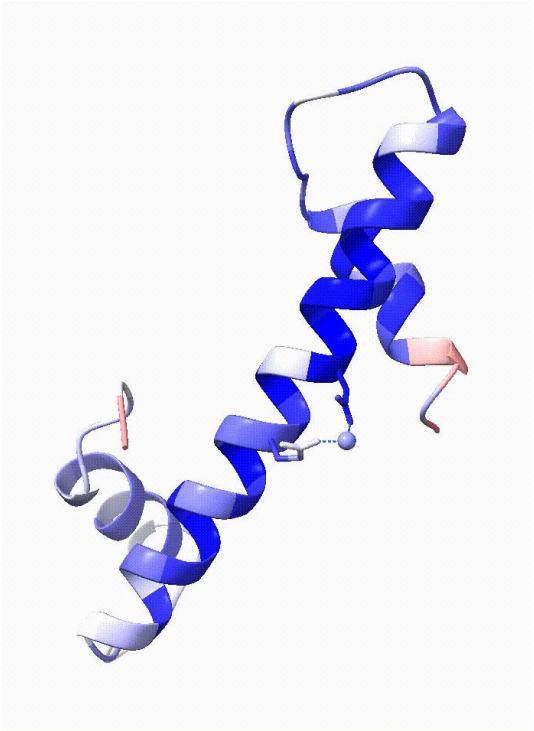
Advancing metal-binding protein predictions with deep learning

Update - Jingkai LAN

Supervisor: Thomas Lemmin
Co-supervisor: Giulia Peteani

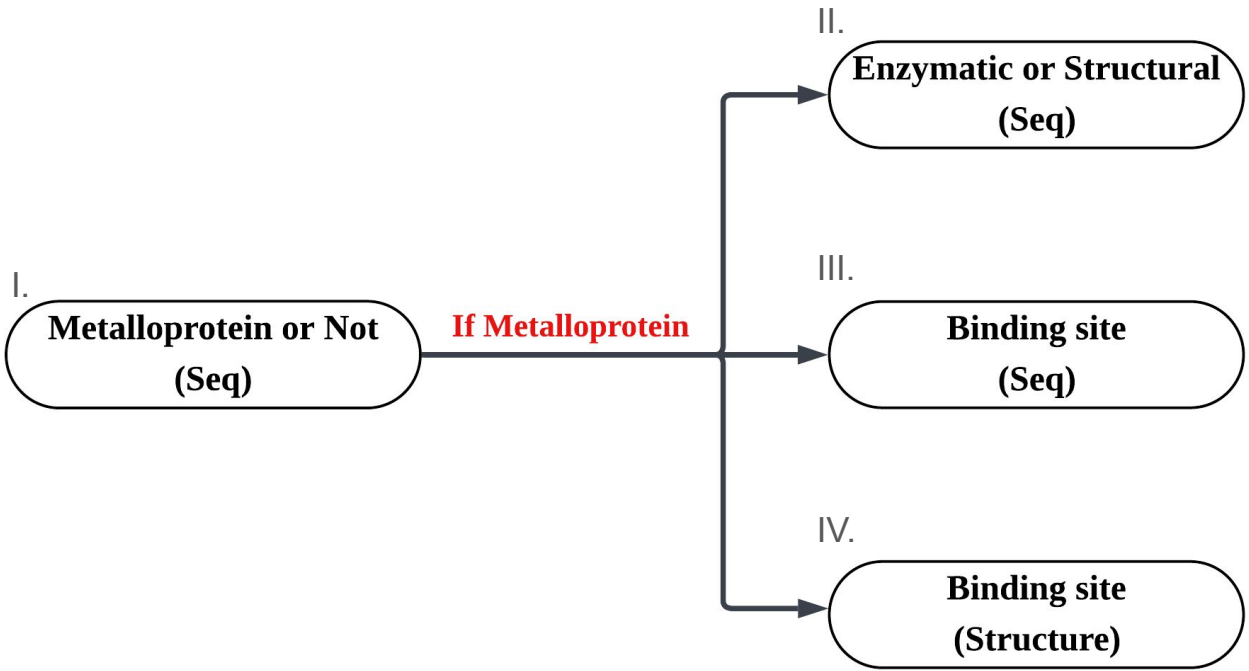
Introduction

- Metalloproteins are **abundant** and perform many essential **biological functions**.
- **Complex interactions** occur within the **coordination sphere**.
- Importance of understanding **metal ion roles** and **structure–function relationships**.



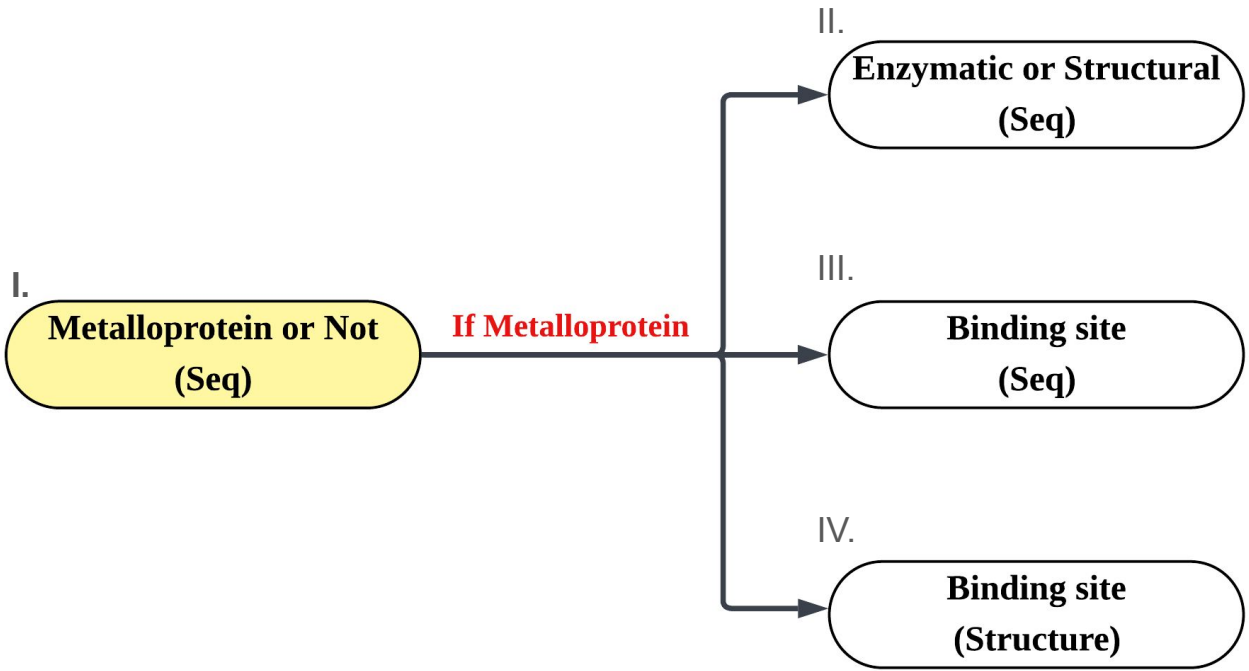
Project goal

Leverage **deep learning** to investigate and model **key properties** of metalloproteins.

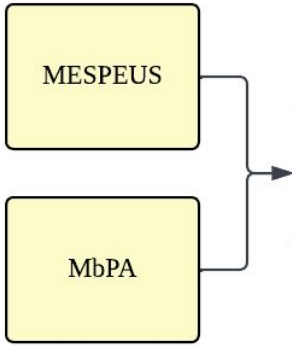


Project goal

Leverage **deep learning** to investigate and model **key properties** of metalloproteins.

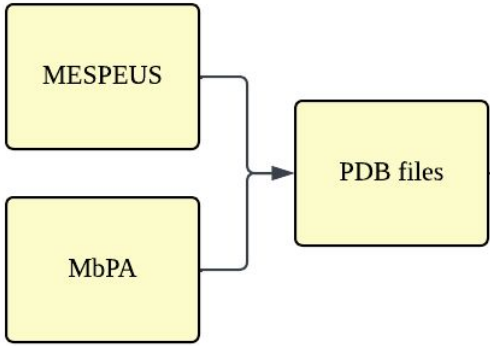


Dataset creation



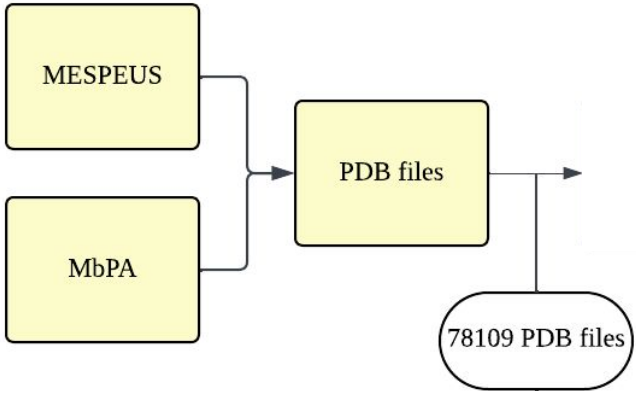
Geng-Yu Lin, Yu-Cheng Su, Yen Lin Huang, Kun-Yi Hsin, MESPEUS: a database of metal coordination groups in proteins, Nucleic Acids Research, Volume 52, Issue D1, 5 January 2024, Pages D483–D493, https://doi.org/10.1093/nar/gkad1009
Jinzhao Li, Xiang He, Shuang Gao, Yuchao Liang, Zhi Qi, Qilemuge Xi, Yongchun Zuo, Yongqiang Xing, The Metal-binding Protein Atlas (MbPA): An Integrated Database for Curating Metalloproteins in All Aspects, Journal of Molecular Biology, Volume 435, Issue 14, 2023, https://doi.org/10.1016/j.jmb.2023.168117.

Dataset creation: Positive set

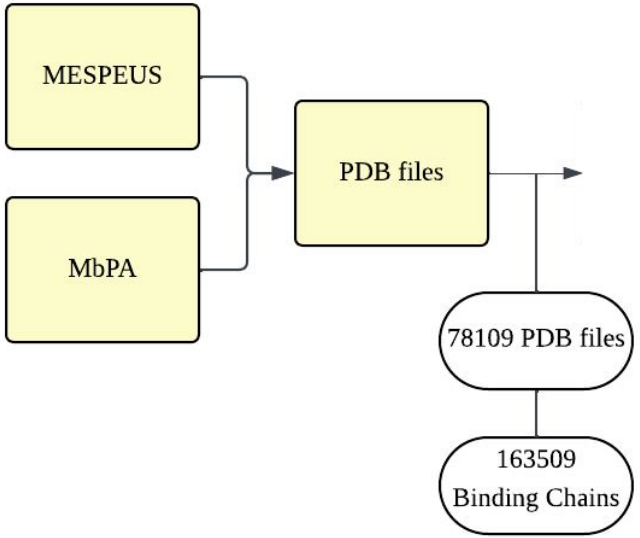


H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank (2000) Nucleic Acids Research 28: 235-242 https://doi.org/10.1093/nar/28.1.235.

