

TSUBAME 共同利用 令和 4 年度 学術利用 成果報告書

深層学習を用いた分子動力学シミュレーションの高速化
Acceleration of molecular dynamics simulation using deep learning

林 周斗
Shuto Hayashi

東京医科歯科大学 難治疾患研究所
Medical Research Institute, Tokyo Medical and Dental University
<https://www.shimamlab.info>

タンパク質が持つ機能を解析するため、分子動力学シミュレーションによりタンパク質立体構造のダイナミクスを解析する研究が広く行われている。しかしながら、分子動力学シミュレーションには大きな計算コストを必要とするという問題点があり、長時間・大規模なシミュレーションを行うにはスーパーコンピュータを用いても膨大な時間を要する。そこで本課題では、分子動力学シミュレーションを高速かつ高精度に模倣する深層ニューラルネットワークを開発した。ベンチマークデータセットを用いた検証により、我々の手法が任意のタンパク質に対して高速かつ高精度にシミュレーションを実行できることを示した。

To analyze the functions of proteins, molecular dynamics simulations are widely used to study the dynamics of protein structures. However, molecular dynamics simulations have the drawback of requiring significant computational resources, and even with the use of supercomputers, conducting long and large-scale simulations can be time-consuming. In this project, we developed a deep neural network that can mimic molecular dynamics simulations with high speed and accuracy. Through validation using benchmark datasets, we demonstrated that our method can perform fast and accurate simulations for any given protein.

Keywords: molecular dynamics simulation, deep learning, E(3)-equivariant neural network, graph neural network, protein structure

背景と目的

タンパク質が持つ機能はその立体構造によって決定づけられており、タンパク質の立体構造を精査することはその機能を知る上で重要な手がかりとなる。これまで X 線結晶構造解析、核磁気共鳴分光法、電子顕微鏡などの実験手法により、さまざまなタンパク質の立体構造情報が解析されてきた。しかしながら、これらの手法によって得られる立体構造は一時刻点における静止像であるため、タンパク質立体構造の動的変化を捉えることはできない。実際、薬学分野においては、薬・標的タンパク質間の相互作用を見極めるために立体構造の動的変化を解析することが重要視されている。

これまでタンパク質立体構造の動的変化を解析する研究が数多くなされてきたが、それらの中でも特に広く用いられているのが分子動力学シミュレーションである。しかしながら、分子動力学シミュレーションには大きな計算コストを必要とするという問題点があり、スーパーコンピュータを用いたとしても現実的な時間内で計算で

きるのは長くともマイクロ秒～ミリ秒スケールのシミュレーションまでである。

そこで本課題では深層学習をベースとしたデータ駆動型のアプローチを用いることで分子動力学シミュレーションを高速に行う手法を開発することを目的とする(図1)。

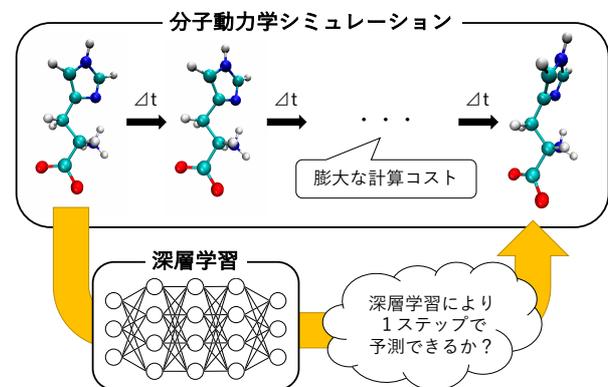


図1: 本研究の背景

概要

本課題ではタンパク質の全原子座標、速度、力、原子種を入力として、一定時間後の全原子座標、速度、力を高速・高精度に予測する深層ニューラルネットワーク(LAMDA)を学習する。学習データセットとして、実際のタンパク質分子動力学シミュレーションデータを用いる。ネットワークの構築、および学習には PyTorch を用いる。

