

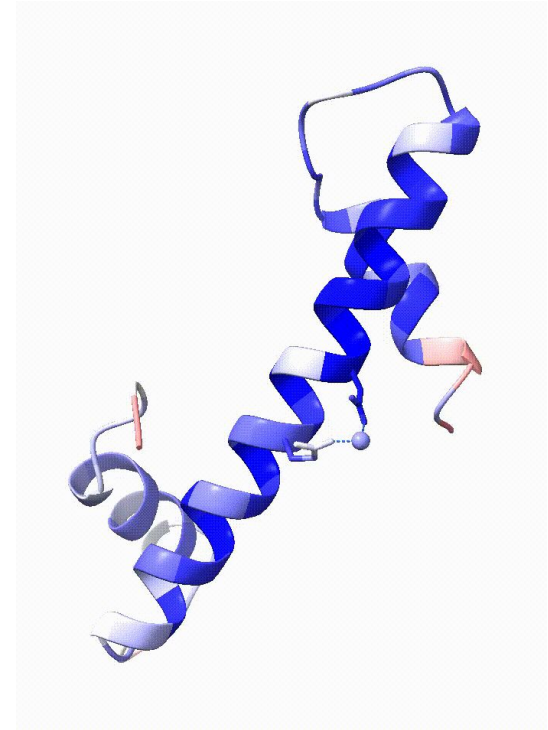
Advancing metal-binding protein predictions with deep learning

Update - Jingkai LAN

Supervisor: Thomas Lemmin
Co-supervisor: Giulia Peteani

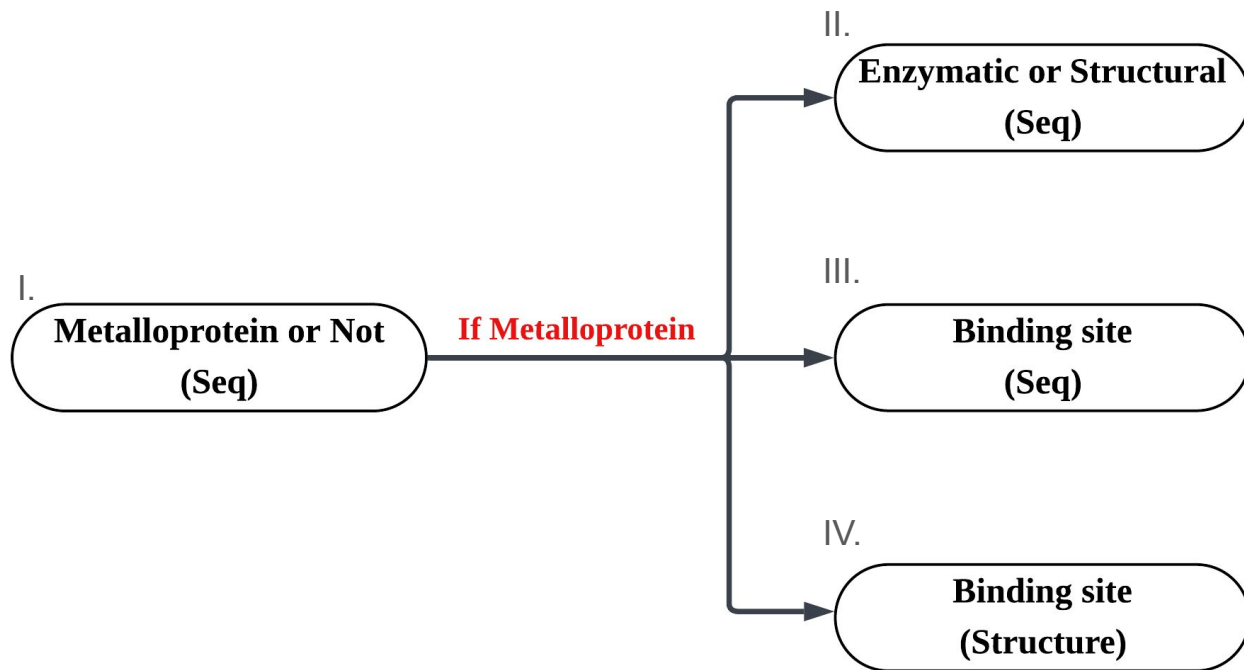
Introduction

- 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**.



