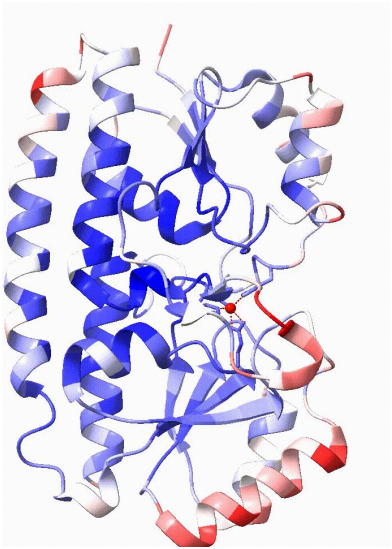
Advancing metal-binding protein predictions with deep learning

Jingkai LAN

Supervisor: Thomas Lemmin
Co-supervisor: Giulia Peteani

Introduction

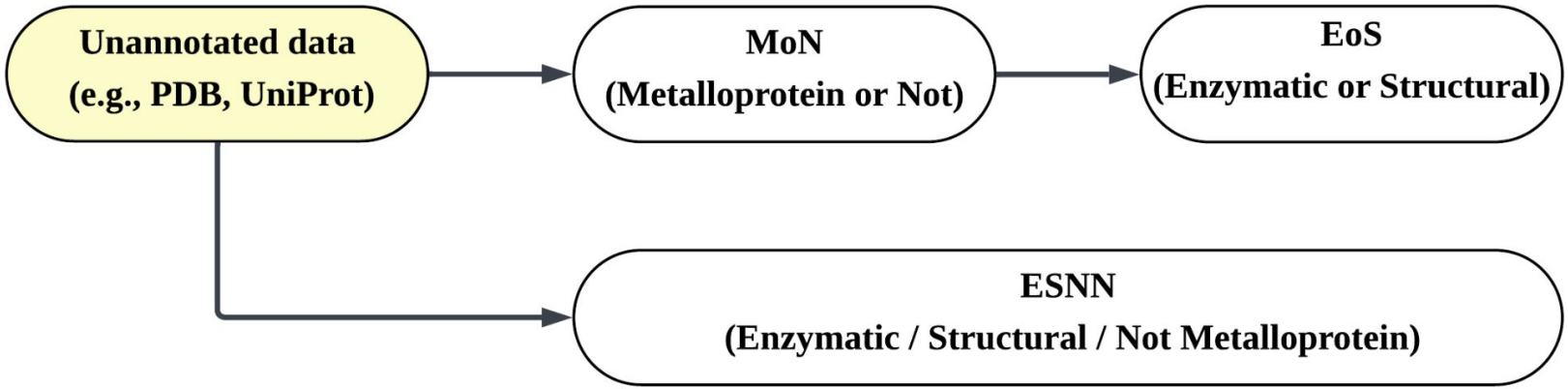
- **Metalloproteins** play essential roles in biological systems, contributing to diverse processes such as **enzymatics, regulation**, and **structural stability**.
- A substantial proportion of **enzymes** are **metal-binding**, with estimates suggesting that nearly **half require metal ion for activity**.
- **Annotation** of metalloproteins remains **incomplete** in major biological **databases** (e.g., PDB, UniProt), limiting comprehensive understanding and systematic studies.



H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank (2000) Nucleic Acids Research 28: 235-242 https://doi.org/10.1093/nar/28.1.235.
The UniProt Consortium , UniProt: the Universal Protein Knowledgebase in 2025, Nucleic Acids Research, Volume 53, Issue D1, 6 January 2025, Pages D609–D617, https://doi.org/10.1093/nar/gkae1010

Project goal

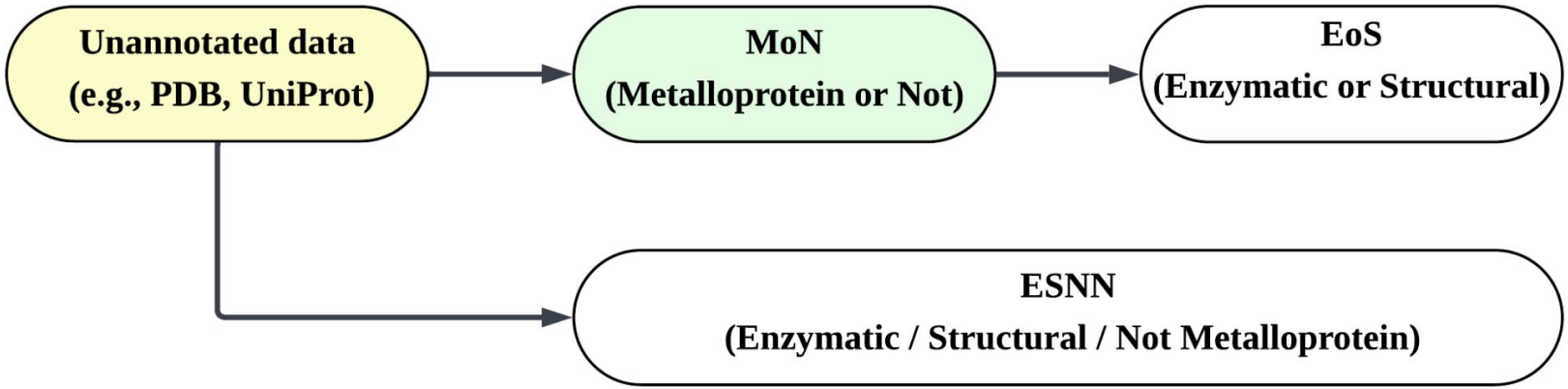
- Training models to perform **classification tasks**, thereby assisting protein annotation.



H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank (2000) Nucleic Acids Research 28: 235-242 https://doi.org/10.1093/nar/28.1.235.
The UniProt Consortium , UniProt: the Universal Protein Knowledgebase in 2025, Nucleic Acids Research, Volume 53, Issue D1, 6 January 2025, Pages D609–D617, https://doi.org/10.1093/nar/gkae1010

Project goal

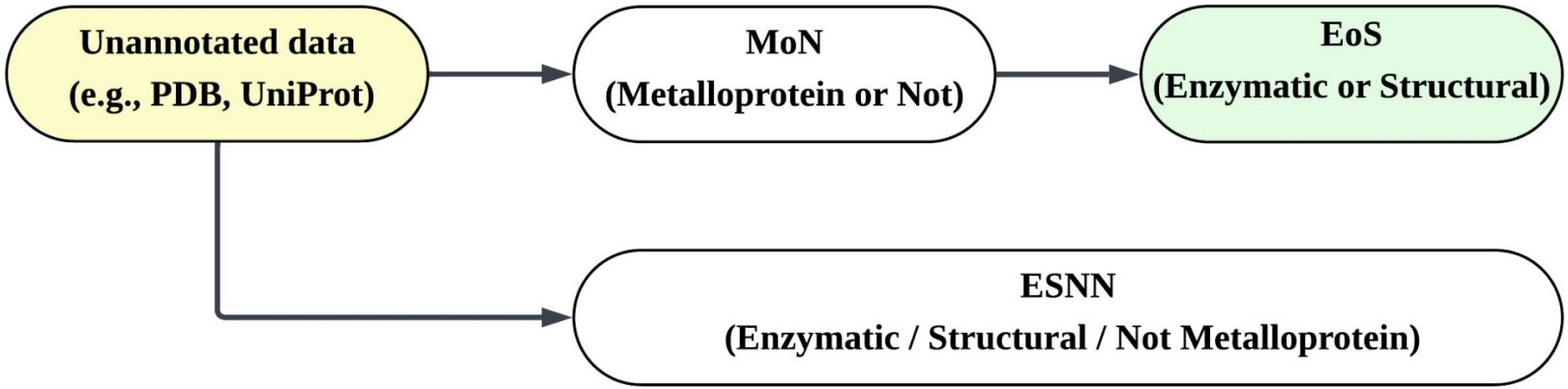
- Training models to perform **classification tasks**, thereby assisting protein annotation.



H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank (2000) Nucleic Acids Research 28: 235-242 https://doi.org/10.1093/nar/28.1.235.
The UniProt Consortium , UniProt: the Universal Protein Knowledgebase in 2025, Nucleic Acids Research, Volume 53, Issue D1, 6 January 2025, Pages D609–D617, https://doi.org/10.1093/nar/gkae1010

Project goal

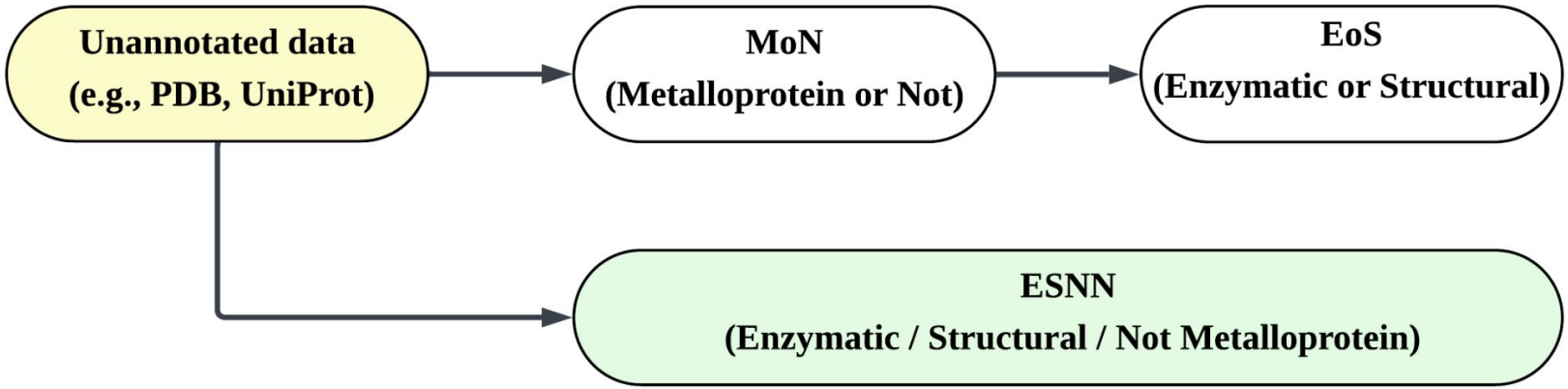
- Training models to perform **classification tasks**, thereby assisting protein annotation.



