# mutTMPredictor

## (I) Datasets

#### (1) The 546 mutations

154 neutral missense mutations;

392 disease-associated missense mutations.

#### (2) The 67584 mutations

38564 neutral missense mutations;

29020 disease-associated missense mutations.

## (Ⅱ) Workflow of mutTMPredictor

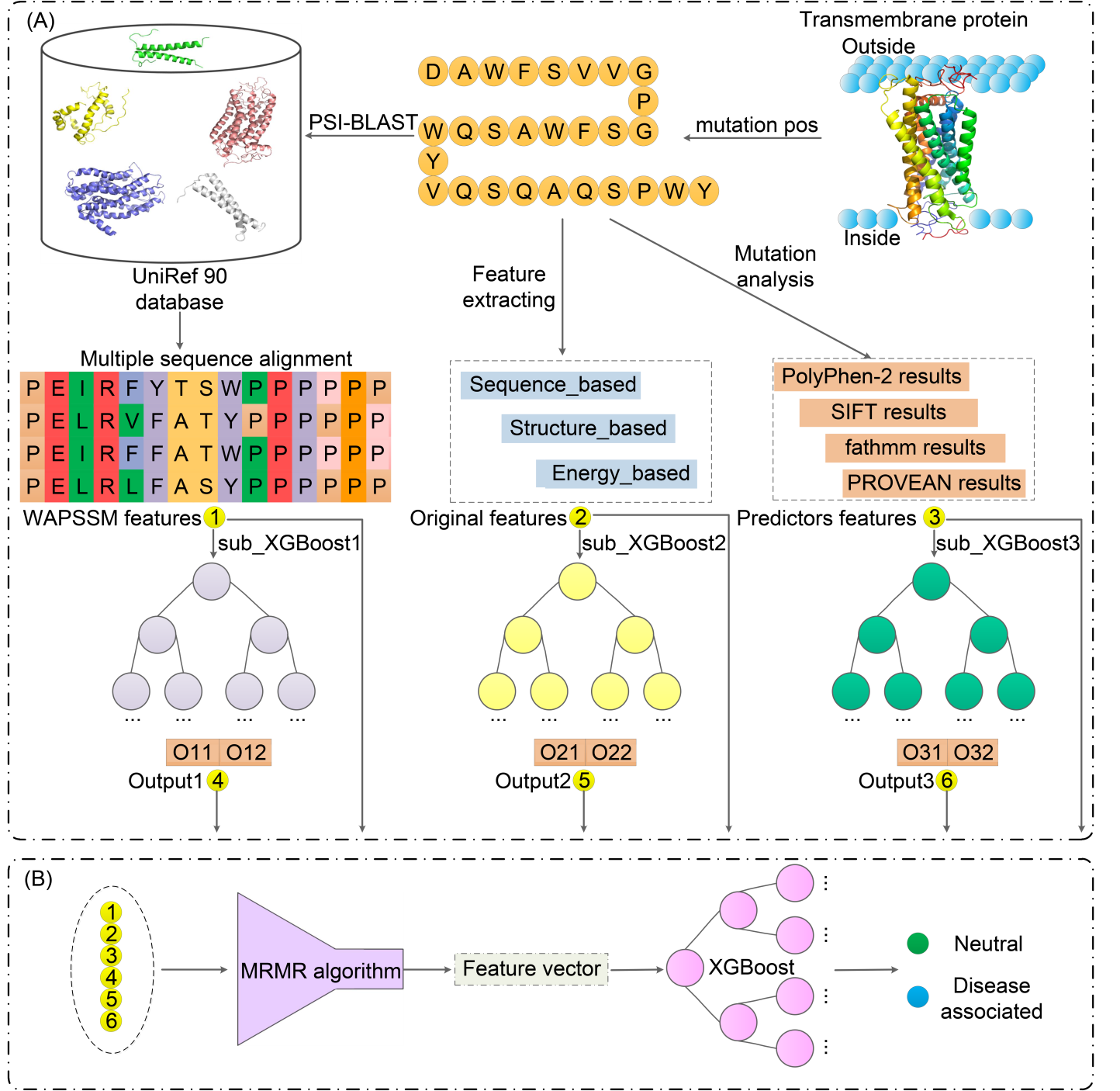


Figure 1. An overall workflow of the proposed MutTMPredictor.

