広島県立総合技術研究所東部工業技術センター研究報告　No.37（2024）
技術ノート

転移学習による小規模データに対する化合物探索技術

渡邉正宗

Study of a method for chemical discovery from small dataset by transfer learning

WATANABE Shoso

機械学習の技術を活用し、材料開発の高速化および高度化を図るマテリアルズ・インフォマティクス（MI）と呼ばれる分野が注目を集めている。しかし材料データを取得するためのコストは実験・計算機実験を問わず高いため、機械学習を適用するためのデータ量が揃わない等、課題も多い。

本研究では、少数データから外挿的な予測を行うための手法として、他のドメインから取得したモデルや特徴表現を目標ドメインの予測に利用する転移学習に注目した。今回は、イソプレンゴムの補強のための表面修飾物質の探索を題材とし、転移元のドメインとして低分子化合物のデータベース QM9 から取得した熱容量データを、転移先のドメインとして分子動力学シミュレーションにより算出したイソプレン分子との相互作用エネルギーを使用した。分子構造の特徴表現にはグラフ畳み込みニューラルネットワークを使用した。

キーワード：転移学習、深層学習、グラフ畳み込みニューラルネットワーク、分子動力学

1. 緒　　言

機械学習の技術を活用し、材料開発の高速化および高度化を図るマテリアルズ・インフォマティクス（MI）と呼ばれる分野が注目を集めている[1)]。しかし材料データを取得するためのコストは実験・計算機実験を問わず高いため、データの量および多様性に乏しくなる場合が多い。そのようなデータセットから内挿的な予測モデルを作成しても、所望の性能を持つ新規材料にたどり着く可能性は低く、機械学習を適用するうえでの課題となっている。

本研究では、少規模データから外挿的な予測を行うための手法として、他のドメインから取得したモデルや特徴表現を目標ドメインの予測に利用する転移学習[2)]に注目した。今回は、イソプレンゴムの補強のための表面修飾物質の探索を題材とし、転移元のドメインとして低分子化合物のデータベース QM9[3)]からランダムに抽出した 5,000 分子の 298.15K における熱容量データを使用した。なお、QM9 は 9 原子以下の炭素、窒素、酸素、フッ素を骨格原子として含む分子に対する第一原理計算の物性データベースである。転移先のドメインとしては、QM9 からランダムに抽出した 32 分子について、イソプレンモノマーとの相互作用エネルギーを使用した。相互作用エネルギーは分子動力学（MD）シミュレーションにより算出した。

分子構造の特徴を表現する手法は種々提案されているが、本研究ではグラフ畳み込みニューラルネットワーク（Graph Convolutional Network、GCNN）を使用した。GCNN はグラフの局所構造に対して畳み込み演算を行う手法である。分子は原子と原子の結合グラフとして表現できるため、GCNN により分子構造を考慮して学習を行うことが可能である。

2023.9.19　材料技術研究部　受理

2. 実験方法

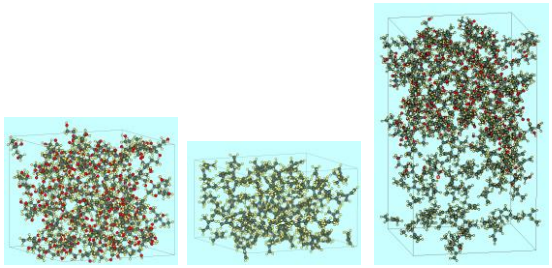
2.1 相互作用エネルギー計算

モデルの作成に Winmostar7[4)]、計算プログラムに LAMMPS[5)]、力場として DREIDING[6)]を用い、 全原子 MD シミュレーションを実施した。**図 1** に相互作用エネルギーの計算モデルを示す。QM9 からランダムに抽出した 32 種類の低分子化合物とイソプレンモノマーとの相互作用エネルギーEは次式に従い求めた。

$$E = E_{tot} - (E_{chem} + E_{ir})\ (kcal/mol) \qquad (1)$$

ここでE_{chem}は低分子化合物 100 分子の系の全エネルギー、E_{ir}はイソプレンモノマー100 分子の系の全エネルギー、E_{tot}は低分子化合物 100 分子とイソプレンモノマー100 分子からなる系の全エネルギーである。
MD シミュレーションは、全て周期境界条件下、Nose-Hoover 法[7)、8)]により 300K に温度制御し、1fs の時間刻みで実施した。

単独分子の全エネルギーは、縦横 30Å、高さ 200Å の異方セルに 100 分子をランダム配置し、Parrinello-Rahman 法[9)]により高さ方向に 200atm の圧力制御を伴う異方圧縮を 100ps 行ったのち、100ps の NVT アンサンブルにて算出した。その後、高さ方向に低分子化合物およびイソプレンモノマーのセルを連結し、同様に 200atm の異方圧縮を 100ps 行ったのち、100ps の NVT アンサンブルにて全エネルギーを算出した。



E_{chem}　　E_{ir}　　E_{tot}

図 1 相互作用エネルギーの計算モデル

2.2 GCNN による熱容量の回帰モデル

QM9 からランダムに抽出した 5,000 件のデータセットのうち、4,000 件を学習データとして使用し、残りの 1,000 件を検証用のデータとして使用した。

分子構造は SMILES（simplified molecular input line entry system）と呼ばれるアルゴリズムによって表現された文字列として読み込み、パッケージ DGL-Life[10]に含まれる関数により、原子の種類と結合の種類によって表現される 74 ビットのバイナリベクトルに変換し、入力とした。

図 2 に本研究の GCNN の構成を示す。入力層、畳み込み層 2 層、プーリング層 1 層、全結合層 3 層、出力層からなる。活性化関数には ReLU、最適化アルゴリズムには Adam を使用し、学習率は 0.001 とした。