Dataset creation: Positive set

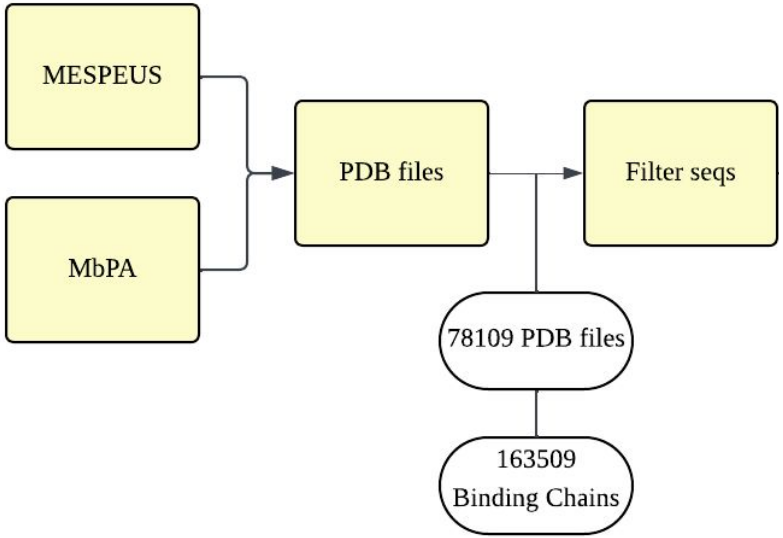


Dataset creation: Positive set

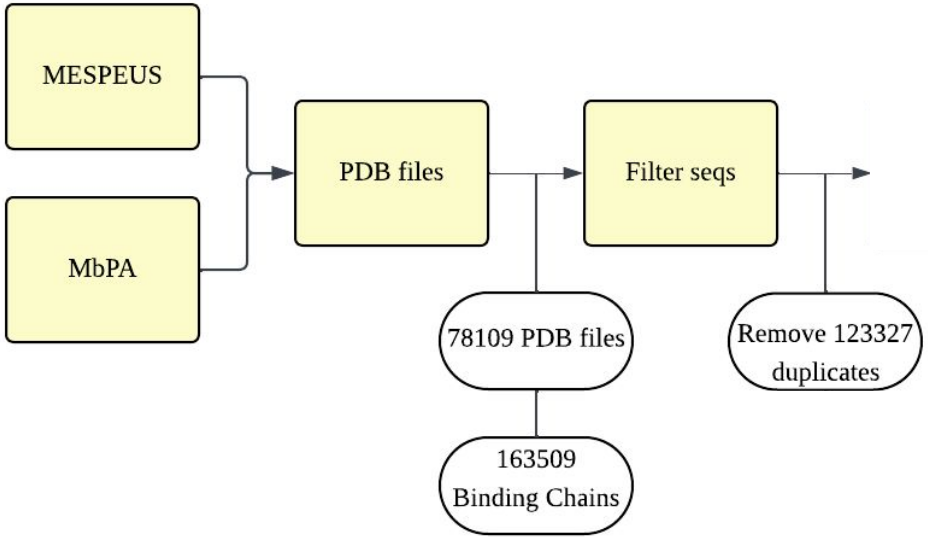



```
REMARK 620
REMARK 620 METAL COORDINATION
REMARK 620 (M=MODEL NUMBER; RES=RESIDUE NAME; C=CHAIN IDENTIFIER;
REMARK 620 SSEQ=SEQUENCE NUMBER; I=INSERTION CODE):
REMARK 620
REMARK 620 COORDINATION ANGLES FOR:  M RES CSSEQI METAL
REMARK 620                              MG A 301  MG
REMARK 620 N RES CSSEQI ATOM
REMARK 620 1 SER A  17   OG
REMARK 620 2 GDP A 302   O1B  92.7
REMARK 620 3 HOH A 405   O    82.8  92.5
REMARK 620 4 HOH A 408   O    91.6  86.0 174.2
REMARK 620 5 HOH A 409   O    87.7 171.9  95.5  85.9
REMARK 620 6 HOH A 436   O   173.2  90.3  90.9  94.7  90.3
REMARK 620 N                   1     2     3     4     5
REMARK 620
REMARK 620 COORDINATION ANGLES FOR:  M RES CSSEQI METAL
REMARK 620                              MG B 301  MG
REMARK 620 N RES CSSEQI ATOM
REMARK 620 1 SER B  17   OG
REMARK 620 2 GDP B 302   O2B  89.4
REMARK 620 3 HOH B 438   O   170.2  96.0
REMARK 620 4 HOH B 439   O    81.7  81.0  91.1
REMARK 620 5 HOH B 444   O    92.5 103.1  94.2 172.9
REMARK 620 6 HOH B 468   O    89.3 173.0  84.3  92.0  83.9
REMARK 620 N                   1     2     3     4     5
```

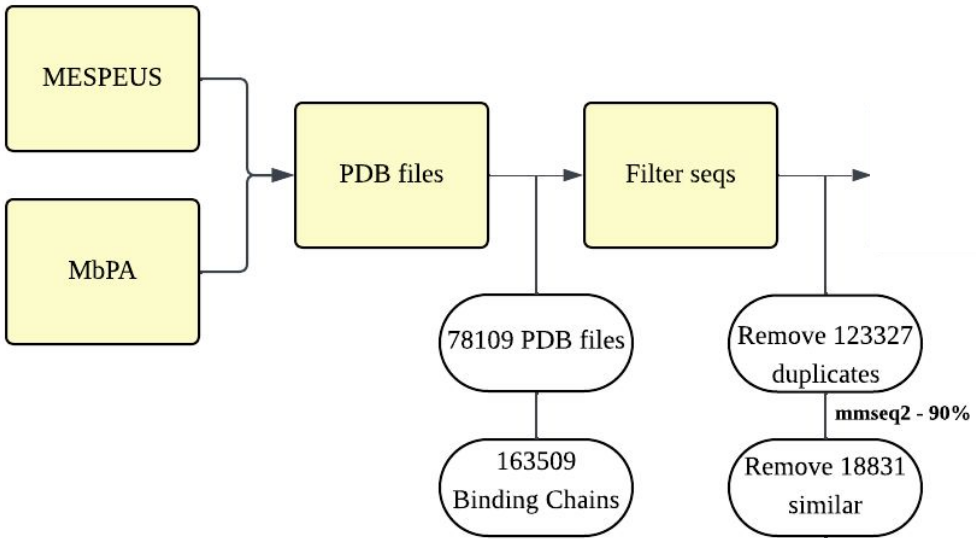
Dataset creation: Positive set



Dataset creation: Positive set

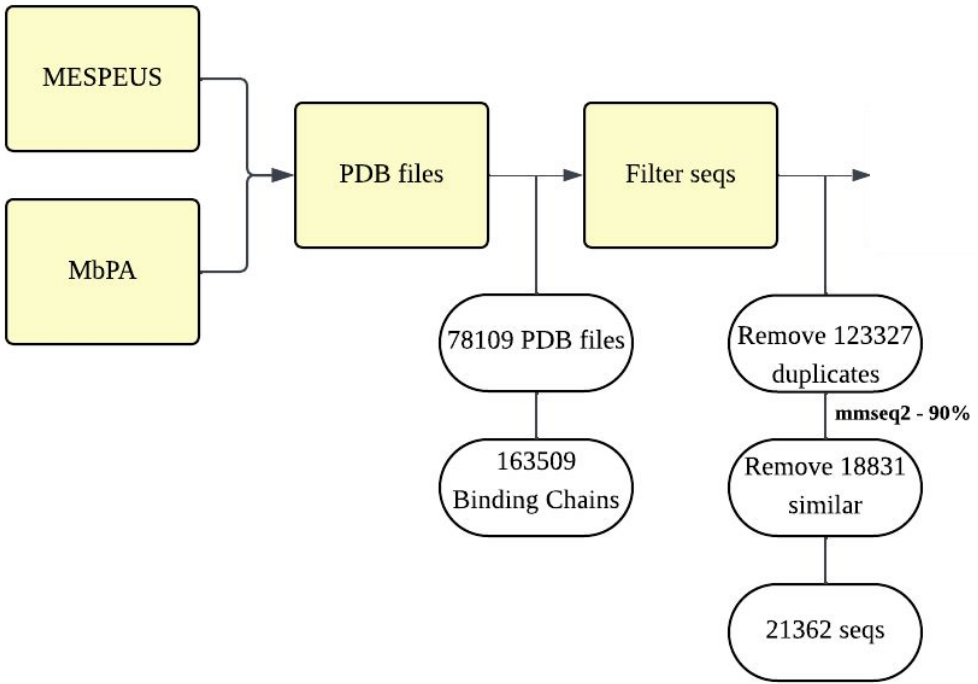


Dataset creation: Positive set

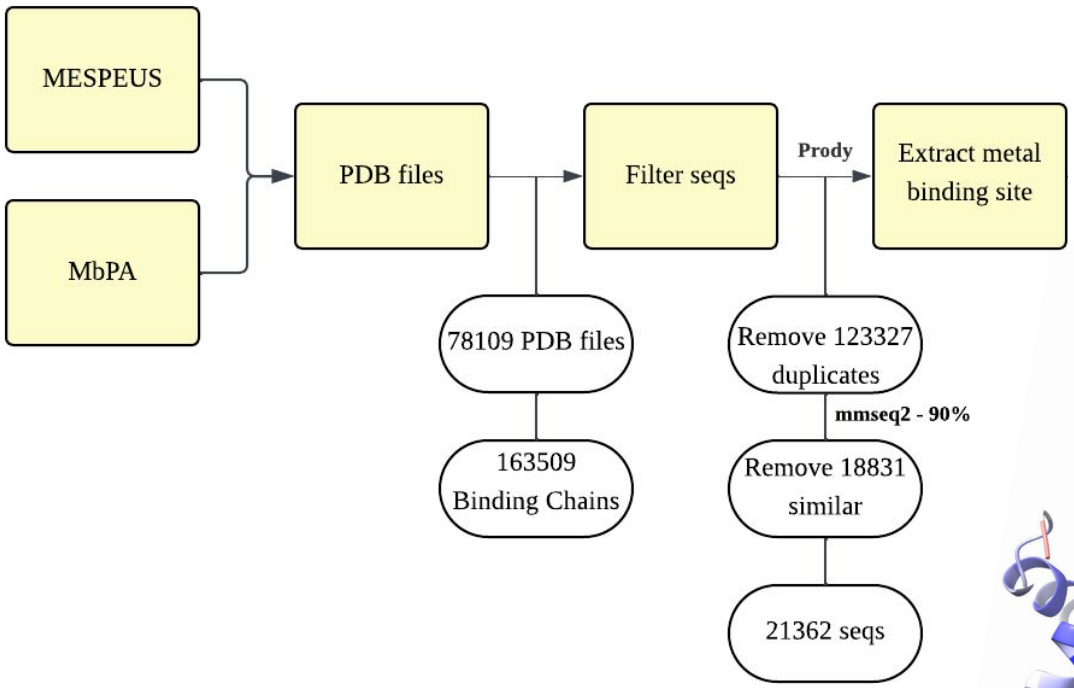


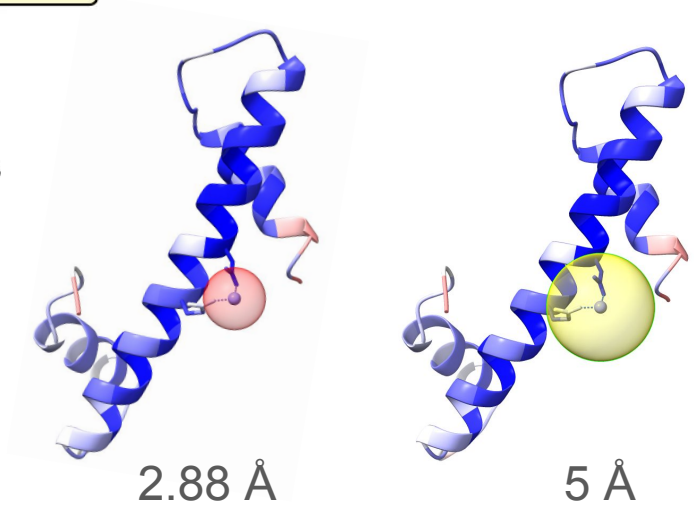
Steinegger, M., Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol 35, 1026–1028 (2017). https://doi.org/10.1038/nbt.3988