Figure 1 is the workflow of the new proposed mutation predictor, named MutTMPredictor. As seen, Figure 1 comprises two parts: part (A) and (B). In Figure 1(A), we extracted four types of features for each mutation, including WAPSSM, Original features, Predictors’ output, and the prediction results of the above three features using XGBoost model. In Figure 1(B), we concatenated the extracted features for further MRMR selection. Then feed the sorted features into XGBoost model for the binary mutation classification.More details could be found in section 2 and section 3.1~3.2 of the manuscript, as below.

## (Ⅲ) Extracting feature matrix

### 1. Position Specific Score Matrix (PSSM)

We adopted the PSI-BLAST (Position-Specific Iterated Basic Local Alignment Search Tool) to generate PSSM information. For more detail information, please refer to:

Schaffer A A, Aravind L, Madden T L, et al. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements [J]. Nucleic Acids Research, 2001, 29(14): 2994-3005.

Software website: https://www.ebi.ac.uk/seqdb/confluence/display/THD/PSI-BLAST.

After calculating the PSSM matrix, we calculated the weight attenuation Position Specific Score Matrix (WAPSSM) features for each protein mutation according to the WAPSSM algorithm.

### 2. Protein sequence-based, structure-based, and energy-based features

Protein sequence-based features mainly include physic-chemical properties, such as hydrophilicity (KUHL950101), amphiphilicity (MITS020101), bulkiness (ZIMJ680102), polarity (GRAR740102), polarizability (CHAM820101), isoelectric point (ZIMJ680104), accessible surface area in a tripeptide (CHOC760101), number of hydrogen bond donors (FAUJ880109), net charge (KLEP840101), radius of gyration of the side chain (LEVM760105), amino acid composition (CEDJ970103), and side-chain contribution to protein stability (TAKK010101), which are collected from AAindex. Protein structure-based features and protein energy-based features mainly were conducted by ICM-Pro and CompoMug

### 3. Four famous existing missense mutation predictors

In the present study, we also utilized the prediction results of four famous protein mutation predictors. Together with the above two types of features, we concatenated them to form the feature vector for mutation representation. The webserver addresses are listed as below.

SIFT webserver: https://github.com/rvaser/sift4g

PolyPhen-2 webserver: http://genetics.bwh.harvard.edu/pph2

fathmm webserver: http://fathmm.biocompute.org.uk/cancer.html

PROVEAN webserver: http://provean.jcvi.org

### 4. Important references

##### [1] AAindex: Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. AAindex: amino acid index database, progress report 2008. Nucleic Acids Res. 2008; 36(Database issue):D202–5.

##### [2] CompoMug: Petr P , Yao P , Ling S , et al. Computational design of thermostabilizing point mutations for G protein-coupled receptors[J]. Elife, 2018, 7:e34729.

##### [3] BorodaTM: P. Popov, I. Bizin, and M. Gromiha, “Prediction of disease-associated mutations in the transmembrane regions of proteins with known 3D structure,” *PloS One*. 14. 7, 2019.

##### [4] MutHTP database: A. Kulandaisamy, S. Binny Priya, R. Sakthivel, S. Tarnovskaya, I. Bizin, P. Hönigschmid, D. Frishman, and M. M. Gromiha, “MutHTP: mutations in human transmembrane proteins,” Bioinformatics. 34. 13. 2325-2326, 2018.

##### [5] SIFT: P. C. Ng, and S. Henikoff, “SIFT: predicting amino acid changes that affect protein function,” Nucleic Acids Research. 31. 13. 3812-3814, 2003.

##### [6] PolyPhen-2: I. Adzhubei, D. M. Jordan, and S. R. Sunyaev, “Predicting functional effect of human missense mutations using PolyPhen-2,” Current Protocols in Human Genetics. Chapter 7. Unit7 20, 2013.

