データ駆動型高分子材料研究における統計的機械学習と分子シミュレーションの融合

南條 舜
総合研究大学院大学 複合科学研究科

1. はじめに

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。

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、たとえば既存のデータベース PoLyInfo[1]に物性値が公開されているホモポリマーの数は 18,000 程度である。その中で、同一の実験条件で測定された特定の物性値のデータに限定した場合、わずか 100 データにも満たない例も存在する[2]。

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。

機械学習のモデルは一般的に内挿的であり、データの存在しない領域の予測性能が大きく低下する。一方、物理法則に基づく分子シミュレーションは、未踏領域の材料特性をある程度予測できることが期待される。そこで、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分子シミュレーションの融合が重要な役割を担うと考えられる。

本研究では高分子材料に目標を定め、データ駆動型材料研究における機械学習と分子シミュレーションの融合技術を創出し、高分子科学の研究者らと共同で概念実証を行うことを目指した。

2. 手法

はじめに、ベイズ最適化[3]・能動学習[4]等の適応的実験計画法を用いて機械学習と分子シミュレーションによる計算機実験を系統的に循環させるワークフローを実装した。図 1 に示すように、現在観測済みの分子シミュレーション結果を訓練データとしてサロゲートモデルを訓練し、獲得関数に基づき次に分子シミュレーションを行う候補を選定する。そして、選定された候補の分子シミュレーションを実施し、結果を訓練データに追加する、というサイクルを繰り返す。

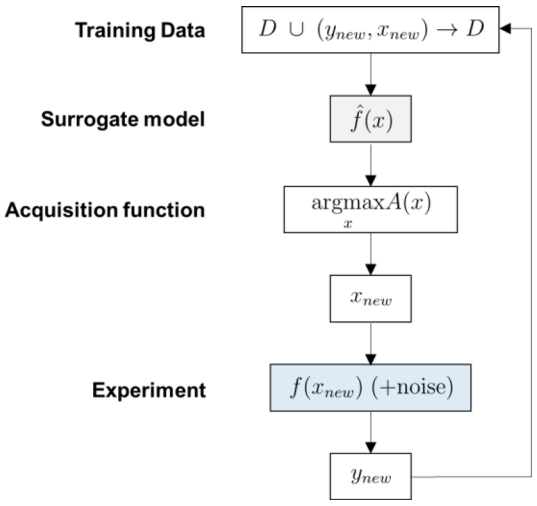

図 1：適応的実験計画法と
分子シミュレーションの融合ワークフロー

本ワークフローにより、現時点でデータが存在しない外挿領域に分子シミュレーションの新しいデータを作り出す。そして、このデータを含めてモデルを訓練することで外挿的予測性能を獲得し、モデルの適応範囲が徐々に拡大することが期待される。

関連する研究として、低分子化合物や単結晶において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[5-6]。しかしながら、高分子材料系では分子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、研究が進んでいない。そこで、近年公開された高分子物性計算の全自動化ライブラリ RadonPy[7]を組み合わせることにより問題解決を図った。RadonPy は全原子古典分子動力学（Molecular Dynamics : MD）シミュレーションによる高分子物性計算を全自動化するオープンソースソフトウェアであり、高分子の繰り返し単位を構成する分子構造、重合度、温度等の計算条件を入力とし、熱物性や光学特性などの 17 種類の物性を自動計算するアルゴリズムが実装されている。そこで、図 1 のワークフローをスーパーコンピュータ SQUID 上で実装することにより、物性計算の自動化・高速化が可能となり、前述の技術的な壁を打ち破ることができた。

3. 実験

実装したワークフローの性能評価の一例として、光学用高分子の探索を実施した。光学用高分子はメガネやカメラレンズ等の様々な製品に用いられる材料であり、その主な要求特性は高屈折率・高アッベ数である。しかしながら、両物性の間には経験的な限界線が知られており、限界線を越える高分子はほとんど存在しないことが知られている[8]。ここで、RadonPy を用いた屈折率・アッベ数の分子シミュレーション結果が文献値を良く再現できることを事前に確認できたので、実装したワークフローを用いたプール型のベイズ最適化を実施することにより、経験的な限界線を越える高分子の発掘を目指した。

本実験では、サロゲートモデルに通常のガウス過程回帰モデルを使用し、モデルの訓練に用いる入力 x は物性計算（分子シミュレーション）のパラメータをカーネル平均埋め込み[9]により固定長化した 170 次元のベクトルを使用した。また、仮想高分子のプールについては確率的言語モデルにより生成された仮想高分子 1 万構造を用いた。さらに、獲得関数については期待超体積改善量[10]を用いた。期待超体積改善量は、予測分布からのサンプルが与えられた際のパレート超体積増加量の期待値であり、本実験のように目的変数が二変数の場合は式(1)から解析的に計算可能である。

$$A(x) = \iint \Delta\mathrm{HVI} \cdot p(y_1|x) \cdot p(y_2|x)\mathrm{d}y_2 dy_1 \quad (1)$$

