



HOKKAIDO UNIVERSITY

Title	Accurate and fast predictions of protein structures and properties via protein language models [an abstract of dissertation and a summary of dissertation review]
Author(s)	許, 仕傑
Degree Grantor	北海道大学
Degree Name	博士(環境科学)
Dissertation Number	甲第16231号
Issue Date	2025-03-25
Doc URL	https://hdl.handle.net/2115/95289
Rights(URL)	https://creativecommons.org/licenses/by/4.0/
Type	doctoral thesis
File Information	XU_Shijie_review.pdf, 審査の要旨



学 位 論 文 審 査 の 要 旨

博士 (環境科学)

氏 名 XU Shijie

審査委員 主査 教授 小野田 晃

副査 教授 Biju Vasudevan Pillai

副査 准教授 加藤 優

副査 グループ長 Mehdi Davari Dolatabadi

(Leibniz Institute of Plant Biochemistry)

学 位 論 文 題 名

Accurate and fast predictions of protein structures and properties via protein language models

(タンパク質言語モデルによるタンパク質の構造と特性の正確かつ迅速な予測)

深層学習によるタンパク質の構造と特性の予測は、遺伝子の塩基配列やタンパク質のアミノ酸配列のデータ量の増加とあわせて、近年発展している。特にタンパク質の構造と特性を正確かつ迅速に予測する手法の開発が重要である。本学位論文では、大規模言語モデル (LLM) と同様の手法を用いてタンパク質の配列を文字として学習したタンパク質言語モデル (PLM) を活用して、タンパク質構造および特性を正確かつ迅速に予測する新たな深層学習の手法を開発した。具体的には、天然変性タンパク質の構造予測、タンパク質の pK_a の予測、金属タンパク質における金属結合部位に関する予測のための新手法を確立した。

序章では、タンパク質科学の背景とタンパク質構造予測について述べている。タンパク質の配列と構造は天文学的な多様性を有している一方で、実験的に蓄積されているタンパク質の構造、特に特性に関するデータ量は限られており、少ないデータを用いて特性などを予測するには革新的な方法論の開発が必要である。深層学習は、この問題を解決する手段となる。本学位論文では、PLMを導入した新たなアーキテクチャーを考案して、構造と特性予測のための手法を以下に示すように開発した。

第2章では、天然状態で無秩序な構造をもつタンパク質 (IDP) を深層学習の利用により予測する手法を開発した。IDPはタンパク質機能の理解に重要であり、疾患の原因となる機能に関与する。従来法ではIDPを予測するために多重配列アラインメント (MSA) が必要であったが、MSAは普遍的に利用できるものではなく、膨大なタンパク質配列データベースに対して網羅的な検索では、時間的な制約のある手法である。そこで、煩雑なMSA再構築プロセスを代替するために、PLMを基盤として、新規かつ効率的で精密なIDP予測手法を開発した。本手法は、最新の手法と比較して優れた予測精度に加えて、高速に動作することが明らかとなり、ゲノムスケールのIDP予測に最適である。

第3章では、タンパク質の配列情報を基に、正確かつ迅速に pK_a を予測する手法を見出した。タンパク質中のアミノ酸側鎖および末端の官能基の pK_a 値は、タンパク質の機能と構造を理解に情報である。しかし、タンパク質環境は多様かつ複雑なために官能基の pK_a は値は環境によ

り変化する。タンパク質 pK_a の実験データ数は限られているため、 pK_a を正確に予測する手法が現在も研究されている。また従来の構造情報に基づく手法では、特に pK_a 値が大幅にシフトした場合の予測精度が不十分であった。そこで、PLMにより導かれる各残基ごとの表現を用いて、Asp、Glu、His、Lys、Cys、Tyr、N末端およびC末端の pK_a 値を予測する革新的な手法を提案した。等電点 (pI) に基づいて転移学習を加えることで、予測精度はさらに向上することが分かった。また本手法は、タンパク質の内部に埋もれる疎水性環境の影響で、 pK_a 値が大きくシフトした残基に対しても、既存手法よりも予測が優れることが示された。本手法は予測速度も数桁向上しており、大規模な予測に適している。

第4章では、タンパク質中の金属結合部位を同定する新手法を開発した。細胞機能に不可欠な金属タンパク質は、産業、医学、環境科学において重要な意味を持つ。既存の塩基配列や構造に基づく方法は、金属結合の位置や金属種の予測で高い精度が得られない。本章では、最初の結合残基同定のための配列モデル、候補部位予測のためのプローブ生成アルゴリズム、部位精密化のための構造モデル、そして冗長性を排除するためのクラスタリングアルゴリズムという多段階アプローチを見出した。この体系的な手法は、PLMに基づく配列モデルの性能の寄与によって、精度と効率の両面で最先端手法を大幅に凌駕している。また、AlphaFold2で予測された構造を用いた際においても、本手法は堅牢であり、金属結合部位を予測できる。

第5章では、分割モデルを用いた系統推定の精度を向上させるために、新しい分割アルゴリズムを提案している。タンパク質あるいはヌクレオチドにおける生物学的配列の不均一性を取り扱うためのモデル化は不十分であった。従来の分割法は、経験的な知識や発見的アプローチに大きく依存していた。本章では、パラメーター化によって分割プロセスを最適化する機械学習を導入したアルゴリズムを開発した。シミュレーションと実際のデータセットの両方から得られた実験結果は、計算効率を損なうことなく系統予測の精度が向上しており、このアルゴリズムの有用性を示している。本手法は、タンパク質配列や塩基配列の分割化において、新戦略といえるものであり、機械学習が系統解析のための重要なツールとなることを示した。

審査委員一同は、以上の成果を高く評価し、また研究者として誠実かつ熱心であり、大学院博士課程における研鑽や修得単位などもあわせ、申請者が博士（環境科学）の学位を受けるのに十分な資格を有するものと判定した。