[BIOINFO 2025]

A sequence-based protein-ligand binding residue prediction model for virtual screening.

Keumseok Kang[1], Mingyeol Kim[1] and Giltae Song[1,2,3,*]

[1]*Division of Artificial Intelligence, Pusan National University*

[2]*Department of Electrical and Computer Engineering, School of Computer Science and Engineering, Pusan National University*

[3]*Center for Artificial Intelligence Research, Pusan National University*

**Corresponding author: gsong@pusan.ac.kr*

Identifying protein-ligand binding residues is fundamental to unlocking molecular recognition and advancing therapeutic development. Accurate identification of these residues plays a critical role in virtual screening by facilitating the discovery of potential drug targets and narrowing down candidate compounds. While structure-based methods generally achieve higher predictive performance, most existing approaches are restricted by the lack of high-quality structural data and the computational cost of large-scale screening. Consequently, sequence-based approaches, have recently gained significant attention due to their scalability and ability to operate without relying on structural information. However, most sequence-based methods rely only on protein information without considering ligand information, which limits their applicability in virtual screening involving diverse ligands. To overcome this limitation, we propose a ligand-aware sequence-based model that incorporates ligand information. The model enhances residue-level representations through advanced embedding techniques and applies contrastive learning to improve the distinction between binding and non-binding residues. Our model consistently demonstrated robust and balanced performance across benchmark datasets, binding affinity evaluations, and virtual screening case studies, highlighting its potential to enhance hit identification in drug discovery pipelines.