# INTISARI

## Analisis Komparasi *Protein Language Modelling* untuk Prediksi Struktur Sekunder Protein

Oleh

Kayla Queenazima Santoso

23/527497/PMU/11657

Prediksi Struktur Sekunder Protein (PSSP) merupakan salah satu permasalahan mendasar dalam bidang biologi komputasi, dengan model-model yang ada saat ini masih menunjukkan kinerja di bawah batas teoretis meskipun telah terjadi kemajuan signifikan. Metode konvensional umumnya mengandalkan informasi sekuens sebagai data utama, namun perkembangan terkini berupa *Protein Language Modelings* (PLMs) memberikan alternatif untuk menangkap pola dan struktural pada protein. Penelitian ini bertujuan untuk melakukan perbandingan sistematis terhadap tiga PLMs yaitu ESM2-T36-3B-UR50D, ProtT5-XL-U50, dan Ankh-large dalam kasus PSSP. Kajian difokuskan pada analisis perbedaan ukuran data atau panjang sekuens, strategi *masking*, serta variasi arsitektur terhadap performa prediksi. Berdasarkan hasil evaluasi eksperimental yang telah dilakukan, didapatkan bahwa optimisasi panjang sekuens melalui eliminasi data outlier mampu meningkatkan kinerja secara signifikan pada seluruh model yang diuji. Arsitektur berbasis T5, dengan data eliminasi serta tanpa strategi *masking*, menunjukkan performa klasifikasi 9-state terbaik dengan pencapaian skor SOV9 sebesar 72,04% dan Q9 sebesar 74,13% pada dataset CB433. Implementasi strategi *masking* memberikan dampak yang bervariasi, yakni pengaruh minimal pada model T5 dan Ankh, namun memberikan peningkatan yang substansial pada model berbasis ESM. Analisis perbandingan arsitektur antara kepala prediksi CNN dan LIFT-SS menghasilkan perbedaan performa yang relatif kecil (Q3: 74,34% berbanding 74,84%) meskipun terdapat perbedaan kompleksitas arsitektur yang cukup signifikan. Temuan penelitian ini menegaskan pentingnya pemilihan PLM yang tepat diikuti optimisasi arsitektur dalam meningkatkan akurasi PSSP.

**Kata Kunci**: Prediksi Struktur Sekunder Protein (PSSP), *Protein Language Modelling*, Panjang Sekuens, Strategi *Masking*.

# ABSTRACT

## A Comparative Analysis of Protein Language Modelling for Protein Secondary Structure Prediction

By

Kayla Queenazima Santoso

23/527497/PMU/11657

Protein Secondary Structure Prediction (PSSP) remains a fundamental challenge in computational biology, with current models performing below theoretical limits despite advances in deep learning approaches. While traditional methods rely primarily on sequence information, recent developments in Protein Language Models (PLMs) offer promising avenues for capturing evolutionary and structural patterns. This study systematically compares three state-of-the-art PLMs—ESM2-T36-3B-UR50D, ProtT5-XL-U50, and Ankh-large for PSSP tasks. The goals are to examine the effects of sequence length optimization, masking strategies, and architectural variations. From the research conducted, experimental evaluation shows that sequence length optimization through outlier removal significantly improves performance across all models. The T5-based architecture, with data cut and without masking strategy, achieved superior 9-state classification performance with SOV9 scores of 72.04% and Q9 scores of 74.13% (CB433). Masking strategies showed differential effects: minimal impact on T5 and Ankh models due to pre-training methodologies, but substantial improvement for ESM-based models. Architectural comparison between CNN and LIFT-SS prediction heads revealed modest performance differences (Q3: 74.34% vs 74.84%) despite significant complexity variations. These results highlight the critical importance of PLM selection and architectural optimization in PSSP, demonstrating that appropriate model choice provides stable baseline performance while targeted architectural modifications can further enhance prediction accuracy.

**Keywords**: Protein Secondary Structure Prediction, Protein Language Modelling, sequences length variation, masking strategy.