A165 · MiLaSol: Modeling Protein Solubility by Mixing Up Multiple Protein Language Models

Presenter: Weiwei Lou — Tufts University

Authors: Weiwei Lou

Session: Day 2 · Fri Oct 2, 2026 · 2:15–4:15 PM

Motivation: Protein solubility is a critical property that significantly impacts therapeutic efficacy and protein reengineering applications. Recent advances in machine learning and deep learning techniques provide unprecedented opportunities to de- velop predictive models for solubility, enabling more efficient protein design and optimization. This work is motivated by the potential of leveraging deep learning to address the solubility prediction challenge and to accelerate protein engineering workflows. Results: Leveraging and combining multiple protein language model representations, our MiLaSol model attains 81% ac- curacy, outperforming prior methods, with the highest Matthews Correlation Coefficient (MCC) score of 0.63 demonstrating balanced performance across both soluble and insoluble proteins. Through simulated annealing optimization coupled with the Raygun model, we also present a computational method to reengineer insoluble protein variants into soluble forms, with predictions confirmed by multiple independent solubility prediction methods. Our results demonstrate the effectiveness of combining machine learning-based solubility prediction with generative optimization for protein engineering.