**2P091**

# ニューラルネットワークを用いた分子の物理量の予測

[1]HPC システムズ株式会社, [2]京都大学
○植野正嗣[1], 谷村吉隆[2]

# Prediction of physical properties of molecules with neural networks

○Seiji Ueno[1], Yoshitaka Tanimura[2]
*[1]HPC Systems Inc., Japan*
*[2] Department of Chemistry, Kyoto University, Japan*

**【Abstract】**In order to predict molecular properties on the basis of existing datasets, we utilize neural networks algorisms. The accuracy of predicted results strongly depends upon a way to construct the structure of neural network to illustrate the physical properties of a molecule. In the present study, we illustrate the performance of convolutional neural networks involving graph convolutional network (GCN) and WeaveNet layers to take into account molecular structures that are independent from the molecular size for prediction of HOMO/LUMO energy levels and some other physical properties of a molecules. Fully connected (FC) layers and recurrent neural network (RNN) layers are also employed to test the description of local features of a molecule calculated from the convolutional neural network approaches. We then constructed a predictor of molecular properties using these networks for the input of atomic and bonding parameters of molecules. Finally, we discuss a role of a molecular structure to predict a molecular nature.

## 【序】

近年計算機と機械学習の手法が発達し、様々な分野で応用されている。とくにニューラルネット（以下、NN）を用いたモデル関数は、様々なタスクにおいてその高い有用性が示されている。分子の性質を予測する予測機を設計するにあたり、その性能は予測器が分子の物理的性質をうまく表現できているかに依存すると考えられる。そのため予測器で使用されているモデル関数がどのような物理を反映しているかの理解は、予測器の性能向上に貢献すると期待される。

当研究では NN 構造や入力特徴量を変化させ、分子の物理量の予測精度にどう影響するかを調査した。

## 【方法 】

小分子のデータセットとして、QM9[1]または PubChemQC[2]を用いて検証を行った。PubChemQC については PubChem の ID が 500000 以下のうち、構成元素などの条件に合致する約 120000 の分子を対象とした。

予測器として、分子１つをグラフとして表現し、そのグラフを入力として、HOMO・LUMO エネルギーレベルや原子ごとの電荷などの物理量を出力とするモデルを設計し、学習させた。

分子を無向グラフとして表現した時、原子をノードとして定義し、結合の種類や距離に応じた重みをもつエッジを定義した。これにより座標に依存した HOMO/LUMO のエネルギーレベルを学習に用いることができる。

NN としては分子グラフの入力に畳み込み層、全結合層、後処理を順に適用するモデルとした。畳み込み層に対してはシンプルな Graph Convolutional Network (GCN) [3]もしくは先行研究で提案されている WeaveNet[4]を用い、比較した。

全結合層については、畳み込み層の出力に対して原子ごとに適用した。単純な全結合層のほか、時系列をモデル化できる RNN の一種である Long Short Term Memory (LSTM) 層を用い、分子全体を 1 つの時系列とするモデル関数を設計した。ただし、これは分子内での時系列のとり方という任意性がある。

畳み込み層・全結合層を通して得られた出力について、すべての構成原子上にわたって平均をとり分子ごとの物理量とした。

また、予測器の NN の構造のほか入力にする原子・結合の特徴量を変化させ、その予測器の精度や回帰の様子を比較した。原子に対しては原子核電荷、元素ラベル、単原子分子としたときの HOMO/LUMO エネルギーレベルなど、結合に対しては単結合・二重結合など通常の化学結合のラベルほか離れた原子間においてもエッジを定義し、特徴量を取捨選択した。

## 【結果・考察】

PubChemQC において CHONPS とハロゲンから構成される分子をデータセットとしたとき、最適化構造については HOMO/LUMO の予測精度は誤差が 1 eV 以下の精度となった。GCN による実装での予測値・真値の散布図を Fig. 1 に示す。

LSTM については Fig. 2 のように導入により予測性能が上がることが確認された。また畳み込み層として、WeaveNet を実装した結果、出力層などの処理やパラメータが異なるので論文ほどの精度は出なかったものの、ある程度の精度を確認した。

当日は分子物性の予測精度が、モデル関数の構造や入力パラメータにどのように依存するかも調べる。

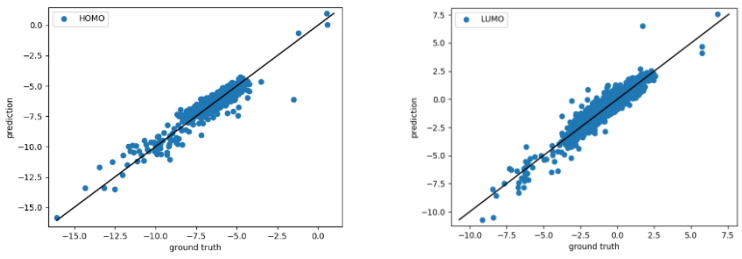


**Fig. 1.** Scatter plots of HOMO (left) and LUMO (right). Horizontal: ground truth, vertical: predicted value with a learned GCN.

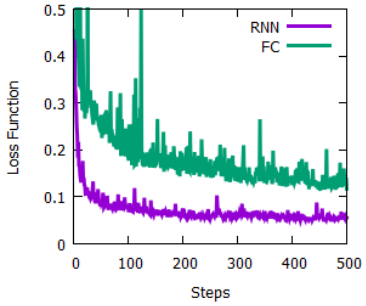


**Fig. 2.** Comparison of loss function of LUMO prediction between GCN with RNN layers and FC layers.

## 【参考文献】


[1] L. Ruddigkeit, R. van Deursen, L. C. Blum, J.-L. Reymond, Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17, J. Chem. Inf. Model. **52**, 2864–2875, 2012.

[2] Maho Nakata and Tomomi Shimazaki, "PubChemQC Project: a Large-Scale First-Principles Electronic Structure Database for Data-driven Chemistry", J. Chem. Inf. Model., 2017, 57 (6), pp 1300-1308.

[3] M. Schlichtkrull et al., arXiv:1703.06103, 2017

[4] Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley, Journal of computer-aided molecular design, 30(8):595–608, 2016.