Distilling Structural Representations into Protein Sequence Models

Jeffrey Ouyang-Zhang, Chengyue Gong, Yue Zhao,
Philipp Krähenbühl, Adam R. Klivans, Daniel J. Diaz
University of Texas at Austin
{jozhang,cygong17,yzhao,philkr,klivans,danny.diaz}@utexas.edu

Abstract

Protein language models, like the popular ESM2, are widely used tools for extracting evolution-based protein representations and have achieved significant success on downstream biological tasks. Representations based on sequence and structure models, however, show significant performance differences depending on the downstream task. A major open problem is to obtain representations that best capture both the evolutionary and structural properties of proteins in general. Here we introduce ***I**mplicit **S**tructure **M**odel* (*ISM*), a sequence-only input model with structurally-enriched representations that outperforms state-of-the-art sequence models on several well-studied benchmarks including mutation stability assessment and structure prediction. Our key innovations are a microenvironment-based autoencoder for generating structure tokens and a self-supervised training objective that distills these tokens into ESM2's pre-trained model. We have made *ISM*'s structure-enriched weights easily available: integrating ISM into any application using ESM2 requires changing only a single line of code. Our code is available at https://github.com/jozhang97/ISM.

AI for New Drug Modalities at NeurIPS 2024.