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、AlphaFold[1)]を始めとする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。例えば、Boltzmann Generator (BG)[2)]は当初、分子動力学(MD)計算のサンプング性能向上を目的として導入された生成モデルであったが、既存の手法 より計算効率が高くなるのは、小規模なタンパク質を扱った場合に限られ、原子数が増えると学習に必要な訓練サンプル数は爆発的に増加してしまう。そこで、本研究では BG の層をよりディープにし、大規模タンパク質にも適用可能に拡張した Deep Boltzmann Generator (DBG) 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。DBG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、訓練データ数を抑える手法として、「敵対的多様体学習(AML)」アルゴリズムを考案した。これは、DBG 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(シミュレーション)と評価を学習と同時に行う。AML を DBG の学習に埋め込むことで、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、DBG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。DBGは1層の特徴変換レイヤと15層のRealNVPレイヤで構成した。学習は25stage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 DBG 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。いずれもコンタクトマップ等の評価から物理的に妥当な構造であることが確認された。また、MD 計算とのサンプリング性能を比較するために Boltzmann 分布を第一、第二主成分上に射影したところ、DBG で生成した場合の方が準安定構造間の広い遷移状態を再現できていることが確認された。また、AML 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た。

参考論文：1) A. W. Senior et al., *Nature* **577** (2020) 706–710.
2) F. Noe et al., *Science* **365** (2019) eaaw1147.