Dataset creation: Positive set



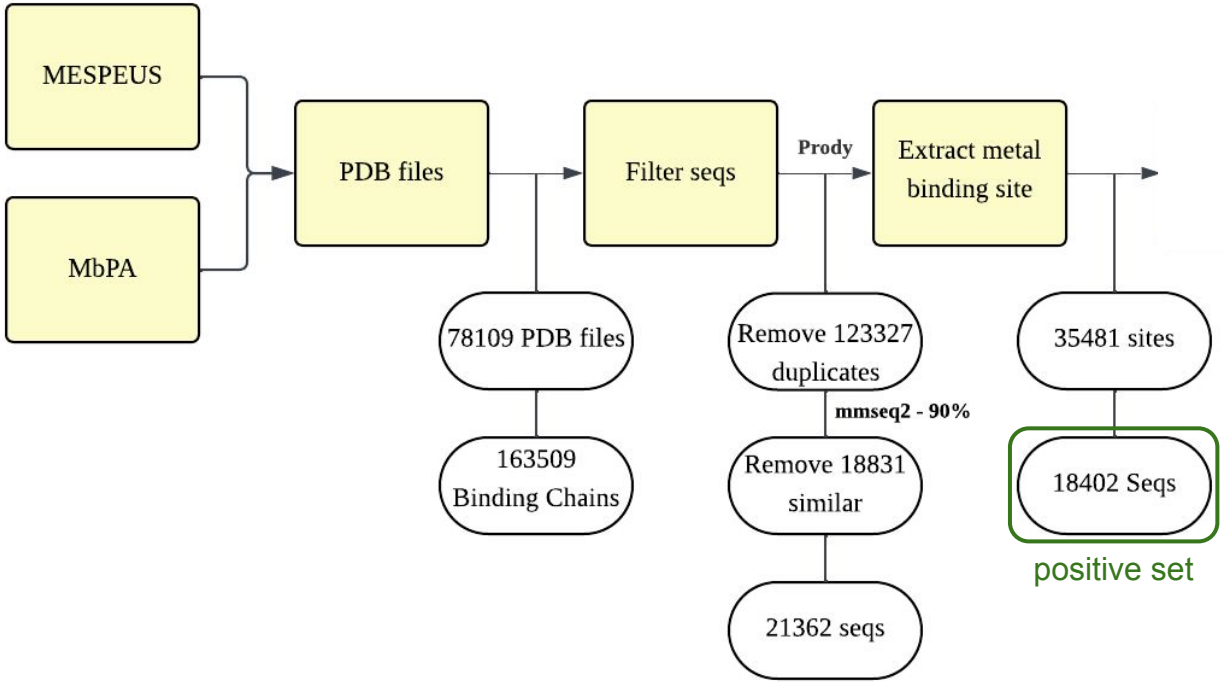
Dataset creation: Positive set



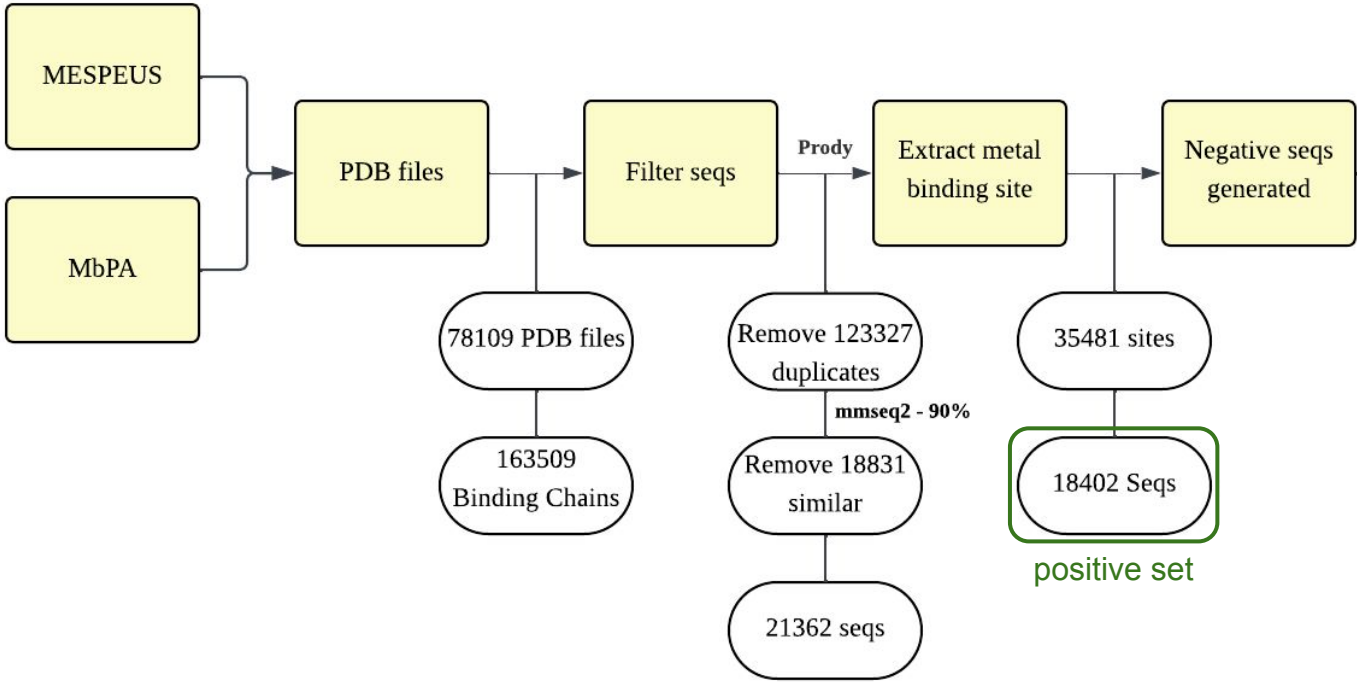


She Zhang, James M Krieger, Yan Zhang, Cihan Kaya, Burak Kaynak, Karolina Mikulska-Ruminska, Pemra Doruker, Hongchun Li, Ivet Bahar, ProDy 2.0: increased scale and scope after 10 years of protein dynamics modelling with Python, Bioinformatics, Volume 37, Issue 20, October 2021, Pages 3657–3659, https://doi.org/10.1093/bioinformatics/btab187

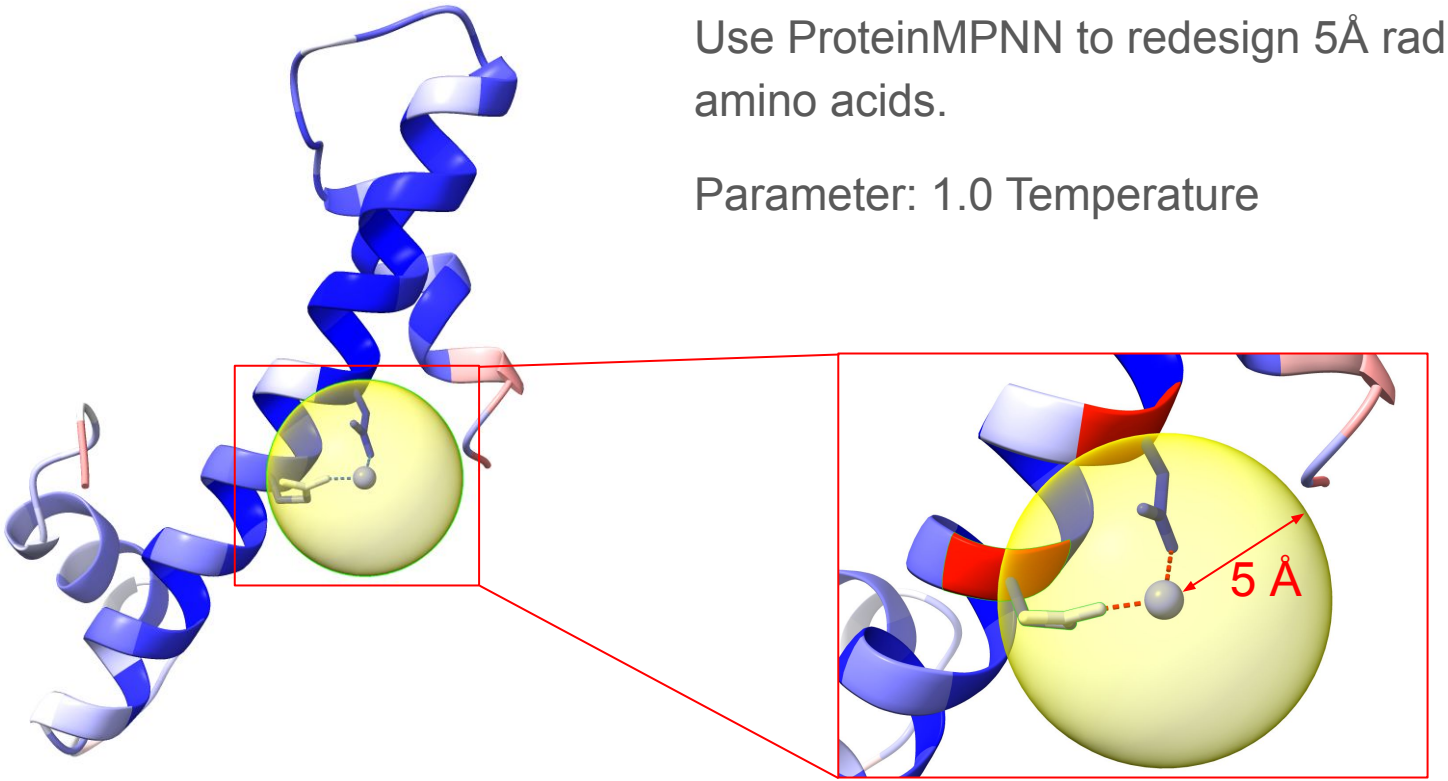
Dataset creation: Positive set



Dataset creation: Negative set



Dataset creation: Negative set

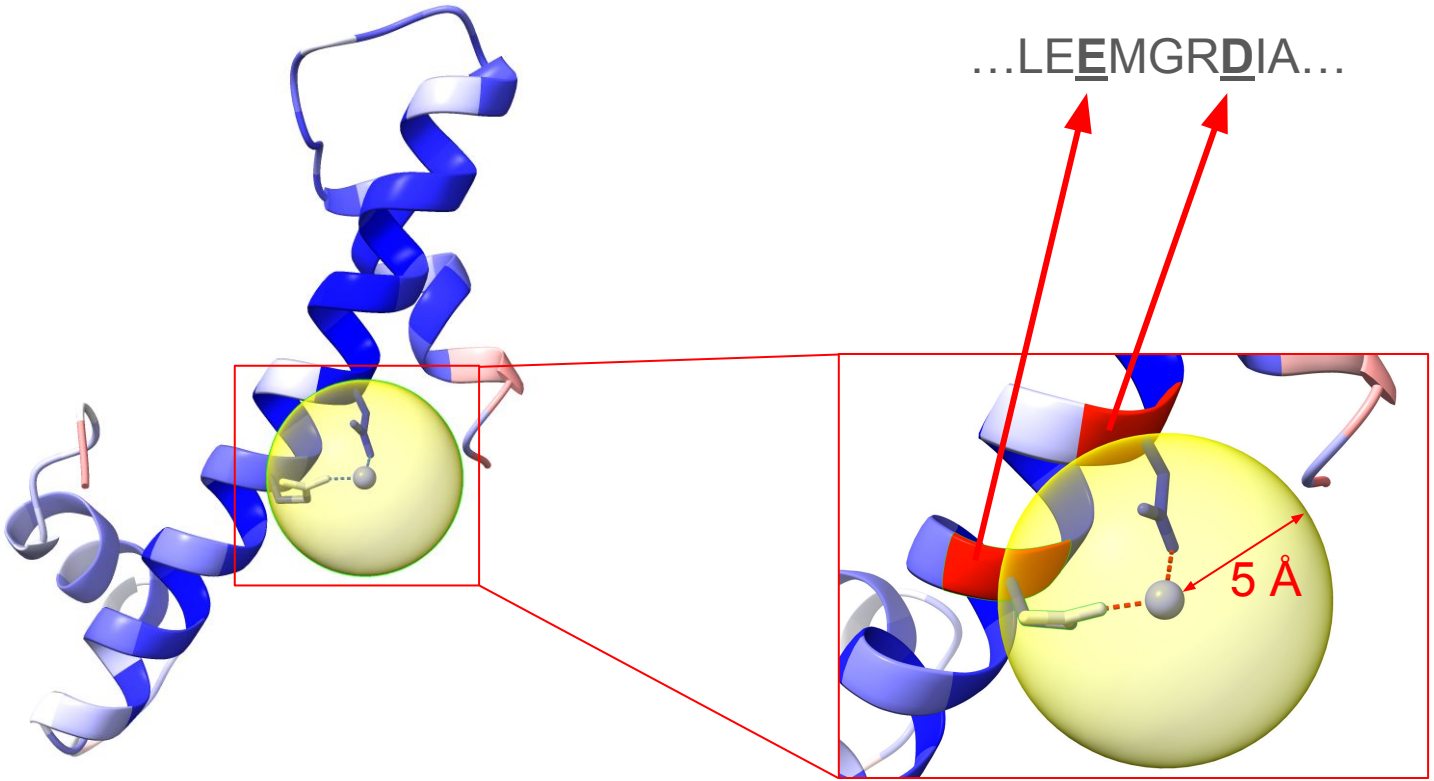


Use ProteinMPNN to redesign 5Å radius amino acids.