H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank (2000) Nucleic Acids Research 28: 235-242 https://doi.org/10.1093/nar/28.1.235.
The UniProt Consortium , UniProt: the Universal Protein Knowledgebase in 2025, Nucleic Acids Research, Volume 53, Issue D1, 6 January 2025, Pages D609–D617, https://doi.org/10.1093/nar/gkae1010

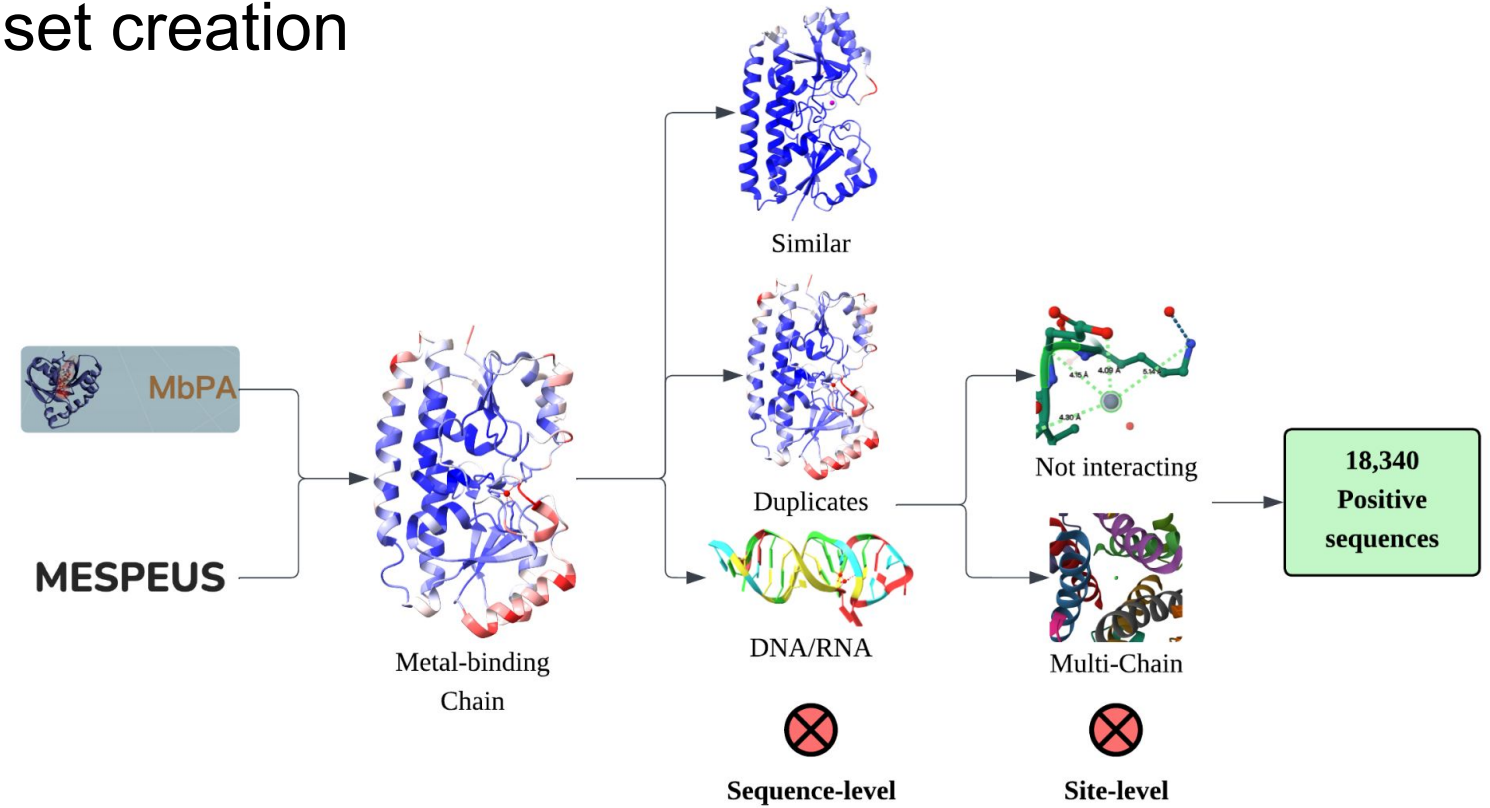
Project goal

- Training models to perform **classification tasks**, thereby assisting protein annotation.



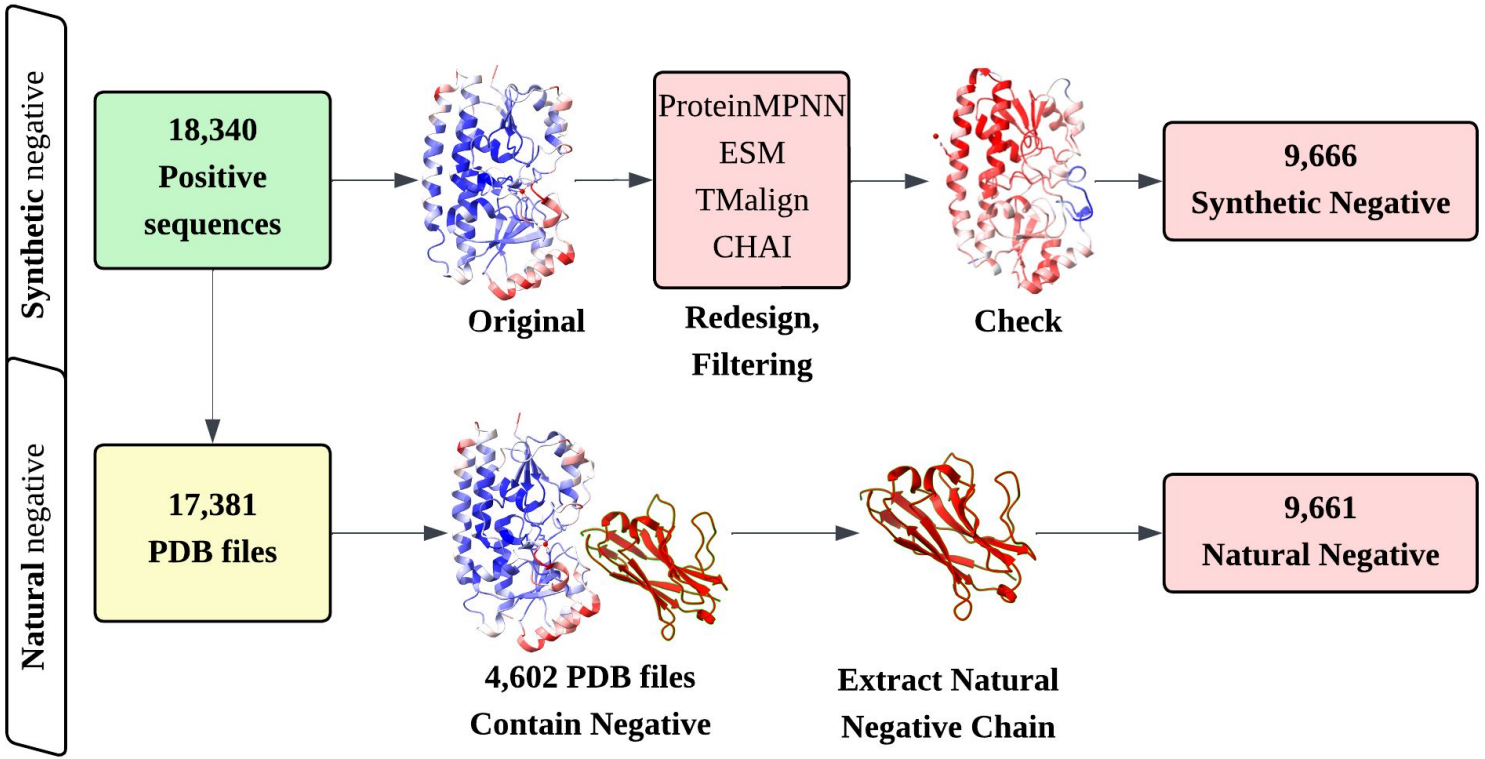
H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank (2000) Nucleic Acids Research 28: 235-242 https://doi.org/10.1093/nar/28.1.235.
The UniProt Consortium , UniProt: the Universal Protein Knowledgebase in 2025, Nucleic Acids Research, Volume 53, Issue D1, 6 January 2025, Pages D609–D617, https://doi.org/10.1093/nar/gkae1010

Dataset creation



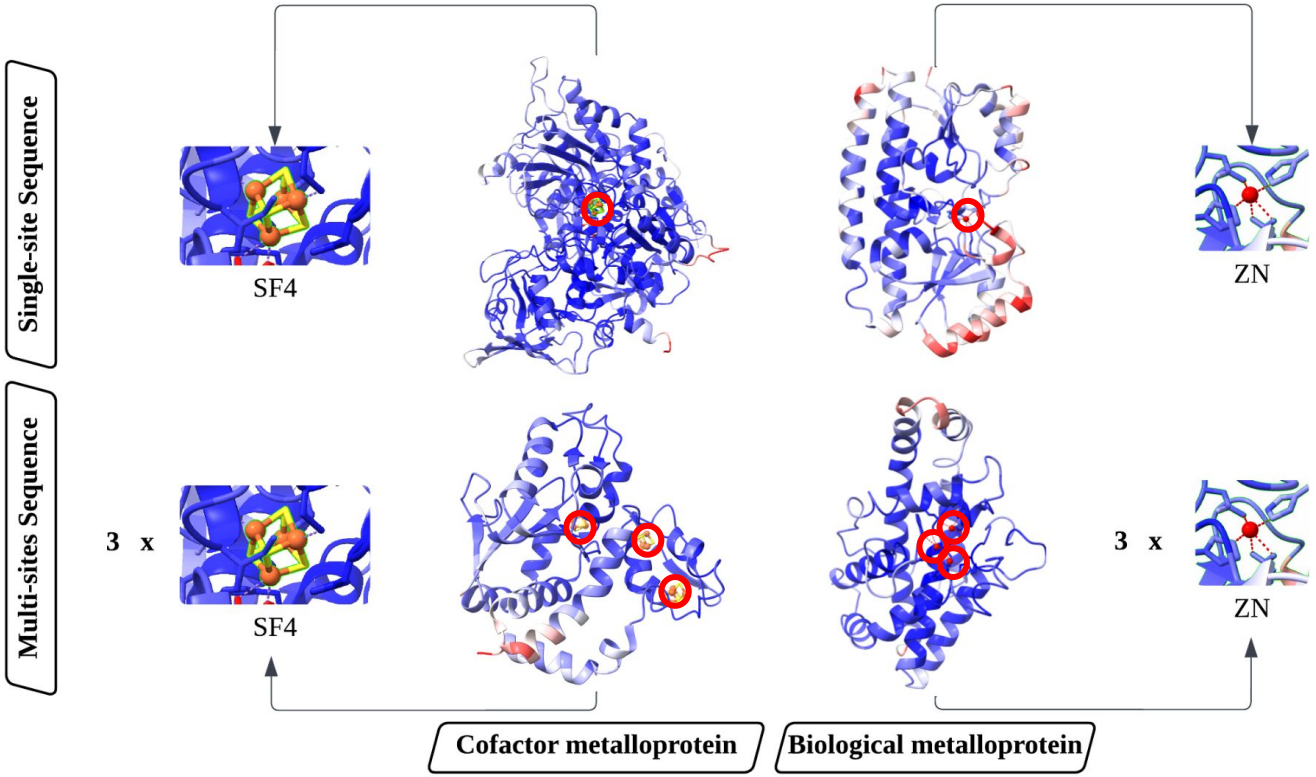
Li J, He X, Gao S, Liang Y, Qi Z, Xi Q, Zuo Y, Xing Y. The Metal-binding Protein Atlas (MbPA): An Integrated Database for Curating Metalloproteins in All Aspects. J Mol Biol. 2023 Jul 15;435(14):168117. doi: 10.1016/j.jmb.2023.168117.
Geng-Yu Lin, Yu-Cheng Su, Yen Lin Huang, Kun-Yi Hsin, MESPEUS: a database of metal coordination groups in proteins, Nucleic Acids Research, Volume 52, Issue D1, 5 January 2024, Pages D483–D493, https://doi.org/10.1093/nar/gkad1009

