PickPocket: Pocket binding prediction for specific ligand families using neural networks.

Benjamin Viart[1*], Claudio Lorenzi[2], María Moriel-Carretero[3], Sofia Kossida[1]

1 IMGT®, IGH, CNRS - Univ Montpellier, Montpellier, France

2 Machine learning and gene regulation, IGH,CNRS - Univ Montpellier, Montpellier, France.

3 Cytoplasmic Control of Genome Stability, CRBM, CNRS - Univ Montpellier, Montpellier, France

* corresponding author

E-mail: benjamin.viart@igh.cnrs.fr

Pocket binding prediction for ligand families

Abstract

Most of the protein biological functions occur through contacts with other proteins or ligands. The residues that constitute the contact surface of a ligand-binding pocket are usually located far away within its sequence. Therefore, the identification of such motifs is more challenging than the linear protein domains. To discover new binding sites, we developed a tool called PickPocket that focuses on a small set of user-defined ligands and

PickPocket: Pocket binding prediction for specific ligand families using neural networks.

uses neural networks to train a ligand-binding prediction model. We tested PickPocket on fatty acid-like ligands due to their structural similarities and their under-representation in the ligand-pocket binding literature.

Our results show that for fatty acid-like molecules, pocket descriptors and secondary structures are enough to obtain predictions with accuracy >90% using a dataset of 1740 manually curated ligand-binding pockets. The trained model could also successfully predict the ligand-binding pockets using unseen structural data of two recently reported fatty acid-binding proteins. We think that the PickPocket tool can help to discover new protein functions by investigating the binding sites of specific ligand families. The source code and all datasets contained in this work are freely available at https://github.com/benjaminviart/PickPocket .

Author Summary :

Most of the protein biological functions are defined by its interactions with other proteins or ligands. The cavity of the protein structure that receives a ligand, also called a pocket, is made of residues that are usually located far away within its sequence. Therefore understanding the

PickPocket: Pocket binding prediction for specific ligand families using neural networks.

complementarity of pocket and ligand is a real challenge. To discover new binding sites, we developed a tool called PickPocket that focuses on a small set of user-defined ligands to train a prediction model. Our results show that for fatty acid-like molecules, pocket descriptors (such as volume, shape, hydrophobicity…) and secondary structures are enough to obtain predictions with accuracy >90% using a dataset of 1740 manually curated ligand-binding pockets. The trained model could also successfully predict the ligand-binding pockets using unseen structural data of two recently reported fatty acid-binding proteins. We think that the PickPocket tool can help to discover new protein functions by investigating the binding sites of specific ligand families.

Introduction :

One of the main tasks of bioinformatics is to associate biological roles to proteins using the always increasing biological data (1,2). To predict the function of a protein based on its sequence, computational methods look for sequence patterns in biological databases of known and already annotated proteins. Homology search (3,4), motif search (5) and functional

PickPocket: Pocket binding prediction for specific ligand families using neural networks.

domain search (6,7) are the most common methods, among the many available tools. Other strategies exploit different data types, such as gene expression (8) or even a combinatorial approach (9,10).

Most of the protein biological functions occur through interactions with other proteins or ligands (11). The residues making the protein contact surface and cavity shape are often located far away within the protein sequence. Therefore, the identification of such motifs is more difficult. Fortunately, the quantity of protein structures and models available in the Protein Data Bank (PDB) archive (12) has increased rapidly (13), providing abundant data for structural analyses. The study of ligands and cavities or pockets to which they bind is of particular interest, especially for drug discovery.

Different algorithms exist to compute the structure of ligand-binding pockets and to predict binding sites using geometric criteria, such as SURFNET (14), APROPOS (15), Q-SiteFinder (16), LIGSITEcsc (17), ConCavity (17,18), fpocket (19),DEPTH (20) or PDBinder (21) among others. Artificial intelligence also has been useful in this field with the development of algorithms for convolutional neural networks, such as

PickPocket: Pocket binding prediction for specific ligand families using neural networks.

FRSite (22), DeepSite (23) and DeepDrug3D (24) or of tools based on the random forest algorithm, such as P2Rank (25).

