5ZJ-02

対照学習を用いたGNNによる化合物特性予測

青木 滉志郎　　Kengkanna Apakorn　　大上 雅史
東京工業大学 情報理工学院 情報工学系

1　序論

創薬プロセスの効率化を目的として，薬剤候補となる化合物の物性や，標的分子に対する活性（これらを化合物特性と呼ぶ）を機械学習によって予測するバーチャルスクリーニング技術が活用されている．化合物は原子と結合情報を含んだグラフによって表現できるため，入力をグラフ情報と捉えたグラフニューラルネットワーク (Graph Neural Network, GNN) による予測モデルの精度が比較的高いことが知られている [1]．一方で，目的変数となる特性の値が得られているデータが少ないケースも多く，その場合は予測モデルがうまく構築できない．そのため，大量に存在するラベルの無い化合物から学習を行う事前学習が注目されている [2]．特に，言語処理や画像処理分野においては，自己教師あり学習による事前学習手法の一つである対照学習 (contrastive learning) が注目されている．対照学習は，似た化合物の埋め込み表現を近付け，異なる化合物の埋め込み表現を遠ざけるように学習を行う．目的変数のラベルが少ない状況において，大規模なデータセットの教師あり学習に匹敵する性能が期待できる．

そこで本研究では，対照学習を用いたGNNによって高精度な化合物特性予測を実現することを試みた．対照学習においては正例/負例埋め込みを構成するaugmentの方法，GNNではアーキテクチャやグラフ表現の方法などに任意性があり，予測性能の高いモデルの構築に寄与する要素の検証も行った．

2　手法

2.1　提案手法の概要

ラベルの無い化合物データを用いてGNNによる対照学習を行い，学習したパラメータをラベル付きデータでファインチューニングして特性予測モデルを構築した．その上で，対照学習でのaugmentの方法を変化させ，予測精度を評価した．モデルにはGraph Convolutional Network (GCN) [3] を使った対照学習手法の実装である MolCLR [4] を用いた．

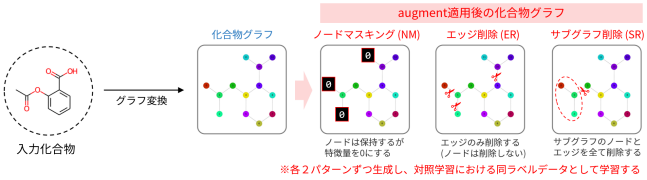

図1　化合物グラフとaugment方法

2.2　化合物のaugment

化合物の原子と共有結合をグラフで表現した化合物グラフに対して以下のaugmentを適用し，各化合物ごとに図1のようなグラフ変換を行った．全ての化合物に関してそれぞれのaugmentによる変換を行った2つのグラフを生成し，それらを同ラベルとした対照学習を行った．

1) **ノードのマスキング** (NM)　グラフを構成するノードの25%をランダムにマスキングする
2) **エッジの削除** (ER)　グラフを構成するエッジの25%をランダムに削除する
3) **サブグラフの削除** (SR)　グラフの全ノードの25%分のノードを持つサブグラフをランダムに削除する

2.3　GNNによる畳込みと対照損失の計算

augmentグラフを5層の畳込み層を持つGCNで更新し，平均プーリングでグラフ単位での特徴ベクトルを求めた．さらに特徴ベクトルの次元を削減するためにGCNの出力を多層パーセプトロンで変換した．得られたベクトルに対して，ベクトル間のNT-Xent（正規化温度スケールクロスエントロピー）損失を算出した．以上の操作を100エポック繰り返し，同じ化合物から生成されたグラフペアの損失（対照損失）が小さくなるようにモデルを学習した．

Chemical property prediction by graph neural network with contrastive learning
Koshiro Aoki, Apakorn Kengkanna & Masahito Ohue, Department of Computer Science, School of Computing, Tokyo Institute of Technology

表 1　分類予測での化合物特性予測結果 (ROC-AUC ↑)

	BBBP	BACE	CYP2C8	Hepatotoxicity	AmesMutag
事前学習なし	0.659	0.778	0.861	0.757	0.872
augment **なし**	0.722	0.786	0.887	0.742	0.868
ER	**0.728**	**0.790**	0.892	0.764	0.874
SR	0.714	0.783	**0.896**	**0.795**	**0.878**
NM+ER	0.714	0.787	0.864	0.794	**0.878**

表 2　回帰予測での化合物特性予測結果 (RMSE ↓)