Dataset creation

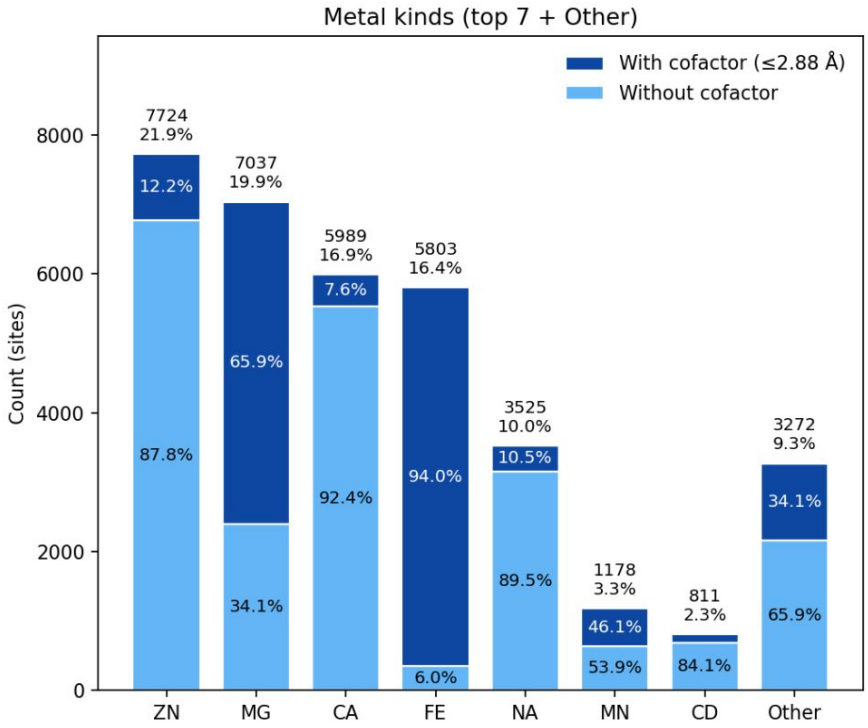


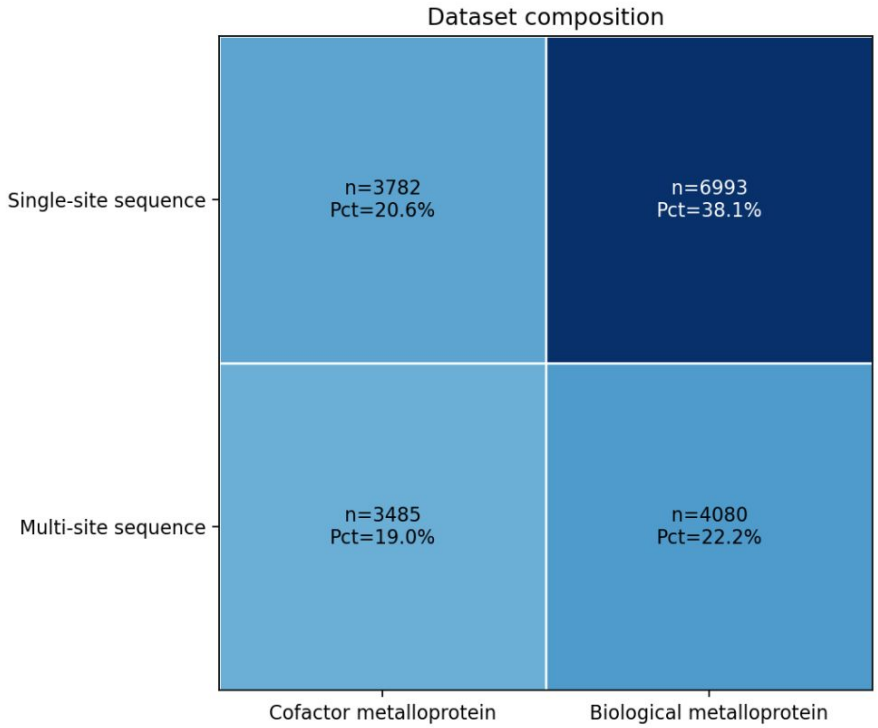
A A Aptekmann, J Buongiorno, D Giovannelli, M Glamoclija, D U Ferreiro, Y Bromberg, mebipred: identifying metal-binding potential in protein sequence, Bioinformatics, Volume 38, Issue 14, July 2022, Pages 3532–3540
Zeming Lin et al. ‚Evolutionary-scale prediction of atomic-level protein structure with a language model.Science379,1123-1130(2023).DOI:10.1126/science.ade2574
Y. Zhang, J. Skolnick, TM-align: A protein structure alignment algorithm based on TM-score, Nucleic Acids Research, 33: 2302-2309 (2005).
Chai Discoveryet al. ‚Chai-1: Decoding the molecular interactions of life. October 2024

Dataset distribution

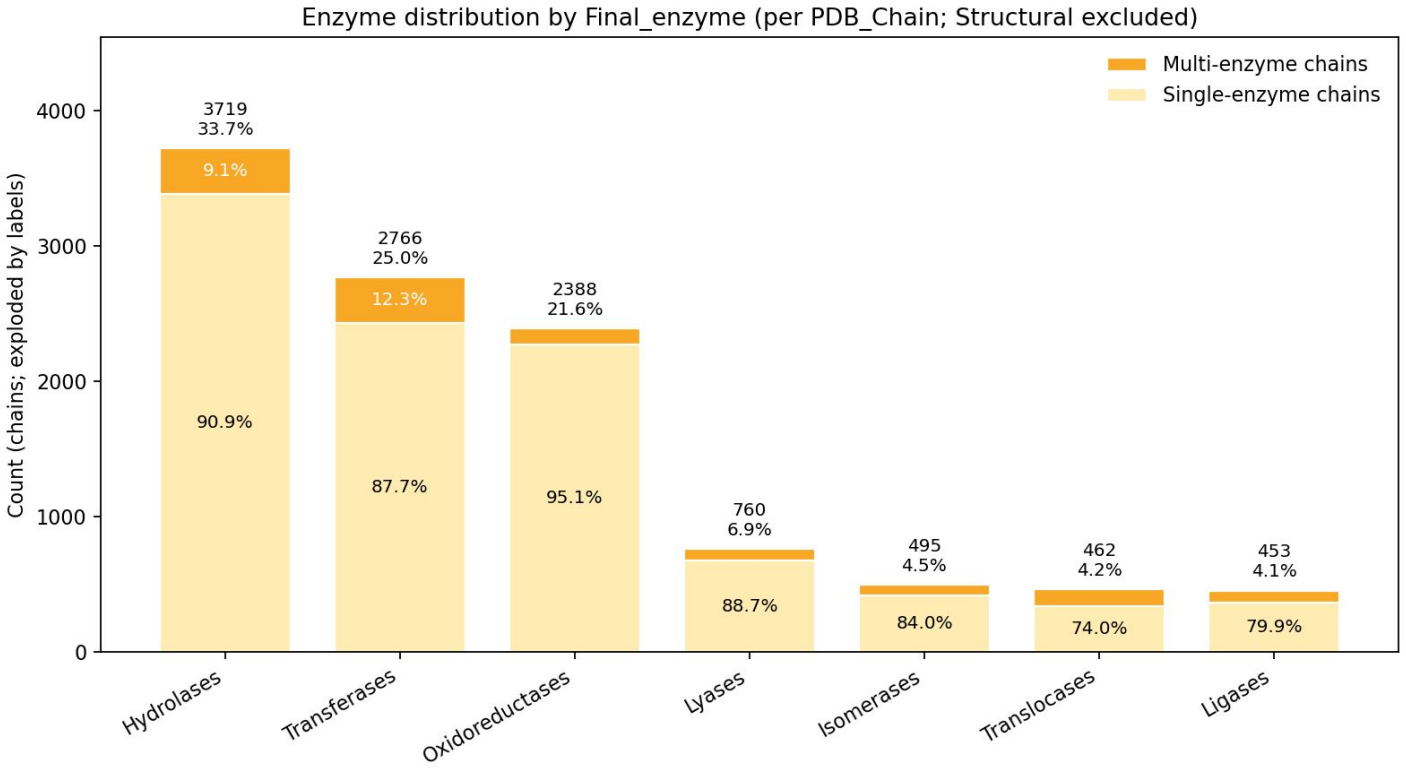


Dataset distribution



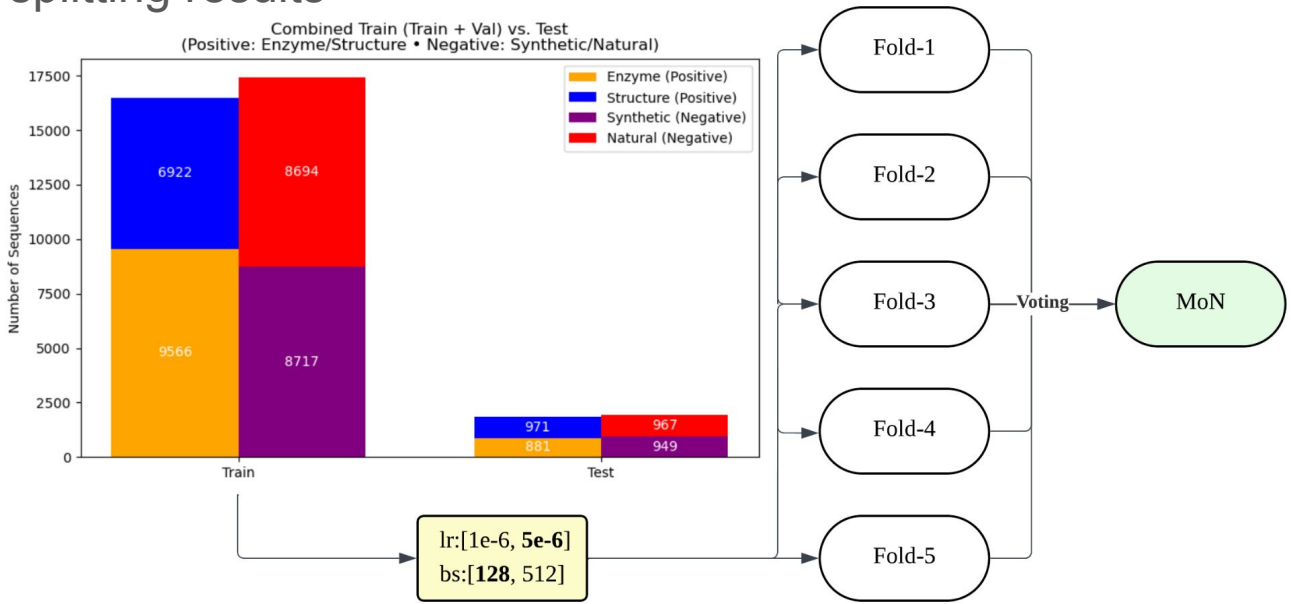


Dataset distribution



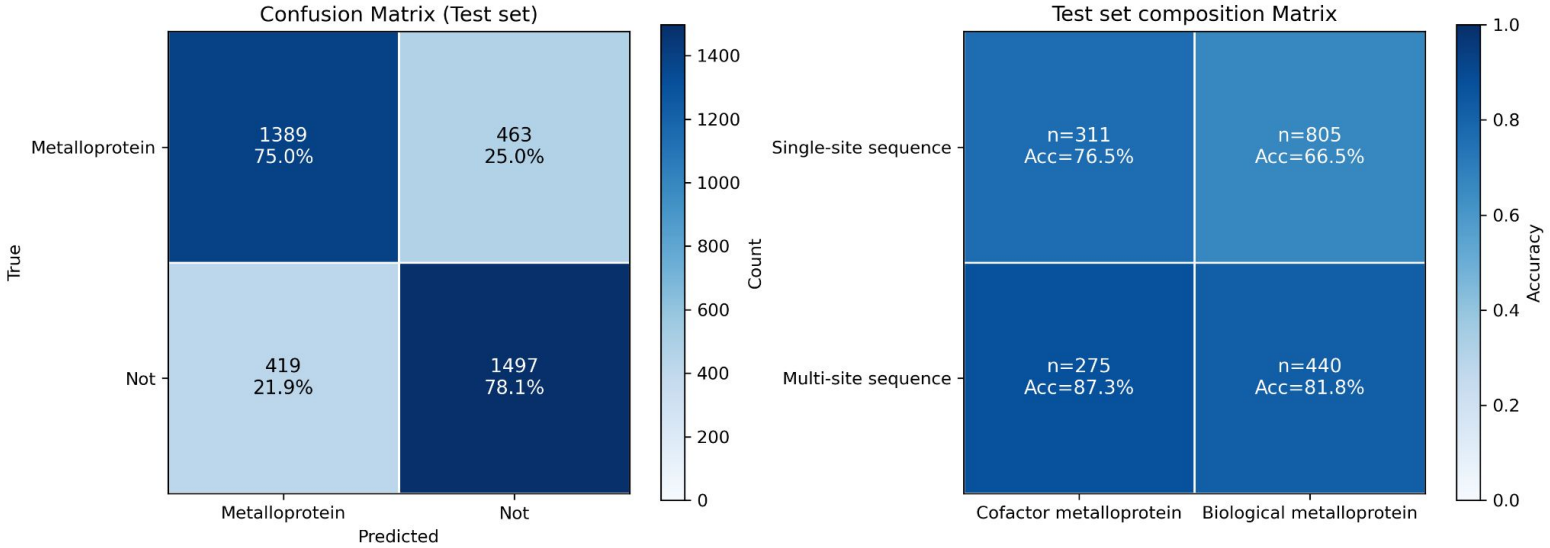
MoN - Meltalloprotein or Not

- Merge Positive and Synthetic Negative → MMseq2 at 30 pident (9:1)
- MMseq2 Natural Negative set
- Merge splitting results