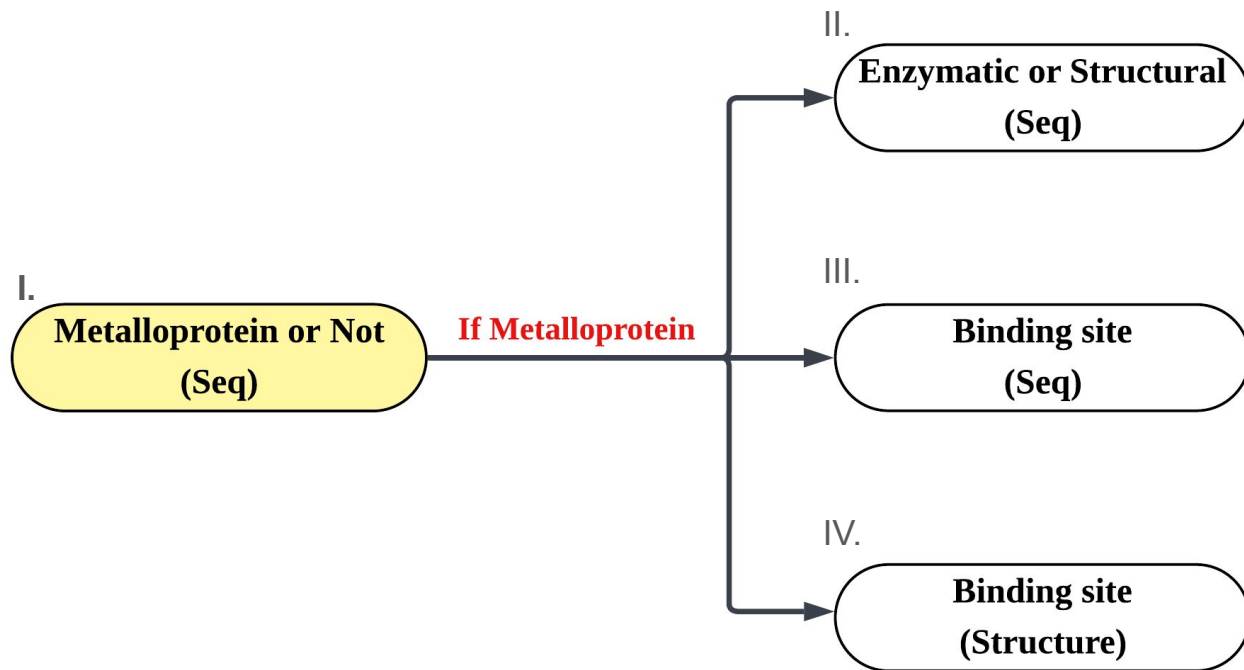
Project goal

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

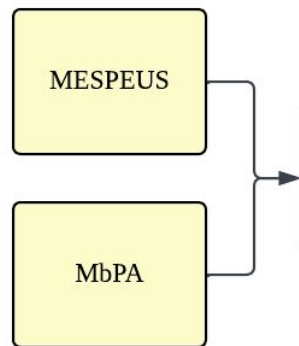


Project goal

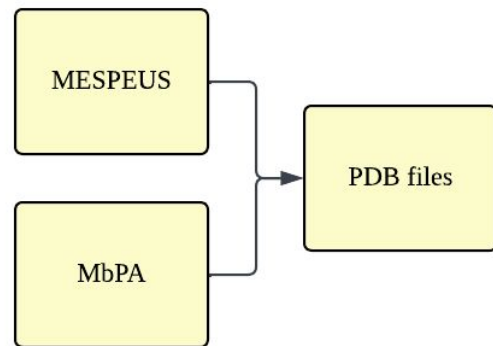
Leverage **deep learning** to investigate and model **key properties** of metalloproteins.



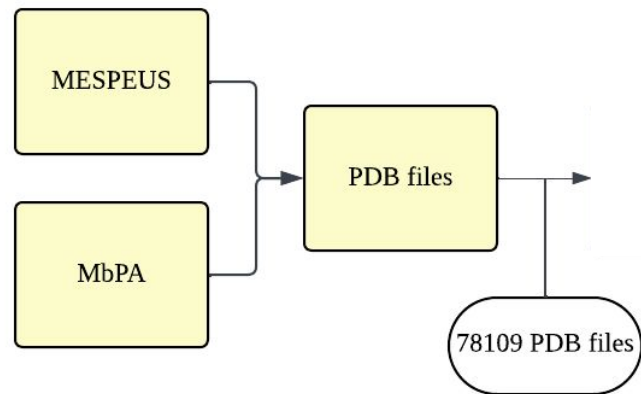
Dataset creation



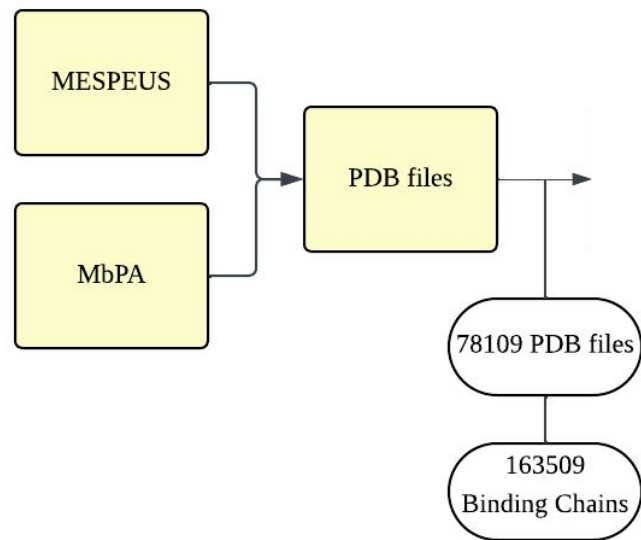
Dataset creation: Positive set



Dataset creation: Positive set



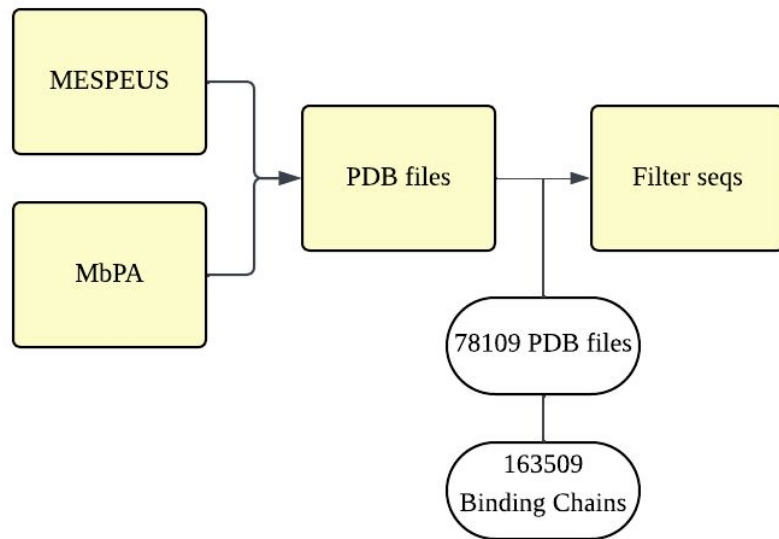
Dataset creation: Positive set



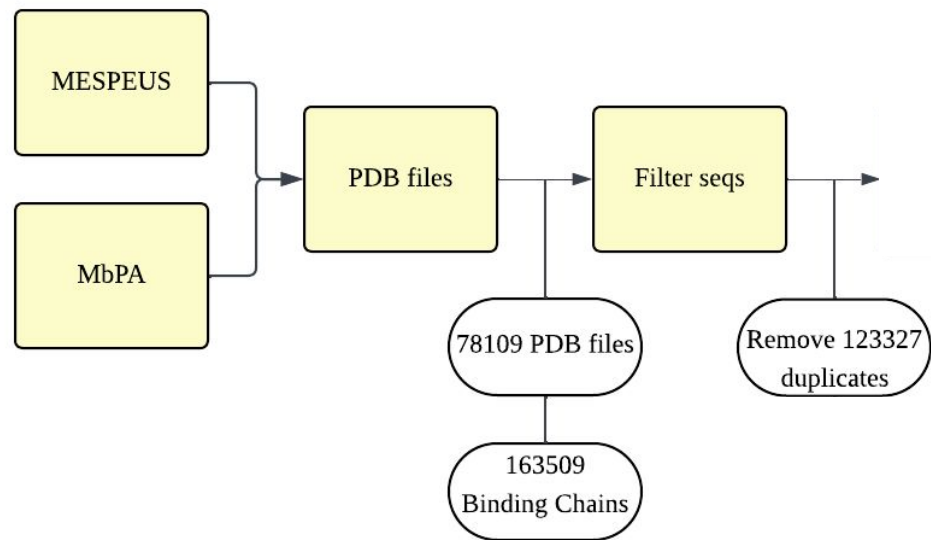
```

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 01B 92.7
REMARK 620 3 HOH A 405 0 82.8 92.5
REMARK 620 4 HOH A 408 0 91.6 86.0 174.2
REMARK 620 5 HOH A 409 0 87.7 171.9 95.5 85.9
REMARK 620 6 HOH A 436 0 173.2 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 02B 89.4
REMARK 620 3 HOH B 438 0 170.2 96.0
REMARK 620 4 HOH B 439 0 81.7 81.0 91.1
REMARK 620 5 HOH B 444 0 92.5 103.1 94.2 172.9
REMARK 620 6 HOH B 468 0 89.3 173.0 84.3 92.0 83.9
REMARK 620 N 1 2 3 4 5
  
