

From residues to pockets: binding site prediction with protein language models

David Hoksza

david.hoksza@matfyz.cuni.cz

*Faculty of Mathematics and Physics
Charles University, Prague, Czech Republic*

Nearly all approaches to identify new small molecules, such as virtual screening, fragment growing, or pocket-designed generative models, begin by determining where on the protein the molecule binds. Because experimental characterization can't keep pace with the rapid accumulation of new proteins, *in silico* methods for locating binding sites have become routine. Traditionally, structure-based methods, which evaluate the geometry and physicochemical properties of the protein surface, have led this field. Sequence-based approaches, limited to conservation profiles and local windows, have underperformed in *de novo* scenarios because residues lining pockets are often scattered along the chain.

Protein language models challenge this view. They produce residue-level predictions from sequence data alone, without structural input, matching the performance of the structure-based methods. Their main limitation is granularity: they label individual residues rather than identify entire pockets, which cannot be docked directly. Here, we introduce a method that reconstructs these predictions spatially, recovering residues missed by the per-residue model and clustering the solvent-accessible surface to define distinct, ranked pockets. This approach is competitive with structure-based methods but uses structural information only as a geometric scaffold.

This has an important consequence: sequence-based models are unaffected by protein conformation, avoiding the bias seen in methods trained on ligand-bound cavities and making them suitable for detecting cryptic sites that are absent in the unbound form. Therefore, we will also compare sequence- and structure-based predictions across the entire structural proteome, highlighting binding sites that current structural methods miss and that represent untapped potential for therapeutic applications.