MoN - Meltalloprotein or Not

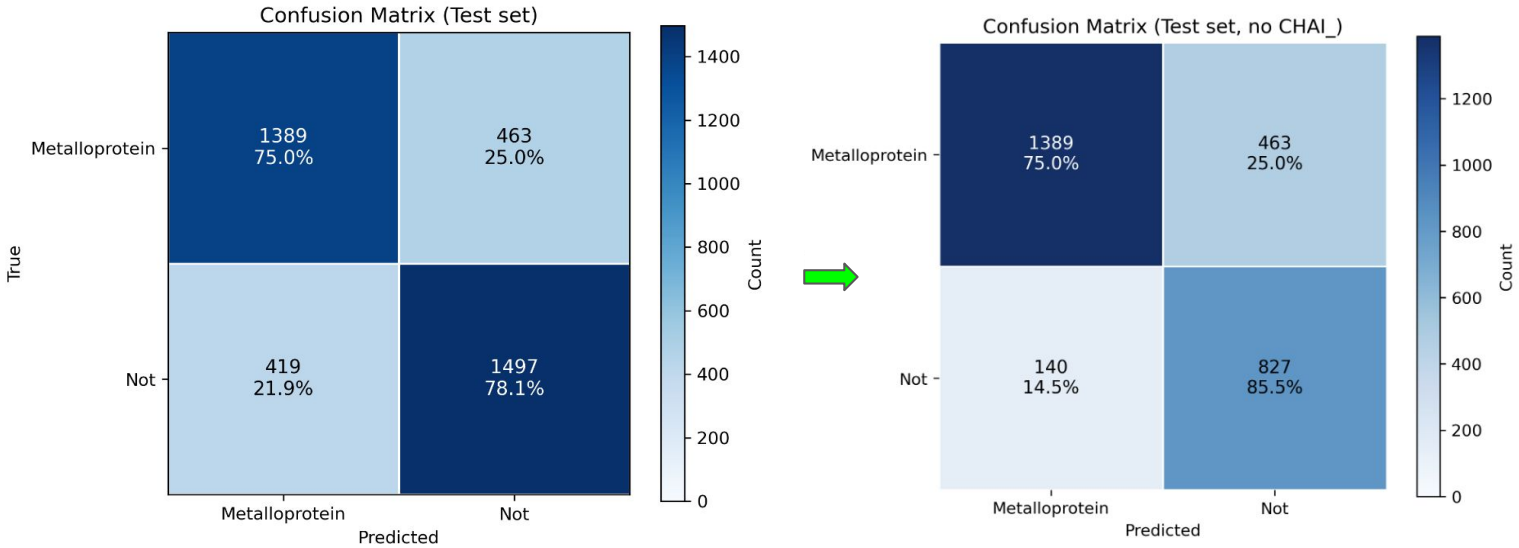
Model	Accuracy	Precision	Recall	F1 Score
Fold 1	75.11%	75.91%	72.30%	74.06%
Fold 2	75.90%	76.43%	73.70%	75.04%
Fold 3	75.72%	76.34%	73.33%	74.80%
Fold 4	73.94%	73.72%	73.00%	73.36%
Fold 5	74.95%	75.11%	73.33%	74.21%
MoN (Voting)	**76.59%**	**76.83%**	**75.00%**	**75.90%**



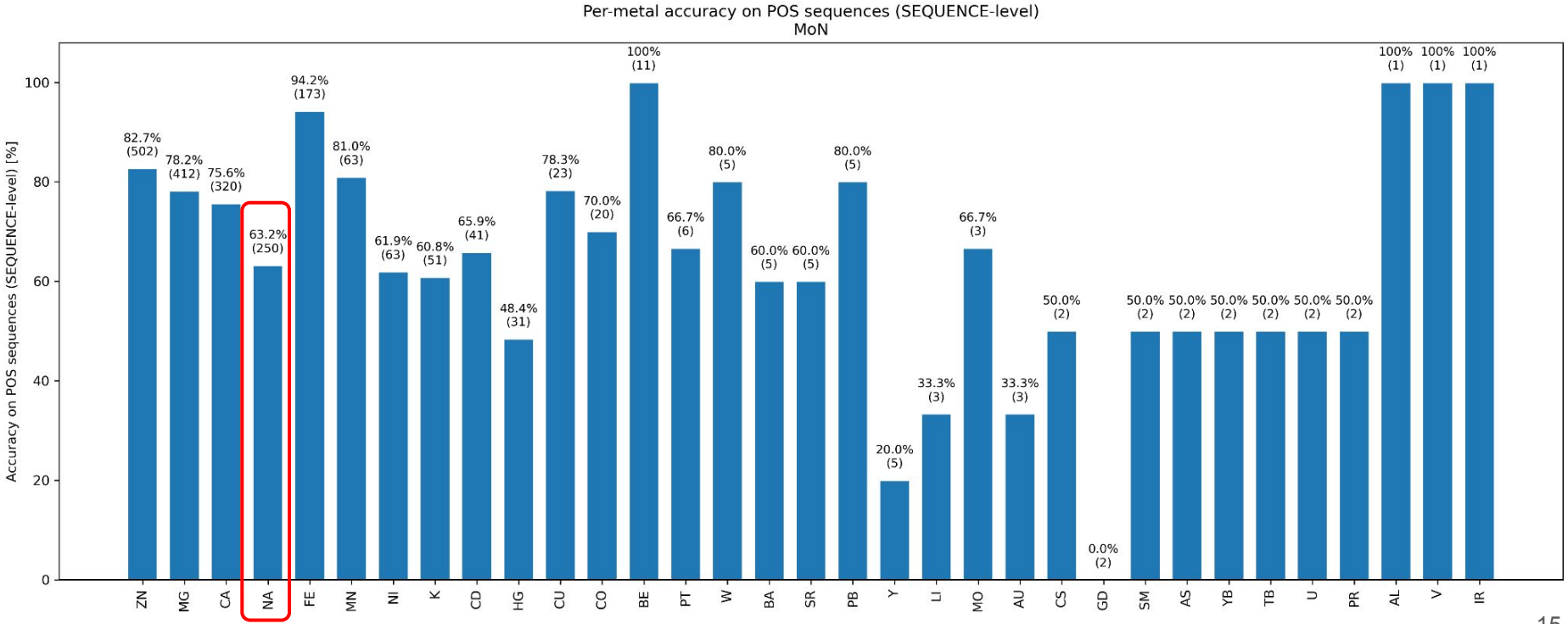
MoN - Meltalloprotein or Not

If we **evaluate** the model only on the **test set** that **contains Natural Negatives**:

Model	Accuracy	Precision	Recall	F1 Score
MoN (Voting)	**76.59%**	**76.83%**	**75.00%**	**75.90%**
MoN (Natural test)	**+ 2.02% = 78.61%**	**+14.01% = 90.84%**	**+0% = 75.00%**	**+6.27% = 82.17%**

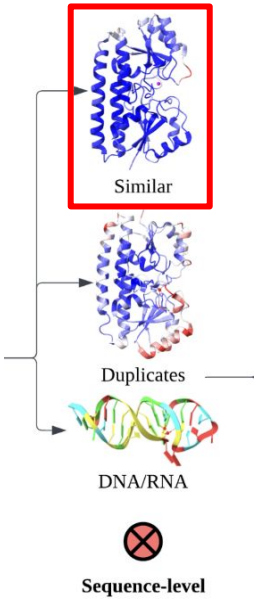


MoN - Meltalloprotein or Not

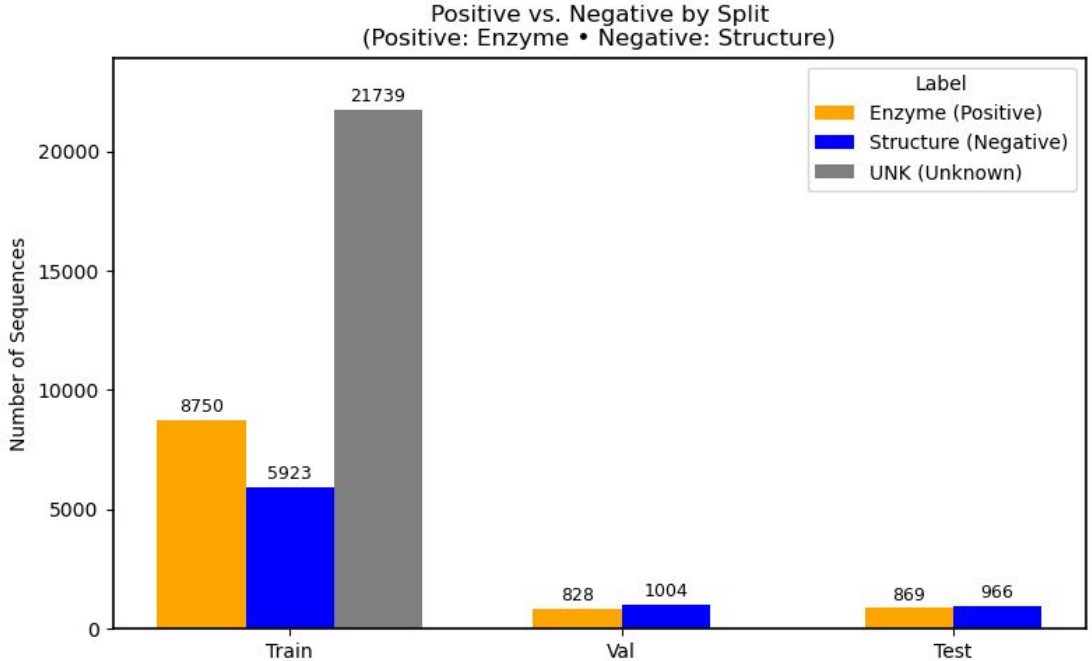


EoS - Enzymatic or Structural

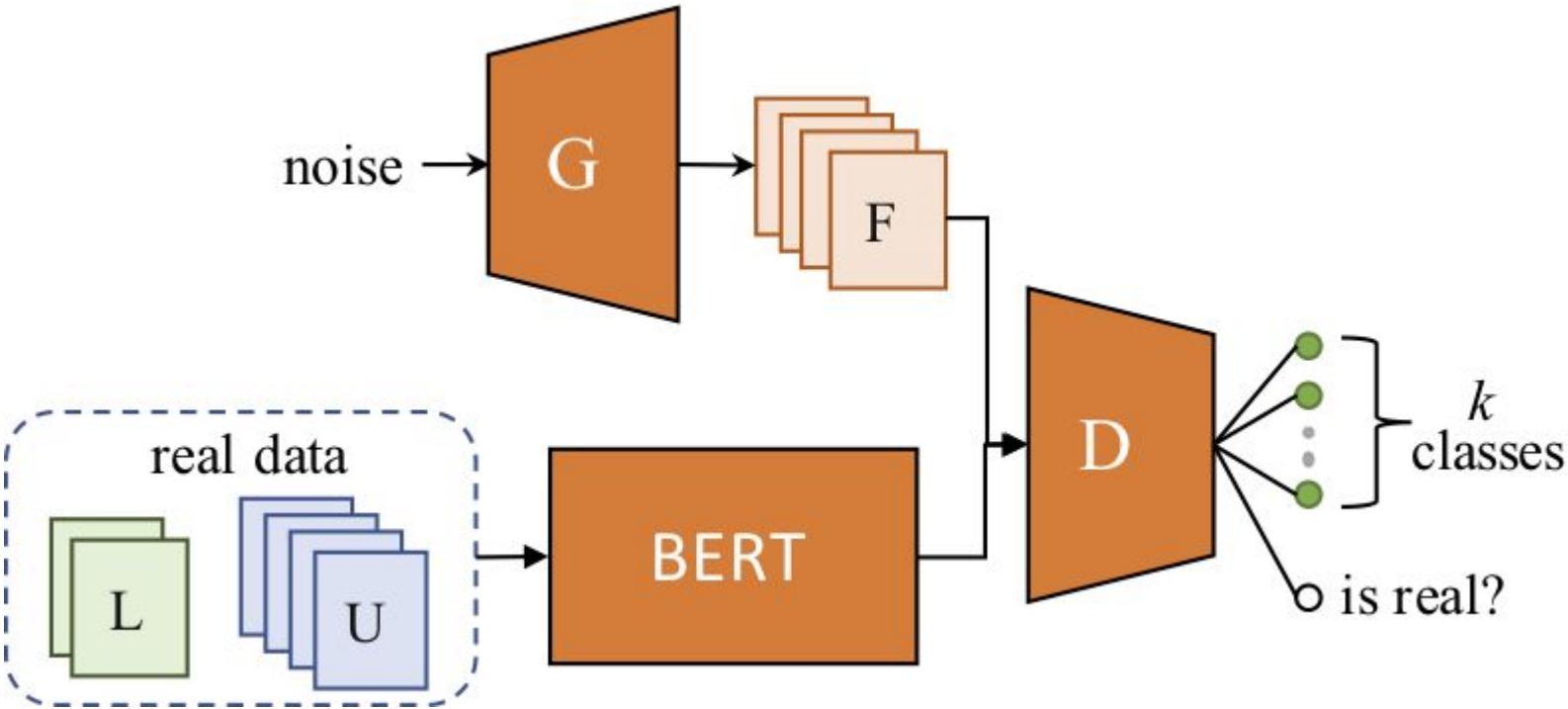
- MMseq2 Positive set at 30 pident (8:1:1)
- Using the previously discarded similar sequences as unlabeled data.