With the progressive increase in structure availability, the need to store and compare ligand-binding pocket data has led to the creation of dedicated databases, such as the Computed Atlas of Surface Topography of proteins (CASTp) (26,27) and the PoSSuM database (28). The collection of all pockets present in a single organism is called a pocketome and has its dedicated database: Pocketome (29). The extensive knowledge of all the pocket structures and their comparison are valuable for drug designers. For special needs, the tool PocketPipe can be used to analyze the pocketome of a single organism (30). Recently, comparative analyses of binding sites have gained momentum due to their capacity to reveal ligand-binding similarities among proteins, regardless of their evolution (31,32). As different proteins can evolve to bind to the same ligand type (33) the accurate classification of binding sites has become an important tool for designing drugs and predicting their possible side effects through unwanted binding (34,35).

PickPocket: Pocket binding prediction for specific ligand families using neural networks.

One limitation of the available tools is that they are tailored for drug design or for the analysis of large pockets, thus excluding the possibility of executing other tasks, such as determining the specific ligand-pocket binding complementarity. Yet, to discover new binding sites, the reverse approach needs to be possible: to focus on a small set of specific ligands and to take into account their molecular and structural specificity.

In this work, we developed a tool called PickPocket to generate a dataset of pockets that interact with a specific user-defined ligand family. The workflow consists of different R (36) and python (36,37) programs organized using bash scripts. From the pocket dataset, we computed a descriptive matrix of all pockets that contains structural information on the cavity and the residue secondary structure. Then, we trained a neural network multilayer perceptron (38) to predict whether a cavity is a 'true' pocket (i.e. can interact with ligands of the family under study) or a 'false' pocket.

To test PickPocket, we decided to use fatty acid-like ligands due to their structural similarities and their under-representation in the ligand-pocket binding literature (39). The source code and all datasets contained

PickPocket: Pocket binding prediction for specific ligand families using neural networks.

in this work are freely available at

https://github.com/benjaminviart/PickPocket.

Results :

The input data for the PickPocket tool consisted of 42 fatty acid-like ligands (Supplementary Information Table 1 : Detail of ligands used as input for PickPocket), 301 structures containing one of the selected ligands and 242 structures containing no ligand. For each pocket, a descriptive matrix is computed that gathers 21 features including structural information of the cavity as well as secondary structure (Table 1). In order to ensure the quality of the training data, the pockets labels were manually checked using pymol (40). A few mis-annotations were detected and corrected. The final training matrix contained 339 ligand-binding (true) pockets and 1401 empty ones (false).

Table 1 : Descriptors used to train the predictive model

PickPocket: Pocket binding prediction for specific ligand families using neural networks.

Pocket score	Charge score
Druggability score	Local hydrophobic density score
Number of vertices	Number of apolar alphaspheres
Mean radius of alpha sphere	Proportion of apolar alpha sphere
Mean alpha sphere solvent accessibility	Solvent accessible surface
Mean B factor	Number of alpha helices
Hydrophobicity score	Number of coils
Polarity score	Number of strands
Volume score	Number of turns
Real volume	Number of bridges
	Number of 310 helices

We obtained the best results using a neural network multilayer perceptron classifier with an architecture of (15, 10, 5). To avoid overfitting, we trained the model using a 5-fold cross-validation. Furthermore, in order to reduce problems associated with unbalanced classes, we downsampled the largest groups according to the smallest one. The model displayed an Area Under the Curve (AUC) of 97.2% (ROC curve in Figure 1). The model accuracy was 93.4%.

PickPocket: Pocket binding prediction for specific ligand families using neural networks.

Figure 1: ROC curve for pocket prediction after 5-fold cross-validation. The Area Under the Curve (AUC) of the best model was 97.2%.

