

2024 年度年次報告書

次世代 AI を築く数理・情報科学の革新

2023 年度採択研究代表者

飯田 慎仁

北里大学 未来工学部

助教

拡散モデルによる蛋白質の立体構造集団の生成

研究成果の概要

近年、機械学習モデルによって高精度に立体構造予測が可能となり、蛋白質の構造に基づいた機能解明が飛躍的に進んでいる。しかしながら、蛋白質は熱により揺らいでおり、機能に重要であるが、この構造揺らぎは機械学習モデルでは依然として予測が困難である。さらに、実験的に構造決定が困難であるペプチドや天然変性蛋白質(IDP)に対しては、そもそもモデルを構築することができていない。また、IDP は数百残基からなるものもあり、この長さの天然変性蛋白質に対して、全原子分子動力学(MD)法を行なっても、正確なデータを取得することは困難である。IDP は球状蛋白質と比較し、1次配列上で遠距離の領域と空間的に相互作用し安定化しにくいという考えから、短いペプチドの立体構造揺らぎのデータから、より長いペプチドの立体構造揺らぎの生成を行えるのではという考えに至った。このことより、ペプチドの立体構造揺らぎの生成に対して、長さ方向の外挿性を持たせた拡散モデルを構築することを 2024 年度の目標とした。

ペプチドの立体構造生成を行う拡散モデルに外挿性を持たせるためには、外挿性を取り込んだ方法を考慮する必要がある。そのために、ペプチドの立体構造揺らぎのデータを、IDP に特有な事前分布に基づいて合成し、教師データを構築・学習することも行なった。この合成データセットは、シミュレーションでは困難な長いペプチドに対する立体構造揺らぎの情報を含めることができ、シミュレーションからは得られない、より長いペプチドの立体構造揺らぎのデータを生成することができると考えている。予備実験であるが、MD 法から得られた短いペプチドのデータだけで学習した拡散モデルは、より長いペプチドの立体構造の生成に失敗(壊れた構造が生成される)するのに対し、合成データセットも含めた学習により得られたモデルはより長いペプチドに対して、より良い品質での構造生成を行うことができた。