~21k sequences

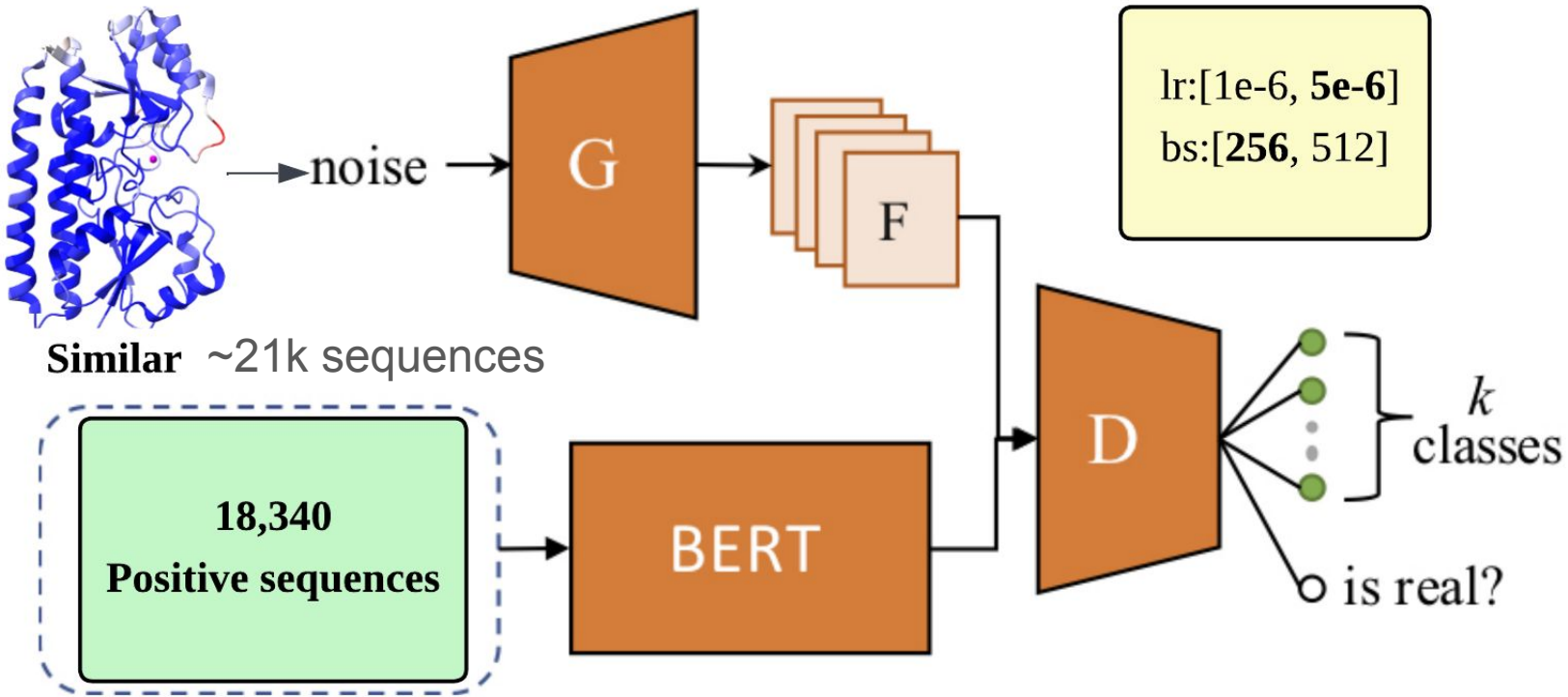


EoS - Enzymatic or Structural



Danilo Croce, Giuseppe Castellucci, and Roberto Basili. 2020. GAN-BERT: Generative Adversarial Learning for Robust Text Classification with a Bunch of Labeled Examples. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2114–2119, Online. Association for Computational Linguistics.

EoS - Enzymatic or Structural



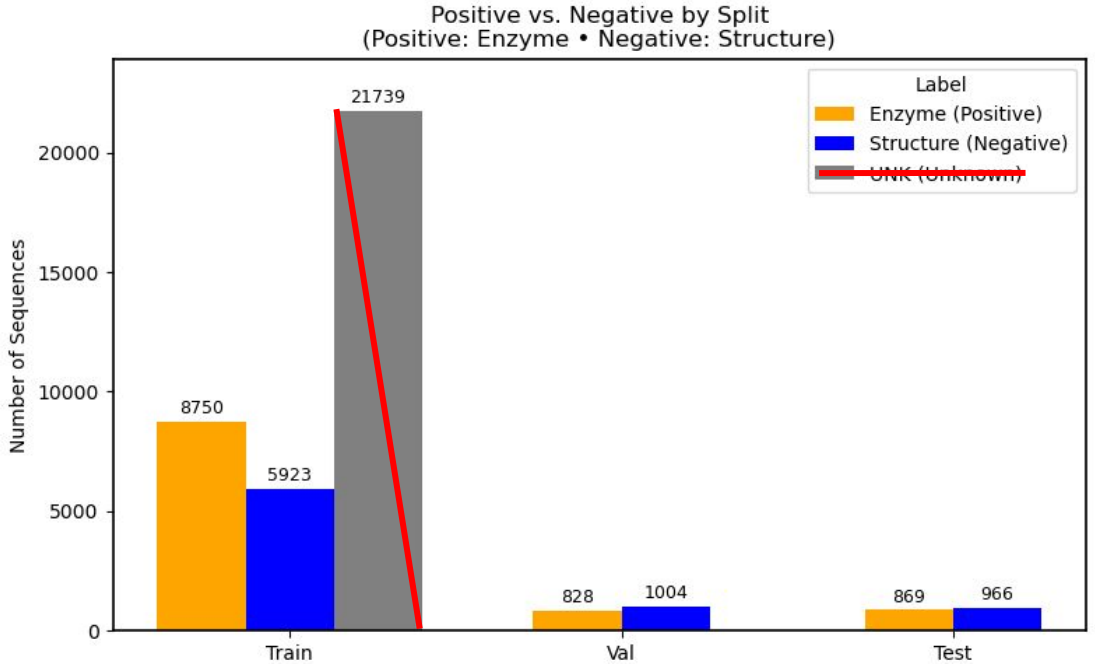
Danilo Croce, Giuseppe Castellucci, and Roberto Basili. 2020. GAN-BERT: Generative Adversarial Learning for Robust Text Classification with a Bunch of Labeled Examples. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2114–2119, Online. Association for Computational Linguistics.

EoS - Enzymatic or Structural

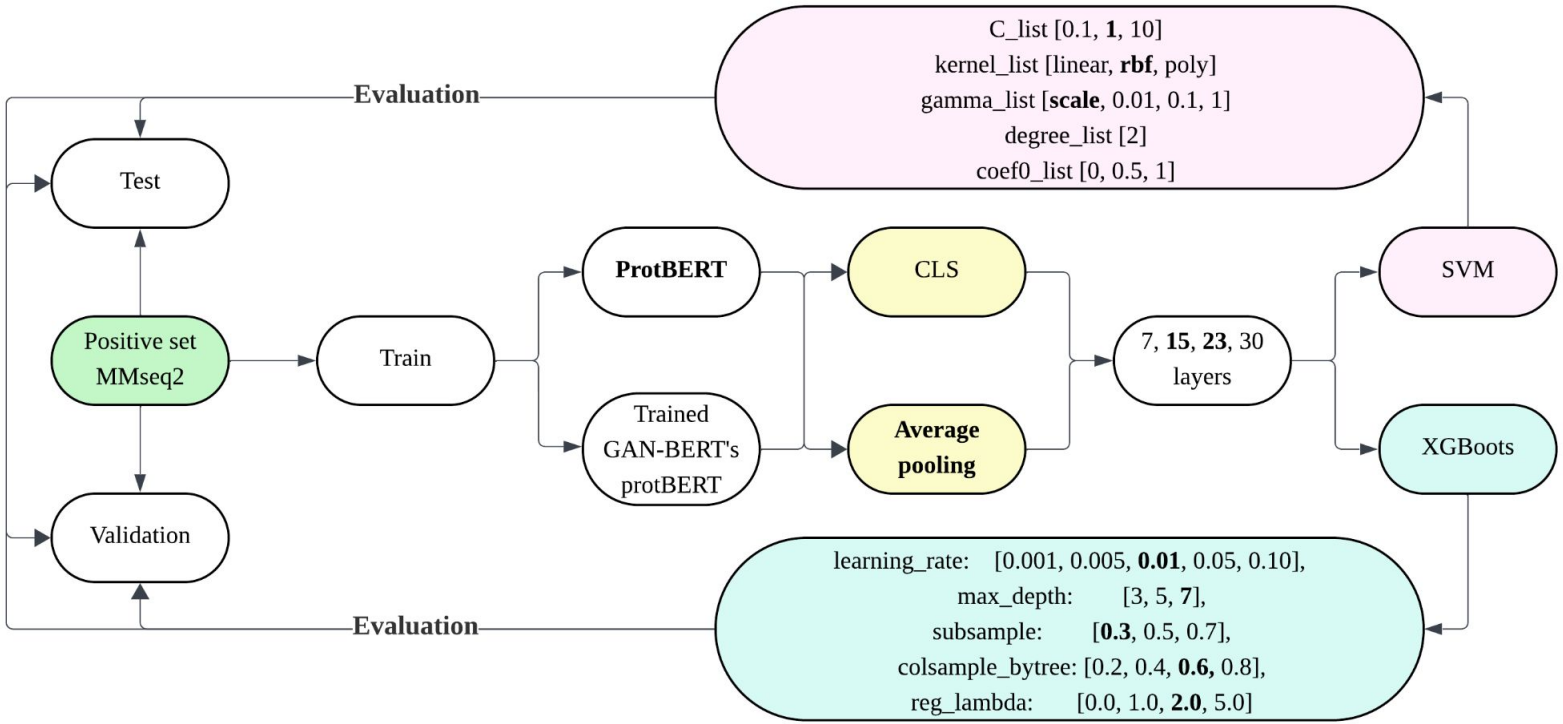
- MMseq2 Positive set at 30 pident (8:1:1) - same as before
- Using **class weights** to handle **class imbalance**.
- Remove unlabeled data.
- $$w_c = \frac{N}{K \times n_c}$$

