POCKETMÓN
SBI-PYTHON PROJECT
BIBLIOGRAPHIC RESEARCH

AUTHORS

Marta García · marta.garcia38@estudiant.upf.edu
Karim Hamed · karim.hamed01@estudiant.upf.edu
Ivon Sánchez · ivon.sanchez01@estudiant.upf.edu

upf. Universitat Pompeu Fabra Barcelona

BIBLIOGRAPHIC RESEARCH

Geometry: PocketDepth: A new depth based algorithm for identification of ligand binding sites in proteins

PocketDepth, a geometry-based algorithm designed to predict ligand binding sites in proteins with high accuracy. The method is rooted in a novel application of a **depth factor**, which quantifies how central a region (or grid cell) is within a potential pocket based on how often it is intersected by paths drawn from the protein surface. This concept of "depth" adds a new dimension to pocket prediction by not only considering how far a site is buried but also how densely connected it is to its surrounding structure, offering insight into how crucial certain regions are for ligand interactions.

The protein structure is first mapped onto a three-dimensional grid, and each cell is classified as internal, surface, or external. Using surface atoms, the algorithm generates "grid bars" connecting pairs of atoms within a certain distance. These paths help determine the depth factor of each cell. Cells with high depth values are then clustered using a flexible, modified version of the DBSCAN algorithm, allowing the algorithm to adapt its sensitivity based on the depth and spatial proximity of the grid cells. This flexibility enables PocketDepth to identify both deeply buried and surface-exposed pockets effectively.

The authors tested the method on the **PDBbind dataset**, comprising over a thousand protein-ligand complexes. Using two sets of parameters—“deeper” for more stringent pocket definition and “surface” for broader coverage—they found that the deeper set offered better pocket delineation, while the surface set improved overall coverage. Combining both yielded a high prediction accuracy of **96.5%**, with **over 41% of correct pockets ranked first**, and **94% within the top 10 ranks**. These results demonstrate not only the method’s precision but also its robustness across proteins of varying structure and function.

Moreover, PocketDepth performed well when benchmarked against other established pocket prediction tools like **CastP**, **LigsiteCSC**, **Q-SiteFinder**, and **LigandFit**, often identifying more accurate or better-ranked binding sites. Impressively, it also showed strong performance on **apo-proteins**, which are not bound to any ligands in their crystal structures, further underscoring its utility in practical applications where the ligand is unknown. Overall, PocketDepth represents a significant advance in structure-based function prediction and drug discovery by combining geometric rigor with algorithmic flexibility.

Machine Learning (Neural network): Predicting locations of cryptic pockets from single protein structures using the PocketMiner graph neural network

PocketMiner, a machine learning model based on **graph neural networks (GNNs)** that predicts where cryptic pockets are likely to form, using only a single, ligand-free protein structure as input. PocketMiner was trained on data from MD simulations that captured cryptic pocket opening events, enabling it to learn the dynamic tendencies of residues from static structural features.

The model utilizes a **Geometric Vector Perceptron-based GNN (GVP-GNN)**, which encodes both chemical and geometric information about residues and their neighbors. By focusing on the likelihood of each residue contributing to pocket formation in a 40 ns simulation window, PocketMiner achieves **high predictive accuracy (ROC-AUC: 0.87)** while being more than **1000 times faster** than earlier methods like CryptoSite. The authors validated its performance on a curated dataset of 39 experimentally confirmed cryptic pockets, finding that PocketMiner could accurately locate binding regions even in rigid proteins or those with subtle conformational changes.

Perhaps most notably, PocketMiner was applied to over **10,000 proteins in the human proteome**, revealing that more than **half of the proteins lacking obvious pockets** likely possess cryptic ones. This vastly expands the pool of potentially druggable targets. The paper also demonstrates specific use cases, such as uncovering cryptic pockets in **PIM2** and **WNT2**, proteins implicated in cancer signaling pathways. These results suggest a new pipeline: use PocketMiner for initial screening, followed by focused simulations and drug docking on promising targets. This work not only accelerates cryptic pocket discovery but also exemplifies how simulation-informed machine learning can overcome long-standing challenges in structural biology and drug design.