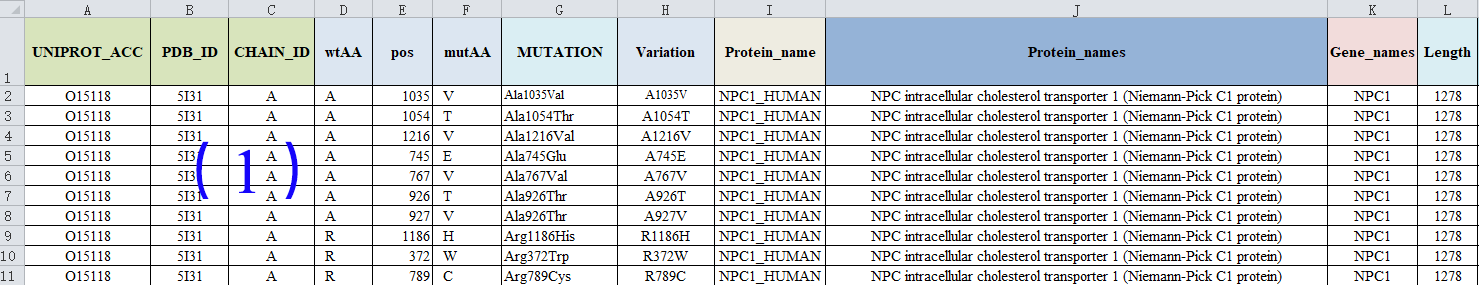
##### [7] fathmm: H. A. Shihab, J. Gough, D. N. Cooper, P. D. Stenson, G. L. Barker, K. J. Edwards, I. N. Day, and T. R. Gaunt, “Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models,” *Human Mutation*. 34. 1. 57-65, 2013.

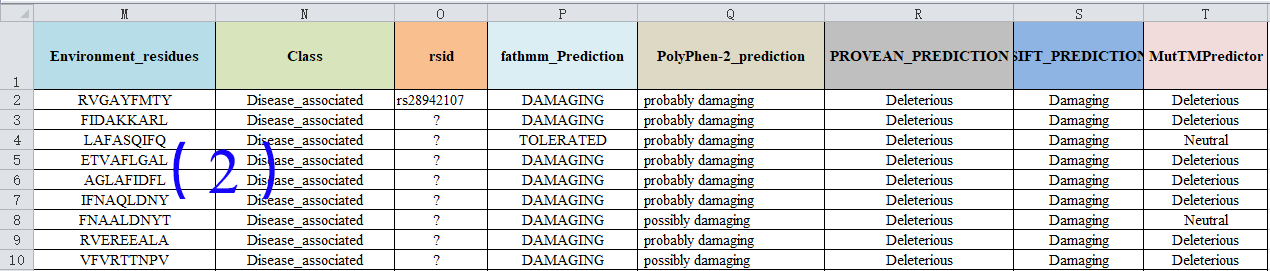
##### [8] PROVEAN: Y. Choi, and A. P. Chan, “PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels,” Bioinformatics. 31. 16. 2745-2747, 2015.

## (Ⅳ) Mutation Prediction

### (1) 546 mutations dataset

As for the 546 mutation dataset, we upload the **EXCEL FILE** which contains the detailed information for each mutation, as listed in "(2) Detailed information of 546 mutations and predictors' output", displayed in mutTMPredictor webserver (<http://csbio.njust.edu.cn/bioinf/muttmpredictor>). Besides, we also displayed the meaning of each row of the EXCEL FILE in "The rows of 546 mutations excel file" part of the webserver, as bellows:





### (1) 67584 mutations dataset

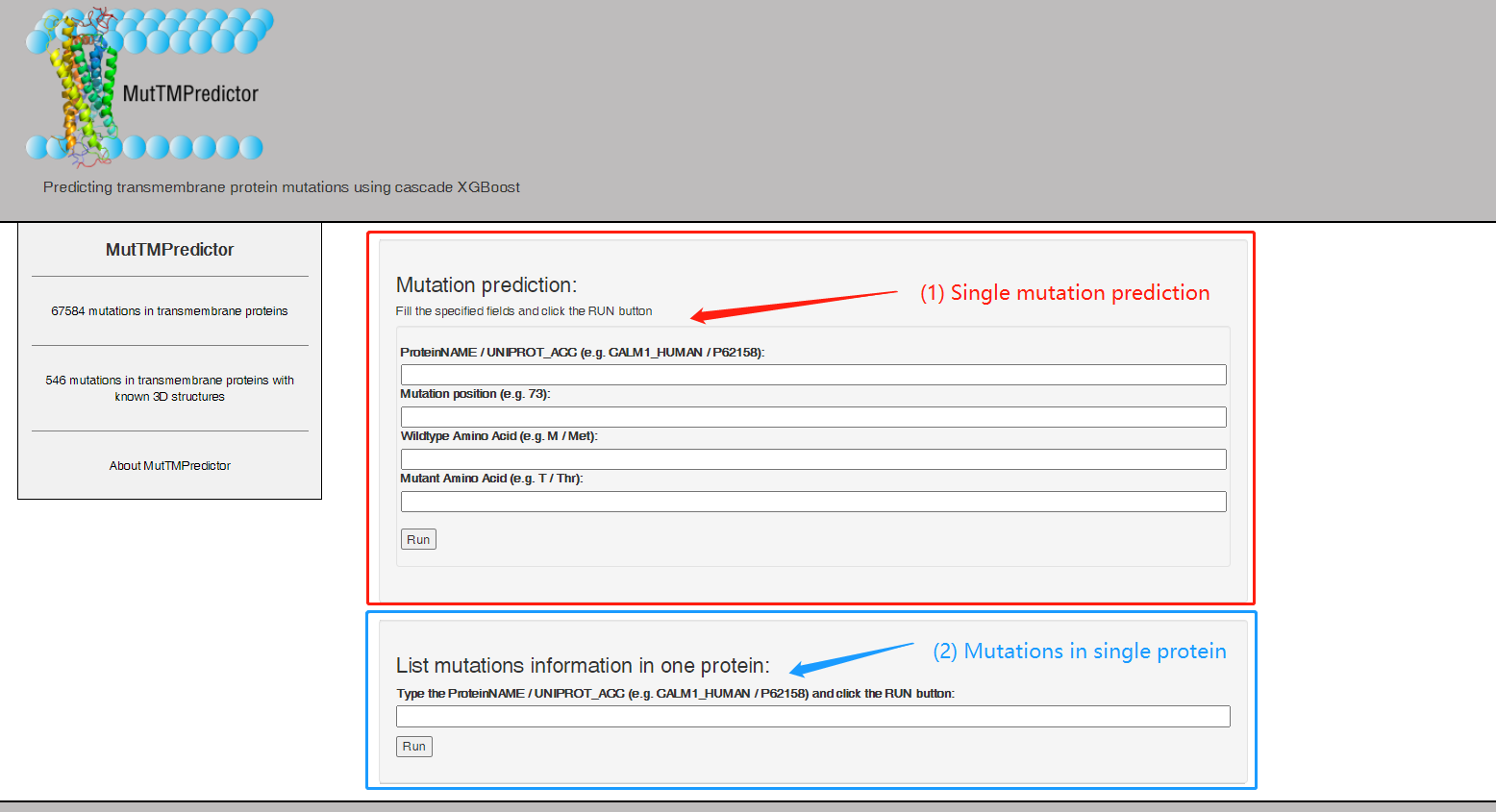
### Single mutation prediction

Users provide the UINPROT\_ACC/proteinName, mutant position, wild\_type amino acid, and mutant amino acid information, and the result page will display matched prediction result, specifically including the “UNIPROT\_ACC”, “ProteinNAME”, “ProteinLENGTH”, “wtAA”, “pos”, “mutAA”, “Variation”, “Environment\_residues”, “Class”, “rsid”, “fathmm\_Prediction”, “PolyPhen-2\_prediction”, “PROVEAN\_PREDICTION”, “SIFT\_PREDICTION”, and “MutTMPredictor” items of the corresponding mutant.

### Mutations in single protein

Users provide the UINPROT\_ACC/proteinName, and the result page will display the predicted resutls of multiple mutations in the input protein. Specifically the results include “UNIPROT\_ACC”, “ProteinNAME”, “ProteinLENGTH”, “wtAA”, “pos”, “mutAA”, “Variation”, “Environment\_residues”, “Class”, “rsid”, “fathmm\_Prediction”, “PolyPhen-2\_prediction”, “PROVEAN\_PREDICTION”, “SIFT\_PREDICTION”, and “MutTMPredictor” items of the protein mutations.

The details of single mutation prediction and mutations in single protein is listed as below.



## (Ⅴ) Contact

If you are interested in our work, OR, if you have any suggestions/questions about our work, PLEASE contact with us. E-mail: gfang0616@njust.edu.cn