

機械学習によるタンパク質リガンド結合 3 次元構造の予測
Predicting 3D protein-ligand docking based on machine learning

京都大学 薬学研究科・薬学部 ジョン・リー (John Lee)

研究成果概要

本研究では、京都大学化学研究所スーパーコンピュータシステムを利用し、Mixture of Experts (MoE) アーキテクチャを深層学習分子ドッキングモデルに適用する研究を行った。

分子ドッキングは、低分子医薬品候補が標的タンパク質にどのように結合するかを予測する計算手法であり、構造ベースのバーチャルスクリーニングにおいて中心的な役割を担う。従来のドッキング手法は物理ベースのスコアリング関数と配座探索に依存しており、計算コストが高く、柔軟な標的に対する精度にも限界がある[1]。実験的に決定されたタンパク質-リガンド複合体構造で訓練された深層学習モデルは、より高速な推論と結合ポーズ予測の改善を可能にする。AlphaFold 3 [2]は、単一の深層学習フレームワークが専用ドッキングツールに匹敵する精度を達成できることを実証した。しかし、現行のモデルはスケーリング特性が不十分であり、モデルサイズの増大に伴い性能が頭打ちになる傾向があり、訓練データに含まれないタンパク質ファミリーへの汎化能力も低いままである[1]。

MoE は、学習されたゲーティングネットワークを通じて各入力を専門化されたエキスパートサブネットワークの部分集合にルーティングすることで、サンプルあたりの計算コストを一定に保ちながら総パラメータ数を拡張できる手法である。このパラダイムは、Vision Transformer [3]や拡散モデル[4]においてその有効性が実証されており、最近の UMA モデル[5]では線形エキスパートの混合を原子スケールのモデリングに適用し、14 億パラメータのモデルでありながら、構造あたり約 5000 万のアクティブパラメータのみで動作することが示されている。本研究では、分子ドッキングにおいて各エキスパートが異なる結合部位の幾何学的特徴、リガンドの化学的性質、またはタンパク質ファミリーに専門化することで、同等の密なモデルと比較してスケーリング効率と分布外汎化性能の両方を改善できると仮説を立てた。予備実験の結果、訓練過程においてエキスパートの専門化が確認され、MoE を組み込んだモデルは密なベースラインと比較して精度と計算コストのトレードオフが改善されることが示された。

これらのモデルの訓練および評価には、大規模な実験とハイパーパラメータの最適化のために大規模な GPU 計算資源が必要であった。化学研究所スーパーコンピュータ