```

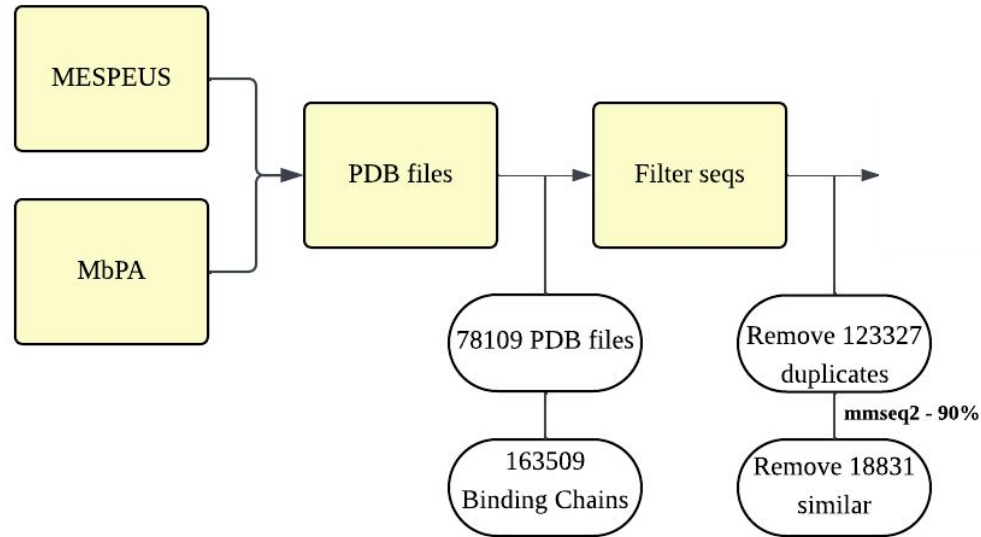

Dataset creation: Positive set



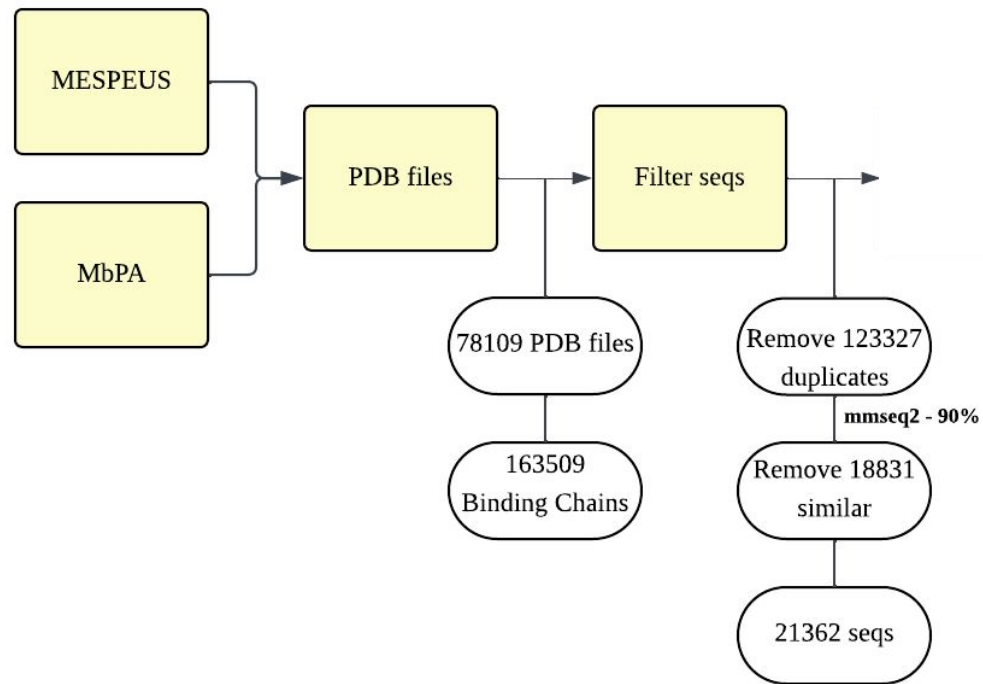
Dataset creation: Positive set



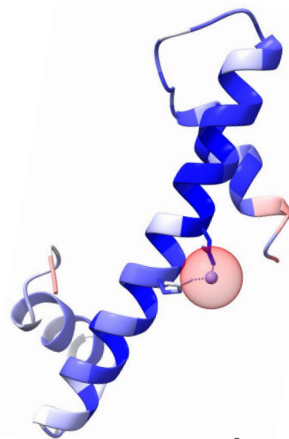
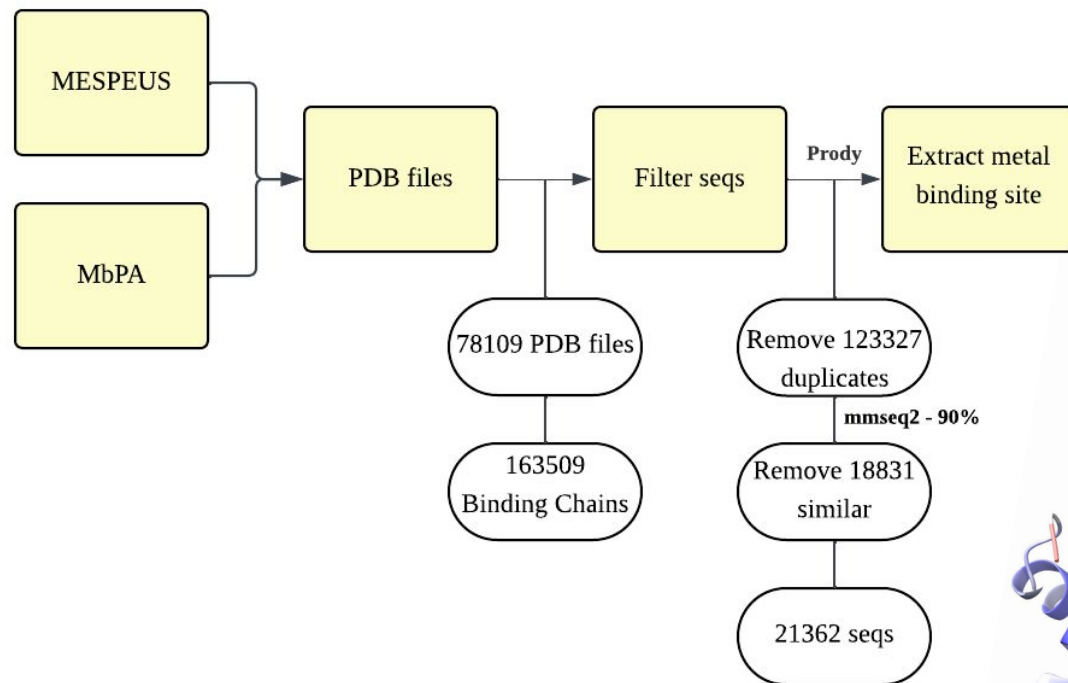
Dataset creation: Positive set