なお、ベイズ最適化の一回のサイクルにつき、獲得関数の上位 10 個の高分子の分子シミュレーションを実施した。

4. 結果

図 2 に実装したワークフローを用いて収集された高分子の屈折率およびアッベ数の分子シミュレーション結果の推移を示す。ベイズ最適化のサイクルを繰り返すにつれて、経験的な限界線を越える高分子の数が徐々に増える様子を確認できた。ここで、経験的な限界線を越える高分子の構造パターンを解析した結果、約 4 割が部分構造に硫黄原子を含むことが明らかとなった。さらに、その中でスルホニル基（-SO_2-）を含む構造が複数確認された。過去の合成実験の知見[11-12]によるとスルホニル基の分子屈折と分子分散の比が大きいことから、高分子の部分構造にスルホニル基を導入することにより屈折率とアッベ数をともに向上できることが実証されている。そのため、今回の実験で発掘された

スルホニル基を有する高分子を実際に合成した場合においても、経験的な限界線を越えることが期待される。以上、材料科学の事前知識を活用しないデータ駆動型手法により、実験結果がほとんど存在しない領域に存在する高分子の候補を発掘することができた。

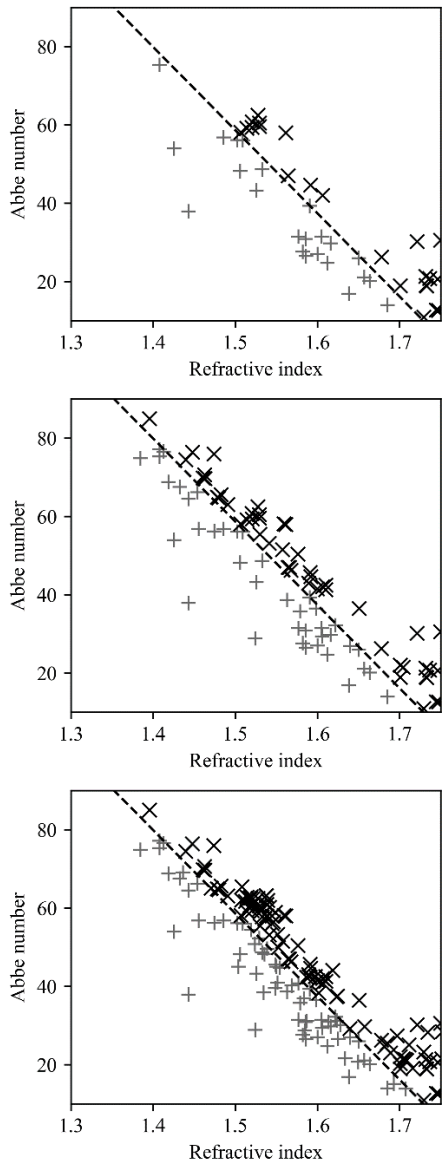

図 2：ベイズ最適化のサイクルを N 回繰り返すことにより収集された高分子の分子シミュレーション結果。上段は N=5、中段は N=10、下段は N=20 の場合を表す。点線は文献[8]に記載された屈折率とアッベ数の経験的な限界線を表す。

5. おわりに

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を解決するための一つの方法として、機械学習と分子シミュレーションの融合ワークフローを実装した。今回の報告では屈折率とアッベ数を対象とした計算機実験結果に焦点を絞ったが、RadonPy を用いて計算可能な 17 物性、もしくは RadonPy の計算結果から転移可能な物性であれば適用できるため、本研究の学術的価値としては汎用性が高いことが考えられる。

今後は高分子科学の研究者らと協同し、今回の実験において発掘された高分子、もしくはその類似高分子が現実世界においても経験的な限界線を越えることを実証したいと考えている。

参考文献

(1) S. Othuka, et al., International Conference on Emerging Intelligent Data and Web Technologies., 22-29, (2011).
(2) W. Stephen, et al., Npj Computational Materials., **5**, 66, (2019).
(3) E. Brochu, et al., arXiv preprint arXiv:1012.2599 (2010).
(4) DA. Cohn, et al., Journal of artificial intelligence research, **4**, 129-145 (1996).
(5) S. Ju, et al., Physical Review X, **7**, 021024 (2017).
(6) G. Agarwal, et al., Chemistry of Materials, **33**, 8133-8144, (2021).
(7) Y. Hayashi, et al., Npj Computational Materials., **8**, 222, (2022).
(8) S. Ando, et al., Japanese journal of optics, **44**, 298-303, (2015).
(9) K. Muandet, et al., Foundations and Trends® in Machine Learning, **10**, 1-141 (2017).
(10) K. Yang, et al., Journal of Global Optimization, **75**, 3-34 (2019).
(11) R. Okutsu, et al., Macromolecules, **41**, 6165-

6168 (2008).
(12) Y. Suzuki, et al., Macromolecules, **45**, 3402-3408 (2012).