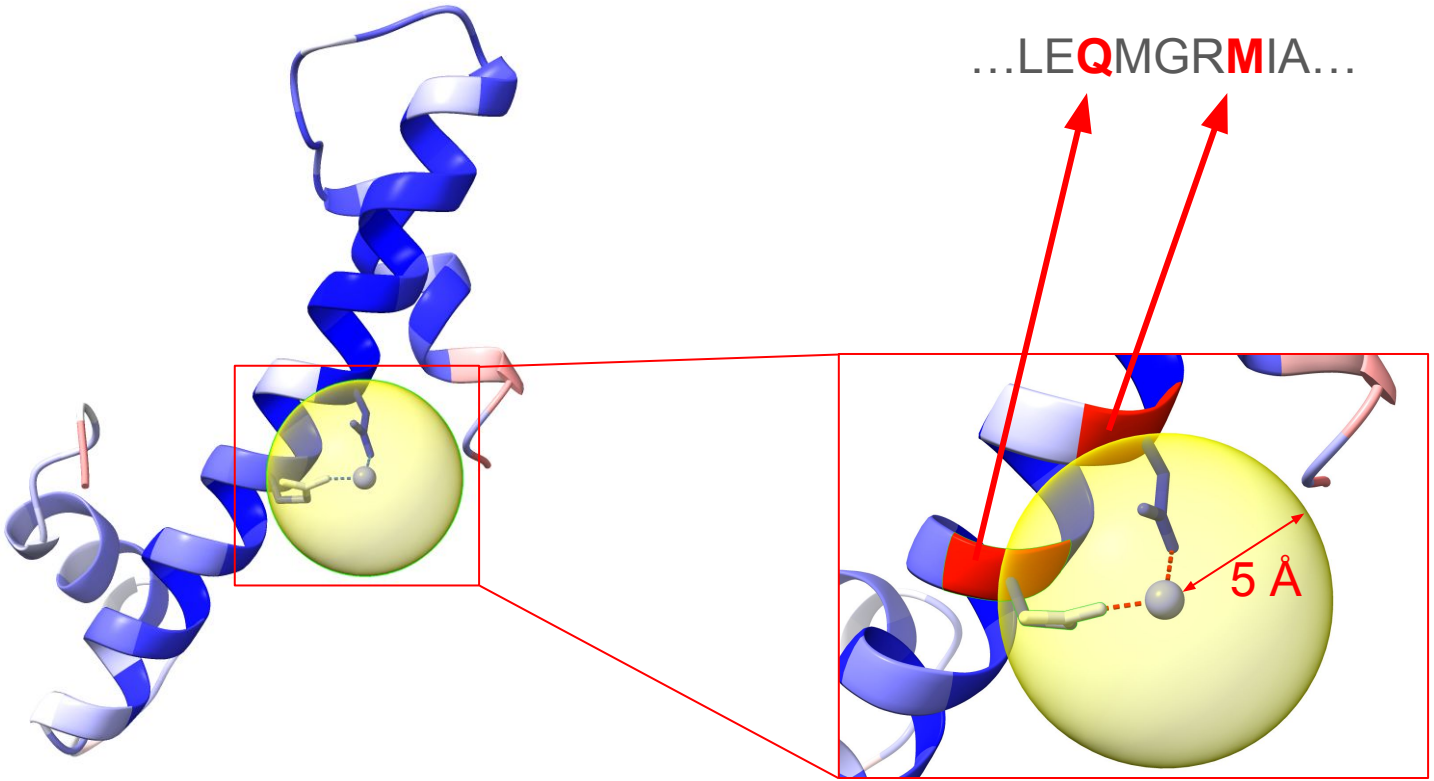
Parameter: 1.0 Temperature

J. Dauparas et al. ,Robust deep learning–based protein sequence design using ProteinMPNN.Science378,49-56(2022).DOI:https://www.science.org/doi/10.1126/science.add2187

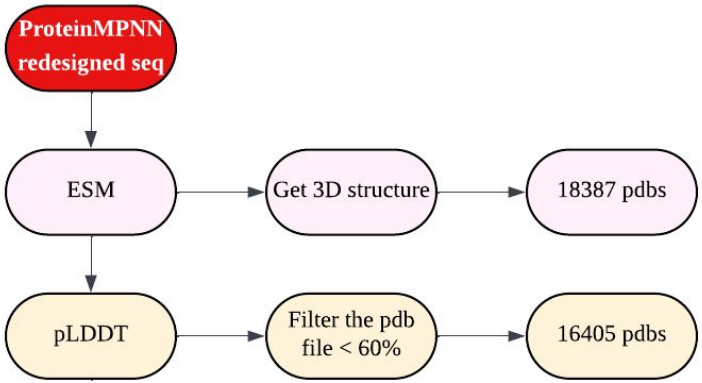
Dataset creation: Negative set



Dataset creation: Negative set

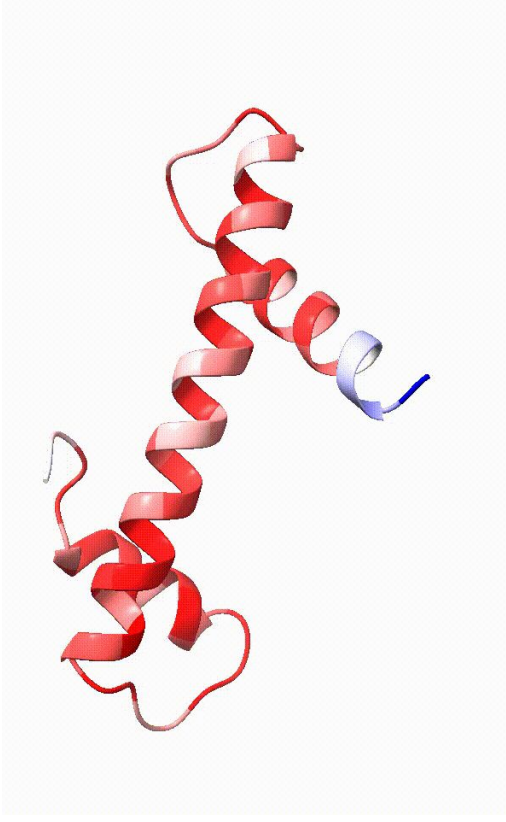


Validation of ProteinMPNN redesigned sequences



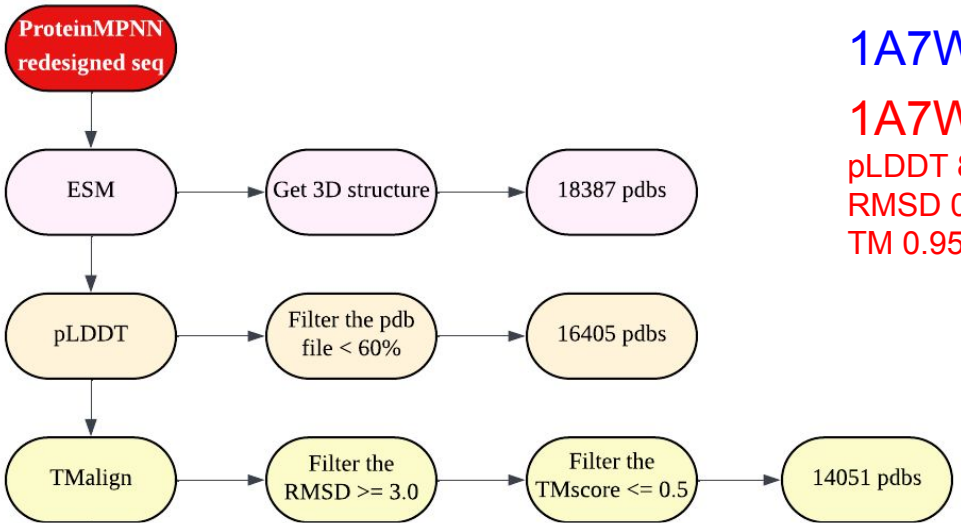
1A7W ESM

pLDDT 87.08



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

Validation of ProteinMPNN redesigned sequences



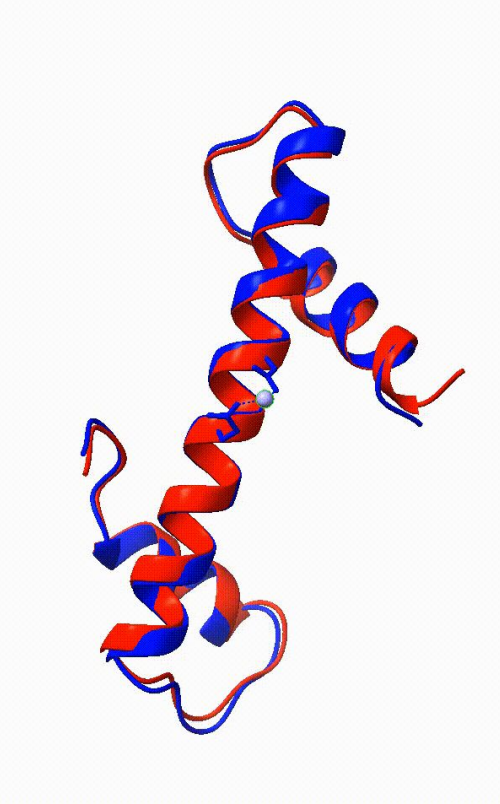
1A7W

1A7W ESM

pLDDT 87.08

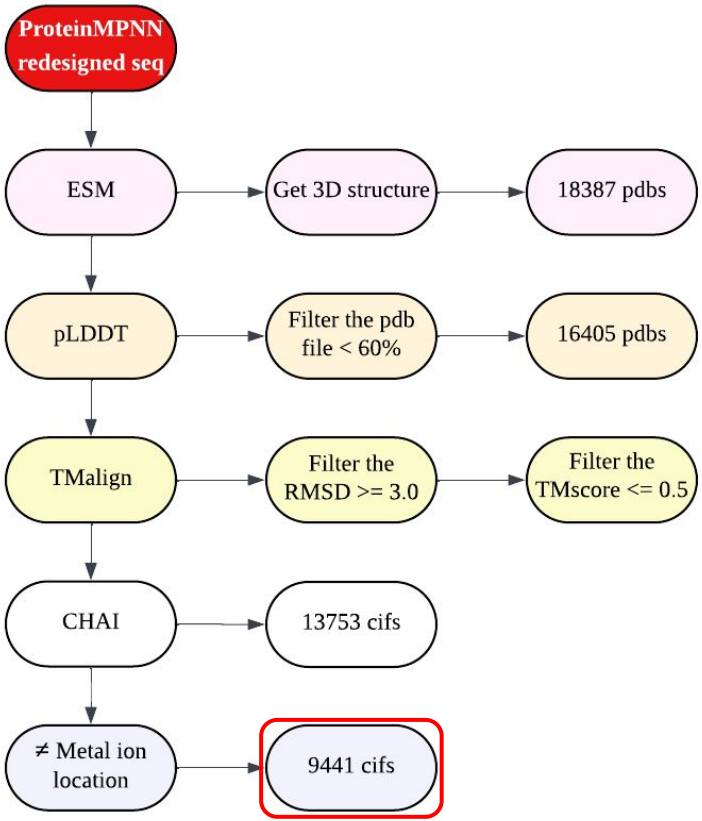
RMSD 0.59

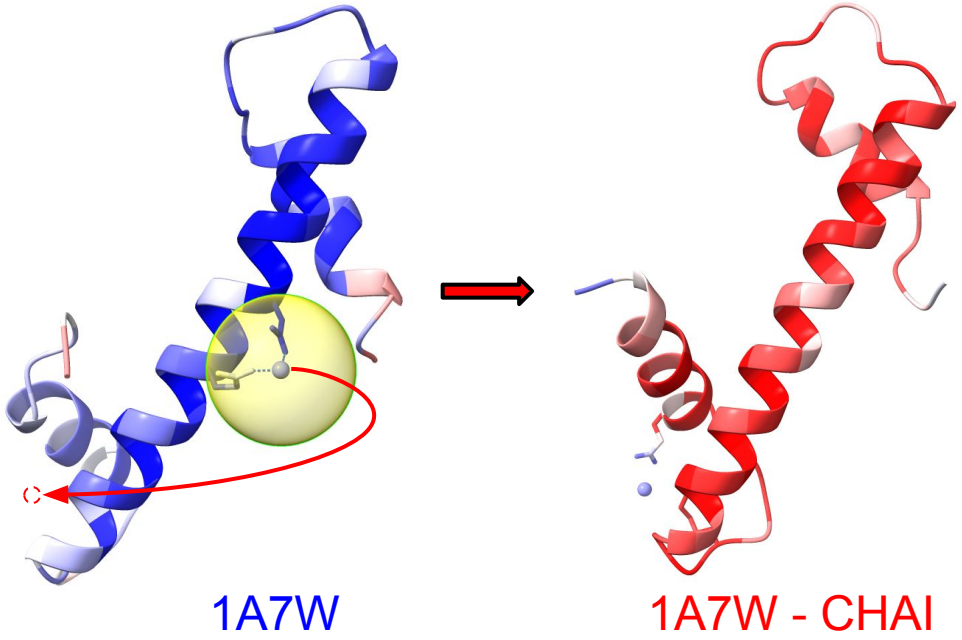
TM 0.95



Y. Zhang, J. Skolnick, TM-align: A protein structure alignment algorithm based on TM-score, Nucleic Acids Research, 33: 2302-2309 (2005).