To demonstrate that PickPocket can predict ligand binding on unseen data, we selected the two most recent (at the time of writing) human protein structures containing fatty acid(s) from the PDB archive: perdeuterated human myelin protein P2 (PDBID=6S2M), which contains two possible fatty acids in the same pocket (vaccenic acid and palmitic acid), and human angiopoietin-like 4 (PDBID=6U1U), which contains palmitic acid. Pockets with a score ≥0.5 were considered positive and were colored in red, the others were in random colors. For both structures, PickPocket correctly identified the fatty acid binding cavity. Careful analysis of the 6S2M structure showed that the cavity fatty acid occupied two pockets (red and blue in Figure 2A). The red pocket, which is the deep part inside the protein and contained the carboxyl part, had a score of 0.78. The blue pocket,

PickPocket: Pocket binding prediction for specific ligand families using neural networks.

which is at the opening of the cavity and contained the fatty acid tail, had a score of 0.03. The fatty acid-binding cavities were large and, as illustrated in this case, fpocket tended to consider them as more than one pocket. As both pockets corresponded to the same cavity and the red pocket had a score well-above the threshold, we considered that PickPocket discovered the fatty acid-binding cavity of the structure. For 6U1U (Figure 2B), the red pocket, corresponding to the palmitic acid cavity, received a score of 0.69.

Fig 2. Prediction details for the recently published structures 6S2M (A) and 6U1U (B). Each set of colored balls represents a pocket. Only the relevant parts of the structure are shown here. Some parts of the protein were set transparent to facilitate the visualization, and some negative pockets distant from the fatty acid were hidden to help visibility. The red pockets are correctly predicted, while the blue pocket, which corresponds to the palmitic acid tail, was incorrectly categorized as false.

Table 2 : Prediction results

PickPocket: Pocket binding prediction for specific ligand families using neural networks.

Structure	Pocket color	Prediction Score
6S2M(A)	RED	0.78
6S2M(A)	BLUE	0.03
6S2M(A)	WHITE	0.18
6S2M(A)	YELLOW	0.2
6S2M(A)	PURPLE	0.04
6U1U(B)	RED	0.69
6U1U(B)	YELLOW	0.0
6U1U(B)	PURPLE	0.0
6U1U(B)	LIGHT GREEN	0.19
6U1U(B)	GREEN	0.0

Prediction scores for the pockets of structure 6S2M and 6U1U. Score >0.5 are considered positives and colored green. Color corresponds to Figure 2. Complete prediction results can be found in supplementary information (SI Table 2 : Full prediction results for 6S2M and 6U1U PDB structures).

PickPocket: Pocket binding prediction for specific ligand families using neural networks.

Discussion and perspectives

Thanks to the increased availability of structural data and the improvement of protein 3D modeling, new pocket-based methodologies can now be developed to help protein function discovery. Our results show that combining pocket descriptors and the residue secondary structure is sufficient to train a model and to predict the pockets for specific ligand families with high accuracy, including when using unseen data. We believe that careful analysis of protein ligand families and their corresponding binding pockets coupled with the high prediction capacity of neural networks is the way forward to close the gap between protein sequences and their functions. PickPocket aims to simplify the procedure for building a ligand family dataset and train a model to recognize the corresponding cavity. The ligand selection and the corresponding structure still need to be done manually, but these steps are made easier by the PDB ligand research tool. Other databases, such as PoSSuM (28), may be used to create the dataset.

PickPocket: Pocket binding prediction for specific ligand families using neural networks.

PickPocket automatic labelling of 'true' and 'false' pockets is very fast, but still needs manual checking because sometimes it makes mistakes when pockets are very close to each other. The training data quality is extremely important for good predictions. Therefore, we strongly recommend users to manually check the generated dataset. Looking in detail at the results for the 6S2M structure, fpocket considered that it had two different pockets because the fatty acid binding cavity is very deep and narrow. This can be corrected by changing the alpha-sphere radius or the maximum distance between pockets. One of the challenges we faced was to tune the fpocket parameters in order to have a big enough pocket size without merging different cavities. PickPocket easy tuning of these parameters allows users to adapt the input to the ligand specificity.

In order to cluster ligand protein complexes, Deepdrug3D and FRsite use an atom-based voxelization. This step also allows generating a compatible input for convolutional neural networks. On the other hand, our methodology uses matrix properties that are faster to generate, but contain less information. We also chose to use the fpocket software, although DeepSite is more accurate against the sc-PDB database of binding sites

PickPocket: Pocket binding prediction for specific ligand families using neural networks.

(40). However, fpocket is fast, and pocket descriptors data is easily retrieved from output files.

