

プロジェクト名(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: 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/>として立ち上げ、挑んでいる [1]。

## 2. 具体的な利用内容、計算結果など

### 2.1 コンフォーマーサーチ

今年度も、分子の最適なコンフォーマー探索についての計算を行った。昨年度と合わせて、**16,070,930** 分子の計算が終わった。(参考: PubChem の CHNOPSFCI 原子を含んだ 500 分子量以下の分子数は 6700 万個程度あった)

- PubChem プロジェクト [2]から、各分子の SDF および、3D 構造付きの PubChem3D [3]のデータを 2020 年 11 月 6 日午前 8 時 9 分頃(日本時間)ダウンロードした。PubChem3D は分子力場を用いていくつかのコンフォーマーをすでに計算したものである。分子種は、約 1 億種類あった。
- 1.の計算対象の分子の Isomeric SMILES 表記、InChI 表記および、PubChem3D の最も小さなコンフォーマー ID での分子内原子の座標などの情報を抽出した。
- PubChem の CID(分子に振られた自然数の ID)を小さい順に 25,000 分子ずつまとめた。
- 一単位について、分子のコンフォーマーサーチ計算は分子量の小さい順に行った。
- 各分子について、PubChem3D に掲載されている分子の各原子座標、および Open Babel[4]の Isomeric SMILES を用いて生成させた各原子座標、これら二種

類をそれぞれ初期座標として GFN2-xTB[5]を用いて分子構造最適化を行った。GFN2-xTB レベルでの局所安定構造でのエネルギーの低い方を次のステップでのコンフォーマーサーチの初期構造として用いた。

- CREST[6]を用いてコンフォーマーサーチを行った。CREST は分子の立体配置(conformer および rotamer)を探索するプログラムである。まず、GFN2-xTB 等を用いて、メタダイナミクスを行う。メタダイナミクスはポテンシャルにガウス型のバイアスを与えて、構造の変化を促しつつ行う分子動力学である。次に、メタダイナミクス中にエネルギーの低いコンフォーマーが見つければ、最小のものが見つかるまで反復的にコンフォーマーサーチを行う(6kcal/molの閾値内でサーチする)。半経験的とはいえ、比較的精密な分子計算を行うダイナミクス計算を伴うため、次に示すように、一分子のコンフォーマーサーチにはかなり時間がかかる。

### 2.2 NeurIPS2022 のワークショップの結果

<https://ogb.stanford.edu/neurips2022/workshop/> の一部で、PubChemQC[1]の 300 万分子のデータを用いた、分子の HOMO-LUMO gap を分子のグラフ構造から予想するコンペを行った。現在の順位は

<https://ogb.stanford.edu/docs/lsc/leaderboards/#pcqm4mv2>

から参照できる。リーダーボードからは、Transformer[6]とよばれる自然言語処理の分野で出てきた深層学習モデルで、これが分子のプロパティ予測にも適していることが分かる。また、0.08eV 以内の誤差で HOMO-LUMOgap を予測できた。

## 3. 参考文献

- [1] J. Chem. Inf. Model. 2020, 60, 12, 5891-5899, J. Chem. Inf. Model., 2017, 57 (6), pp 1300-1308.
- [2] [Nucleic Acids Res.](#) 2016 Jan 4; 44(Database issue): D1202-D1213.
- [3] Journal of Cheminformatics volume 3, Article number: 32 (2011).
- [4] O'Byole et al., [J. Cheminf.](#) 2011, 3:33.
- [5] J. Chem. Theory Comput. 2019, 15, 3, 1652-1671.
- [6] P. Pracht, F. Bohle, S. Grimme, Phys. Chem. Chem. Phys., 2020, 22, 7169-7192. DOI: 10.1039/C9CP06869D
- [7] Vaswani et al., Advances in Neural Information Processing Systems, 2017, 5998-6008.

2022 年度 利用研究成果リスト

**【雑誌に受理された論文】**

**【会議の予稿集】**

**【口頭発表】**

**【ポスター発表】**

**【その他(著書、プレスリリースなど)】**

<https://ogb.stanford.edu/neurips2022/> において Workshop を行った。