Validation of ProteinMPNN redesigned sequences





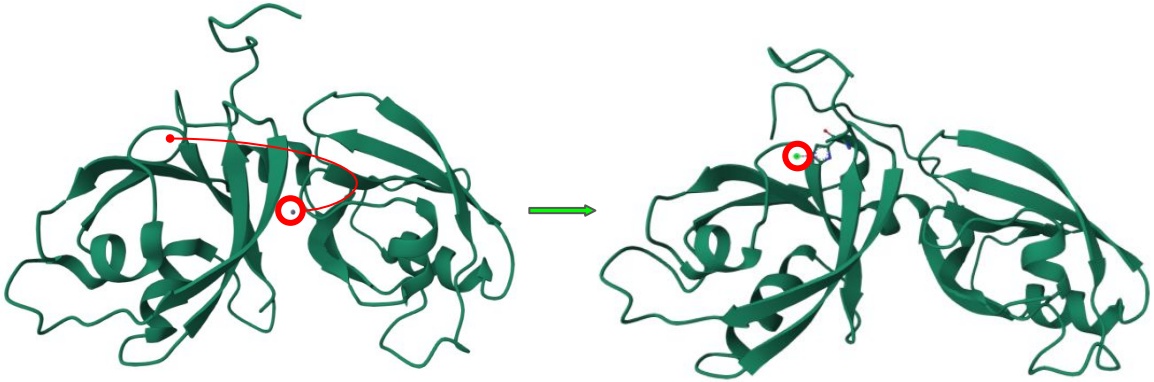
Chai Discoveryet al. ‚Chai-1: Decoding the molecular interactions of life. October 2024, doi: https://doi.org/10.1101/2024.10.10.615955

More examples..