Deep learning: DeepDTA: deep drug–target binding affinity prediction

DeepDTA, a deep learning model that predicts the **binding affinity** between drugs and protein targets using only their **sequence information**. Traditional approaches to drug–target interaction (DTI) prediction often frame the task as a binary classification problem (binding or not binding), thereby neglecting the **continuous nature of binding strength**, which is critical in drug discovery. DeepDTA addresses this limitation by treating the problem as a **regression task**, predicting real-valued affinity scores such as pKd and KIBA, which reflect how strongly a drug binds to a target.

The innovation of DeepDTA lies in its use of **1D sequence representations**—SMILES for drugs and amino acid sequences for proteins—and **convolutional neural networks (CNNs)** to learn informative features directly from these sequences. The architecture includes two separate CNN modules: one for drugs and one for proteins. These are followed by fully connected layers that integrate the learned features and output a continuous affinity prediction. By avoiding reliance on 3D protein-ligand structures or hand-crafted features, the method is applicable to a much wider range of targets and compounds.

The authors validate DeepDTA on two benchmark datasets: **Davis** (kinase inhibitors with dissociation constants) and **KIBA** (integrated kinase bioactivity data). In both datasets, DeepDTA performs competitively or better than state-of-the-art methods like **KronRLS** and **SimBoost**, especially in the larger KIBA dataset. Interestingly, while CNN-based protein encoding performed modestly on its own, combining CNN-derived representations of both drugs and proteins yielded the **highest prediction accuracy**, outperforming models based on predefined similarity matrices.

Overall, DeepDTA demonstrates the power of **sequence-based deep learning** in modeling complex biochemical interactions without relying on structural data. The approach opens new avenues for large-scale, structure-free prediction of drug efficacy, particularly valuable in early-stage virtual screening or when structural information is unavailable. The authors also suggest that incorporating recurrent architectures like LSTM might further improve protein sequence modeling, hinting at future improvements.

CNN: DeepPocket: Ligand Binding Site Detection and Segmentation using 3D Convolutional Neural Networks

DeepPocket, a deep learning framework designed to improve the accuracy of ligand binding site (LBS) prediction in proteins. DeepPocket integrates **geometry-based pocket detection** with **3D convolutional neural networks (CNNs)** to address limitations in traditional methods that rely on handcrafted scoring functions or probe-based energy calculations. The framework leverages Fpocket, a geometry-based tool, to identify candidate pockets via Voronoi tessellation and alpha-sphere clustering. These pockets are then rescored using a 3D CNN that analyzes local structural and physicochemical features within a 16Å³ grid around each pocket's center, replacing Fpocket's scoring function to enhance ranking accuracy.

A key innovation of DeepPocket is its **segmentation module**, which employs a high-resolution 3D CNN to predict the voxel-level shape of top-ranked pockets. This enables precise localization of subpockets and binding residues, providing actionable insights for drug design. The framework was trained on the **scPDB v.2017** dataset (17,594 binding sites) using 10-fold cross-validation to prevent data leakage. Testing across multiple benchmarks—including **COACH420**, **HOLO4k**, and the novel **SC6K** dataset

(2,378 recent PDB structures)—demonstrated DeepPocket’s superior performance over state-of-the-art tools like P2Rank, Kalasanty, and Fpocket.

DeepPocket achieved **93.4% recall** on the SC6K dataset, highlighting its ability to generalize to novel protein structures. The segmentation module matched ground-truth pocket shapes (Volsite masks) with a **Dice coefficient of 0.72** on HOLO4k and accurately identified binding residues within 4Å of ligands (precision: 0.81, recall: 0.76). These results underscore the framework’s utility in structure-based drug discovery, particularly for identifying druggable subpockets. The authors emphasize that combining geometric detection with deep learning not only improves ranking accuracy but also provides detailed structural insights, streamlining the drug design pipeline.

OUR CHOICE

Deep learning with Convolutional Neural Networks (CNNs) was chosen because it allows the model to learn intricate relationships from data that are infeasible for traditional machine learning algorithms. CNNs excel in computer vision tasks by modeling the 3D protein structure as a voxelized image, which enables the network to automatically extract and learn relevant features for binding site detection. Unlike classical methods that rely on handcrafted features or template-based approaches that require extensive databases, CNNs can identify complex patterns directly from the protein's 3D structure, leading to improved accuracy and generalization across diverse protein structures.