Enzymatic weight = 0.8384

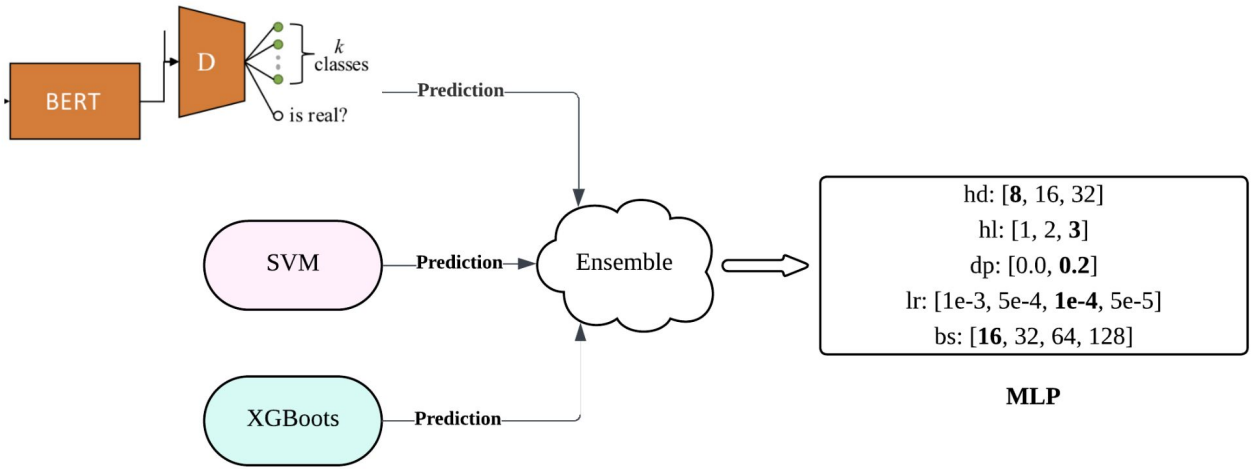
Structural weight = 1.2386



EoS - Enzymatic or Structural



EoS - Enzymatic or Structural



Model	Accuracy	Precision	Recall	F1 Score
GANBERT	72.75%	71.23%	71.23%	71.23%
SVM	73.24%	72.08%	71.00%	71.54%
XGBoost	73.19%	70.12%	**75.60%**	72.76%
EoS(Ensemble)	**74.11%**	**72.23%**	73.65%	**72.93%**

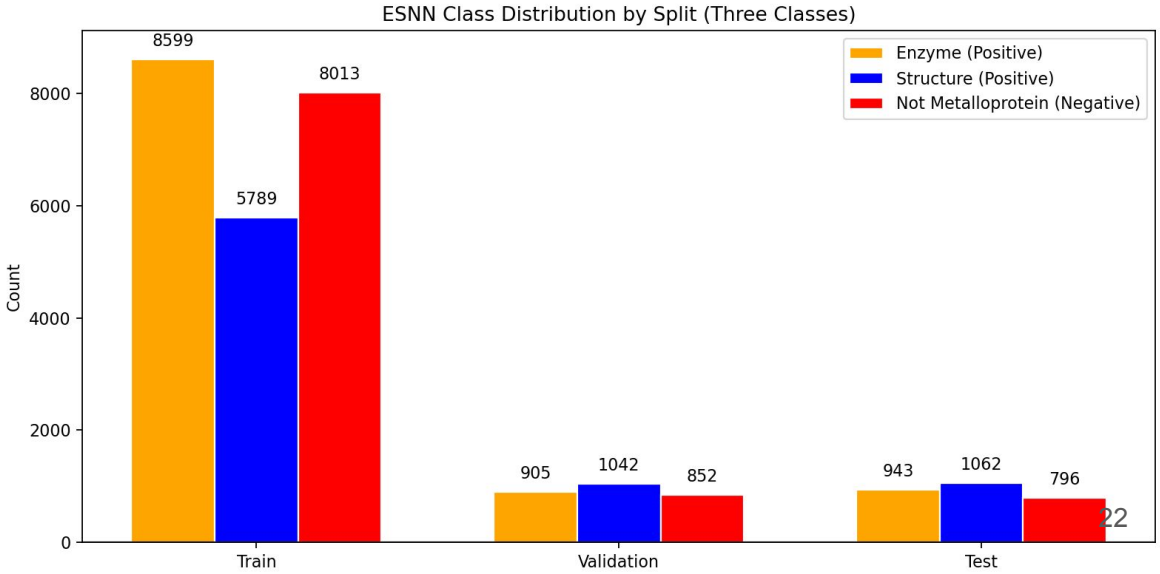
ESNN - Enzymatic, Structural or Natural Negative

- Merge Positive and Natural Negative ⟹ MMseq2 at 30 pident
- Using **class weights** to handle **class imbalance**.
- $$w_c = \frac{N}{K \times n_c}$$

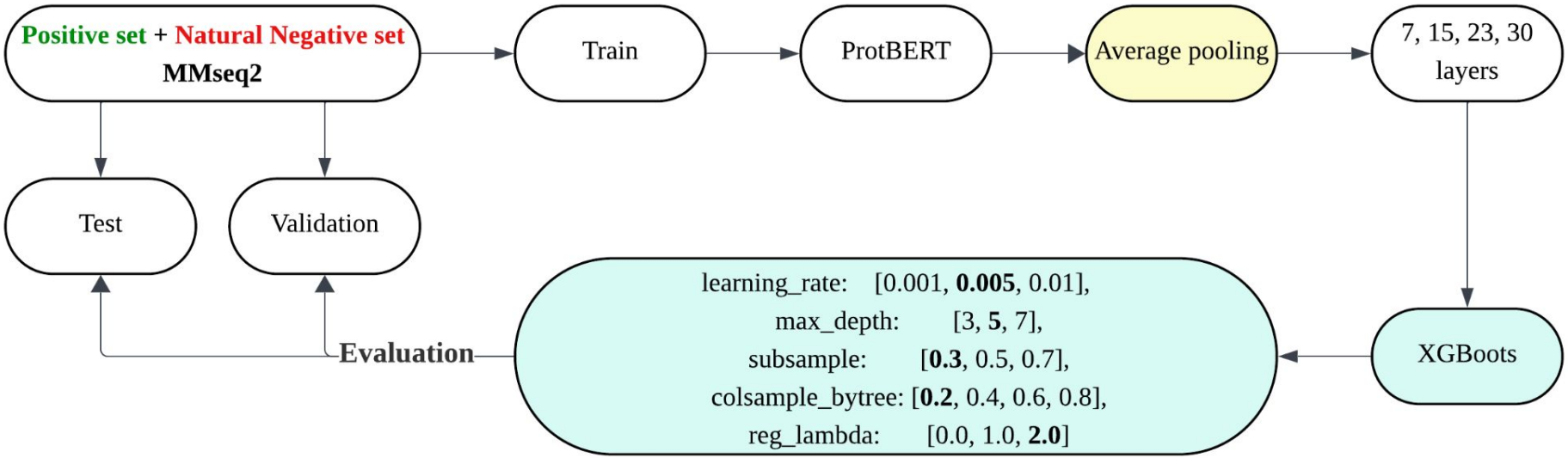
Enzymatic weight = 0.8683

Structural weight = 1.2899

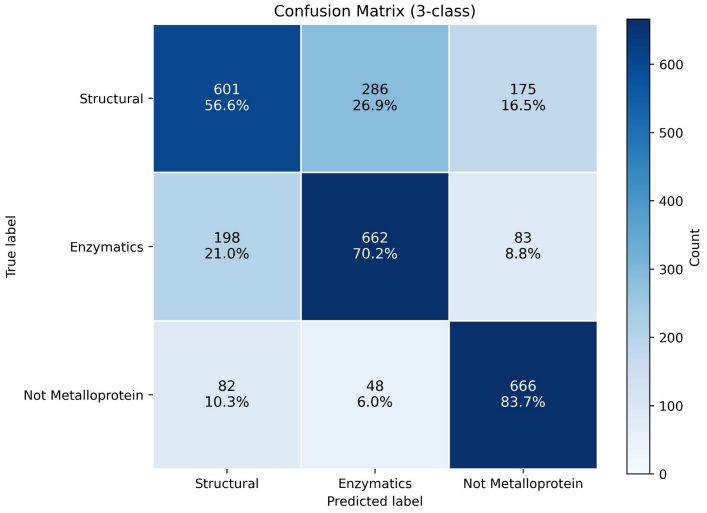
Natural Neg weight = 0.9319



ESNN - Enzymatic, Structural or Natural Negative



ESNN - Enzymatic, Structural or Natural Negative



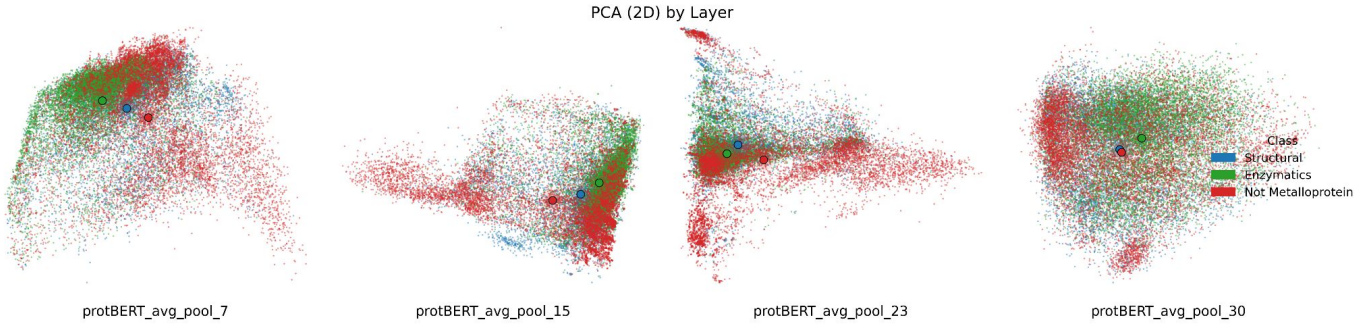
3-class performance

Accuracy: 68.87%

Presicion: 68.92%

Recall: 70.15%

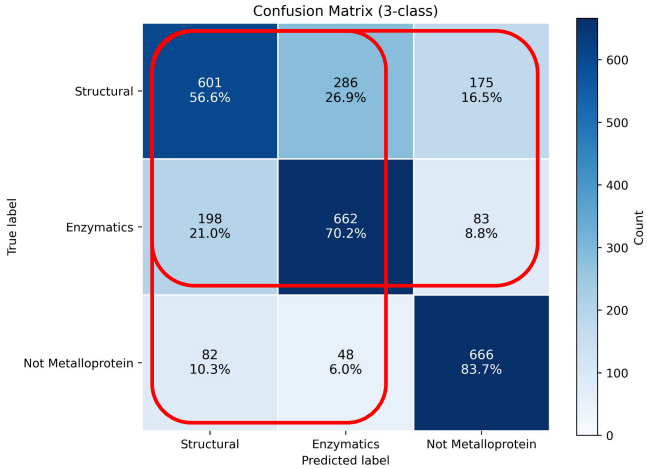
F1 score: 69.20%



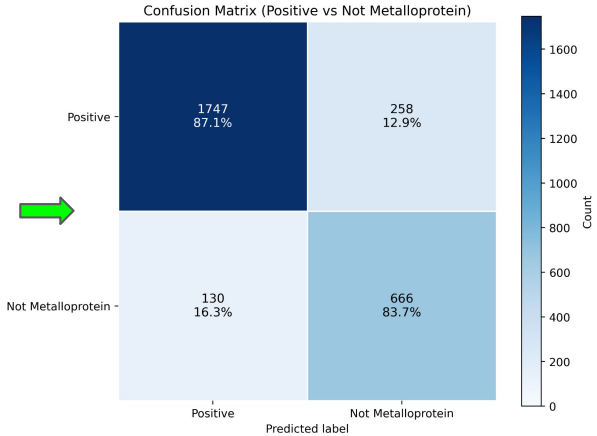
ESNN - Enzymatic, Structural or Natural Negative

If we only evaluate the 3-class ESNN on its performance in metal-binding protein prediction?