結果および考察

タンパク質の全原子座標、速度、力、原子種を入力として、64 フェムト秒後の全原子座標、速度、力を予測するモデル(LAMDA)を学習した。LAMDA を学習させるため、AlphaFold2^[1,2]によって推定されたヒトタンパク質立体構造を初期構造として MD シミュレーションを実行し、そのトラジェクトリデータを利用した。

性能評価のため、ベンチマークデータセットに対して分子動力学シミュレーションを適用して得られたトラジェクトリと LAMDA の予測を比較した。図2はあるタンパク質のある水素原子とある炭素原子におけるトラジェクトリ(黒丸)と LAMDA の予測(赤丸)である。これらの原子は非線形な動きをしているにも関わらず、LAMDA が 64 フェムト秒後の座標を正確に予測できていることが分かる。

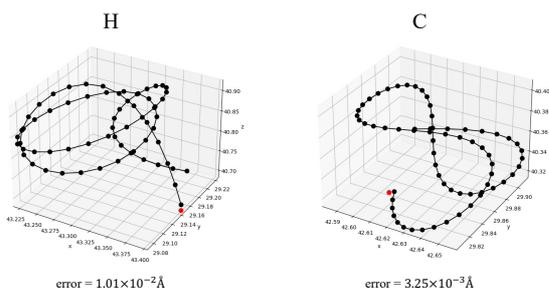


図2: MD によるトラジェクトリ(黒)と LAMDA による予測(赤)

また、ベンチマークデータセット中の全 2172 タンパク質の全水素原子と全炭素原子に対して、タンパク質の全原子数と予測誤差をプロットした(図3)。ほとんどの原子に関しては 10^{-2} Å スケールの誤差であり、誤差の大きさはタンパク質の大きさに依存しないことが分かる。

最後に、MD シミュレーションと LAMDA の実行時間

の比をプロットした(図4)。平均して約 20 倍の高速化を達成した。

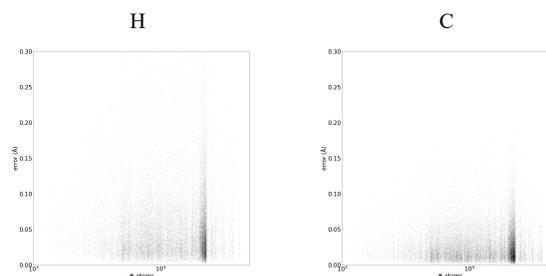


図3: LAMDA の予測誤差

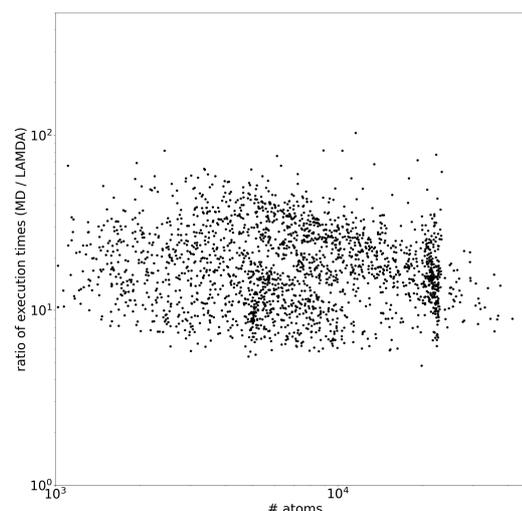


図4: MD シミュレーションと LAMDA の実行時間比

まとめ、今後の課題

本課題では分子動力学シミュレーションを模倣するニューラルネットワーク(LAMDA)を学習し、64 フェムト秒後の原子座標を高速かつ高精度に計算することができた。長時間のシミュレーションを行うには LAMDA を繰り返し適用する必要があるが、安定して実行することは現状困難である。今後は安定した長時間シミュレーションを実現するための技術開発に取り組んでいく予定である。

参考文献

1. Jumper, J et al. Highly accurate protein structure prediction with AlphaFold. Nature (2021).
2. Varadi, M et al. AlphaFold Protein Structure Database: massively expanding the structural

coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research* (2021).