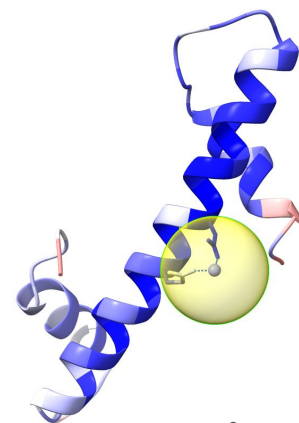
Dataset creation: Positive set



Dataset creation: Positive set

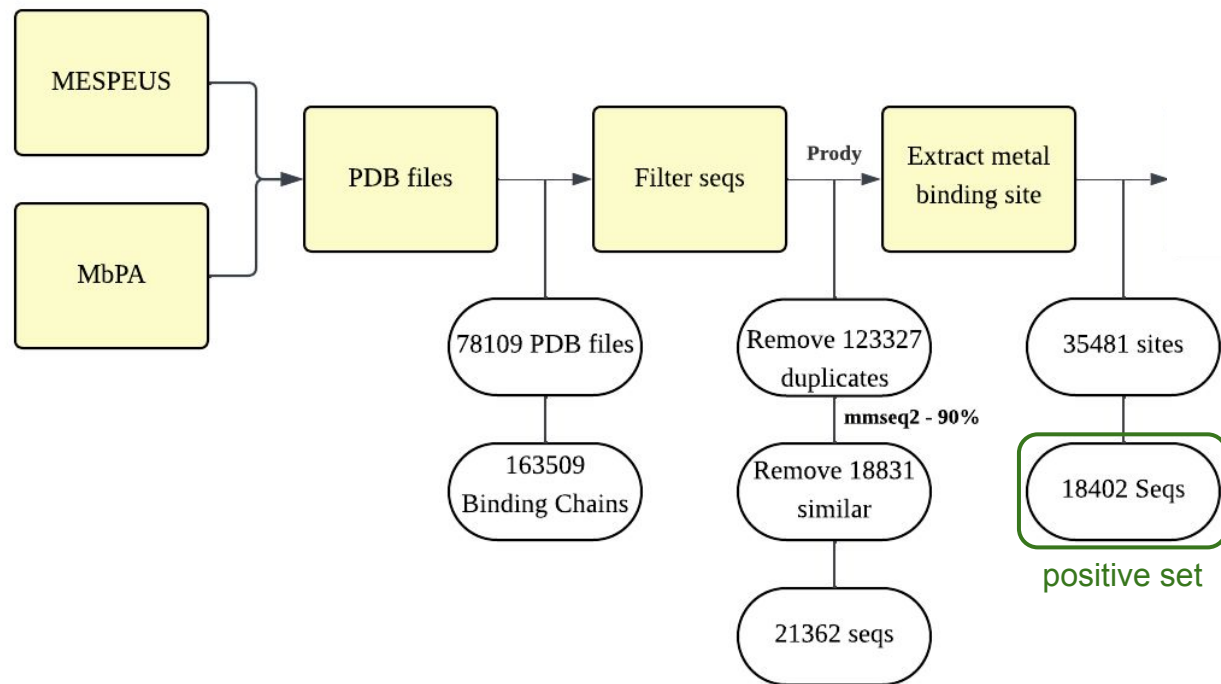


2.88 Å

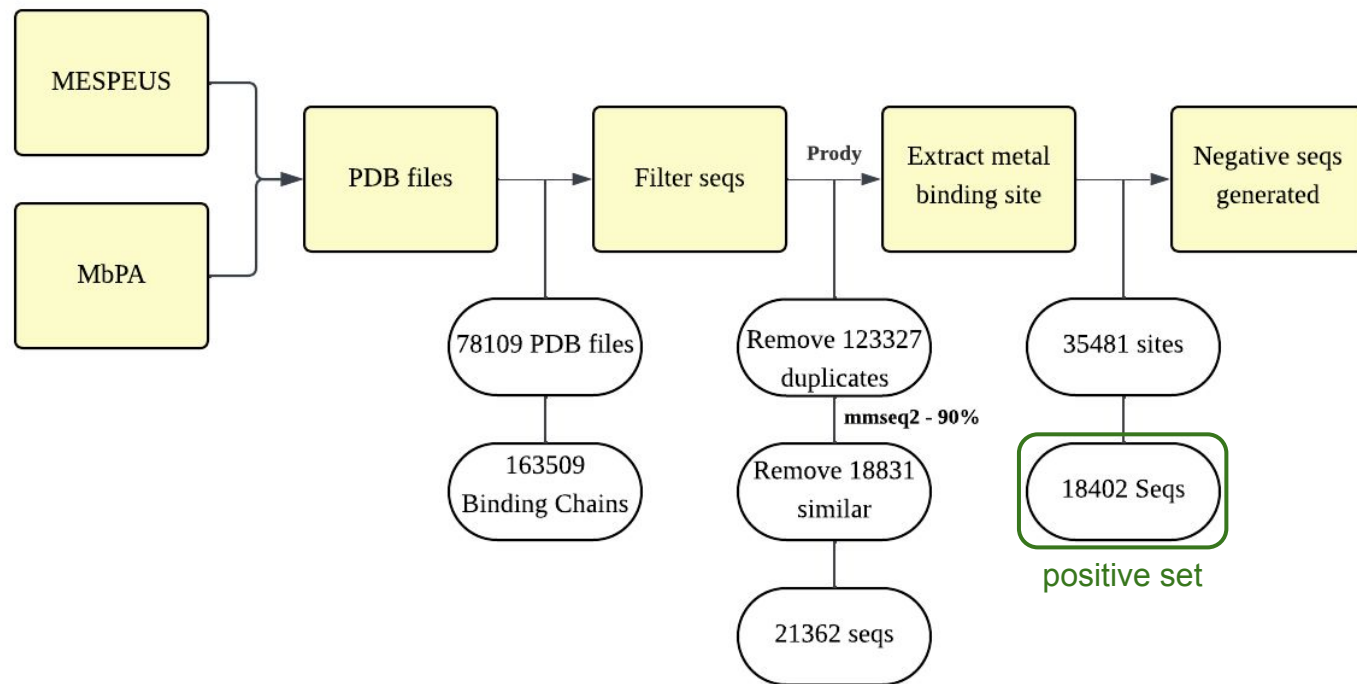


5 Å