PickPocket can help to discover new protein functions by investigating the binding sites of a specific ligand family. The results we obtained prove that for fatty acid-like molecules, pocket descriptors and secondary structure are enough to obtain predictions with >90% accuracy. Thanks to its high prediction accuracy, PickPocket can be used as a tool for *in silico* screens, and should boost novel research.

Material and Methods

PickPocket methodology can be divided into five steps (Figure 3). First a selection of ligands and structure, second the combination of fpocket and Stride to generate the descriptive matrice of the pockets, third the selection of 'true' pocket (pocket that binds the selected ligands), fourth training the desired model and last computing the prediction on the full PDB pocket dataset.

PickPocket: Pocket binding prediction for specific ligand families using neural networks.

To select a set of fatty acid-like ligands present in the PDB we used the ligand search tool. We selected ligands having this SMILES 'CCCCCCCCCCCCCCCCCCCCCCCCCCCC(O)=O' as superstructure and a molecular weight superior to 100.0 g/mol.

This resulted in a list of mostly fatty acids and other molecules composed of aliphatic chains including some with alcohols.From the structures containing our ligands we only selected the one with human proteins, resolved using X-ray technique and excluding DNA and RNA molecules. A representative subset at 70% sequence identity was then used as the input. A set of random structures, using the same criteria (human, X-ray, redundancy), not containing any of the previously selected ligands was used as negative data. All pockets were computed using fpocket (19) and all secondary structure information with STRIDE (41).

The 'true' pockets containing ligands were identified using euclidean distance. For each pocket in each structure, if any voronoi vertices from a pocket and any ligand atom distance is inferior or equal to 1 angstrom, then the pockets are labelled as 'true'. All other pockets in the PDB file are by default categorized as 'false'.

PickPocket: Pocket binding prediction for specific ligand families using neural networks.

To train the predictive model PickPocket offers the possibility to use random forest, support vector machine algorithms or neural networks implemented in Python. The neural network is configured to test multiple architectures and automatically select the model with the best accuracy. The Python program included the following libraries: scikit-learn(41), numpy (42) and pandas (43).

Once the desired model it can be saved and used to make predictions on unseen data. A file containing all the Protein Data Bank pockets (all pockets from all structures) can be found in the data attached to the software.

Figure 3 : methodology workflow. Five steps of the PickPocket methodology for specific ligands binding prediction.

PickPocket: Pocket binding prediction for specific ligand families using neural networks.

Funding

Bibliographie

1. Hawkins T, Kihara D. FUNCTION PREDICTION OF UNCHARACTERIZED PROTEINS [Internet]. Vol. 05, Journal of Bioinformatics and Computational Biology. 2007. p. 1–30. Available from: http://dx.doi.org/10.1142/s0219720007002503

2. Consortium TU, The UniProt Consortium. The Universal Protein Resource (UniProt) 2009 [Internet]. Vol. 37, Nucleic Acids Research. 2009. p. D169–74. Available from: http://dx.doi.org/10.1093/nar/gkn664

3. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403–10.

PickPocket: Pocket binding prediction for specific ligand families using neural networks.

4. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997 Sep 1;25(17):3389–402.

5. Bairoch A. The PROSITE database, its status in 1995 [Internet]. Vol. 24, Nucleic Acids Research. 1996. p. 189–96. Available from: http://dx.doi.org/10.1093/nar/24.1.189

6. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, et al. InterPro, progress and status in 2005. Nucleic Acids Res. 2005 Jan 1;33(Database issue):D201–5.

7. Coggill P, Finn RD, Bateman A. Identifying protein domains with the Pfam database. Curr Protoc Bioinformatics. 2008 Sep;Chapter 2:Unit 2.5.

8. Troyanskaya OG. Integrated Analysis of Microarray Results [Internet]. Methods in Molecular Biology. 2007. p. 429–37. Available from: http://dx.doi.org/10.1007/978-1-59745-304-2_27

PickPocket: Pocket binding prediction for specific ligand families using neural networks.

9. Rzhetsky A, Seringhaus M, Gerstein M. Seeking a new biology through text mining. Cell. 2008 Jul 11;134(1):9–13.

10. Si L, Yu D, Kihara D, Fang Y. Combining gene sequence similarity and textual information for gene function annotation in the literature [Internet]. Vol. 11, Information Retrieval. 2008. p. 389–404. Available from: http://dx.doi.org/10.1007/s10791-008-9053-0

11. Petrey D, Chen TS, Deng L, Garzon JI, Hwang H, Lasso G, et al. Template-based prediction of protein function. Curr Opin Struct Biol. 2015 Jun;32:33–8.

