2020 年度　利用報告書

課題名（タイトル）： 第一原理大規模データと機械学習を応用し欲しい物性から分子を設計する

利用者氏名： 中田真秀

理研における所属研究室名： 開拓研究本部 柚木計算物性物理研究室

1.　本課題の研究の背景、目的

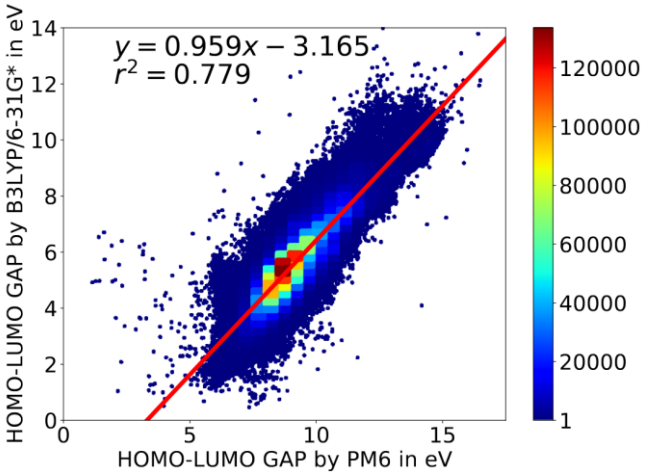
量子化学計算は非常に発展し、実験と理論計算は化学にとっての両輪となっている。ただ、量子化学の現状として、分子を決定すれば正確な物性、分子構造など手に入れられるが、ある特定の物性がほしい、ある特定の構造を持った分子を作りたい、となると非常に難しくなる。いわゆる逆問題を解かねばならないからだ。計算化学の逆問題は一般に非常に難しいが、これを解決する一つとして、網羅的な分子データベース、および検索システムの構築、機械学習による補完で挑んでいる。昨年度から今年度にかけてPubChem[1]に登録されている分子について中性、カチオン、アニオン、スピン反転状態についてそれぞれ PM6 法により分子の構造最適化計算を行い、以前 B3LYP/6-31G*[3]と比較した。さらに中性分子について、PM6 で得られた分子構造の上により精密な計算法である B3LYP/6-31G*による電子構造計算を行った。

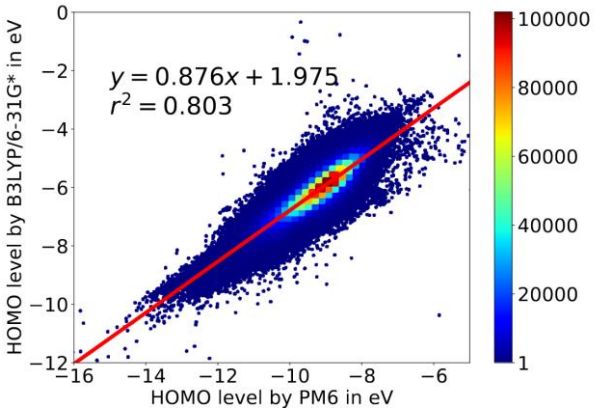
2.　具体的な利用内容、計算結果など

1. PubChem プロジェクトから 2016 年にダウンロードした。この中には全体で 91,679,247 分子存在し、さらに、分子量が 1000 以下かつ中性の 88,886,036 分子が存在した。これらを計算対象とした。
2. Isomeric SMILES 表記から Open Babel [3]を使い分子の初期構造を求めた。
3. Gaussian09 を用い、初期構造から分子の構造最適化計算を PM6 法で行い、計算が成功した分子については、同じ分子のカチオン、アニオン、スピン反転状態の構造最適化計算を行い、さらに、サブセットである、CHONのみを含む分子量 500 以下の分子のデータベース、同様に CHNOPS500 以下、CHNOPSFCl, CHNOPSClNaKMgCa500 以下について作成し、公開した
4. PM6 計算による中性分子の構造最適化は 86,213,135 分子成功した。これらについてより精密な電子構造を求めるため、GAMESS を用いて B3LYP/6-31G*計算を行った。これについて、99.8%の 86,059,366 分子について成功した。

3.　結果

1. [3] で得られた 260 万分子程度の HOMO/HOMO-LUMO gap を plot し、線形回帰計算を行った。





決定係数は、0.779, 0.803 出会った。機械学習の方法を使って [6]、決定係数を 0.99 程度まで改善できると予測される。これにより計算が数 10-数 100 倍程度高速化されるだろうと考えられる。

5. 参考文献

[1] Nucleic Acids Res. 2016 Jan 4; 44(Database issue): D1202-D1213.

[2] J. Chem. Inf. Model. 2020, 60, 12, 5891-5899

[3] J. Chem. Inf. Model., 2017, 57 (6), pp 1300-1308

[3] O'Byole et al., J. Cheminf. 2011, 3:33

[4] http://pubchemqc.riken.jp/pm6_datasets.html

[5] J Mol Model. 2007 Dec; 13(12): 1173-1213.

[6] J.Chem.Inf.Model.2017,57,11-21

2020 年度　利用研究成果リスト

【雑誌に受理された論文】

J. Chem. Inf. Model. 2020, 60, 12, 5891−5899 ″PubChemQC PM6: Data Sets of 221 Million Molecules with Optimized Molecular Geometries and Electronic Properties″