2021 年度　利用報告書

**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：開拓研究本部**　柚木計算物性物理研究室

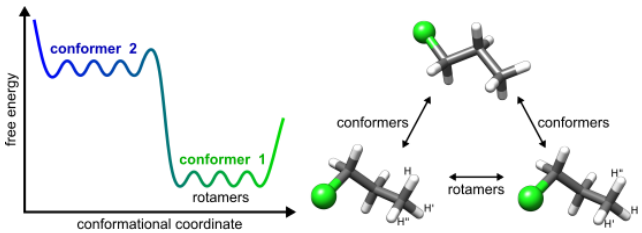
1.　本課題の研究の背景、目的

化学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://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

今年度は、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。

- PubChem プロジェクト [2]から、各分子の SDF および、3D 構造付きの PubChehm3D [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、主に 3 つの情報を抽出した。
- 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。これを一単位として計算を行った。
- 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かれば、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。




- HOKUSAI BWMPC の 5%程度の計算機資源を用い、全体の 10 分の 1 程度、1000 万分子程度の計算が終了した。分子量は最大で 200 程度であった。尚、PubChem の CHON 元素を含む分子で 300 分子量以下のものは約 1700 万分子、同様に CHNOPS500分子量以下のものは約 4600 万分子あった。

5. 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

2021 年度　利用報告書

2021 年度 利用研究成果リスト

【会議の予稿集】

1. OGB-LSC: A Large-Scale Challenge for Machine Learning on Graphs

*Weihua Hu · Matthias Fey · Hongyu Ren · **Maho Nakata** · Yuxiao Dong · Jure Leskovec*NeurIPS | 2021

Thirty-fifth Conference on Neural Information Processing Systems

1. “Molecule3D: A Benchmark for Predicting 3D Geometries from Molecular Graphs”, Zhao Xu, Youzhi Luo, Xuan Zhang, Xinyi Xu, Yaochen Xie, Meng Liu, Kaleb Dickerson, Cheng Deng, **Maho Nakata**, Shuiwang Ji, https://arxiv.org/abs/2110.01717 .

【口頭発表】

“PubChemQC PM6: A dataset of 221 million molecules with optimized molecular geometries and electronic properties”, **Nakata Maho**, Maeda Toshiyuki, Shimazaki Tomomi, Hashimoto Masatomo, 2021 International Chemical Congress of Pacific Basin Societies, Dec 17, 2021 - Dec 22, 2021.

【その他（著書、プレスリリースなど）】

KDD Cup 2021 team member; https://ogb.stanford.edu/kddcup2021/ and https://kdd.org/kdd2021/

◎分子の性質、スパコンで計算 理研と千葉工大チーム 薬や電池開発の土台に スーパーコンピューター 理化学研究所熊本日日新聞 2022 年 02 月 11 日　朝刊　18 面 3 段 1 枚

実験を行わず時間短縮 分子の性質 計算で解明へ 薬や太陽電池開発 土台に、中国新聞 2022 年 02 月 20 日　朝刊　113 面 3 段 1 枚 写図表

▽〈科学〉分子の性質 計算で迫る*理化学研究所と千葉工大*スパコン 2 億個以上の情報解析【縮小北海道新聞 2022 年 02 月 19 日　夕刊 2 面 3 段 1 枚 写図