# Cryptic binding site detection with protein language models

Vít Škrhák, David Hoksza

Protein-ligand binding sites play a vital role in cellular function and drug discovery. A particularly challenging subset, cryptic binding sites (CBSs), only form after significant conformational changes, making them difficult to detect in unbound (apo) structures. Existing structure-based methods for binding site prediction often fail in such cases due to their dependence on particular protein conformations.

In this work, we explore the use of protein language models (pLMs) - deep learning models trained on amino acid sequences - to identify CBSs from sequence alone, as sequence information is inherently conformation-independent. Starting with a transfer learning approach, we construct a baseline predictor and subsequently evaluate multiple fine-tuning methods to enhance predictive power. One such approach involves multitask learning, where the model simultaneously predicts binding site residues and estimates local flexibility, an important factor in the CBS formation.

Our models are fine-tuned primarily on CryptoBench, a large benchmark dataset containing CBSs. However, we also consider additional data sources. Our experiments lead to consistent improvements in predictive performance across standard evaluation metrics, including more than a 2% increase in AUC. We also conduct an error analysis to better understand model limitations and implement a post-processing step to enhance prediction smoothness.

Our work opens a new direction for CBS prediction and contributes to the effort of improving early-stage drug discovery tools.