Dataset creation: Positive set



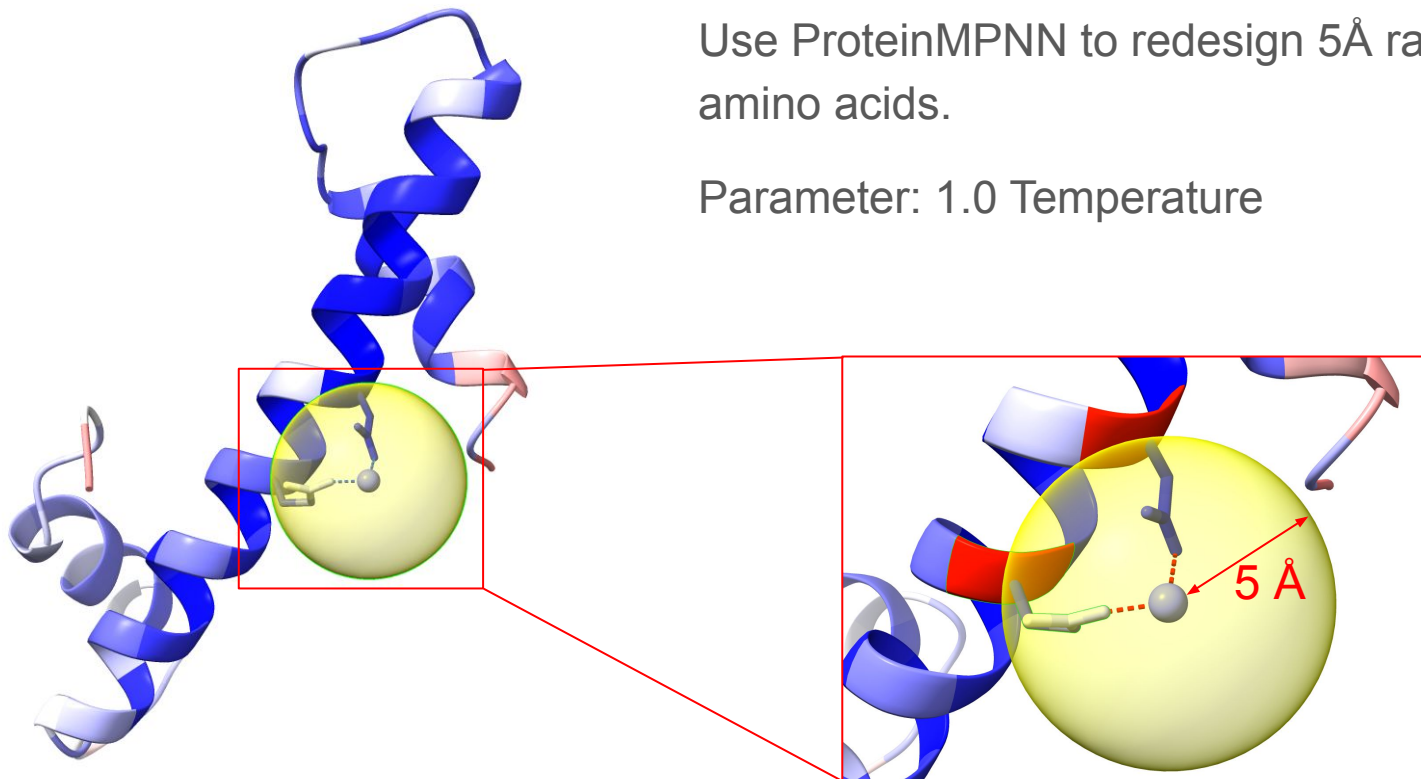
Dataset creation: Negative set



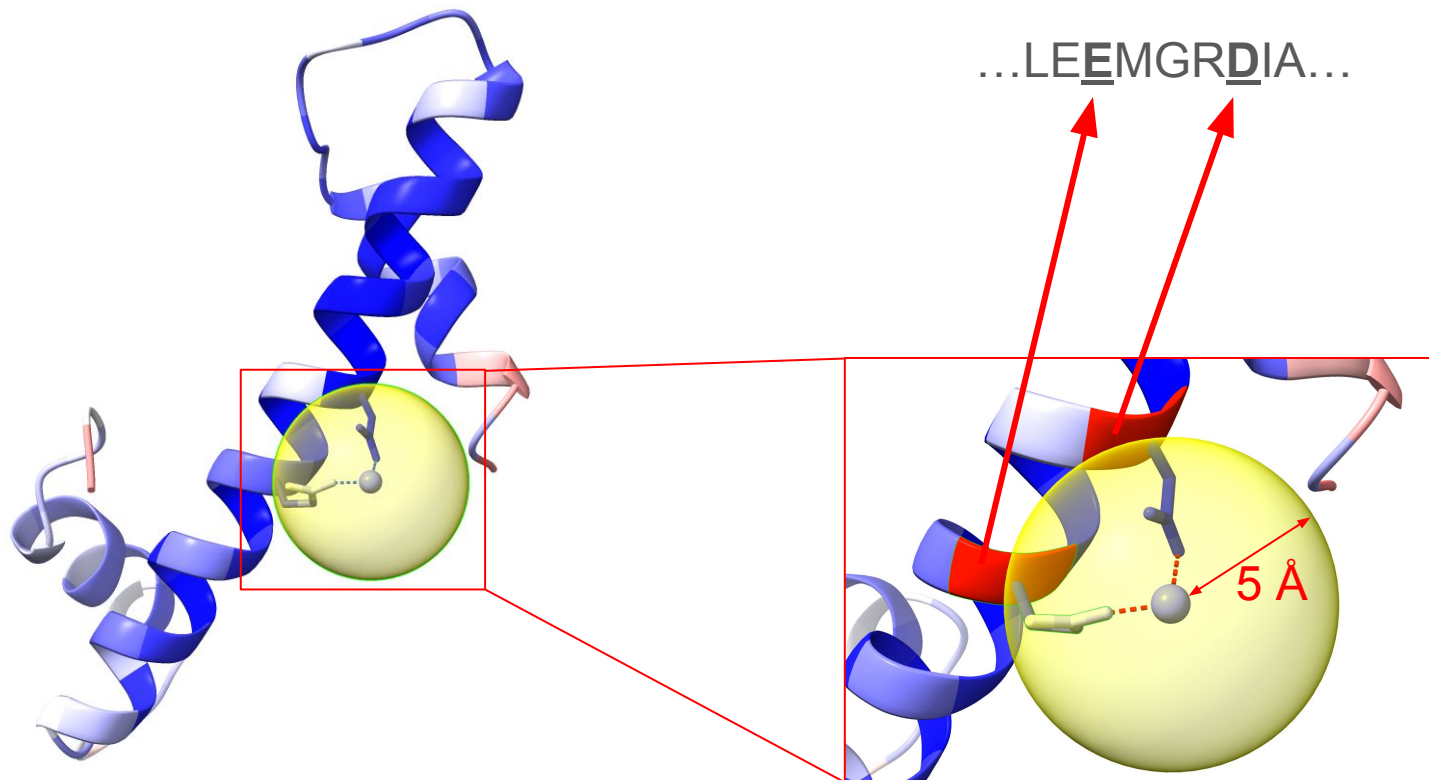
Dataset creation: Negative set

Use ProteinMPNN to redesign 5Å radius amino acids.