original PDB

CHAI prediciton

4kfv_A

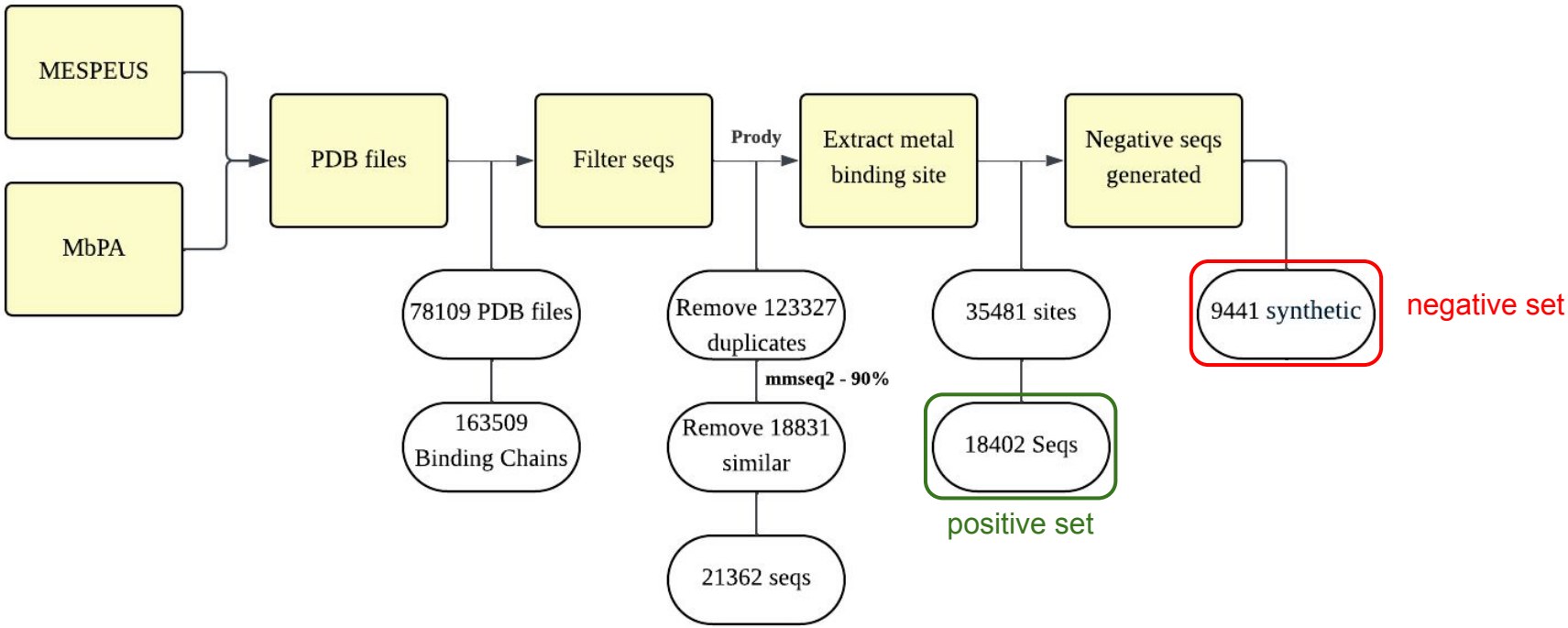


4kfv_A

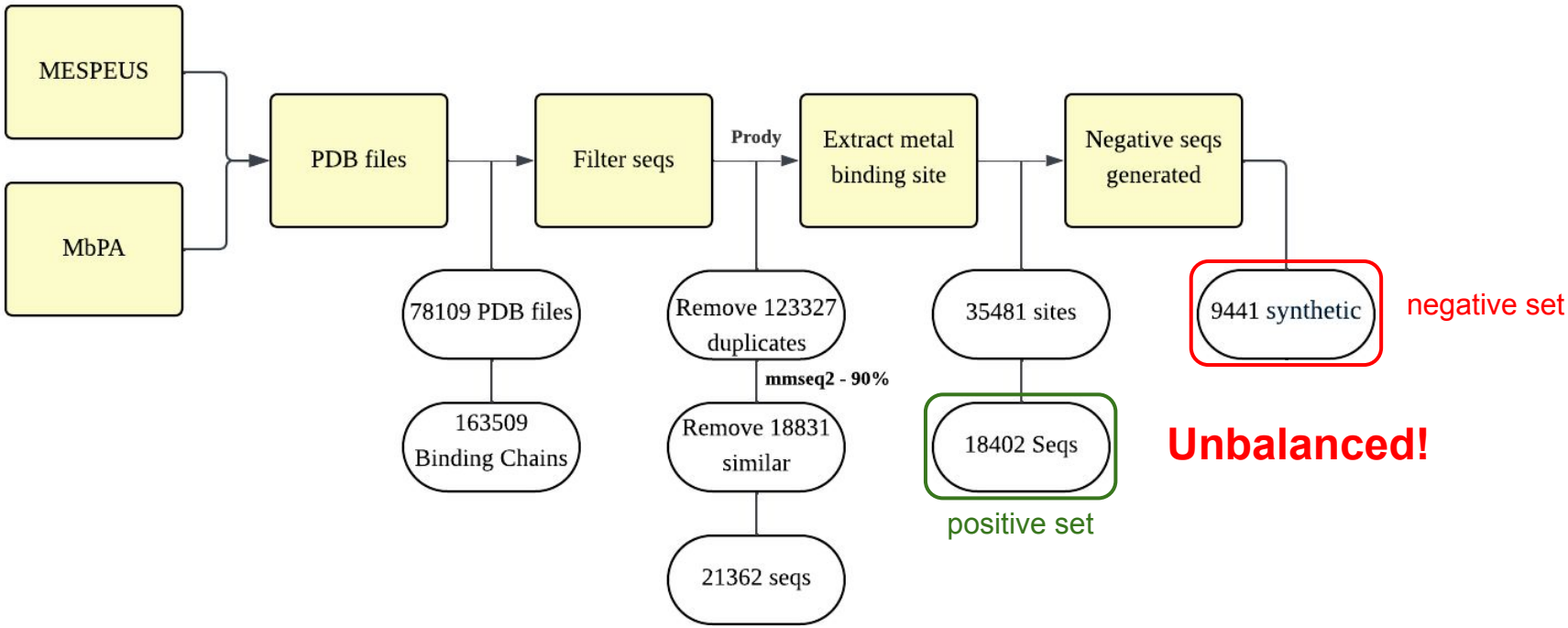
5zhf_A

5zhf_A

Dataset creation: Negative set

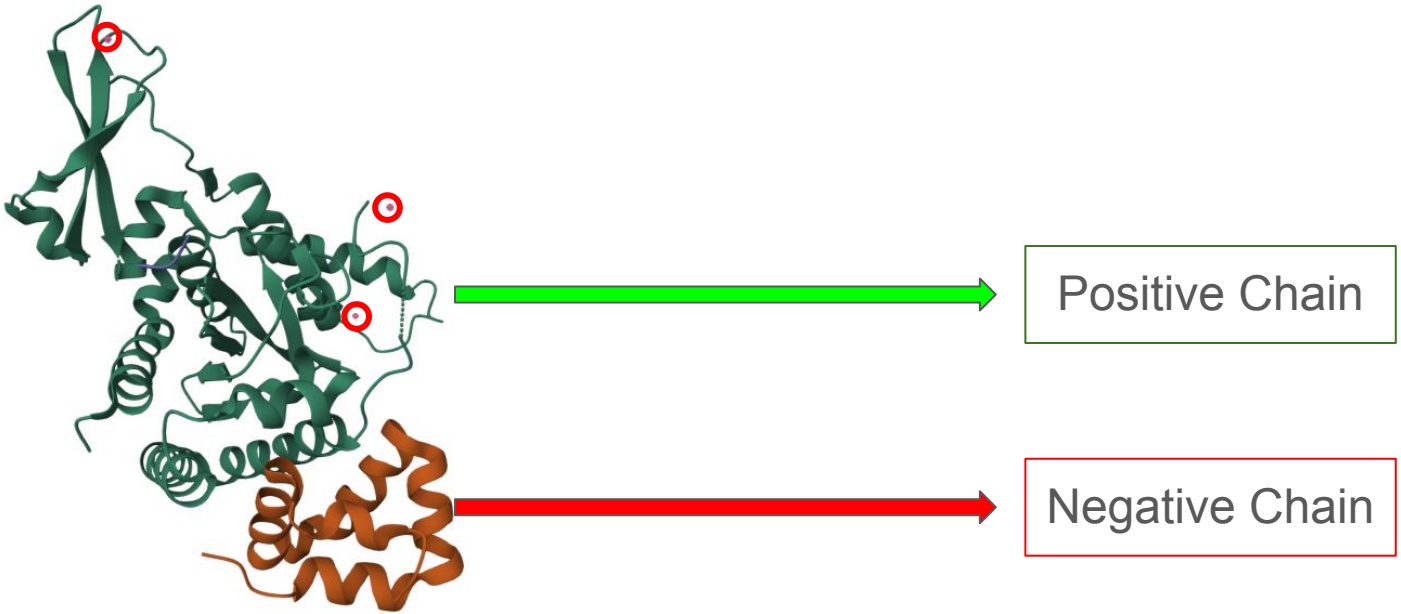


Dataset creation: Negative set



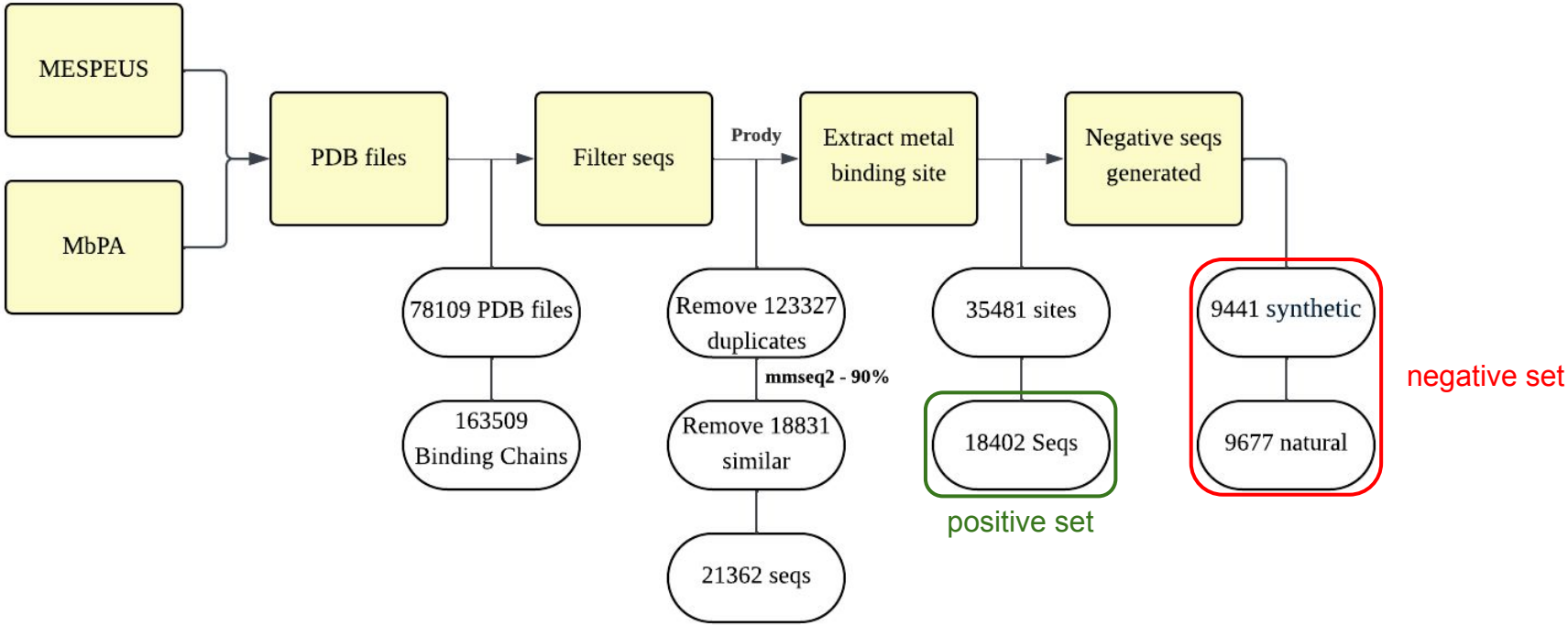
Second approach to get negative sequences

Identify heteromeric protein complexes within the positive set and use their non-binding chains to expand negative set.

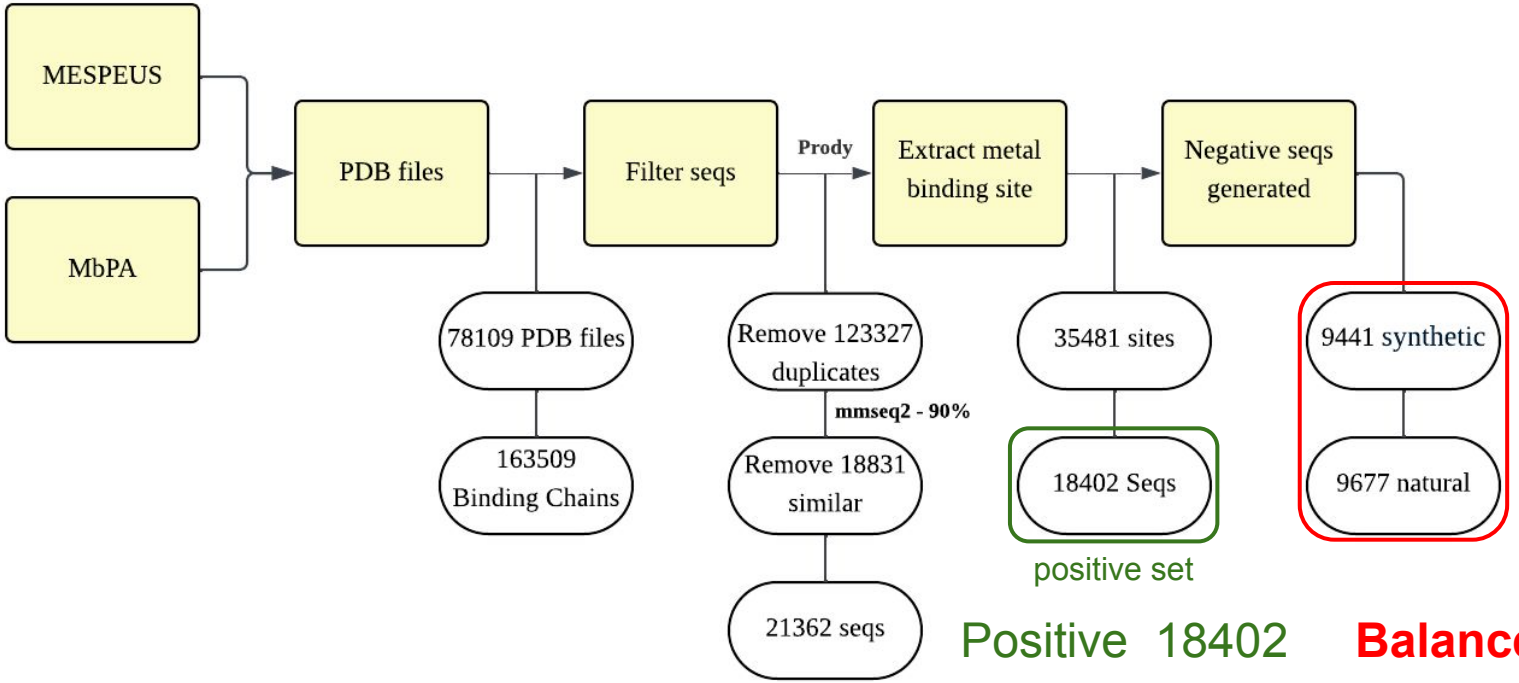


A A Aptekmann, J Buongiorno, D Giovannelli, M Glamoclija, D U Ferreiro, Y Bromberg, mebipred: identifying metal-binding potential in protein sequence, Bioinformatics, Volume 38, Issue 14, July 2022, Pages 3532–3540, https://doi.org/10.1093/bioinformatics/btac358

Dataset creation: Negative set



Dataset creation: Negative set



Positive 18402 **Balanced!**

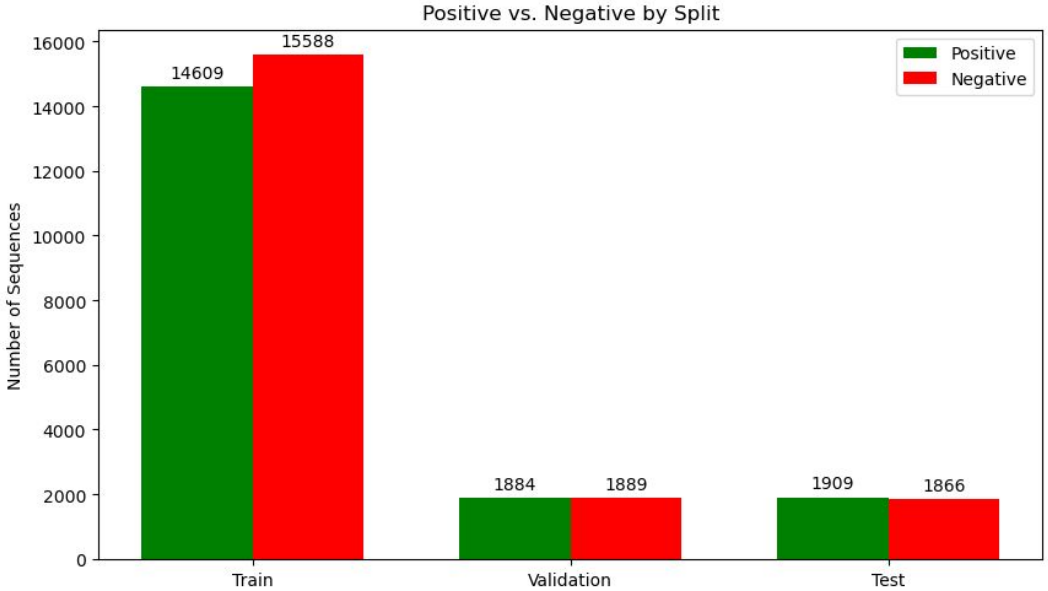
Negative 9441+9677 = 19118

Dataset split

Merge the positive and negative sets.

Cluster at **30 pident** with mmseq2.

Split the dataset base on clusters → **8:1:1** to Train, Validation, Test set



Training

Model: Training ProtBERT model with adapters.

Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, Michal Linial, ProteinBERT: a universal deep-learning model of protein sequence and function, Bioinformatics, Volume 38, Issue 8, March 2022, Pages 2102–2110, https://doi.org/10.1093/bioinformatics/btac020

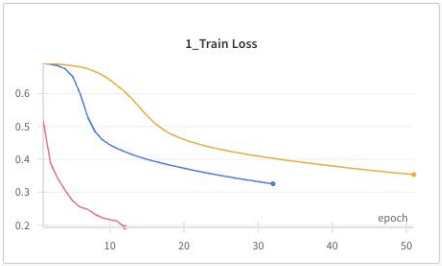
Training - Parameter search

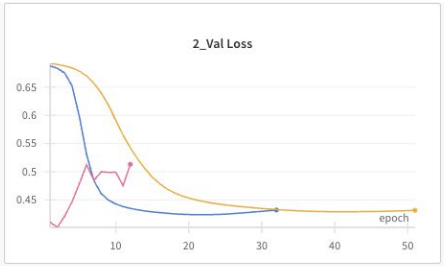
Model: Training ProtBERT model with adapters.

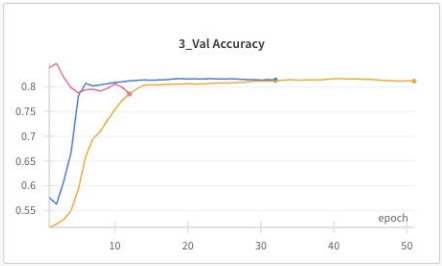
learning_rate: [1e-4, 5e-4, 1e-5, 5e-6, 1e-6]

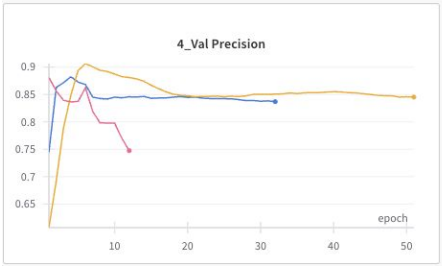
batch_size: [64, 128, 256, 512]

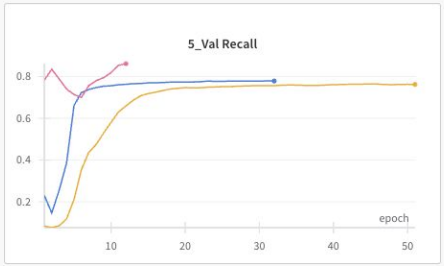
dropout_p: [0.3, 0.5]

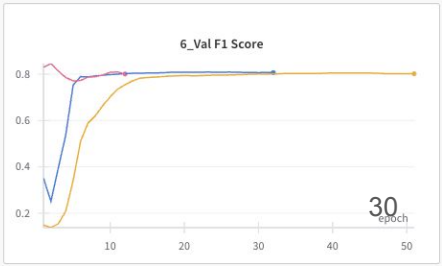












Training - Parameter search

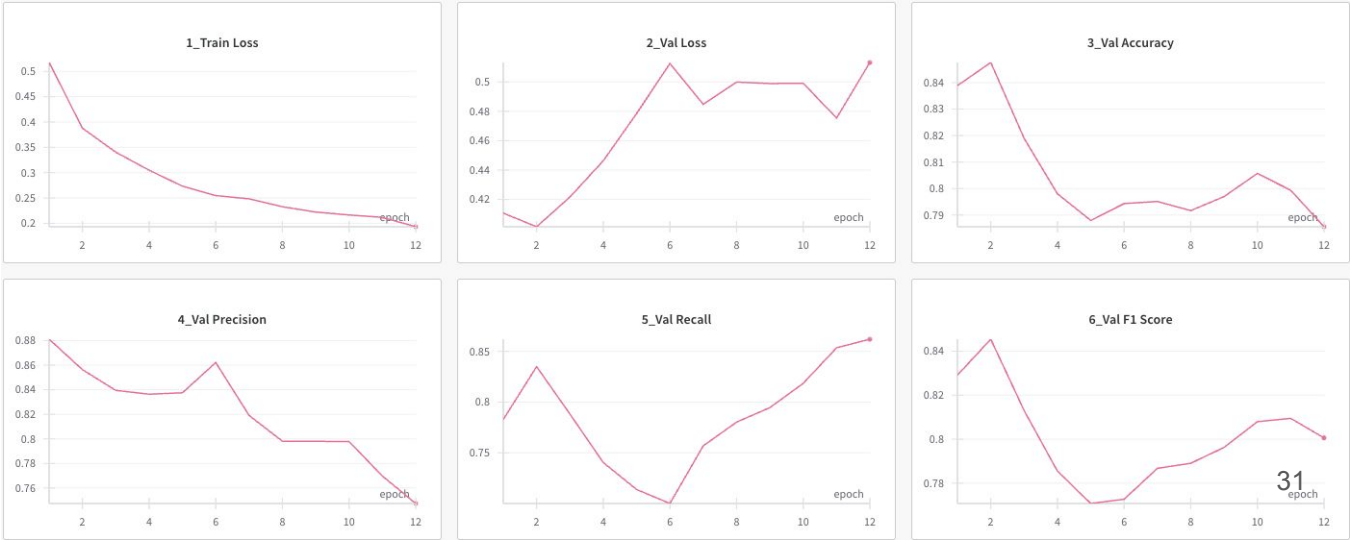
Model: Training ProtBERT model with adapters.

learning_rate: [1e-4, **5e-4**, 1e-5, 5e-6, 1e-6]

batch_size: [64, **128**, 256, 512]

dropout_p: [0.3, **0.5**]

Best performing model



Training - Parameter search

Model: Training ProtBERT model with adapters.

learning_rate: [1e-4, **5e-4**, 1e-5, 5e-6, 1e-6]

batch_size: [64, **128**, 256, 512]

dropout_p: [0.3, **0.5**]

Best performing model

After 2 epochs, model overfits!

