機械学習による周辺環境を考慮した原子の分類とその解析

（早大先進理工[1]・早大理工総研[2]）○大村 拓登[1]、中嶋 裕也[2]、清野 淳司[1,2]
Classification of atoms considering surrounding environment by machine learning and its analysis ([1]Advanced Science and Engineering, Waseda University, [2]Waseda Research Institute for Science and Engineering) ○Takuto Ohmura[1], Yuya Nakajima[2], Junji Seino[1,2]

In this study, we performed clustering by machine learning for the QM9 dataset, which contains 133,885 molecules composed of light elements, using atom-centered structural and electronic descriptors that consider the surrounding environment. Furthermore, the results were applied to validate accuracies of quantum chemical methods in calculations of NMR chemical shifts.
Keywords : Machine Learning; Classification; Structural and Electronic Descriptors; Quantum Chemical Calculation; Cheminformatics

【緒言】密度汎関数理論（DFT）や近年開発が進んでいる機械学習ポテンシャルにおいて、分子系や物性により計算精度が異なるため、どのようなデータセットで検証するかは重要である。これまで QM9[1)]などの膨大なデータセット全体の統計的な検証や、GMTKN55 などの化学的観点から分類されたデータセットによる検証が行われてきた。本研究では、大規模なデータセットに対してクラスタリングを行うことで、機械学習により化学的観点から分類し、より詳細な精度検証を行うことができる手法を開発した。

【クラスタリング】本稿では、周辺環境を考慮した炭素原子に対するクラスタリングに着目する。データセットとして、QM9 内の 133,885 分子からランダムに 5,000 分子（30,782 原子）を抽出した。各原子に対する記述子として、周辺原子との距離や角度などに関する 3 個の構造的記述子と、DFT 計算による NBO 解析などを用いた 46 個の電子的記述子を使い、特徴量削減により、8 個の記述子を選別した。クラスタリング手法として、用いたデータの特徴を考慮して DBSCAN を用いた。その結果、混成軌道の違いやヘテロ原子が隣接しているかなど、おおよそ定性的に化学的な直感と一致する 28 個のクラスターを得た。

【精度検証への適用】分子物性やエネルギーに対してクラスター毎の精度検証を行うことで、より詳細な精度情報が得られる。本稿では NMR 化学シフトを対象とした DFT 手法の精度検証を示す。参照として MP2/cc-pVDZ（δ^{MP2}）を、検証する DFT 汎関数として SVWN（LDA）、PBE（GGA）、wB97XD（hybrid-GGA）（δ^{DFT}）を用いた。図 1 に上記の 5,000 分子に対する参照値からの差（$\delta^{DFT}-\delta^{MP2}$）の箱ひげ図を示す。横軸は自然電荷の平均の大きさでソートした後のクラスターID を示す。この結果、クラスター毎で誤差の分布が異なり、それぞれの分布幅は全体のそれと比較して小さいことが確認された。本手法は分子系・物性値毎で計算手法の選択に有用な情報となりうることがわかった。

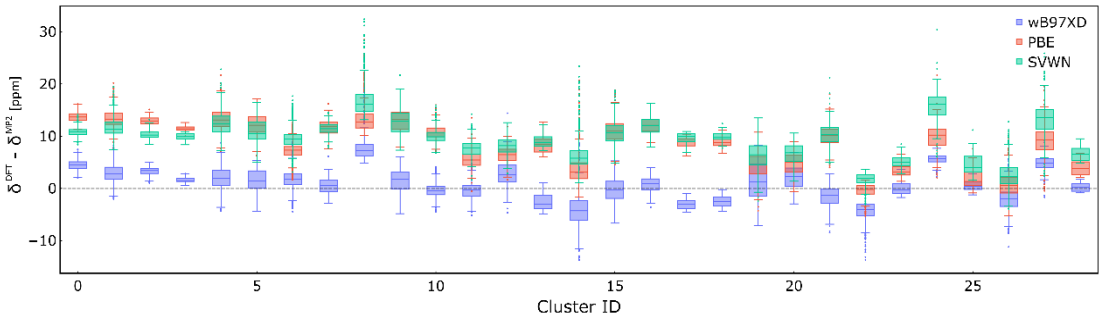

Fig. 1. Results of accuracy validation of DFT against MP2 in each cluster.

1) L. Ruddigkeit, R. van Deursen, L. C. Blum, J. L. Reymond, *J. Chem. Info. Model.* **2012**, 52, 2864–2875