

課題名(タイトル):

第一原理計算による大量の分子計算と機械学習モデルの構築

利用者氏名:○中田真秀

理研における所属研究室名: 柚木計算物性物理研究室

## 1. 本課題の研究の背景、目的

化学における分子の種類はそれこそ無限にあり、我々にとって重要な分子だけでも億単位はあると思われる。それらの正確なデータの蓄積、分析は化学にとって重要である。さて、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の構築、および検索システムの構築、機械学習による補完などのプロジェクト PubChemQC プロジェクト <http://nakatamaho.riken.jp/pubchemqc.riken.jp/> [1,2,3]として挑んできた。

また、近年の機械学習の発達により化学データへの需要が高まっている。これに資することができれば幸いである。

## 2. PubChemQC

$\omega$  B97X-D/def2-SVPD//B97X-D/def2-SVPD

データセットの構築の状況報告

### 2.1 初期コンフォーマー生成

2020 年 11 月 06 日に PubChem から PubChem Compound [4]をダウンロードした。2021 年から 2022 年にかけて、コンフォーマーサーチは以下の手順で実施した。合計約 1800 万分子程度の計算についてコンフォーマーサーチが成功した。

- Open Babel [5]を用いて初期コンフォーマーを生成した。
- CREST プログラム[6]により網羅的なコンフォーマーサーチを行った。この際、CREST は分子計算には GFN2-xTB というタイトバインディング DFT 法を用い、コンフォーマーサンプリングにはメタダイナミクス、NVT MD、及び遺伝的アルゴリズムを適用する。

### 2.2 分子の構造最適化

分子の構造最適化は、GAUSSIAN 16C02 を用い、 $\omega$  B97X-D/def2-SVP レベルで行なった。

### 2.3. IR、ラマン分光計算

分子構造最適化が成功した分子について、同様の計算条件  $\omega$  B97X-D/def2-SVP レベルでの振動解析計算、およびラマン分光計算を行った。

### 2.3 分子の励起状態

2.2 で求めた分子構造について、励起状態を TDDFT/ $\omega$  B97X-D/def2-SVPD レベルで一重項 10 個、三重項 10 個ずつ求めた。

### 2.4 分子のイオン化状態

2.2.で求めた分子について、 $\omega$  B97X-D/def2-SVPD レベルで陽イオン状態、陰イオンを計算した。

### 2.5 最適化でのトラジェクトリーについて

量子化学データセットのレビューによれば、[8]機械学習による、分子の MD ポテンシャル作成には、構造最適化中の初期構造から最終的な最適化された構造までのトラジェクトリー(コンフォーマー)とその各点での力が重要となるため、これも保存するようにした。

### 2.6 その他

MO 係数、AO 積分、全エネルギー(核間反発、運動エネルギー)、双極子、四重極子モーメント、など様々なプロパティを一括で msgpack を主体としたデータベースを構築中である。

### 2.7 現在までに計算終了した分子

2025-02-26 現在、HPCI の資源を含め、750 万分子程度  $\omega$ B97X-D レベルの構造最適化および推定 1 億程度のコンフォーマー計算が終了している。図1は、計算の推移である。計算効率を良くするため、分子量の小さな分子から計算を開始したが、500 万分子を超えた 2024-08 頃からは、原子

数の少ない分子の計算に切り替えた。

## 2.8 使用した計算リソース

2024 年度について、2025-3-10 までに利用した計算リソースは 1800 万コア時間であった。HPCI のリソースは 3100 万コア時間であった。単純計算で 275 万分子程度、RB230009 のリソースで行えたと考えられる。

## 2.9 データ容量

2025-03-10 日現在、生データは約 100Ti となっている。

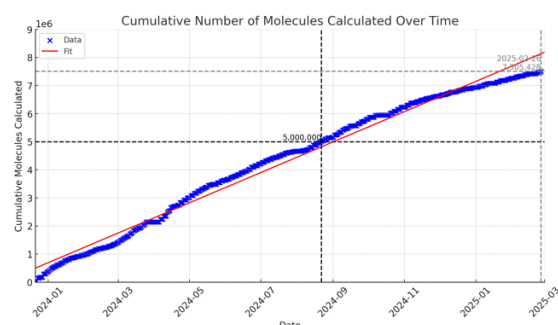


図 1 : 計算の推移 (2024-01-2025-03)

