統計数理研究所 75 周年記念事業（2019）

機械学習による新物質の発掘

Machine learning for accelerated materials discovery

データ科学研究系 吉田 亮（Ryo Yoshida）

キーワード：マテリアルズインフォマティクス，機械学習，ベイズ推論，転移学習

1. マテリアルズインフォマティクス

MI の問題の多くは，順問題と逆問題の形式に帰着する．順問題の目的は, 系の入力 S に対する出力 Y の予測である．物性予測の文脈では，入力は物質（分子，組成，結晶等），出力は物性値（エネルギー，電子状態等）に相当する．これまでの材料研究では，第一原理計算や分子動力学計算等の理論計算が順方向の予測を担ってきた．このタスクをデータ科学のモデルに代替させることが，MI の中心的課題のひとつである．これに対し，逆問題では文字通り逆方向の予測を行う．すなわち，出力 Y の値（例えば目標物性）を設定した上で，それを達成する入力 S の状態（構造）を予測する．データ科学の観点からみると，これらの計算は，物質構造の“表現・学習・生成”を行うことに相当する．記述子と呼ばれる特徴ベクトルを通して物質の構造を“表現”し，データのパターンから構造から物性の数学的写像を“学習”する．さらに，計算機を用いて所望の物性値を有する物質を“生成”し，有望な候補物質を炙り出す．対象となる入力 S は，分子，組成，結晶，混合物，プロセス，合成経路等，問題に応じて多様な形式をとりうる．

2. 機械学習によるハイスループットスクリーニング

構造と物性の関係を表す実験や理論計算のデータから，物質 S の物性 Y の予測モデル f(S) を導くことが目的である．記述子と呼ばれる特徴ベクトルを通して物質の構造を“表現”し，データのパターンから構造から物性の数学的写像を“学習”する．記述子は MI における最も基本的な要素技術である．入力 S の形式が多様であることから，対象領域ごとに独自に研究が展開している．膨大な数の候補物質のライブラリを作製した上で，訓練済みモデルを用いてスクリーニング実験を実施する．実験や理論計算に比べて機械学習のモデルは圧倒的に計算コストが低いため，膨大な数の候補物質を対象とする物性評価を行うことができる．

3. 転移学習

機械学習の他の応用領域に比べて，材料研究のデータ数は圧倒的に少ない．データ科学が本格的に導入されて間もないこともあり，データベースの整備は発展途上の段階にある．とりわけ，研究対象が最先端に近づくにつれて，スモールデータの傾向はより顕著になる．スモールデータに対する解決策として，転移学習と呼ばれるアプローチが有望視されている．転移学習では，あるタスクの訓練済みモデルを別のタスクに再利用する．我々は，XenonPy.MDL という訓練済みモデルライブラリを開発している（Yamada et al. (2019)）．低分子化合物，高分子，無機結晶等，様々な物質に対する>105 の物性推算モデルが収録されている．図 1 は，ニューラルネットワークの転移学習に基づくポリマー定圧熱容量（Cp）の予測を例示したものであ

る．第一原理計算で低分子化合物の化学構造と定容熱容量（Cv）の関係を表す 133,805 個のデータを取得し（Ramakrishnan et al. (2014)），ソースモデルを導いた．ソースモデルの部分ネットワークを用いて特徴量の計算を行い，高分子材料データベース PoLyInfo（Otsuka et al. (2011)）に登録されている 58 個の Cp のデータを用いて予測モデルを構築した．ソースタスクの学習過程で Cv と Cp に共通する縮約特徴量を獲得し，これを用いることでたった 58 個のデータからポリマー Cp の予測モデルを導くことに成功した．

4. ベイズ推論による物質構造の設計

我々は，ベイズ推論や機械学習を方法論の基軸とし，新物質の創製を目的に設計と合成を対象とする機械学習の手法とソフトウェアを開発してきた（Ikebata et al. (2017)）．実験やシミュレーションから得られるデータを用いて，物質の構造から物性の順方向の予測モデルを構築する．これに条件付き確率のベイズ則を適用し，物性から構造の逆方向のモデルを導き，このモデルから仮説物質を発生させることで，所望の物性を有する埋蔵物質を炙り出す．確率的言語モデルに基づく構造生成器や機械学習の様々な技術を結集させて構築した確率推論のアルゴリズムである．現在，この手法を用いて様々な材料系を対象に実証研究を進めている（Wu et al. (2019) など）．機械学習で物質構造の設計図が描き，大量の埋蔵物質を発掘する．これが本研究のグランドチャレンジである．

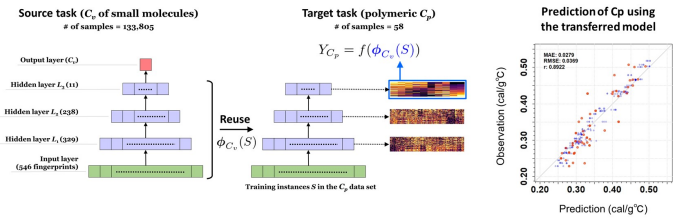

図 1．転移学習によるポリマー定圧熱容量（Cp）の予測

参 考 文 献

Ramakrishnan, R., Dral, P. O., Rupp, M. and von Lilienfeld, O. A. (2014). Quantum chemistry structures and properties of 134 kilo molecules, Scientific Data, 1, 140022.

Otsuka, S., Kuwajima, I., Hosoya, J., Xu, Y. and Yamazaki, M. (2011). PolyInfo: Polymer database for Polymeric Materials Design, 2011 International Conference on Emerging Intelligent Data and Web Technologies, 22-29.

Yamada, H., Liu, C., Wu, S., Koyama, Y., Ju, S., Shiomi, J., Morikawa, J. and Yoshida, R. (2019). Transfer learning: accelerated materials discovery with small data.

Ikebata, H., Hongo, K., Isomura, T., Maezono, R. and Yoshida, R. (2017). Bayesian molecular design with a chemical language model, Journal of Computer-Aided Molecular Design, 31(4), 379-391.

Wu, S., Kondo, Y., Kakimoto, M., Yang, B., Yamada, H., Kuwajima, I., Lambard, G., Hongo, K., Xu, Y., Shiomi, J., Schick, C., Morikawa, J. and Yoshida, R. (2019). Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm.