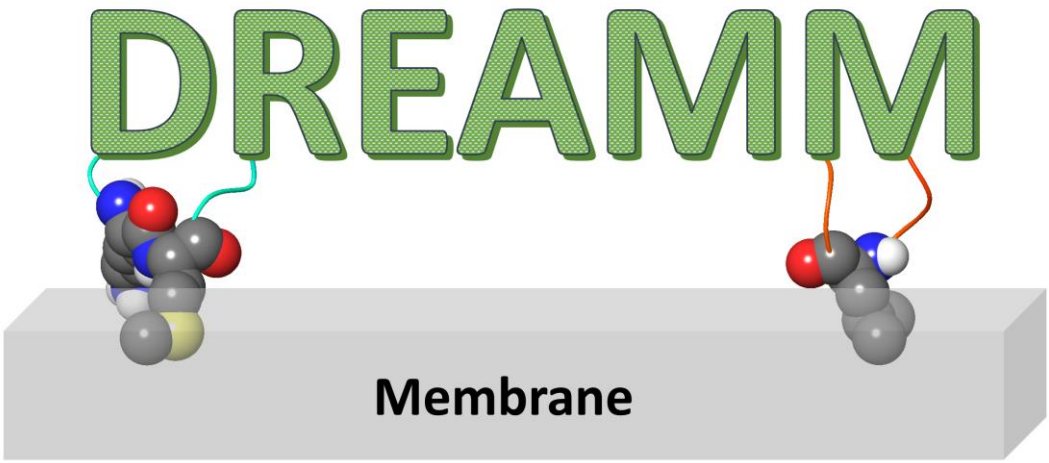
DREAMM

Drugging pRotein mEmbrAne Machine learning Method



Alexios Chatzigoulas & Zoe Cournia

Biomedical Research Foundation

Academy of Athens

https://dreamm.ni4os.eu

Contents

1. Methodology

DREAMM implements machine learning in order to predict the membrane-penetrating residues, and optionally to predict binding sites near the predicted membrane-penetrating residues in protein conformational ensembles. To predict the membrane-penetrating residues many machine learning classifiers have been trained and the final decision is taken using a voting classifier judging by the majority rule.

When a .pdb structure is inputted in DREAMM, firstly it is prepared with HTMD (1) and then the feature extraction begins, generating various physicochemical and biochemical features. These features include the secondary structure definition using DSSP (2), the solvent-accessible surface area using FreeSASA (3), the residue and Cα depth using MSMS (4), the Wimley-White whole-residue interface and octanol hydrophobicity scales (5, 6), the charges using PDB2PQR (7, 8), the conservation score using HHblits (9), the squared fluctuations using PRODY (10, 11), the number of nearby amino acids, and others. Furthermore, in order to take into consideration the surrounding amino acid properties of each residue, the mean values of the aforementioned features are calculated, for each residue and the residues lying in a distance of Cα - Cα 7 Å. In addition, the ProtDCal tool is facilitated (12), which calculates many thermodynamics, topographic, and property-based features.

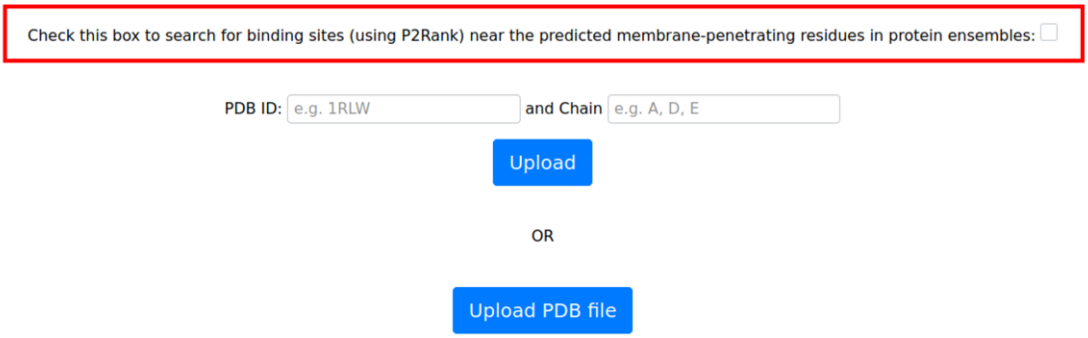
When the feature extraction is completed, the voting classifier predicts if the residues are membrane-penetrating or not. To reduce the false positive non hydrophobic amino acids, DREAMM keeps only those which lie in a COM-COM distance of 14 Å from at least one of the predicted hydrophobic amino acids. The results are displayed in the web-server and visualized with JSmol (13, 14).

Moreover, the user may choose to search for binding sites near the predicted membrane-penetrating residues. To take into account the dynamic nature of proteins we search for binding sites in conformational ensembles. After the user input the protein conformational ensemble, the P2Rank stand-alone open source software package (15) is used to predict binding sites in each conformation separately. The PyMOL (16) scripts that produce 3D visualizations, generated from P2Rank, are modified to display only the predicted membrane-penetrating residues and the binding pockets near the membrane-penetrating residues, in a distance of 5 Å, which are then clustered based on their center coordinates. The final results can be downloaded through the web-server. For more details regarding the methodology please refer to our paper.

2. Input

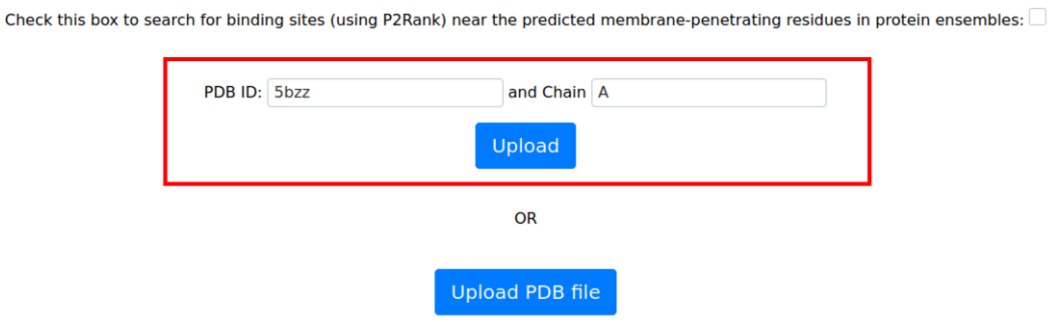
a) Without binding site prediction

The user first chooses to use the option of searching for binding sites near the predicted membrane-penetrating residues.



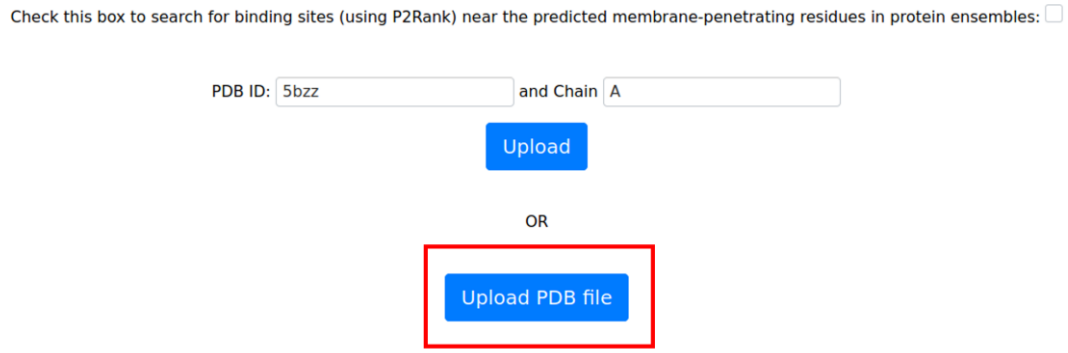
i. PDB ID

In case that the user does not want to search for binding sites, the user may input the PDB ID and the chains and hit the upload button. If the “Chain” field is empty the whole structure will be used to predict membrane-penetrating residues.



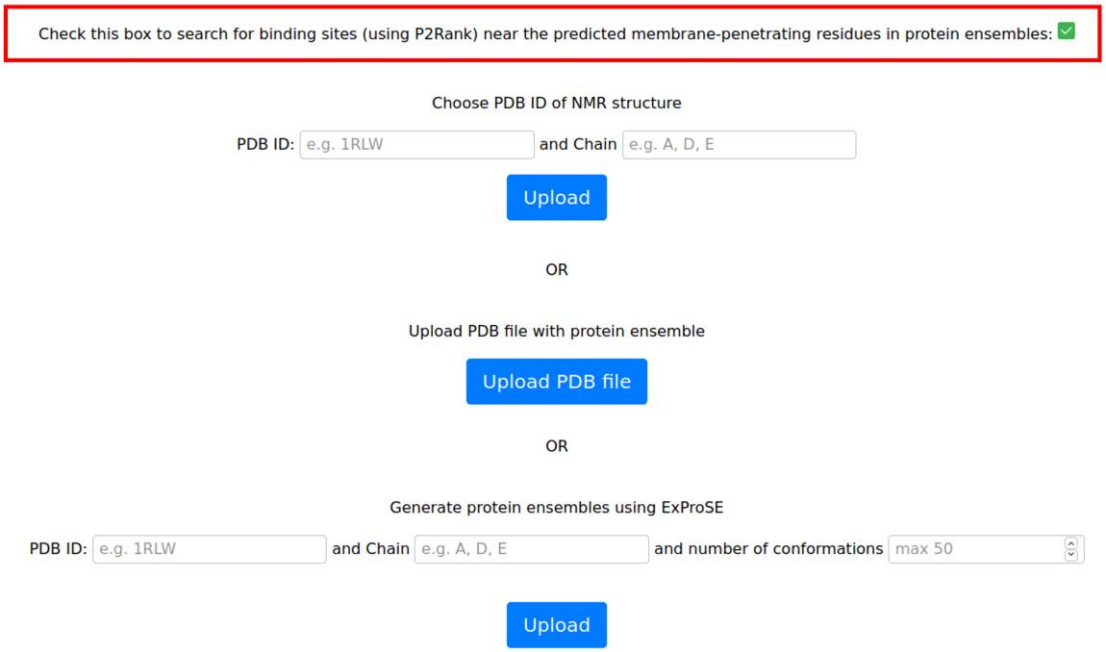
ii. Upload PDB

The user may also choose as an alternative to upload his/her own structure by hitting the Upload PDB file. Once the user chooses his/her structure the prediction automatically starts.



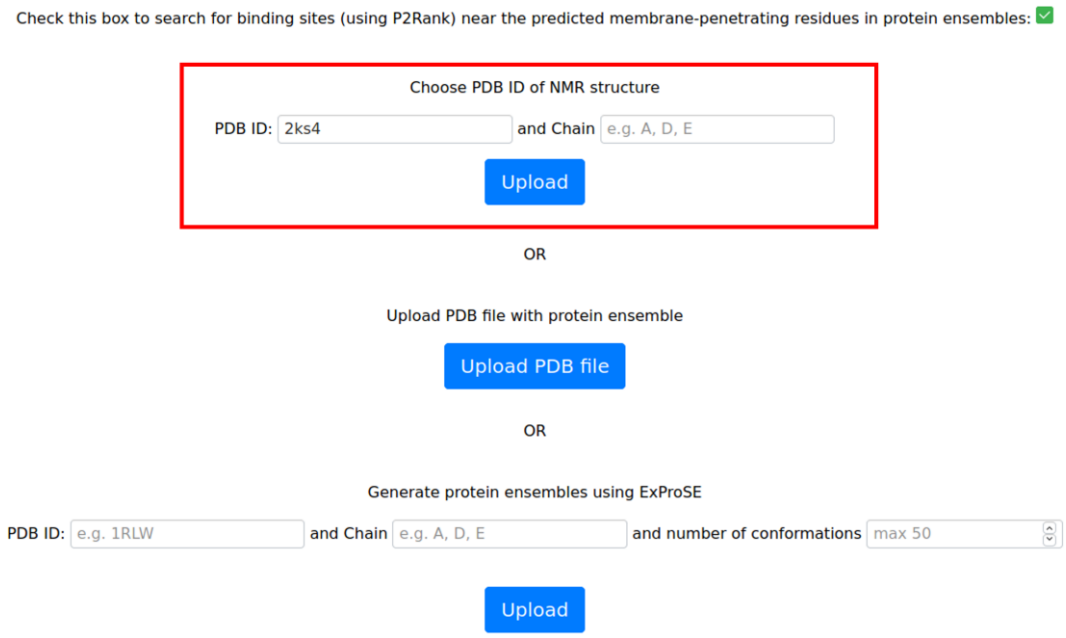
b) With binding site prediction

In case that the user want to search for binding sites near the predicted membrane-penetrating residues he/she has to check the appropriate box. Once it is checked, new options will appear. The binding site prediction is performed in protein conformational ensembles and the membrane-penetrating residues prediction will be performed in the first model of the ensemble.



i. NMR PDB ID

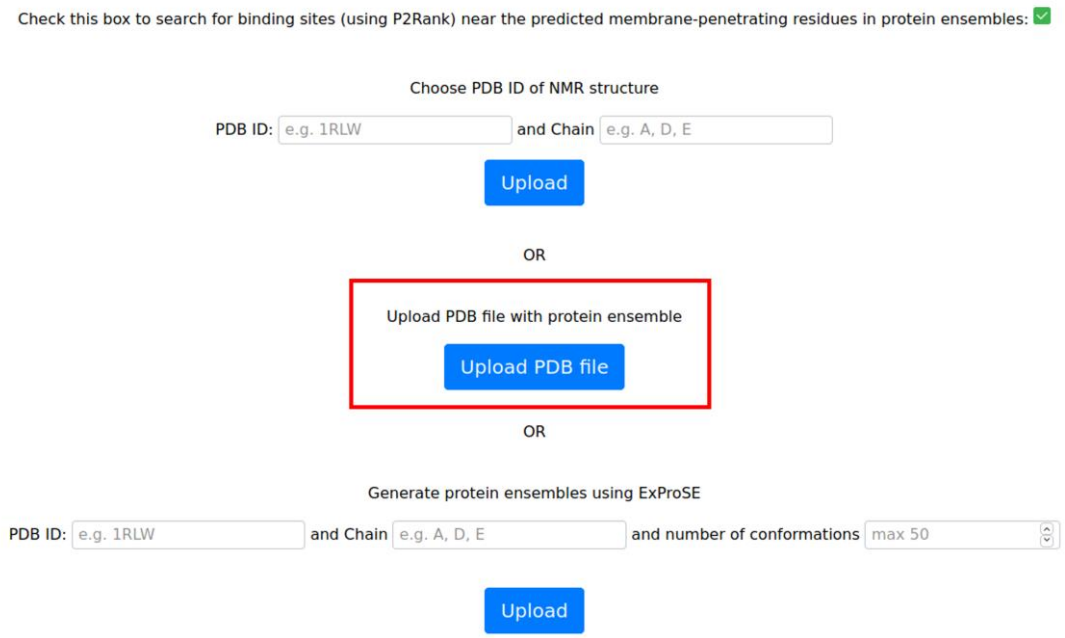
The first option is to use the PDB ID of an NMR structure, choose the chains, and hit the upload button. If the “Chain” field is empty the whole structure will be used to predict membrane-penetrating residues and binding sites.



Hint: If the PDB is not an NMR structure, DREAMM will proceed with the calculation and the predictions even if it is an X-ray structure.

ii. Upload PDB with conformational ensemble

The user may also choose as an alternative to upload his/her own protein conformational ensemble by hitting the Upload PDB file. Once the user chooses his/her structure the prediction automatically starts.

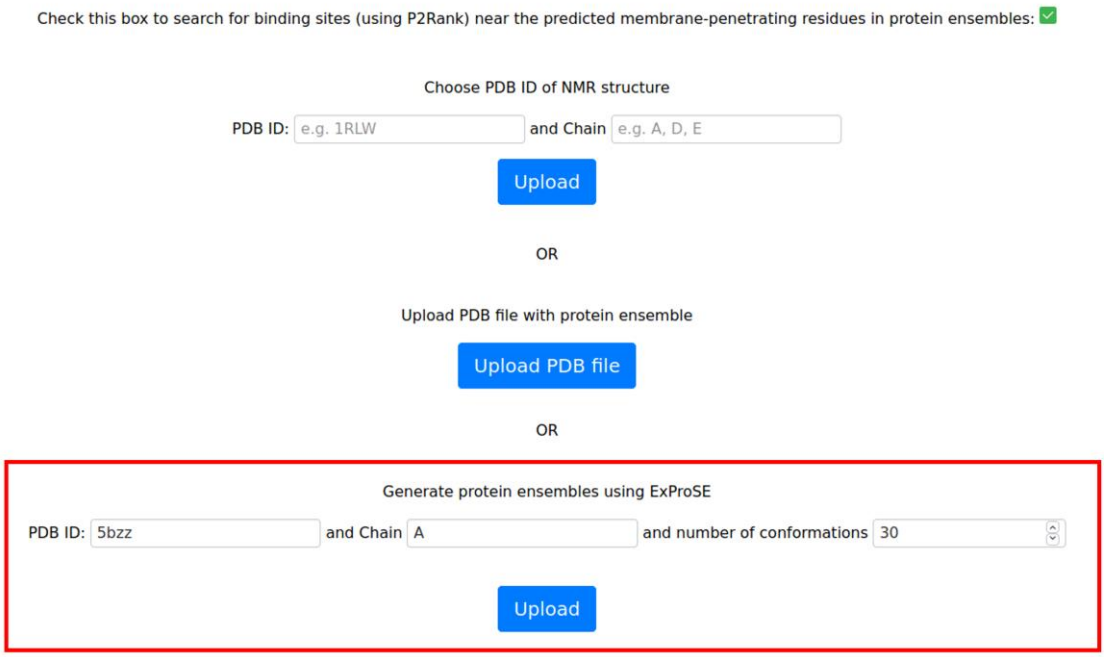


Important note: The user has to upload his/her protein conformational ensemble in a .pdb file with the models divided by MODEL/ENDMDL!

Hint: If the PDB contains only one model, DREAMM will proceed with the calculation.

iii. Generate conformational ensemble with ExProSE

In case that an NMR structure and a protein conformational ensemble are not available, the user may choose to create a protein conformational ensemble using ExProSE. The user has to input the PDB ID, the chains, the number of the additional conformations to generate, and hit the upload button. If the “Chain” field is empty the whole structure will be used to predict membrane-penetrating residues and generate conformational ensembles. The limit for the number of conformations is 50.



Important note: The membrane-penetrating residue prediction will be carried out in the initial PDB and not in one of the generated conformations.

Important note: The binding site prediction is also performed in the initial structure, so if you choose i.e., 20 conformations, the output will have 21.

3. Output

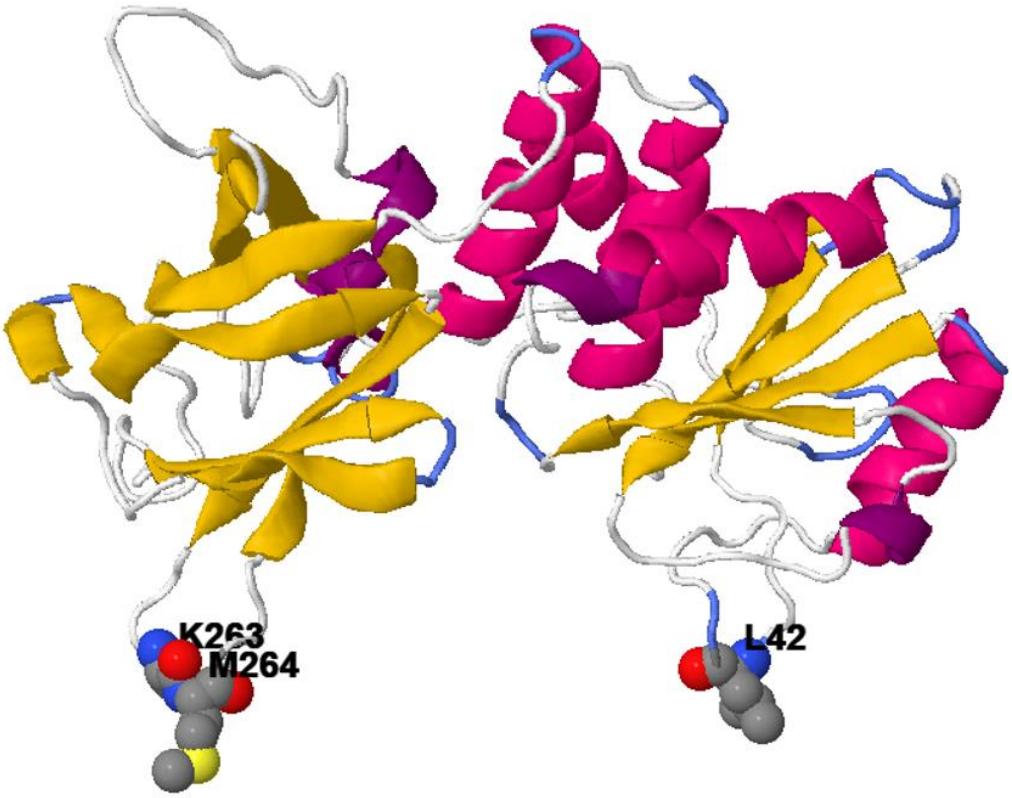
a) Display results

As a result the predicted membrane-penetrating residues are displayed in the format:

The residues: “Chain name” “resid” “resid” … “resid” “Chain name” “resid” “resid” … “resid” … are predicted to insert the membrane

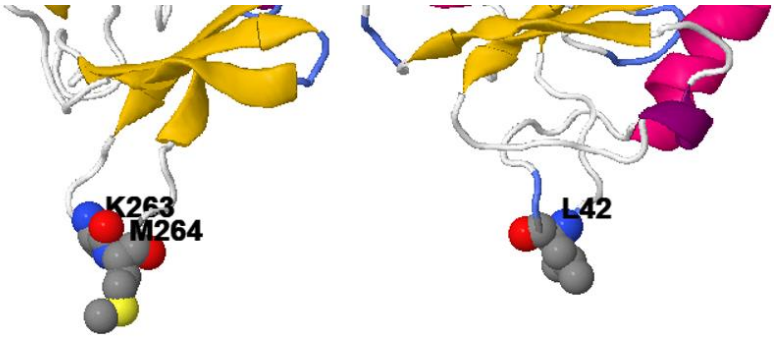
and visualized with JSmol(13, 14). The protein is represented in secondary structure, and the membrane-penetrating residues in CPK.

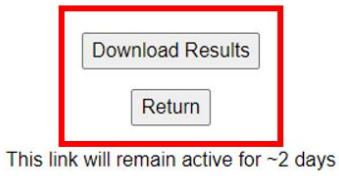
The residues: A 42 263 264 are predicted to insert the membrane.



b) Download results

Moreover, the user may download the results by hitting the “Download Results” button, or to return to the main page for performing further predictions by hitting the “Return” button. The buttons are placed just below the JSmol visualization.



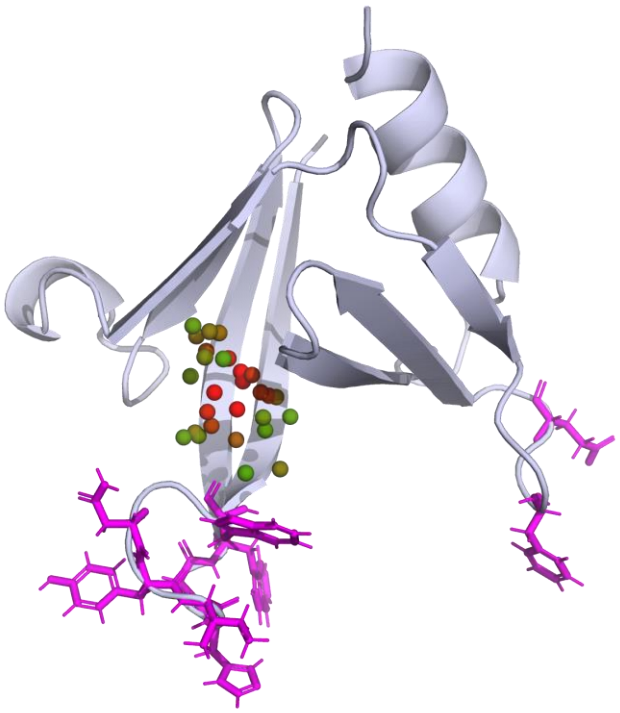


The downloaded results vary depending from the choice of predicting binding sites or not. If not, a .csv file will be downloaded including the prediction, and specifically, the chain, the resid, the one-code amino acid letter, and if the prediction is near N- or C-terminal, or near missing loops, which might be a possible false prediction (0 for no, and 1 for yes).

chain	resnum	Amino acid	broken_chain
A	42	L	0
A	263	K	0
A	264	M	0

If the user's choice was to predict binding sites, a .zip file will be downloaded including:

1. the abovementioned .csv file,
2. the ExProSE results (if used) in the "pdbs" folder,
3. the prepared structures in the "pdbs/prepared" folder,
4. the P2Rank predictions in the "pdbs/prepared/pockets" folder,
5. the PyMOL visualizations in the "pdbs/prepared/pockets/visualizations" folder,
6. and the results from the binding site clustering in the "pdbs/prepared/pockets/results" folder,



In the above example of the first model from the NMR structure 2RSG, only one of the 3 binding sites predicted by P2Rank is displayed in PyMOL, as it is the only one in a distance of 5 Å from the predicted membrane-penetrating residues.

If you encounter problems please feel free to contact us at:
a.chatzigoulas@gmail.com

4. Bibliography

1. Doerr S, Harvey MJ, Noe F, De Fabritiis G. HTMD: High-Throughput Molecular Dynamics for Molecular Discovery. J Chem Theory Comput. 2016;12(4):1845-52.
2. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983;22(12):2577-637.
3. Mitternacht S. FreeSASA: An open source C library for solvent accessible surface area calculations. F1000Res. 2016;5:189.
4. Sanner MF, Olson AJ, Spehner JC. Reduced surface: an efficient way to compute molecular surfaces. Biopolymers. 1996;38(3):305-20.
5. Wimley WC, White SH. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. Nat Struct Biol. 1996;3(10):842-8.
6. Wimley WC, Creamer TP, White SH. Solvation energies of amino acid side chains and backbone in a family of host-guest pentapeptides. Biochemistry. 1996;35(16):5109-24.
7. Dolinsky TJ, Nielsen JE, McCammon JA, Baker NA. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. Nucleic Acids Res. 2004;32(Web Server issue):W665-7.
8. Dolinsky TJ, Czodrowski P, Li H, Nielsen JE, Jensen JH, Klebe G, et al. PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. Nucleic Acids Res. 2007;35(Web Server issue):W522-5.
9. Remmert M, Biegert A, Hauser A, Soding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods. 2011;9(2):173-5.
10. Bakan A, Meireles LM, Bahar I. ProDy: protein dynamics inferred from theory and experiments. Bioinformatics. 2011;27(11):1575-7.
11. Bakan A, Dutta A, Mao W, Liu Y, Chennubhotla C, Lezon TR, et al. Evol and ProDy for bridging protein sequence evolution and structural dynamics. Bioinformatics. 2014;30(18):2681-3.
12. Ruiz-Blanco YB, Paz W, Green J, Marrero-Ponce Y. ProtDCal: A program to compute general-purpose-numerical descriptors for sequences and 3D-structures of proteins. BMC Bioinform. 2015;16:162.
13. Hanson RM. Jmol–a paradigm shift in crystallographic visualization. Journal of Applied Crystallography. 2010;43(5):1250-60.
14. Hanson RM, Prilusky J, Renjian Z, Nakane T, Sussman JL. JSmol and the Next-Generation Web-Based Representation of 3D Molecular Structure as Applied to Proteopedia. Israel Journal of Chemistry. 2013;53(3-4):207-16.
15. Krivak R, Hoksza D. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. J Cheminform. 2018;10(1):39.
16. Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 2.0 2015.