	FreeSolv	ESOL	Lipo	HumanPPB	AqSolDB
事前学習なし	4.522	1.473	0.770	0.151	1.261
augment **なし**	3.640	**1.441**	0.757	0.149	1.158
ER	3.112	1.508	0.752	0.146	**1.118**
SR	5.724	1.473	**0.750**	0.145	1.153
NM+ER	**2.112**	1.456	0.753	**0.144**	1.144

2.4　ファインチューニング

ファインチューニングは事前学習モデルと同一のアーキテクチャを用いた．対照学習で学習された GNN のパラメータを用いて，特徴量を更新して学習を行った．ファインチューニングでは 300 エポックの学習を行い，モデルの予測性能を検証した．

2.5　データセット

対照学習におけるラベル無しデータには ChemBERTa [6] で用いられた PubChem-10M データセットの化合物を用いた．予測対象となるラベル付きデータは，化合物特性予測データセット集である MoleculeNet [7] から BBBP, BACE, FreeSolv, ESOL, Lipo の 5 種を，その他の創薬関連データセットから CYP2C8 [8], Hepatotoxicity [9], AmesMutag [10], HumanPPB [11], AqSolDB [12] を扱った．対照学習ではデータを 95 : 5 の割合でランダムに分割した．またファインチューニングでは MoleculeNet タスクでは 80 : 10 : 10 の割合で scaffold 分割を行い，その他は 80 : 10 : 10 のランダム分割を行った．

3　結果

表 1 および表 2 に，対照学習による事前学習を用いた予測モデルの精度を示す．分類タスクでは ROC-AUC 値を，回帰タスクでは RMSE 値を評価指標に用いた．事前学習しない場合や，対照学習でない通常の事前学習（augment なし）と比較して，ESOL 以外のデータセットで対照学習による予測精度の向上が確認できた．今回検証した中ではデータセットによって最高精度を得た augment 手法は異なっていたが，ノードマスキングとエッジ削除を組合せた augment (NM+ER) が総合的に精度が高かった．

4　結論

本研究では, 化合物特性予測における GNN を使った対照学習の効果を検証するため，事前学習の有無での性能の比較と対照学習における augment 手法による性能の比較を行った．その結果，事前学習とグラフの augment が化合物特性予測の精度向上に寄与することを示した．

一般的な化合物の特性予測ベンチマークデータセット以外でも検証を行ったことで，データセットに依らず効果がある可能性が示唆されたが，一方で改善の小さかったデータセットにおける予測結果の吟味や，ハイパーパラメータの選択など，より詳細な検証を進める必要があり，今後の課題としたい．

謝辞 本研究は，JST 創発的研究支援事業 (JPMJFR216J)，JST ACT-X (JPMJAX20A3) の支援を受けて行われた．

参考文献

[1] Xu K, *et al.* How powerful are graph neural networks? In *ICLR 2019*, 2019.
[2] Chen T, *et al.* A simple framework for contrastive learning of visual representations. In *ICML 2020*, 149, 1597–1607, 2020.
[3] Kipf TN & Welling M. Semi-supervised classification with graph convolutional networks. In *ICLR 2017*, 2017.
[4] Wang Y, *et al.* Molecular contrastive learning of representations via graph neural networks. *Nat Mach Intell*, 4, 279–287, 2022.
[5] Kim S, *et al.* PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res*, 47, D1102–D1109, 2019.
[6] Chithrananda, S., *et al.* ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. *arXiv preprint arXiv:2010.09885*, 2020.
[7] Wu Z, *et al.* MoleculeNet: a benchmark for molecular machine learning. *Chem Sci*, 9, 513–530, 2018.
[8] Zhang X, *et al.* In Silico Prediction of CYP2C8 Inhibition with Machine-Learning Methods. *Chem Res Toxicol*, 34, 1850–1859, 2021.
[9] He S, *et al.* An in silico model for predicting drug-induced hepatotoxicity. *Int J Mol Sci*, 20, 1897, 2019.
[10] Hansen K, *et al.* Benchmark dataset for in silico prediction of Ames mutagenicity. *J Chem Inf Model*, 49, 2077–2081, 2009.
[11] Lou C, *et al.* IDL-PPBopt: A Strategy for Prediction and Optimization of Human Plasma Protein Binding of Compounds via an Interpretable Deep Learning Method. *J Chem Inf Model*, 62, 2788–2799, 2022.
[12] Sorkun MC, *et al.* AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds. *Sci Data*, 6, 143, 2019.