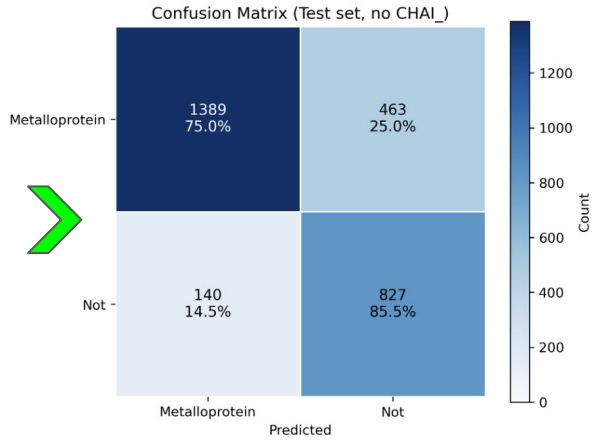
Model	Accuracy	Precision	Recall	F1 Score
MoN (Natural test)	78.61%	90.84%	75.00%	82.17%
Metal/Not task - ESNN (Natural test)	+ 7.53% = 86.14%	+2.23% = 93.07%	+12.13% = 87.13%	+7.84% = 90.01%



ESNN

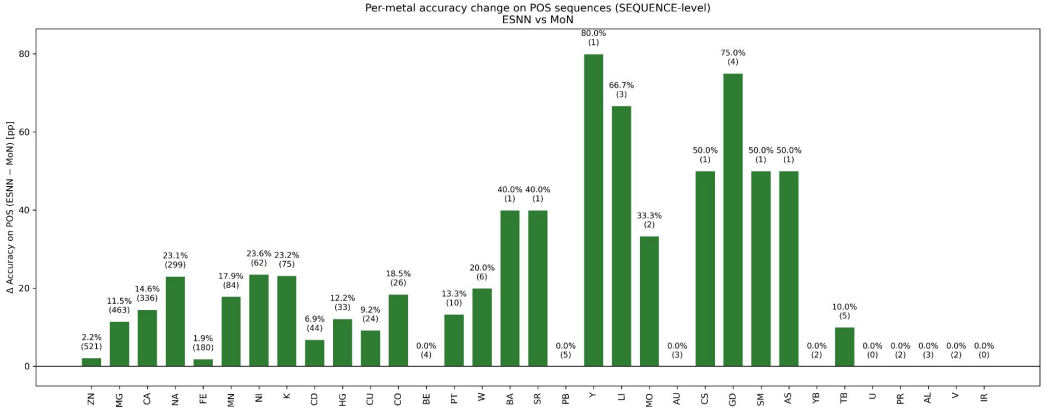


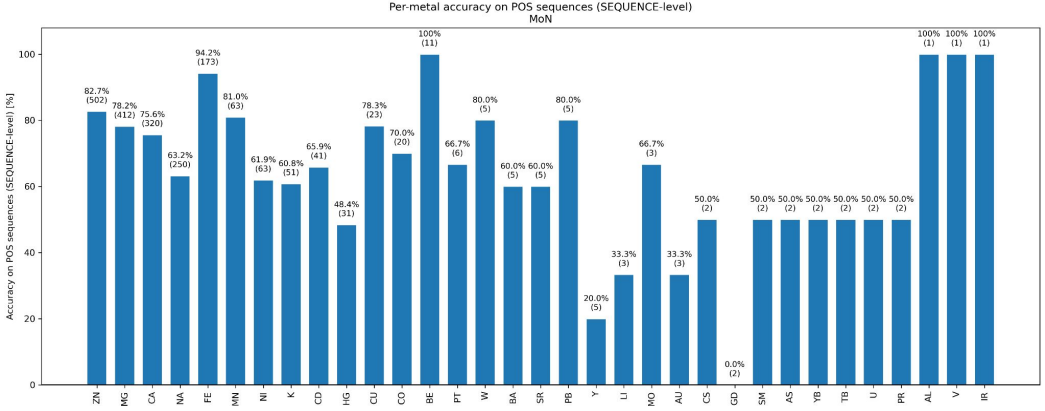
ESNN



MoN

ESNN - Enzymatic, Structural or Natural Negative





Metal	Sample Number	MoN	ESNN	difference
Zn	502 / 521	82.67%	84.84%	+2.17%
Mg	**412 / 463**	**78.16%**	**89.63%**	**+11.48%**
Ca	**320 / 336**	**75.62%**	**90.18%**	**+14.55%**
Na	**250 / 299**	**63.2%**	**86.29%**	**+23.09%**
Fe	173 / 180	94.22%	96.11%	+1.89%
Mn	**63 / 84**	**80.95%**	**98.81%**	**+17.86%**

Limitation

- The dataset was not filtered by **experimental method** or **resolution**.
- Although our dataset contains 44 types of metal ions, the **top 5** account for **85% of all entries**.
- The model still exhibits **bias** based on the distribution of **metal ion types**.
- The synthetic negative sequence is too less in number.

Outlook

- Continuous growth of proteomics and structural biology will provide **more high-quality data**.
- Incorporating **higher-resolution structural** details and experimental validation can **improve model reliability**.
- With advances in AI methods may enable more efficient and **biologically realistic designs** of mutated metal-binding sites.
- Such progress will **reduce the risk of generating overly artificial sequences** that fail downstream screening.

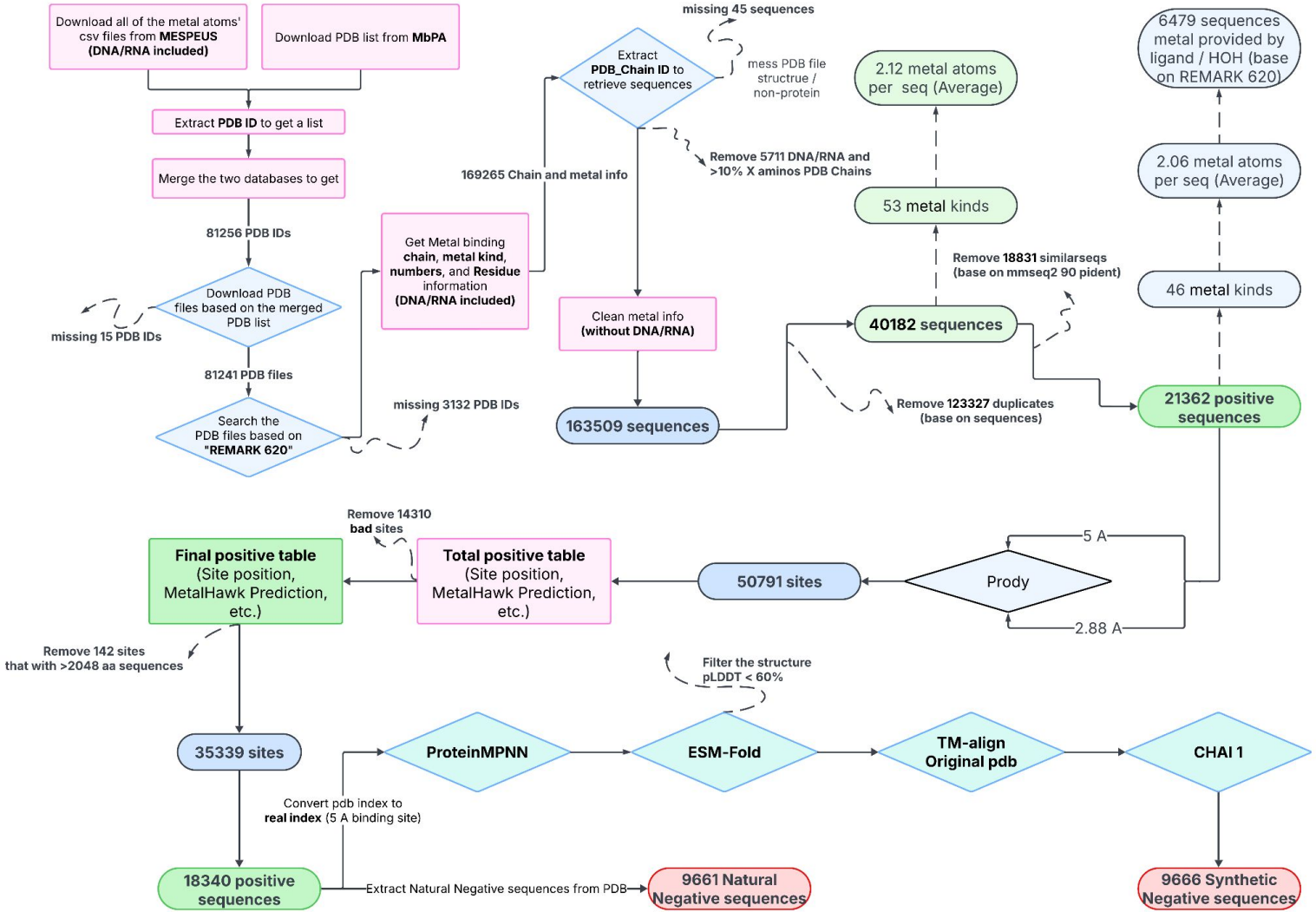
Conclusion

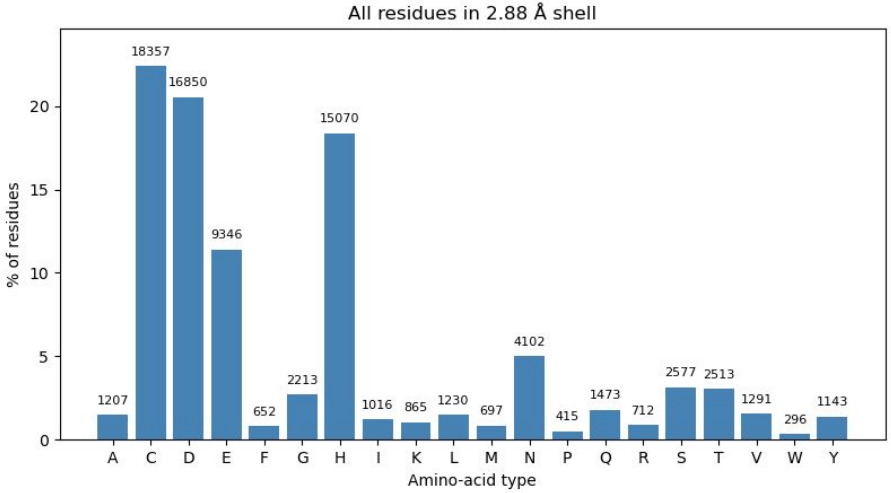
- A **high-quality metalloprotein dataset was built** through rigorous data filtering and a novel negative set generation strategy.
- Based on this dataset, we developed **sequence-based** prediction models addressing both **metalloprotein identification** and **enzymatic activity prediction**.
- This work provides a more comprehensive **framework for the annotation of metalloproteins**.

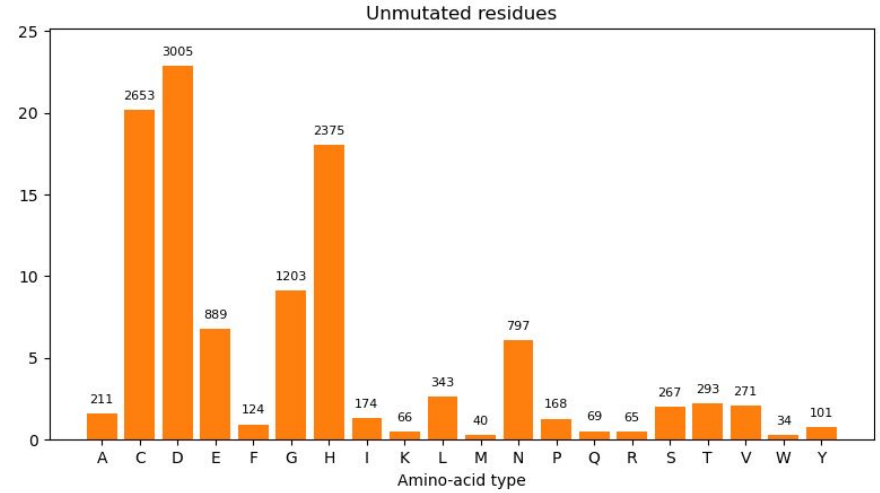
Thanks for your attention!