Evaluation

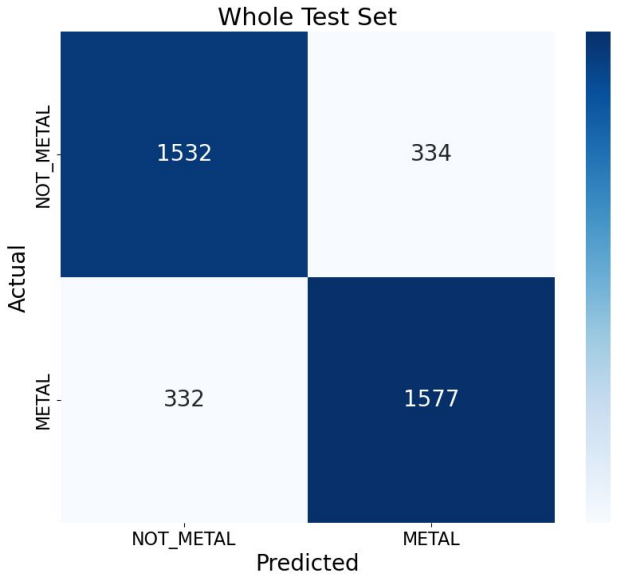
Metrics:

Accuracy: 82.36%

Precision: 82.52%

Recall: 82.61%

F1 Score: 82.57%



Evaluation - only on real sequences

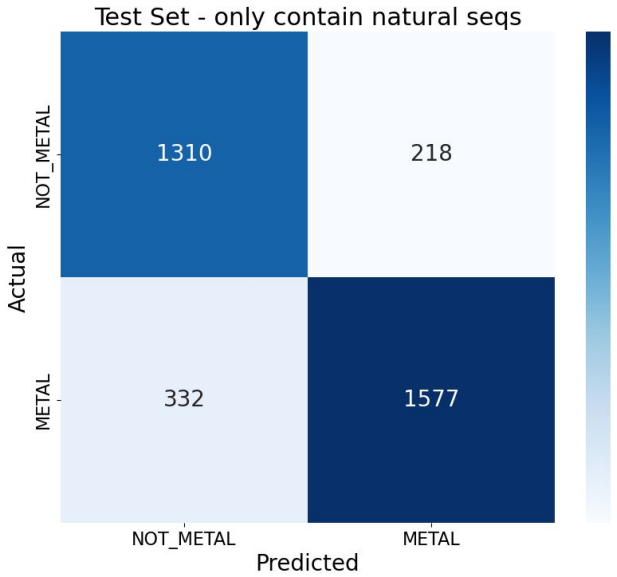
Metrics (excluding ProteinMPNN redesigned sequences):

Accuracy: 82.36% ⟶ **84.00%**

Precision: 82.52% ⟶ **87.86%**

Recall: 82.61%

F1 Score: 82.57% ⟶ **85.15%**

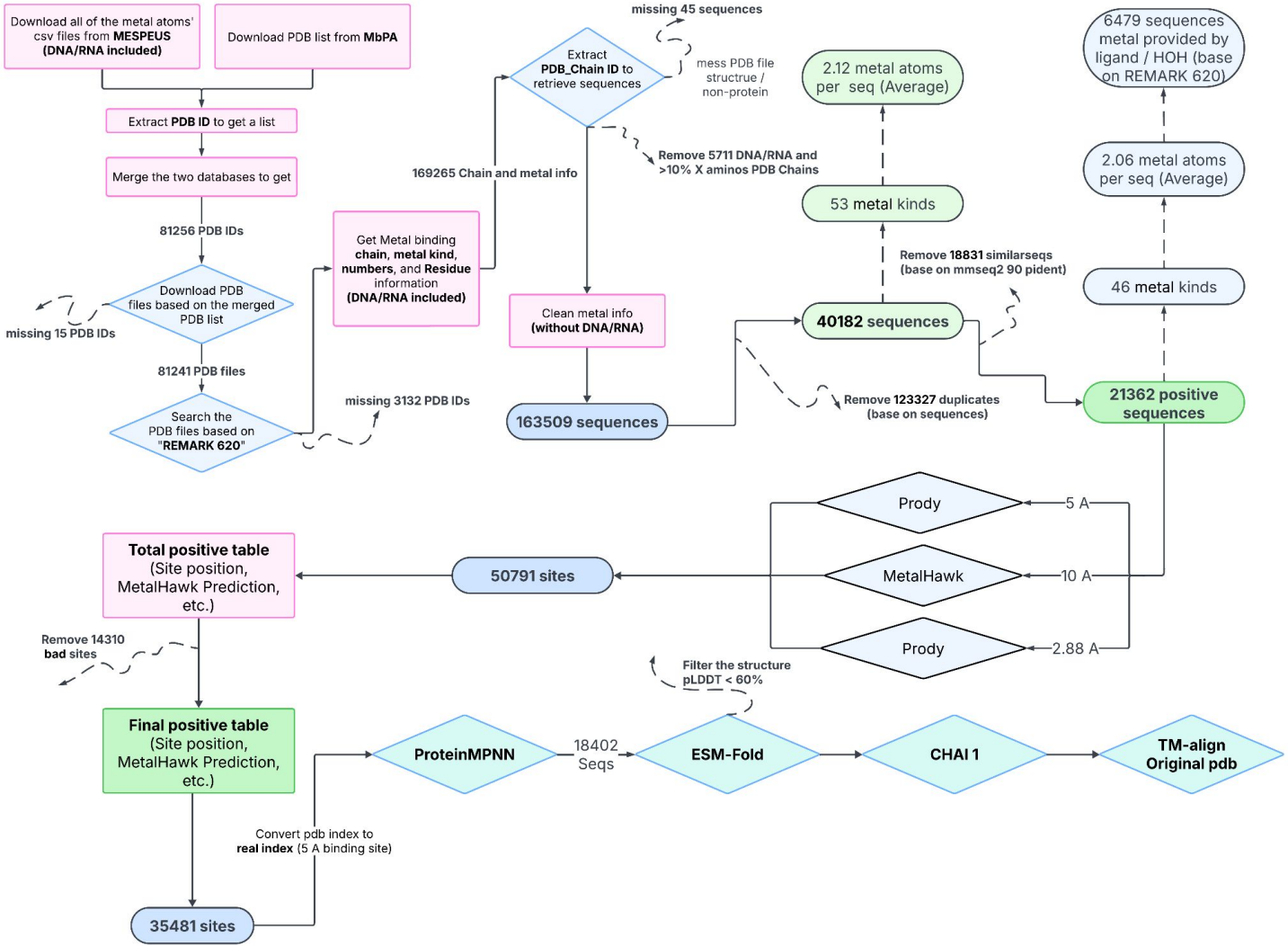


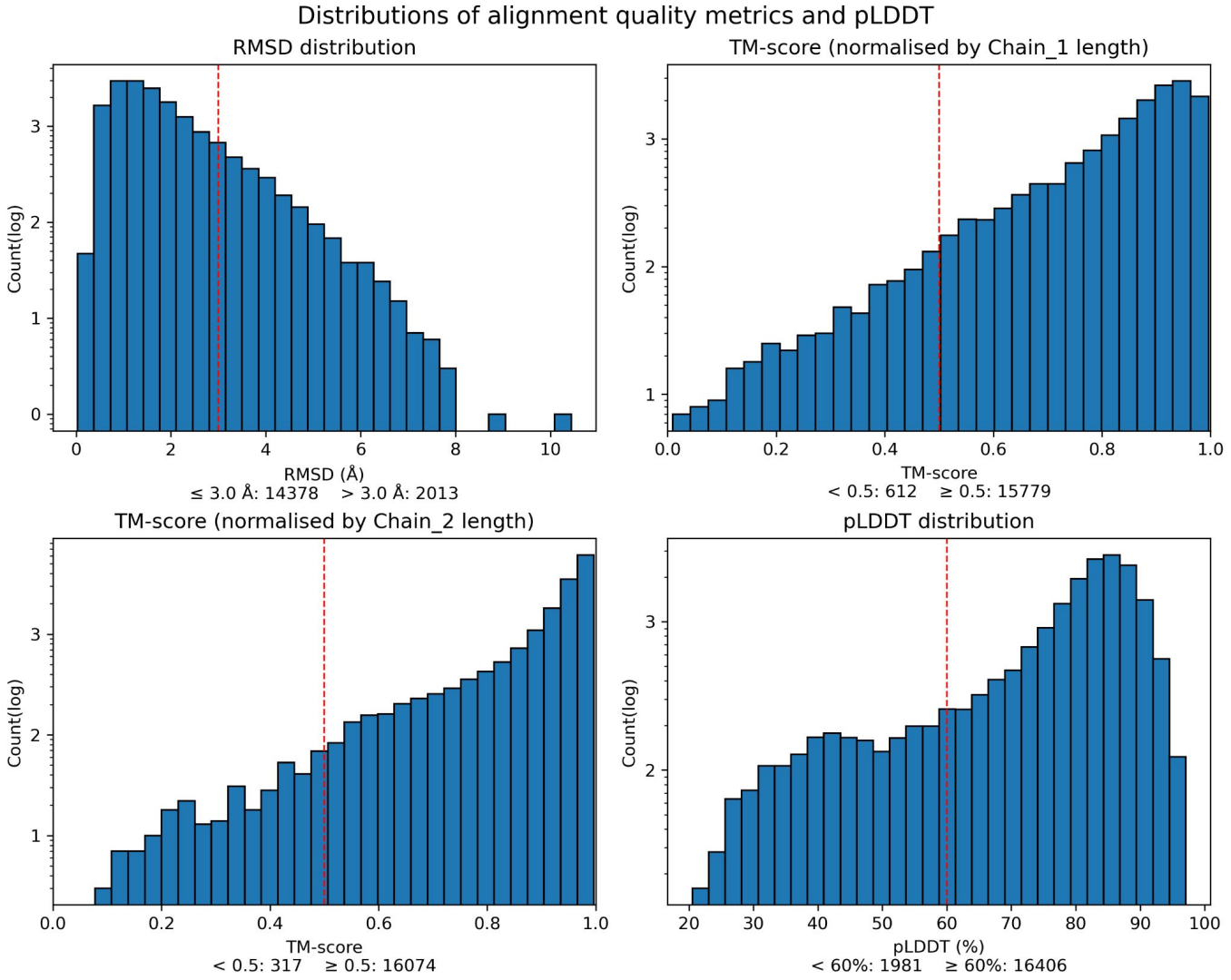
Outlook

1. Train the model with only synthetic data, and only with natural data, to compare the performances.
2. Implement the remaining 3 models.

