TSUBAME 共同利用 令和 2 年度 学術利用 成果報告書

利用課題名： 多様なデータを活用する深層学習モデルの検証

英文： Development of Computational Science and Technology for Functional Property Analysis

利用課題責任者 後藤 仁志
First name Surname Hitoshi Goto

所属 豊橋技術科学大学
Affiliation Toyohashi University of Technology
URL https://www.tut.ac.jp/

邦文抄録（300 字程度）

医薬候補化合物の活性や，新規機能材料の物性の予測にはグラフニューラルネット（GNN），野菜市場価格の予測に RNN/LSTM 等の機械学習モデルを適用するなどして，比較的高い予測精度を実現している。本申請課題では，さらに予測精度を向上させるために，多様なデータを大規模に収集し，正当な精査を適用し，最先端の機械学習モデルの検証を行った。令和 2 年度は活性予測精度の向上に関する取り組みを主として計算を行った。成果として，ハイパーパラメータ変更による予測精度への影響や，ターゲットタンパク質の違いによる予測の性質を捉えることができた。

英文抄録（100 words 程度）

Recurrently high prediction accuracy is achieved by applying graph neural networks (GNN) to predict the activity of drug candidate compounds and physical properties of new functional materials, and machine learning models such as RNN / LSTM to predict vegetable market prices. It has been realized. In this application, in order to further improve the prediction accuracy, we collected various data on a large scale, applied proper scrutiny, and verified the most advanced machine learning model. In this year, calculations were mainly made for efforts to improve the accuracy of activity prediction. As a result, we were able to grasp the effect of hyperparameter changes on prediction accuracy and the nature of prediction due to differences in target proteins.

Keywords: DNN, GNN, RNN/LSTM, QSAR

背景と目的

深層学習法は物性/活性予測にも有効であることが示されており，さらに時系列データの予測にも効果を発揮することが分かってきた。我々はこれまで医薬候補化合物の活性や，新規機能材料の物性の予測にはグラフニューラルネット（GNN），野菜市場価格の予測に RNN/LSTM 等の機械学習モデルを適用し，予測精度のさらなる向上を目指している。

令和 2 年度のプロジェクトでは特に活性予測精度の向上ならびに予測機序の解明を中心として計算を行い，成果として予測精度の向上を確認し，ターゲットタンパク質の違いによる予測の性質を捉えることができた。

概要

我々の活性予測に関するプロジェクトでは重大な疾患に関わる 15 種類の標的タンパク質に対する数千もの化合物の分子構造データ（記述子）と活性値から構造-活性モデルを構築し，より高い活性値を有する新たな化合物を探索するための生理活性予測システムを開発している。一般に，全結合多層ニューラルネットワークは広範囲な予測問題に適用できる汎用性の高い機械学習法である。本プロジェクトでは活性予測のために本手法を用いており，ハイパーパラメータの最適化を行うことによってより高精度な予測を目指している。

結果および考察

学習の繰り返し回数である epoch に関しては，以前までは 3000，試行回数を最低 3 回と定義していた。Figure 1 は各ターゲットのデータセットを用いた DNNs 予測において，どの epoch タイミングで最良の R^2（データセットに含まれる実験により求められた活性値とそれに対応するシステムによる予測値間の相関係数）が出現したかを示している。なお，試行回数は 10 回行っており，最大 epoch は 4000 としている。結果として，最良