12. Rose PW, Bi C, Bluhm WF, Christie CH, Dimitropoulos D, Dutta S, et al. The RCSB Protein Data Bank: new resources for research and education. Nucleic Acids Res. 2013 Jan;41(Database issue):D475–82.

13. Brylinski M. Is the growth rate of Protein Data Bank sufficient to solve the protein structure prediction problem using template-based modeling? [Internet]. Vol. 11, Bio-Algorithms and Med-Systems. 2015. p. 1–7. Available from: http://dx.doi.org/10.1515/bams-2014-0024

PickPocket: Pocket binding prediction for specific ligand families using neural networks.

14. Laskowski RA. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. J Mol Graph. 1995 Oct;13(5):323–30, 307–8.

15. Peters KP, Fauck J, Frömmel C. The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. J Mol Biol. 1996 Feb 16;256(1):201–13.

16. Laurie ATR, Jackson RM. Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. Bioinformatics. 2005 May 1;21(9):1908–16.

17. Huang B, Schroeder M. LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. BMC Struct Biol. 2006 Sep 24;6:19.

18. Capra JA, Laskowski RA, Thornton JM, Singh M, Funkhouser TA. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. PLoS Comput Biol. 2009 Dec;5(12):e1000585.

PickPocket: Pocket binding prediction for specific ligand families using neural networks.

19. Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: an open source platform for ligand pocket detection. BMC Bioinformatics. 2009 Jun 2;10:168.

20. Tan KP, Varadarajan R, Madhusudhan MS. DEPTH: a web server to compute depth and predict small-molecule binding cavities in proteins. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W242–8.

21. Bianchi V, Gherardini PF, Helmer-Citterich M, Ausiello G. Identification of binding pockets in protein structures using a knowledge-based potential derived from local structural similarities. BMC Bioinformatics. 2012 Mar 28;13 Suppl 4:S17.

22. Jiang M, Wei Z, Zhang S, Wang S, Wang X, Li Z. FRSite: Protein drug binding site prediction based on faster R–CNN [Internet]. Vol. 93, Journal of Molecular Graphics and Modelling. 2019. p. 107454. Available from: http://dx.doi.org/10.1016/j.jmgm.2019.107454

23. Jiménez J, Doerr S, Martínez-Rosell G, Rose AS, De Fabritiis G. DeepSite: protein-binding site predictor using 3D-convolutional neural networks. Bioinformatics. 2017 Oct 1;33(19):3036–42.

PickPocket: Pocket binding prediction for specific ligand families using neural networks.

24. Pu L, Govindaraj RG, Lemoine JM, Wu H-C, Brylinski M. DeepDrug3D: Classification of ligand-binding pockets in proteins with a convolutional neural network. PLoS Comput Biol. 2019 Feb;15(2):e1006718.

25. Krivák R, Hoksza D. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. J Cheminform. 2018 Aug 14;10(1):39.

26. Tian W, Chen C, Lei X, Zhao J, Liang J. CASTp 3.0: computed atlas of surface topography of proteins. Nucleic Acids Res. 2018 Jul 2;46(W1):W363–7.

27. Dundas J, Ouyang Z, Tseng J, Binkowski A, Turpaz Y, Liang J. CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues [Internet]. Vol. 34, Nucleic Acids Research. 2006. p. W116–8. Available from: http://dx.doi.org/10.1093/nar/gkl282

28. Ito J-I, Ikeda K, Yamada K, Mizuguchi K, Tomii K. PoSSuM v.2.0: data update and a new function for investigating ligand analogs and target proteins of small-molecule drugs [Internet]. Vol. 43, Nucleic Acids

PickPocket: Pocket binding prediction for specific ligand families using neural networks.

Research. 2015. p. D392–8. Available from:

http://dx.doi.org/10.1093/nar/gku1144

29. Kufareva I, Ilatovskiy AV, Abagyan R. Pocketome: an encyclopedia of small-molecule binding sites in 4D. Nucleic Acids Res. 2012 Jan;40(Database issue):D535–40.

