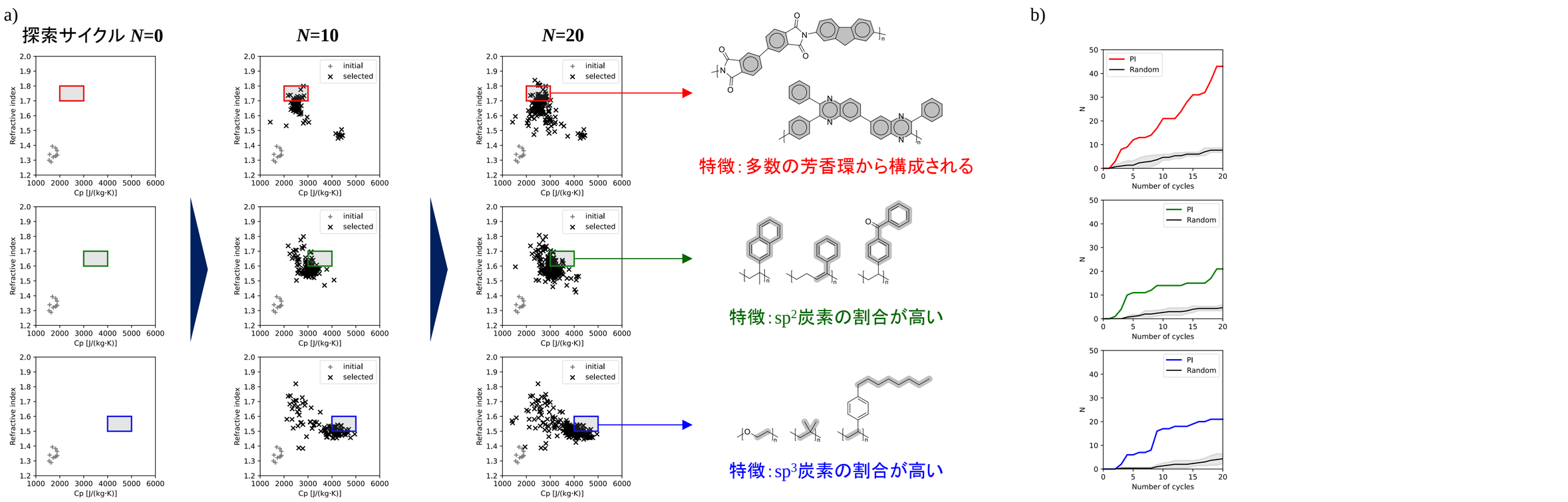

Figure 2. a）探索サイクル数 N=0, 10, 20における物性値の分布.　b）探索サイクル数と目標領域内に物性値が観測されたデータ数のプロット.

5. まとめと今後の予定

本研究ではデータ駆動型高分子材料研究における問題点の解決を目指し，RadonPyを用いた高分子物性自動計算と機械学習の融合に関する基盤技術構築および動作検証を実施した. また，詳細は省略するが，獲得関数にExpected Hypervolume Improvement（EHVI）を用いた探索も実施し，複数物性のパレート解を効率よく取得できることも確認した. 今後は構築した基盤技術を活用し，実際の材料開発における対象物性となる屈折率・アッベ数の経験的な限界線を越えるポリマーの発掘を目指す. また，本研究におけるワークフローではプールデータを固定した場合に焦点を絞ったが，プールデータには存在しない外挿領域に存在するポリマーを発見するための方法論開発にも着手したいと考えている.

[1] 吉田亮. マテリアルズインフォマティクス概説. 統計数理 69:5-33 (2021).

[2] Ikebata, H., Hongo, K., Isomura, T., Maezono, R., Yoshida, R., Bayesian molecular design with a chemical language model. Journal Computer-Aided Molecular Design. 31:379-391 (2017).

[3] Hayashi Y, Shiomi J, Morikawa J, Yoshida R. RadonPy: automated physical property calculation using all-atom classical molecular dynamics simulations for polymer informatics. npj Comput Mater. 2022;8(1):1-15.

[4] Muandet K, Fukumizu K, Sriperumbudur B, Schölkopf B. Kernel mean embedding of distributions: a review and beyond. FNT in Machine Learning. 2017;10(1-2):1-141.