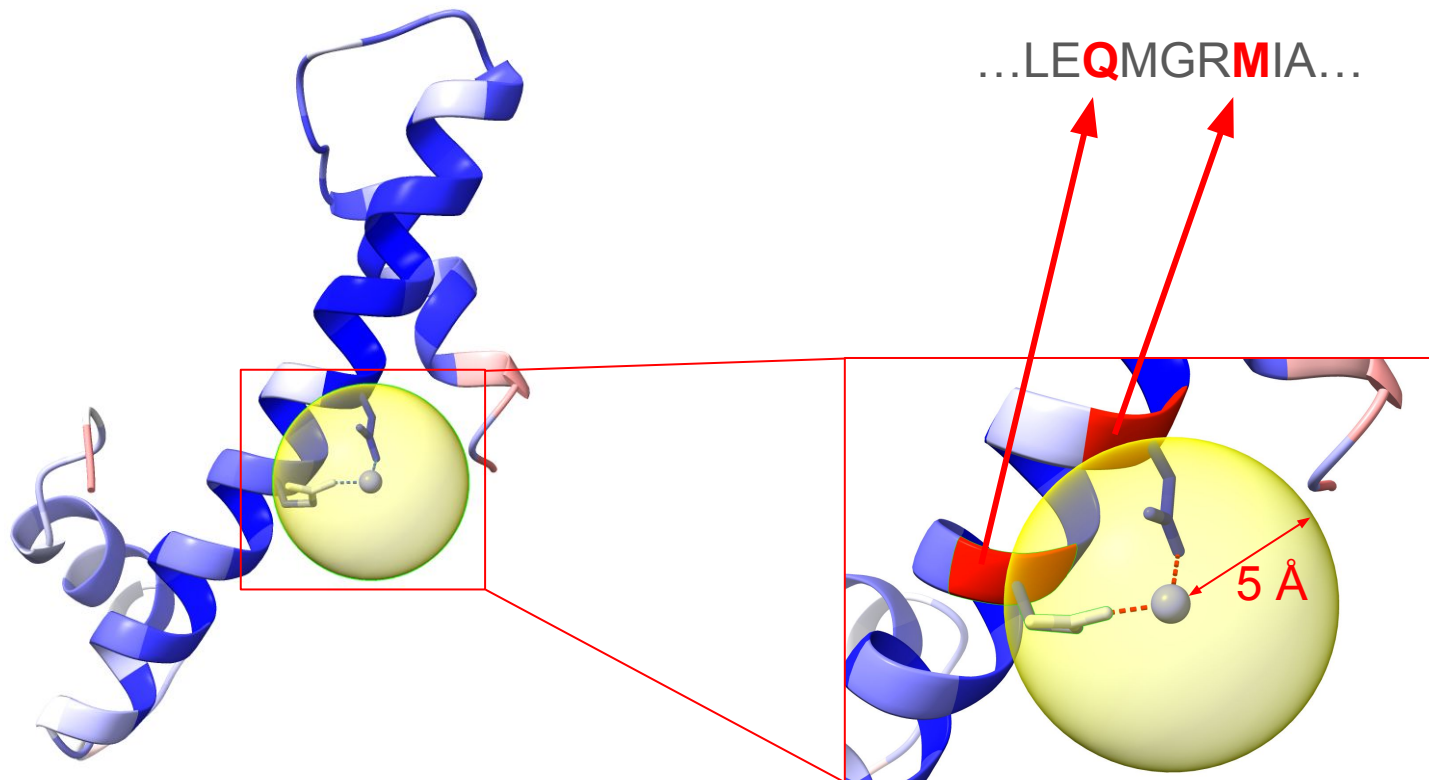
Parameter: 1.0 Temperature



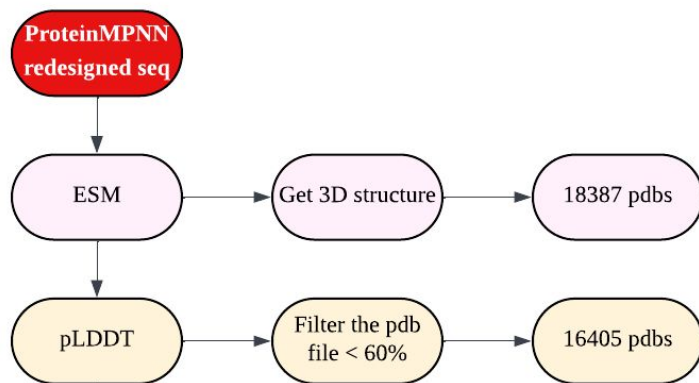
Dataset creation: Negative set



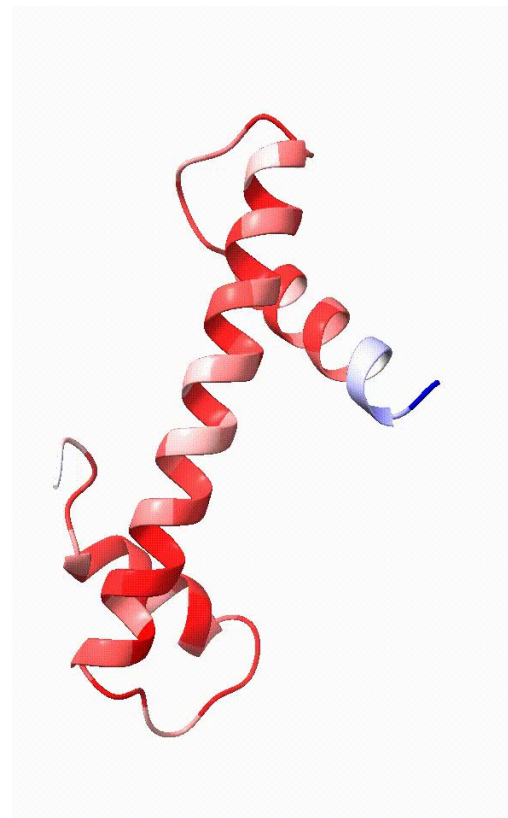
Dataset creation: Negative set



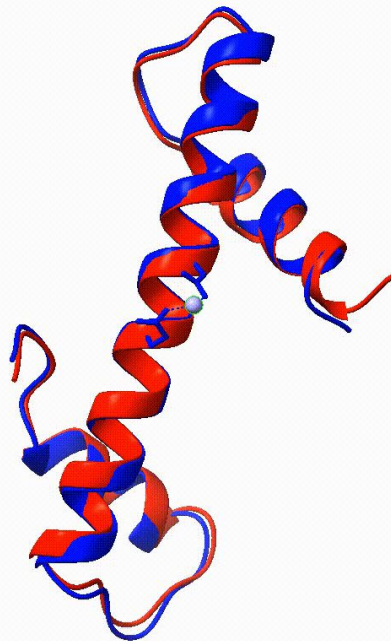
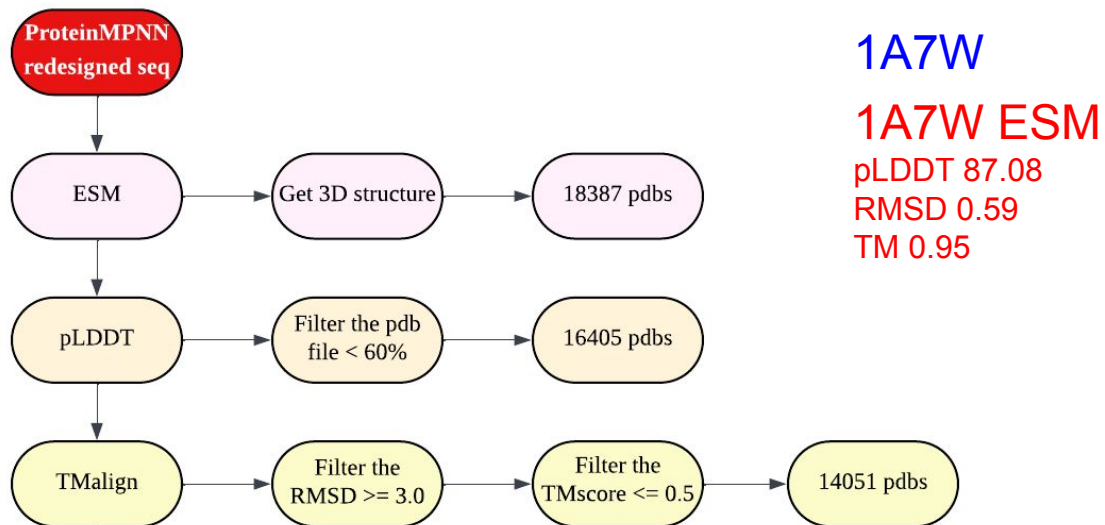
