APPLICATIONS OF PROTEIN LARGE LANGUAGE MODELS FOR NON-MODEL ORGANISMS

DAVID R. NELSON

A2S2 Group, Division of Science, New York University Abu Dhabi, Abu Dhabi, UAE
E-mail: drn2@nyu.edu

Abstract - Large language models (LLMs) trained on protein sequences provide a powerful framework for studying non-model organisms, where homology-based methods recover only a fraction of encoded functions. After briefly introducing our neural-network approaches for predicting cyanobacterial light niches and modeling seaweed multicellularity, I will focus on our pan-microalgal protein LLMs, which classify the functional "dark matter" of microalgal proteomes at scale. I will also discuss using LLMs to predict protein–protein interactions in non-model organisms. Because these models capture complex sequence patterns not evident from alignments or motifs, a central component of this work is an interpretability pipeline that makes learned signals explicit. Using Captum, DeepLIFT, and SHAP, we elucidate the residues, motifs, and biochemical features that drive model decisions, revealing conserved constraints and emergent domain-level signals inaccessible to alignment-based tools. Together, this framework shows how language-inspired AI and transparent interpretability advance functional insight in understudied organisms.