1A01

# PubChemQCプロジェクト：分子データベース構築と機械学習による電子構造の推定

[1]理研ACCC，[2]理研AICS
○中田真秀[1]，島崎智実[2]

# PubChemQC Project:a large-scale first-principles calculation database and estimation of electronic structure by machine learning

○Maho NAKATA[1], Tomomi SHIMAZAKI[2]
[1] *Advanced Center for Computing and Communication, RIKEN, Wako, Japan*
[2] *Advanced Institute for Computational Science, RIKEN, Kobe, Japan*

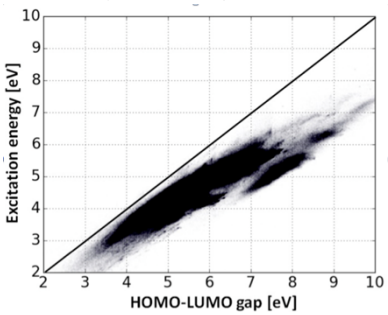
【**Abstract**】Our aim is to create a database using quantum chemical calculations and create new functional molecules from that data (http://ppcdb.org/). In recent years, with speech recognition, image recognition, etc., with the development of machine learning, highly accurate estimation close to humans has become possible. We wan to apply such approach to chemistry. In that case, we need large amounts of molecule data. Fortunately there are about 100 million molecules registered in PubChem (https://pubchem.ncbi.nlm.nih.gov/) and it is freely available. From the end of 2013 to May 2017, the structure optimization and the excited state energy by B3LYP/6-31G* are performed by TDDFT (B3LYP) /6-31+G* and accumulated about 4 million molecules of calculated data (Http://pubchemqc.riken.jp/). Using this huge amount of data, the electronic structure of the molecule was predicted by machine learning. For example, HOMO-LUMO gap had an accuracy of about 0.4 eV.

【**序**】 我々は量子化学計算を使いデータベースを作成し、そのデータから新規機能分子の創出、提案を行うシステム構築を目指している(http://ppcdb.org/)。近年、音声認識、画像認識などでは、機械学習の発達により人間に近い高精度な推定が可能になってきている。これを化学にも応用できないだろうか。その際には大量の分子のデータが必要だが、幸い PubChem(https://pubchem.ncbi.nlm.nih.gov/)に 1 億分子ほど分子構造が登録されており、自由に利用可能である。これに基づき我々は 2013 年末から 2017 年 5 月に至るまで B3LYP/6-31G*による構造最適化および励起状態エネルギーは TDDFT(B3LYP)/6-31+G*で行い、400 万程度の分子の計算データを蓄積してきた(http://pubchemqc.riken.jp/)。この膨大なデータを用い機械学習により分子の電子構造の予想を行った。HOMO-LUMO ギャップは 0.4eV 程度の精度であった。

【**方法(分子の計算)**】

PubChem(https://pubchem.ncbi.nlm.nih.gov/) に登録されている分子をすべてダウンロードし、CID (Compound Identification)、 分子の InChI 表記、分子の SMILES 表記、および分子量の表を作成した。分子量でソートし、小さい分子量をもった分子から計算を始めた。分子の初期座標生成は InChI 表記の Open Babel を用いた。この初期座標を用い、PM3、HF/STO-6G、B3LYP/6-31G* の順で構造最適化(量子化学計算)を行った。計算パッケージには SMASH, FireFly も用いつつ、最終の B3LYP/6-31G*レベルの計算は GAMESS に依った。引き続き、この構造を用い、TDDFT の計算を行い、10 個の低いエネルギー状態の励起状態を求めた。これらのうち正常に計算が終了したものを、http://pubchemqc.riken.jp にアップロードしている。分子については、PubChem に登録されている、純物質(混合物でない、SMILES 表記で"."を含まない物質)、元素の種類としては 6-31G*でサポートされている H-Kr をそれぞれ計算対象とした。

【方法(データベース構築)】
まず得られた分子に対して InChI によるバリデーションを行った。構造最適化によって構造が変化し、分子としての解釈が変化することが有るためである。バリデーションの可否については、PubChem に登録されている InChI と、構造最適化によって得られた InChI の化学式サブレイヤーおよび原子のコネクションレイヤーを比較して一致した分子を可とした。計算結果から、軌道エネルギー、HOMO-LUMO ギャップ、励起エネルギー、双極子モーメントなどを抽出し、PostgreSQL データベースで管理を行った。
【方法(機械学習)】
今回は SMILES 表記の分子から、HOMO-LUMO ギャップなど電子構造を予想することを行った。これには SMILES 表記を RDKit に含まれる 1024bit を分子の特徴ベクトルとした Toplogical fingerprint に変換し、それから scikit-learn ライブラリを用い、support vector machine(SVM)および Gaussian radial basis function (RBF)をカーネルとして機械学習を行った。教師データとしては InChI バリデーションを行った分子で、4.5-6.5eV まで HOMO-LUMO ギャップが一様に分布するよう、ランダムに 2 万分子選んだ。そこから 98 万分子の HOMO-LUMO ギャップの予想を行った。
【結果・考察】
分子計算: 理研 RICC クラスタなどを用い、一日数千分子の計算が可能であった。また、InChI 表記には曖昧さがあり、PubChem データーベースには固体、混合物、ラセミ体なども登録されていたため、量子化学計算をこのような化学のデータベースに直接厳密に適用するのは、できないし、また、意味もない。また、InChI 表記などの限界により金属錯体なども必ずしも正しく表現できない場合があった(例としてはフェロセン)。これらをどう解決するかは今後の課題である。
データベース構築:
左図は、220 万分子について HOMO-LUMO ギャップを横軸に励起エネルギーを縦軸にその相関をプロットしたものである。これから励起状態は HOMO-LUMO ギャップとよい相関に有ることが解った。尚、約 10%の分子がバリデーションに失敗した。分子検索エンジンも http://pccdb.org/ で公開中である。




機械学習 : 下図は HOMO-LUMO ギャップを SVM および Ridge regression によって予測した結果である。RMSE は 0.3〜0.4eV 程度と、SMILES のみを用いて予測したが、TDDFT による結果とよく一致した。詳細は当日発表する。

Table 1. HOMO–LUMO gap predictions based on the machine learning approach.

| Method | Kernel | RMSE [eV] |
|---|---|---|
| SVM regression | RBF | 0.36 |
| | second-order polynomial | 0.39 |
| | third-order polynomial | 0.43 |
| Ridge regression | RBF | 0.37 |
| | second-order polynomial | 0.38 |
| | third-order polynomial | 0.36 |
| | fourth-order polynomial | 0.48 |

【参考文献】
[1] 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.
[2] Nakata Maho, "The PubChemQC project: A large chemical database from the first principle calculations", AIP Conf. Proc. 1702, 090058 (2016).