30. Ansar S, Centre for Bioinformatics, Kamalnayan Bajaj Institute for Research in Vision and Ophthalmology, Vision Research Foundation, SankaraNethralaya, - C, et al. PocketPipe: A computational pipeline for integrated Pocketome prediction and comparison [Internet]. Vol. 15, Bioinformation. 2019. p. 295–8. Available from:

http://dx.doi.org/10.6026/97320630015295

31. Brylinski M. Local Alignment of Ligand Binding Sites in Proteins for Polypharmacology and Drug Repositioning [Internet]. Methods in Molecular Biology. 2017. p. 109–22. Available from:

http://dx.doi.org/10.1007/978-1-4939-7015-5_9

32. Najmanovich RJ. Evolutionary studies of ligand binding sites in proteins. Curr Opin Struct Biol. 2017 Aug;45:85–90.

PickPocket: Pocket binding prediction for specific ligand families using neural networks.

33. Barelier S, Sterling T, O'Meara MJ, Shoichet BK. The Recognition of Identical Ligands by Unrelated Proteins [Internet]. Vol. 10, ACS Chemical Biology. 2015. p. 2772–84. Available from: http://dx.doi.org/10.1021/acschembio.5b00683

34. Ehrt C, Brinkjost T, Koch O. Impact of Binding Site Comparisons on Medicinal Chemistry and Rational Molecular Design. J Med Chem. 2016 May 12;59(9):4121–51.

35. Naderi M, Lemoine JM, Govindaraj RG, Kana OZ, Feinstein WP, Brylinski M. Binding site matching in rational drug design: algorithms and applications. Brief Bioinform [Internet]. 2018 Aug 31; Available from: http://dx.doi.org/10.1093/bib/bby078

36. Tierney L. The R Statistical Computing Environment [Internet]. Lecture Notes in Statistics. 2012. p. 435–47. Available from: http://dx.doi.org/10.1007/978-1-4614-3520-4_41

37. Python Language Reference [Internet]. Python for Bioinformatics. 2009. p. 457–538. Available from: http://dx.doi.org/10.1201/9781584889304.axd

PickPocket: Pocket binding prediction for specific ligand families using neural networks.

38. Schmidhuber J. Deep learning in neural networks: An overview [Internet]. Vol. 61, Neural Networks. 2015. p. 85–117. Available from: http://dx.doi.org/10.1016/j.neunet.2014.09.003

39. Bhagavat R, Sankar S, Srinivasan N, Chandra N. An Augmented Pocketome: Detection and Analysis of Small-Molecule Binding Pockets in Proteins of Known 3D Structure. Structure. 2018 Mar 6;26(3):499–512.e2.

40. Desaphy J, Bret G, Rognan D, Kellenberger E. sc-PDB: a 3D-database of ligandable binding sites—10 years on [Internet]. Vol. 43, Nucleic Acids Research. 2015. p. D399–404. Available from: http://dx.doi.org/10.1093/nar/gku928

