令和 3 年度　　京都大学化学研究所　スーパーコンピュータシステム　利用報告書

生体系物質の原子・電子解析
Atomic and electronic analyses on biological matters

京都大学大学院エネルギー科学研究科　馬渕守

研究成果概要

近年、深層学習を医療・創薬へと応用する試みが注目されている。例えば、患者のゲノム配列や疾患データを用いた疾病予測システムや、アミノ酸配列に基づくタンパク質の第一原理構造予測等は従来の技術を大きく上回る性能を見せている。しかし、タンパク質の構造変化やリガンド結合といった動的情報の予測に関しては、計算速度や学習の安定性がネックとなり実用化が進んでいない。2019 年に提案された Boltzmann Generator (BG)[1]は、分子動力学(MD)計算のサンプング性能向上を目的として提案された生成モデルだが、既存の手法より効率が高くなるのは小規模なタンパク質に限られ、原子数が増えると学習に必要な訓練データ数は爆発的に増加してしまう。そこで本研究では BG の層をより深くするとともに、画像生成等で用いられる ResNet[2]を導入することで大規模タンパク質にも適用可能に拡張した Deep Generative Model (DGM) の開発に取組んだ。

平衡状態のタンパク質構造は Boltzmann 分布に従って生成される。しかし、従来の MD 計算では一度構造がポテンシャル障壁にトラップされると、抜け出すのに時間がかかり、Boltzmann 分布の全体を探索するのに膨大な計算時間を要する。DGM では、複雑な Boltzmann 分布と標準ガウス分布の間の可逆な変換関係を学習し、複数のポテンシャル障壁で仕切られた準安定構造を、潜在空間上の単一極小値へと圧縮する。学習後に、ガウス分布からサンプルした点を逆変換することで一度に多様な構造を生成することができる。本研究では更に、必要な訓練データ数を抑える手法として、「Training with Exploration (TwE) 」を考案した。これは、DMG の学習領域を物理的に正しいタンパク質構造が存在する多様体上に制限し、この多様体上で新たな構造の探索と評価を学習と同時に行う。TwE を DGM の学習に組み込むことで、訓練データを物理的に正しい (＝人為的な Data Augmentation ではない) 方向へと循環させることができる。

本研究では、約 13,000 原子からなるインテグリンに対して、DGM を学習させた。訓練データとして短時間 MD 計算で取得した 100,000 配位の構造を用いた。DGM は 1 層の特徴変換レイヤと ResNet を含む 15 層の RealNVP レイヤで構成した。学習は 23stage に分けて、ハイパーパラメータを段階的に調節しながら行い、最終的に損失関数が一定値へと収束した。学習後の DGM でインテグリン構造を生成したところ、X 線実験構造に類似した構造や、訓練データに元々含まれていない遷移状態も取得された。また、Boltzmann 分布を第一、第二主成分上に射影したところ、DGM で生成した場合の方が MD 計算と比較して準安定構造間の広い遷移状態を生成できていることが確認された。また、TwE で推定した潜在空間の多様体に沿って準安定構造を補間することで、従来の MD 計算の課題であった準安定構造間の最小エネルギー遷移経路の可視化にも成功し、バイアスポテンシャルを用いた MD ベースの補間手法よりも 2 倍以上効率的な探索を実行できた。

参考論文：1) F. Noe et al., *Science* **365** (2019) eaaw1147.
2) K. He et al., *arXiv*:1512.03385 (2015)