Validation of ProteinMPNN redesigned sequences



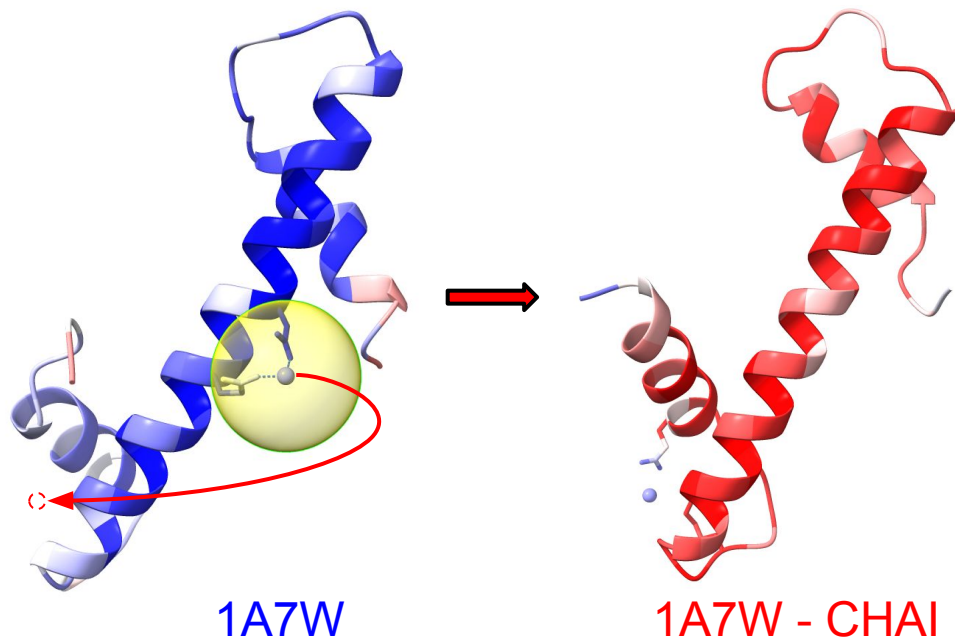
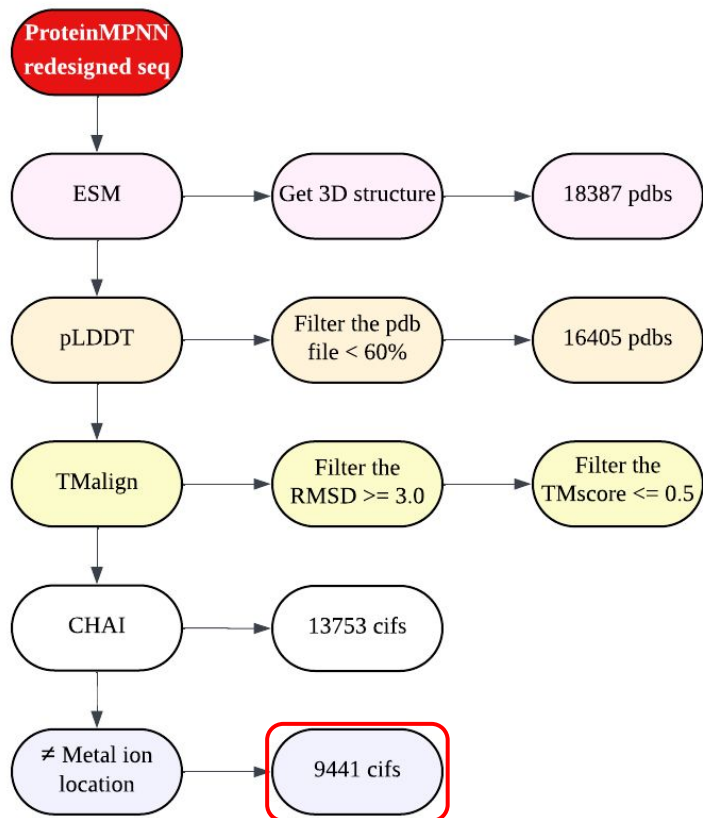
1A7W ESM
pLDDT 87.08



Validation of ProteinMPNN redesigned sequences



Validation of ProteinMPNN redesigned sequences



More examples..