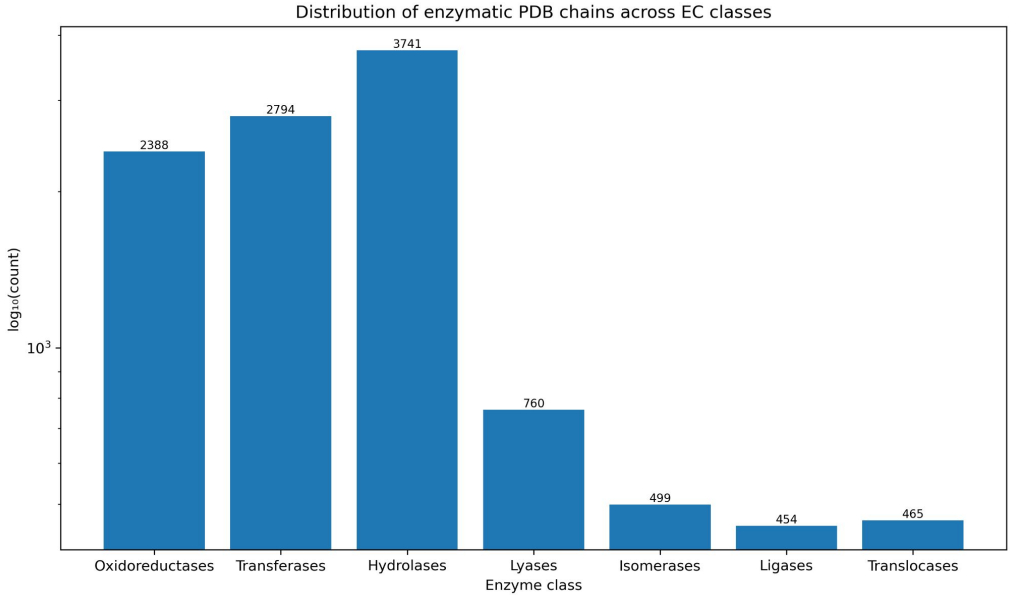
Thanks for your attention!

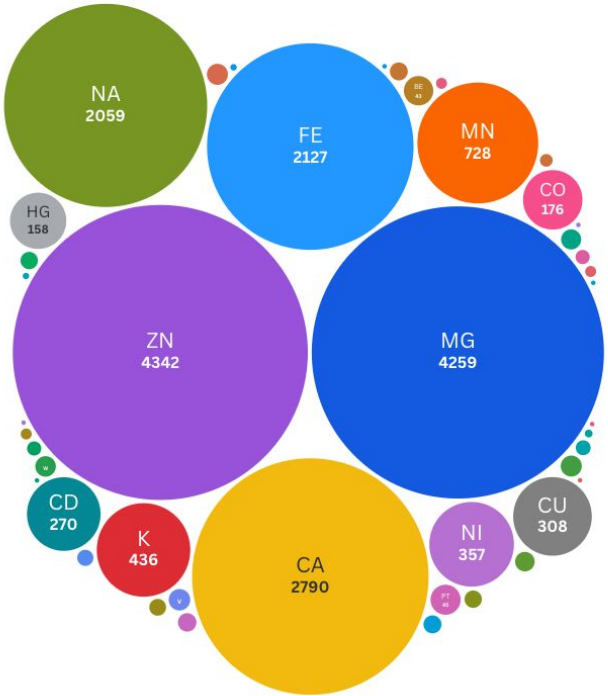
Backup slides


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
MetalHawk
10 A
Prody
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.)
Filter the structure pLDDT < 60%
ProteinMPNN
18402 Seqs
ESM-Fold
CHAI 1
TM-align Original pdb
Convert pdb index to real index (5 A binding site)
35481 sites


Distributions of alignment quality metrics and pLDDT
RMSD distribution
Count(log)
RMSD (Å)
≤ 3.0 Å: 14378 > 3.0 Å: 2013
TM-score (normalised by Chain_1 length)
Count(log)
TM-score
< 0.5: 612 ≥ 0.5: 15779
TM-score (normalised by Chain_2 length)
Count(log)
TM-score
< 0.5: 317 ≥ 0.5: 16074
pLDDT distribution
Count(log)
pLDDT (%)
< 60%: 1981 ≥ 60%: 16406

Distribution/Analysis of our Positive dataset - 18402 seqs





REMARK 620

In the PDB file have a REMARK 620 header, it contain the metal ion's **coordination** information for this protein.

- Manual review
- High confidence

```
REMARK 620
REMARK 620 METAL COORDINATION
REMARK 620 (M=MODEL NUMBER; RES=RESIDUE NAME; C=CHAIN IDENTIFIER;
REMARK 620 SSEQ=SEQUENCE NUMBER; I=INSERTION CODE):
REMARK 620
REMARK 620 COORDINATION ANGLES FOR:  M RES CSSEQI METAL
REMARK 620                              MG A 301  MG
REMARK 620 N RES CSSEQI ATOM
REMARK 620 1 SER A  17   OG
REMARK 620 2 GDP A 302   O1B  92.7
REMARK 620 3 HOH A 405   O    82.8  92.5
REMARK 620 4 HOH A 408   O    91.6  86.0 174.2
REMARK 620 5 HOH A 409   O    87.7 171.9  95.5  85.9
REMARK 620 6 HOH A 436   O   173.2  90.3  90.9  94.7  90.3
REMARK 620 N                   1     2     3     4     5
REMARK 620
REMARK 620 COORDINATION ANGLES FOR:  M RES CSSEQI METAL
REMARK 620                              MG B 301  MG
REMARK 620 N RES CSSEQI ATOM
REMARK 620 1 SER B  17   OG
REMARK 620 2 GDP B 302   O2B  89.4
REMARK 620 3 HOH B 438   O   170.2  96.0
REMARK 620 4 HOH B 439   O    81.7  81.0  91.1
REMARK 620 5 HOH B 444   O    92.5 103.1  94.2 172.9
REMARK 620 6 HOH B 468   O    89.3 173.0  84.3  92.0  83.9
REMARK 620 N                   1     2     3     4     5
```

Filter positive sequences - Remove similar

We run mmseq2 at **90 pident**.

For each cluster, we retain only the centroid.

And remove the other sequences within the same cluster (which share at least 90% similarity with the centroid).

centroid	cluster_members
3cif_B	3cif_B
3l8g_A	3l8g_A
3l8g_A	3esr_A
6jv4_A	6jv4_A
2dcj_A	2dcj_A
2dcj_A	6kka_A
2dcj_A	6kjl_A
5a2h_A	5a2h_A
5a2h_A	1vrk_A
5a2h_A	1rfj_A

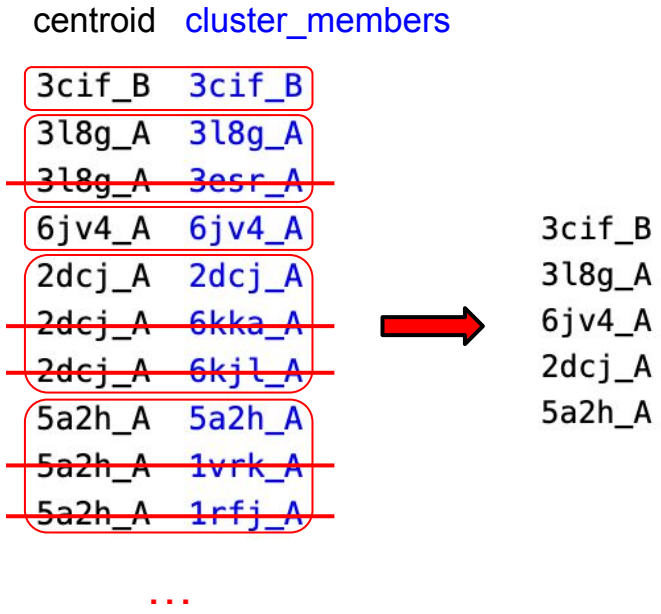
...

Filter positive sequences - Remove similar

We run mmseq2 at **90 pident**.

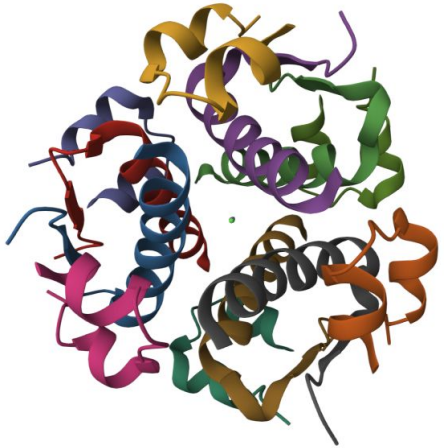
For each cluster, we retain only the centroid.

And remove the other sequences within the same cluster (which share at least 90% similarity with the centroid).



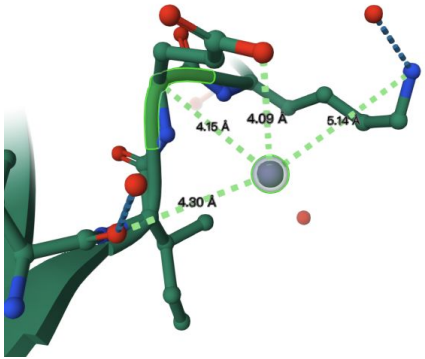
Extract metal binding site - drop bad sites

For the bad sites, we remove them:



1QIY

1. multichain



7CMZ

2. No interacting (>2.88 Å)

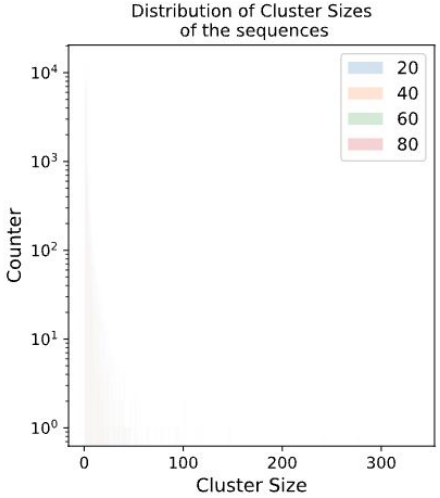
3. No REMARK 620 information

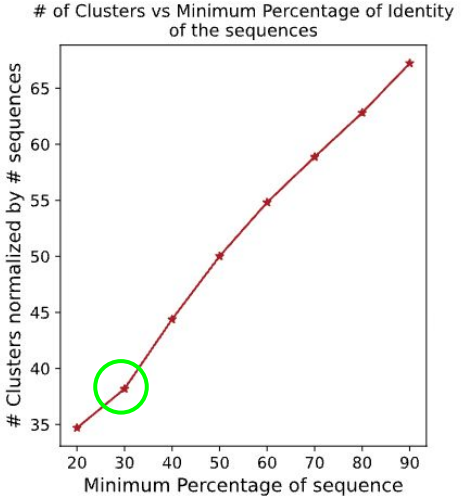
Dataset split

Merge the positive and negative sets.

Choose **30 Pident** as threshold to run mmseq2.

Split the dataset base on clusters -> **8:1:1** to Train, Validation, Test set





Dataset creation: Negative set

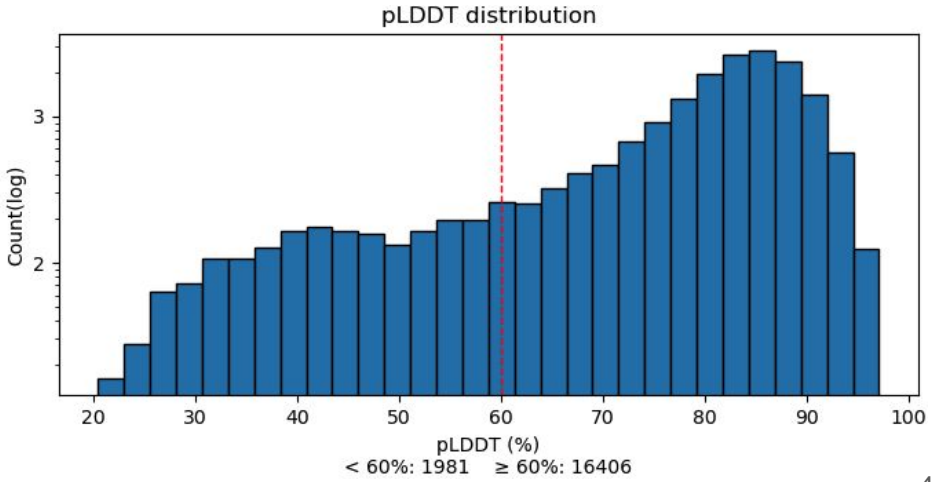
Qualitative evaluation:

1A7W original: …LE**E**MGR**D**IA…

ProteinMPNN: …LE**Q**MGR**M**IA…

ESM simulation → pLDDT calculation

We only keep the structure that **higher than 60%**



Chai Discoveryet al. ‚Chai-1: Decoding the molecular interactions of life. October 2024, doi: https://doi.org/10.1101/2024.10.10.615955