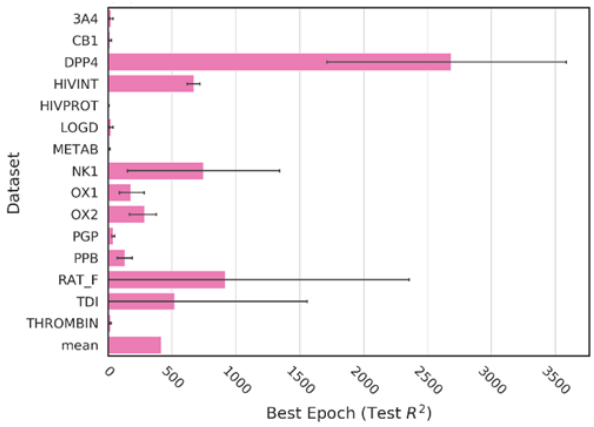

Figure 1 Appearance of Best Value Epoch

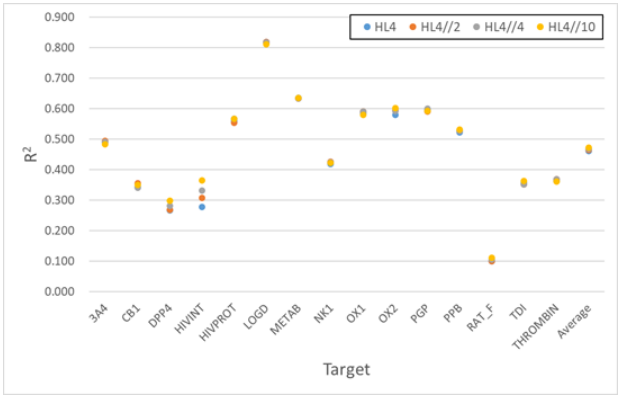

Figure 2 Comparison of Minibatch Size

値の R^2 が出現するタイミングはデータセットによって大きく異なることが分かった。この傾向は予測精度と大きく関係しており，予測精度が良いものは少しの epoch で最良の精度へ到達し，悪いものほど epoch が必要である。これは予測しようとしている対象がどれだけ複雑なものであるかが起因していると考えている。また，学習の繰り返しを最適化することは計算コストを抑えることに繋がり，試行回数を増やすことができるため，予測精度の優位性を明確にすることができる。

これまで我々が行った検証において精度向上に有意であるとされたミニバッチサイズの検証を追加で行った。ミニバッチサイズとは一度の順/逆伝搬において対象とするデータの個数である。従って，本問題においては学習をする際にどれだけ個々の分子の情報を反映するか変化させるパラメータといえる。比較結果を Figure 2 に示す。結果として，活性予測においては，網羅的な分子情報よりも個々の分子がどのように作用するかが重要であることが示された．平均として，ミニバッチサイズをデフォルトの 1/10 にした場合（HL4//10）が最良の結果を示した。ミニバッチサイズは小さくするほど 1 つ 1 つの分子の情報を捉えてネットワークの更新を行っていくようになる。更に，HIVINT においては 0.1 程度の大きな改善が行われている。なぜ HIVINT が突出して改善傾向になるのかについては現在検討を行っている。また，よりミニバッチサイズを小さくすることも検討したが，小さくすると計算コストがそれに従って大きくなるため，現状では行っていない．本結果が意味することは，それぞれの分子特徴を明確に捉える必要があると考えるのが妥当であるため，今後実施予定である各分子の配座異性体情報が予測として重要となりうることが示唆されている。

まとめ、今後の課題

令和 2 年度では特に活性予測精度の向上ならびに予測機序の解明を中心として計算を行い，成果として予測精度の向上を確認し，ターゲットタンパク質の違いによる予測の性質を捉えることができた。学習の繰り返しに関する計算では，対象となるタンパク質ならびに予測しようとする活性値の複雑さによって，最良の精度へ到達するタイミングが大きく変化することが分かった。また，ハイパーパラメータの一つであるミニバッチサイズに関する計算では，今回新たに計算したよりミニバッチサイズを小さくした場合において，最良の予測精度を記録した。従って，本問題においては 1 つ 1 つの分子の特徴を捉えてネットワークを更新していくことが重要であるといえよう。

今後は，引き続き予測ネットワークの最適化を行うとともに，該当データセットから予測することができる予測ネットワークの精度の上限値を探索したいと考えている。これはどの程度まで精度が向上すればネットワークの最適化が十分であるのかを確認するための指標になる。また，新たな試みとして，各入力分子の配座異性体を生成し，三次元的情報を学習に用いる予測を行いたいと考えている。