## 3. 計算結果の一部

主に 2024-03 頃までに計算終了した 100 万分子について、簡単な解析を行なっている。

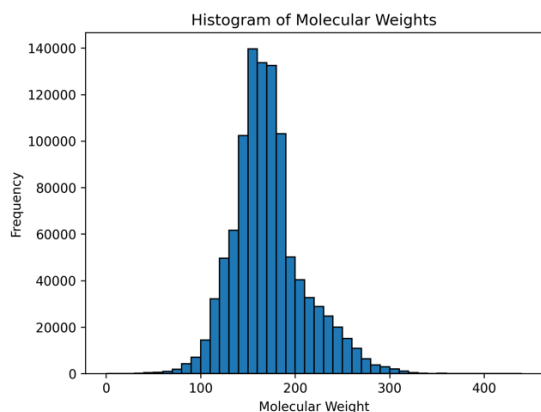


図 2 : 分子量分布

図 2 は分子量分布である。ここでは、150 程度にピークが見られ、多くの重原子が 8-10 程度あることがわかる。他のデータセットと比べ、これでも非常に大きい[7]。

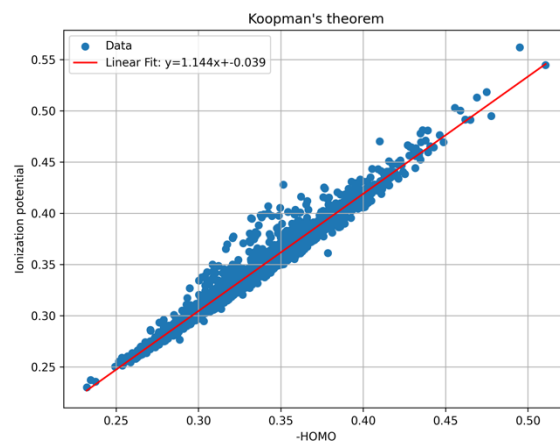


図 3 : Koopmans 定理

図 3 は Koopmans 定理の比較である。一般にイオン化エネルギーの符号を逆にした値と、HOMO エネルギーをプロットし線型回帰も行なった。これらは本来は近似良いはずであるが、比較的ズレも見られる。

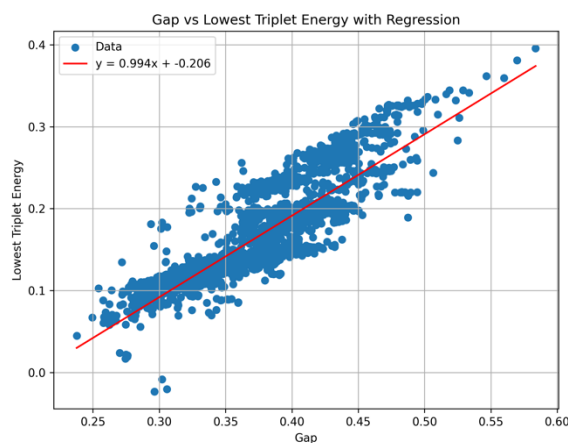


図 4: HOMO-LUMO gap vs triplet excitation

全データは、CC-BY-4.0 ライセンスの下

[https://nakatamaho.riken.jp/pubchemqc.riken.jp/b3lyp\\_p\\_m6\\_datasets.html](https://nakatamaho.riken.jp/pubchemqc.riken.jp/b3lyp_p_m6_datasets.html) で公開中予定である。

## 3. 参考文献

- [1] J. Chem. Inf. Model., 2017, 57 6, 1300-1308.
- [2] J. Chem. Inf. Model. 2020, 60, 12, 5891-5899.
- [3] J. Chem. Inf. Model. 2023, 63, 18, 5734-5754.
- [4] [Nucleic Acids Res.](#) 2016 Jan 4; 44(Database issue): D1202-D1213.
- [5] O'Byole et al., [J. Cheminf.](#) 2011, 3:33.
- [6] S. Grimme, *J. Chem. Theory Comput.*, **2019**, 155, 2847-2862.  
DOI: [10.1021/acs.jctc.9b00143](https://doi.org/10.1021/acs.jctc.9b00143)
- [7] Arif Ullah et al 2024 *Mach. Learn.: Sci. Technol.* **5** 041001 <https://iopscience.iop.org/article/10.1088/2632-2153/ad8f1>