[BIOINFO 2025]

Structure-Guided Motif Densification Improves Protein Language Models

Min Su Yoon[1], Minkyung Baek[1,2*]

[1] *School of Biological Sciences, Seoul National University*
[2] *Interdisciplinary Program in Artificial Intelligence, Seoul National University*
* *Corresponding author: minkbaek@snu.ac.kr*

Recent work shows that protein language models (PLMs) implicitly learn evolutionary statistics of interacting sequence motifs (ISMs). Because this signal reflects what is represented in training databases, PLMs underperform on proteins with few homologs. To address this, we densified valid ISMs by filling empty regions of sequence space. Rather than sampling new folds, we focused on sequences around known structures, since domain-level fold space is already well covered and grows slowly; the bottleneck is sequence under-sampling around existing backbones. We therefore augmented the training set with structure-conditioned designed sequences on known domain backbones using ProteinMPNN. This motif-population step adds sequences that preserve physically plausible local and long-range residue–residue couplings, sharpening the sequence–structure mapping without architectural changes. Fine-tuning ESM2-3B on the augmented dataset yielded two gains without extra supervision: (i) ESMFold predictions for previously poorly modeled proteins improved, giving higher TM-scores, and (ii) zero-shot variant-effect prediction on ProteinGym improved over the ESM2 baseline. These results support that structure-guided motif densification enriches the evolutionary statistics available to PLMs and improves their structural and functional generalization.