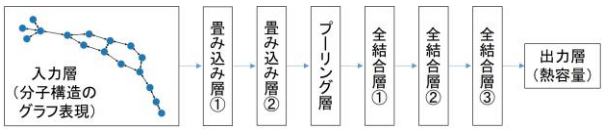


図 2 GCNN の構成

2.3 転移学習による相互作用エネルギーの回帰モデル

MD シミュレーションによりイソプレンモノマーとの相互作用エネルギーを算出した 32 件の化合物のうち、24 件を学習データとして使用し、残りの 8 件を検証用のデータとして使用した。

図 3 に本研究の転移学習の構成を示す。畳み込み層およびプーリング層については熱容量の回帰モデルを流用し、全結合層および出力層について再学習を行うファインチューニングとし、学習率は 0.0001 とした。

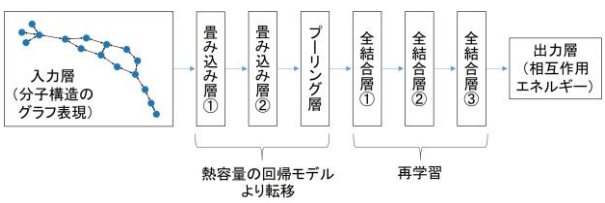


図 3 転移学習の構成

3. 結果と考察

3.1 GCNN による熱容量の回帰モデル

図 4 にエポック数 5,000 における予測値-文献値のプロットを示す。決定係数 R^2 は、学習データで 0.978、検証データで 0.964 となり、高い相関モデルを得ることができた。QM9 のように件数が多く、分子量の範囲が狭く、多様な構造を含んだデータベースに適用する場合、GCNN は分子構造の特徴抽出に有力な手法であることが確認できた。

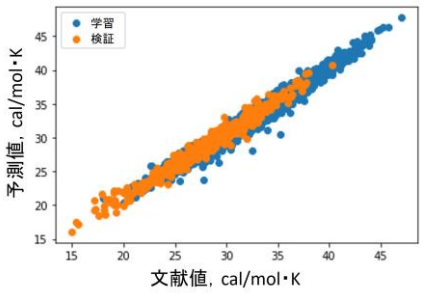


図 4 熱容量の予測値-文献値プロット

3.2 転移学習による相互作用エネルギーの回帰モデル

図 5 にエポック数 10,000 における予測値-計算値のプロットを示す。決定係数 R^2 は、学習データで 0.881、検証データで 0.659 となった。予測精度は転移元のモデルに劣り、また過学習が起きているものの、一定水準の相関を得ることができた。

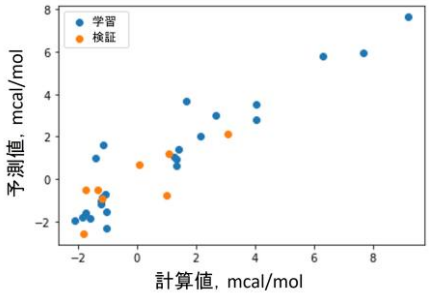


図 5 相互作用エネルギーの予測値-文献値プロット

4. 結　　言

本研究では、MI における課題の 1 つである、データ件数が少ない場合での機械学習の適用手法として、転移学習に着目した。転移元が QM9 のような、限られた骨格および分子量で、かつ大規模なデータセットであれば、GCNN により効果的に分子構造の特徴を抽出することができ、得られたモデルのプーリング層までを流用することで、小規模データに対しても一定水準の予測器を作成することができた。

今回、転移先の目的変数にはMDシミュレーションにより算出したイソプレンモノマーとの相互作用エネルギーを使用した。このようなシミュレーションには相応の計算コストが伴い、数百件～数千件のデータセットを揃えることは困難であるが、転移学習により小規模のデータセットでも有効となるのであれば、シミュレーションにより算出できる物性や性能全てに機械学習が適用できる可能性がある。ただし転移学習の有効性は転移元のデータセットの規模や多様性に左右されるため、今後はQM9の空間以外の物質について適用範囲を拡げることが課題と考えている。

文　　献

1) Ramprasad, R. *et al.*：Machine learning in materials informatics：recent applications and prospects, npj Comput Mater, **3**, 54 (2017).
2) Yamada, H. *et al.*：Predicting Materials Properties with Little Data Using Shotgun Transfer Learning, ACS Central Science, **5**(10), 1717-1730 (2019).
3) Ramakrishnan, R. *et al.*：Quantum chemistry structures and properties of 134 kilo molecules, Sci Data, **1**, 140022 (2014).
4) Winmostar, https://winmostar.com/
5) LAMMPS, https://www.lammps.org/
6) Mayo, SL., Olafson, BD. & Goddard, WA.：DREIDING: a generic force field for molecular simulations, The Journal of Physical Chemistry, **94**(26), 8897-8909 (1990).
7) Hoover, WG.：Canonical dynamics: Equilibrium phase-space distributions, Phys Rev A Gen Phys, **31**(3), 1695-1697 (1985).
8) Nose, S.：A molecular dynamics method for simulations in the canonical ensemble, Molecular Physics, **52**(2), 255-268 (1984).
9) Parrinello, M. & Rahman, A.: Crystal Structure and Pair Potentials: A Molecular-Dynamics Study, Phys. Rev. Lett., **45**(14), 1196-1199 (1980).
10) Li, M. *et al.*：DGL-LifeSci: An Open-Source Toolkit for Deep Learning on Graphs in Life Science, ACS Omega, **6**(41), 27233-27238 (2021).