41. Heinig M, Frishman D. STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins [Internet]. Vol. 32, Nucleic Acids Research. 2004. p. W500–2. Available from: http://dx.doi.org/10.1093/nar/gkh429
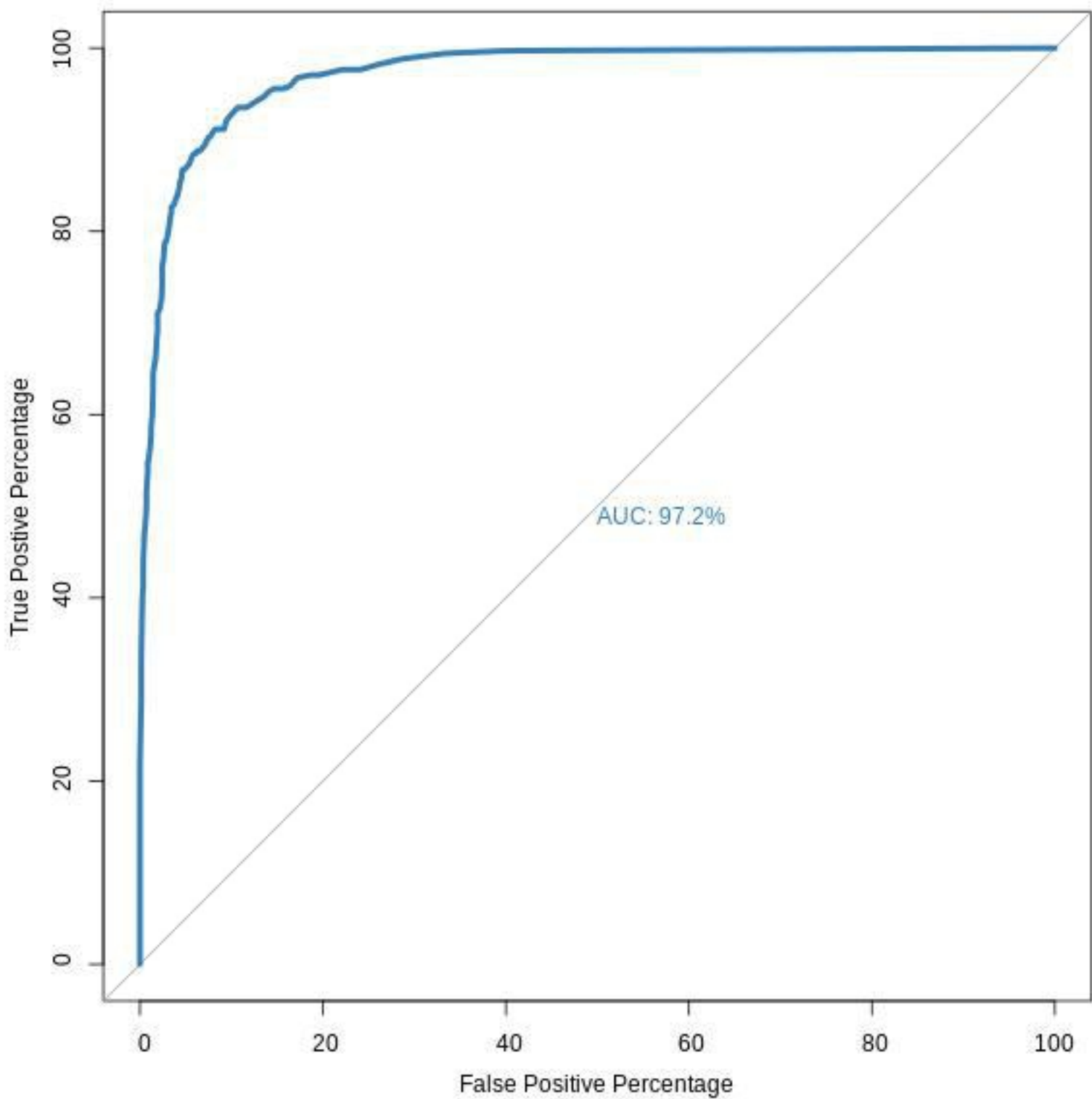
PickPocket: Pocket binding prediction for specific ligand families using neural networks.
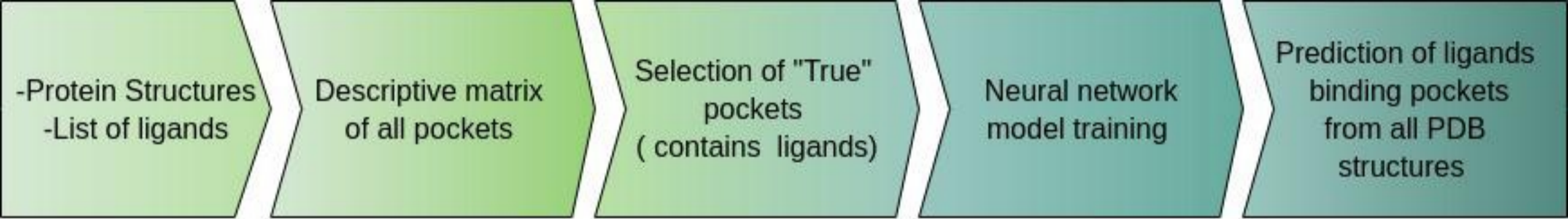
AUC: 97.2%

Figure 1

Figure 2

Figure 3