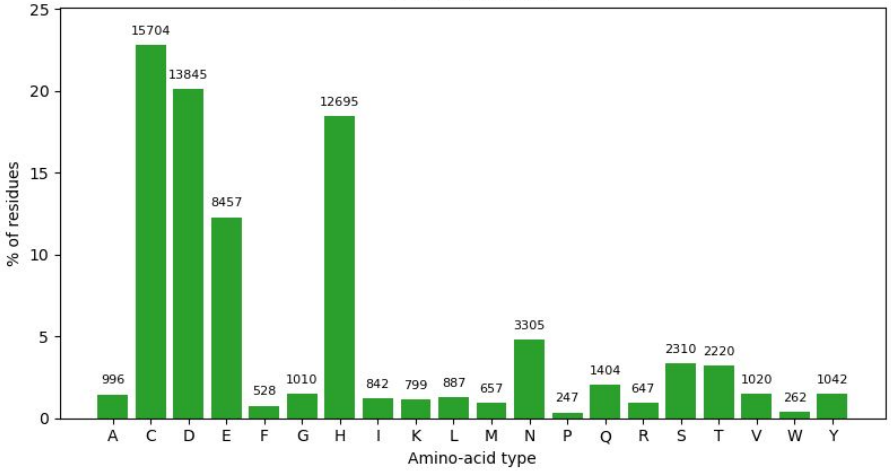
And thanks to all of you :)

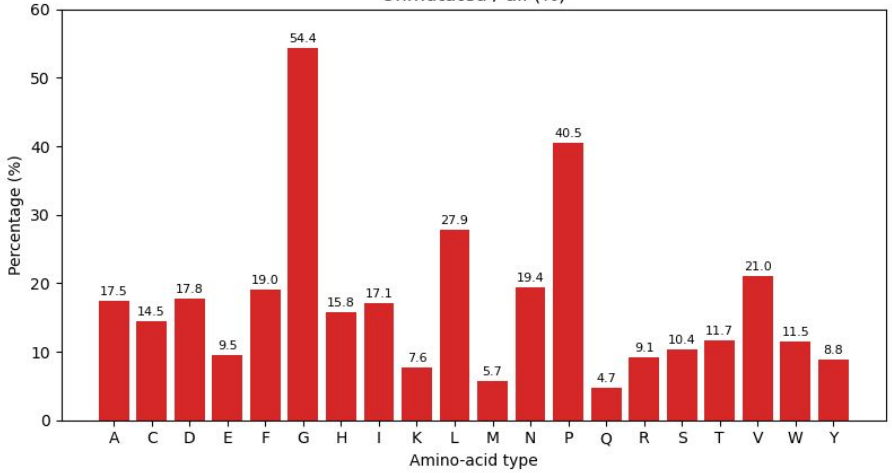
Backup slides

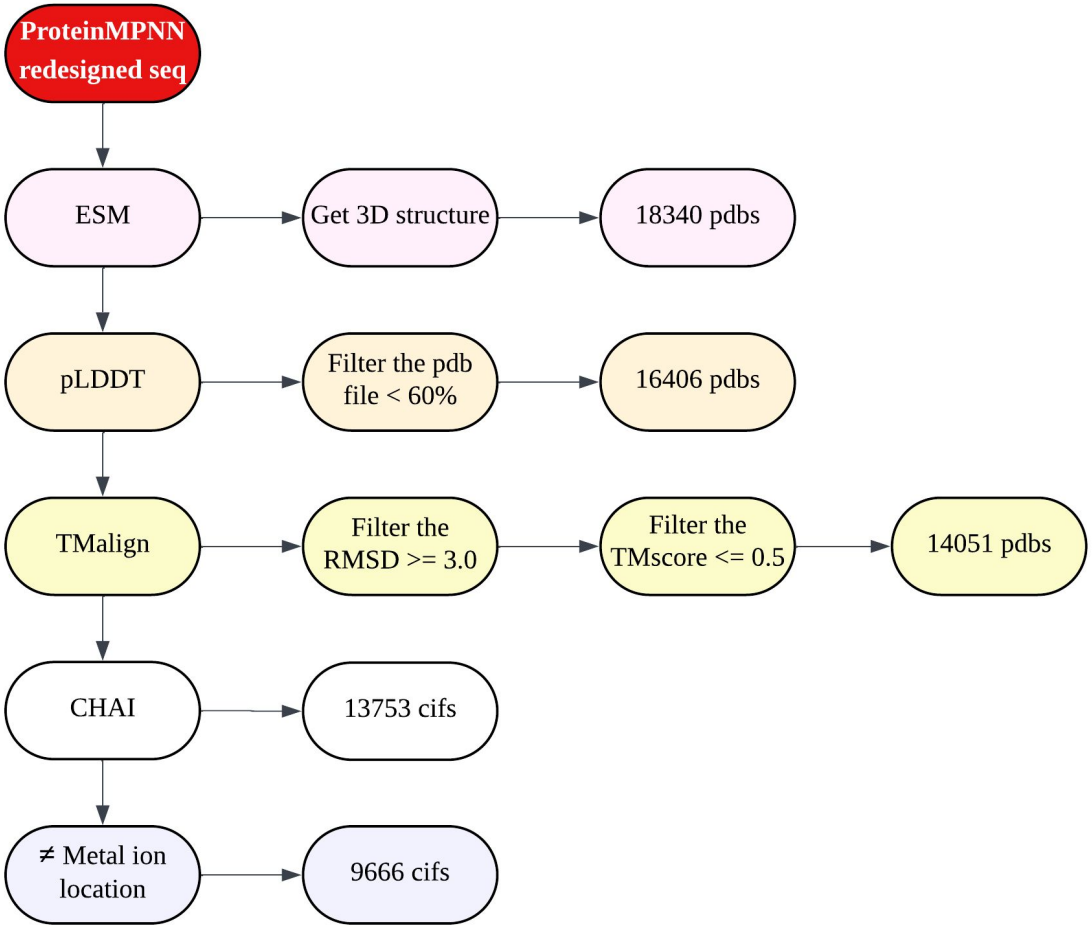

Download all of the metal atoms' csv files from MESPEUS (DNA/RNA included)
Download PDB list from MbPA
Extract PDB ID to get a list
Merge the two databases to get
81256 PDB IDs
Download PDB files based on the merged PDB list
missing 15 PDB IDs
81241 PDB files
Search the PDB files based on "REMARK 620"
missing 3132 PDB IDs
Get Metal binding chain, metal kind, numbers, and Residue information (DNA/RNA included)
169265 Chain and metal info
Extract PDB_Chain ID to retrieve sequences
missing 45 sequences
mess PDB file structrue / non-protein
Remove 5711 DNA/RNA and >10% X aminos PDB Chains
Clean metal info (without DNA/RNA)
163509 sequences
40182 sequences
Remove 123327 duplicates (base on sequences)
Remove 18831 similarseqs (base on mmseq2 90 pident)
53 metal kinds
2.12 metal atoms per seq (Average)
6479 sequences metal provided by ligand / HOH (base on REMARK 620)
2.06 metal atoms per seq (Average)
46 metal kinds
21362 positive sequences
Prody
5 A
2.88 A
50791 sites
Total positive table (Site position, MetalHawk Prediction, etc.)
Remove 14310 bad sites
Final positive table (Site position, MetalHawk Prediction, etc.)
Remove 142 sites that with >2048 aa sequences
35339 sites
Filter the structure pLDDT < 60%
ProteinMPNN
ESM-Fold
TM-align Original pdb
CHAI 1
Convert pdb index to real index (5 A binding site)
18340 positive sequences
Extract Natural Negative sequences from PDB
9661 Natural Negative sequences
9666 Synthetic Negative sequences


All residues in 2.88 Å shell
% of residues
Amino-acid type
A C D E F G H I K L M N P Q R S T V W Y
1207 18357 16850 9346 652 2213 15070 1016 865 1230 697 4102 415 1473 712 2577 2513 1291 296 1143


Unmutated residues
Amino-acid type
A C D E F G H I K L M N P Q R S T V W Y
211 2653 3005 889 124 1203 2375 174 66 343 40 797 168 69 65 267 293 271 34 101


Mutated residues
% of residues
Amino-acid type
A C D E F G H I K L M N P Q R S T V W Y
996 15704 13845 8457 528 1010 12695 842 799 887 657 3305 247 1404 647 2310 2220 1020 262 1042


Unmutated / all (%)
Percentage (%)
Amino-acid type
A C D E F G H I K L M N P Q R S T V W Y
17.5 14.5 17.8 9.5 19.0 54.4 15.8 17.1 7.6 27.9 5.7 19.4 40.5 4.7 9.1 10.4 11.7 21.0 11.5 8.8


ProteinMPNN redesigned seq
ESM
Get 3D structure
18340 pdbs
pLDDT
Filter the pdb file < 60%
16406 pdbs
TMalign
Filter the RMSD >= 3.0
Filter the TMscore <= 0.5
14051 pdbs
CHAI
13753 cifs
≠ Metal ion location
9666 cifs

	A	P	R	F1
SVM_GANBERT_CLS_L07	0.693	0.68	0.662	0.671
SVM_GANBERT_CLS_L15	0.693	0.672	0.685	0.678
SVM_GANBERT_CLS_L23	0.694	0.67	0.697	0.684
SVM_GANBERT_CLS_L30	0.731	0.725	0.696	0.71
SVM_GANBERT_POOL_L07	0.718	0.696	0.72	0.708
SVM_GANBERT_POOL_L15	0.733	0.724	0.705	0.714
SVM_GANBERT_POOL_L23	0.72	0.7	0.716	0.708
SVM_GANBERT_POOL_L30	0.725	0.706	0.719	0.713
SVM_PROTBERT_CLS_L07	0.694	0.683	0.661	0.672
SVM_PROTBERT_CLS_L15	0.696	0.672	0.702	0.687
SVM_PROTBERT_CLS_L23	0.686	0.668	0.667	0.668
SVM_PROTBERT_CLS_L30	0.718	0.718	0.666	0.691
SVM_PROTBERT_POOL_L07	0.718	0.694	0.724	0.709
SVM_PROTBERT_POOL_L15	0.732	0.721	0.71	0.715
SVM_PROTBERT_POOL_L23	0.725	0.715	0.697	0.706
SVM_PROTBERT_POOL_L30	0.721	0.727	0.658	0.691
XGB_GANBERT_CLS_L07	0.687	0.667	0.675	0.671
XGB_GANBERT_CLS_L15	0.697	0.661	0.74	0.698
XGB_GANBERT_CLS_L23	0.692	0.667	0.696	0.681
XGB_GANBERT_CLS_L30	0.732	0.724	0.702	0.713
XGB_GANBERT_POOL_L07	0.725	0.706	0.72	0.713
XGB_GANBERT_POOL_L15	0.728	0.704	0.734	0.719
XGB_GANBERT_POOL_L23	0.724	0.71	0.703	0.707
XGB_GANBERT_POOL_L30	0.721	0.703	0.711	0.707
XGB_PROTBERT_CLS_L07	0.684	0.665	0.672	0.669
XGB_PROTBERT_CLS_L15	0.701	0.673	0.716	0.694
XGB_PROTBERT_CLS_L23	0.687	0.654	0.72	0.686
XGB_PROTBERT_CLS_L30	0.712	0.702	0.681	0.692
XGB_PROTBERT_POOL_L07	0.725	0.71	0.711	0.71
XGB_PROTBERT_POOL_L15	0.729	0.695	0.764	0.728
XGB_PROTBERT_POOL_L23	0.732	0.701	0.756	0.728
XGB_PROTBERT_POOL_L30	0.72	0.729	0.652	0.689

GANBERT Model confusion matrix