Allosteric Pocket Prediction Using Protein Language Models

Victor Pryakhin, Malika Smaïl-Tabbone, Yasaman Karami

Université de Lorraine, CNRS, LORIA, INRIA

Allostery is a process in which the binding of a molecule at one site on a protein induces conformational changes at a distant site, ultimately impacting protein function. Identifying the precise binding sites of allosteric modulators—known as allosteric pockets—is critical for advancing drug discovery and protein design. In this work, we tackle the classification task of predicting allosteric pockets using transfer learning with protein language models (PLMs) by developing models at both the residue and pocket levels.

At the residue level, each amino acid is analyzed solely from its sequence context; however, extreme class imbalance (approximately 30 non-allosteric residues per allosteric residue) results in very low F1 scores. For pocket-level predictions, PLM-derived embeddings are combined with concise 19-feature descriptors—capturing essential physico-chemical properties of pockets identified by Fpocket—and classification is performed using feed-forward neural networks. Preliminary evaluations reveal that overall performance remains modest; while larger models (e.g., ESM3) offer some improvements, the high-dimensional final-layer embeddings are noisy compared to the concise descriptors.

Perspectives include integrating information from multiple PLM layers to capture richer representations and improve generalizability, paving the way toward more accurate allosteric pocket prediction.

Keywords: Allostery, Allosteric Pocket, Protein Language Model, Transfer Learning