original PDB

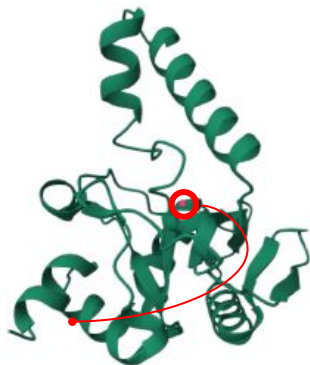
CHAI prediction

4kfv_A



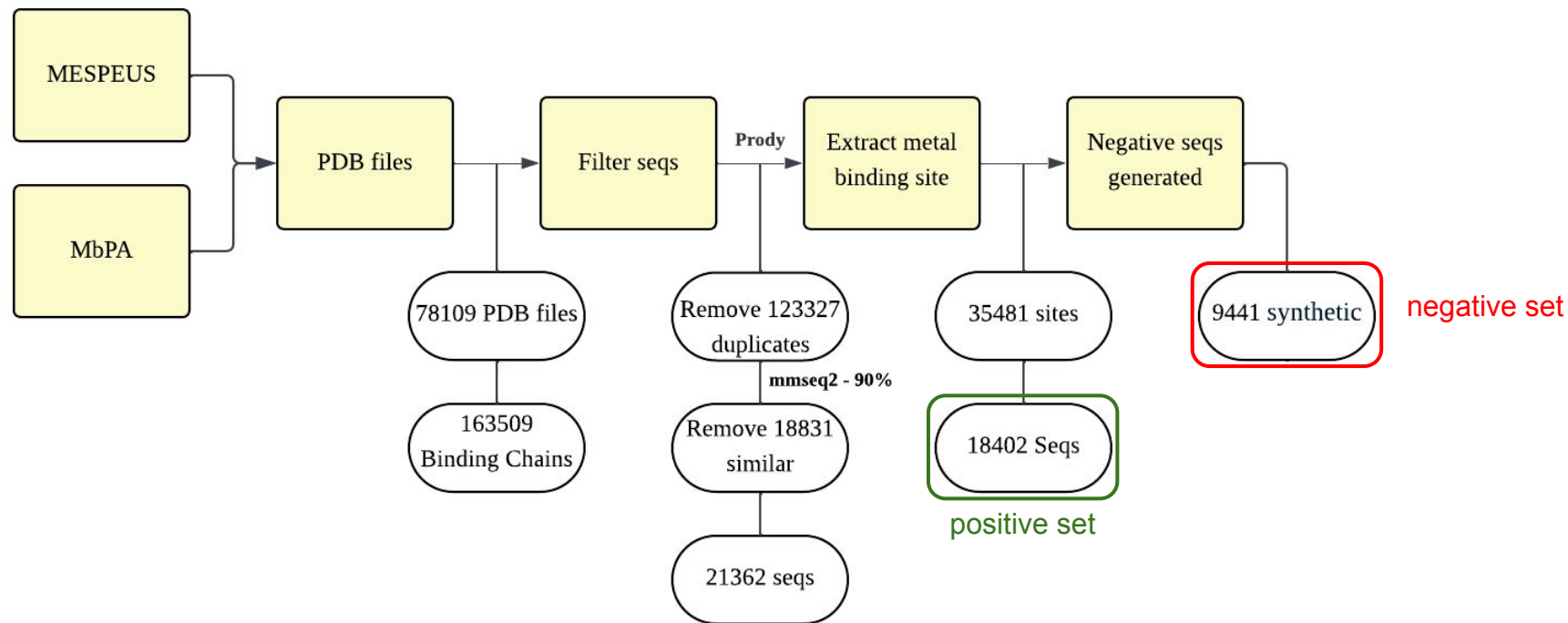
4kfv_A

5zhf_A

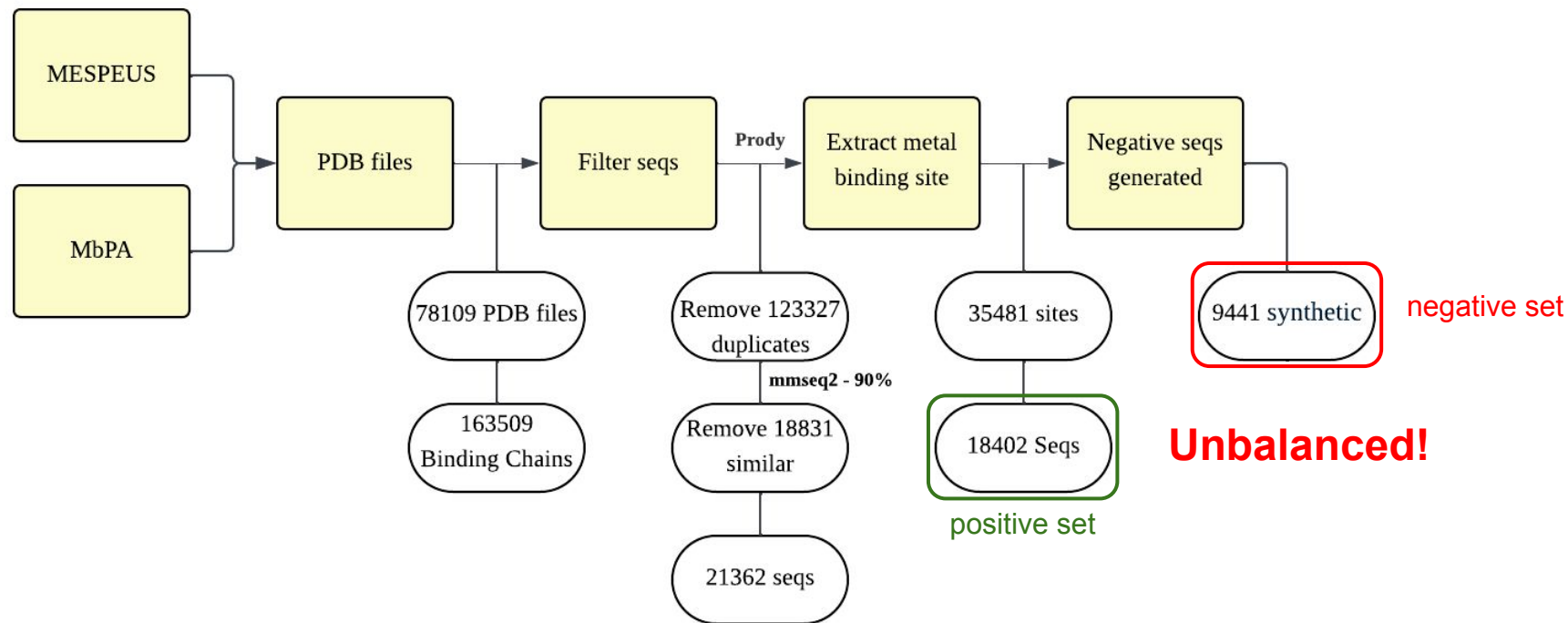


5zhf_A

Dataset creation: Negative set

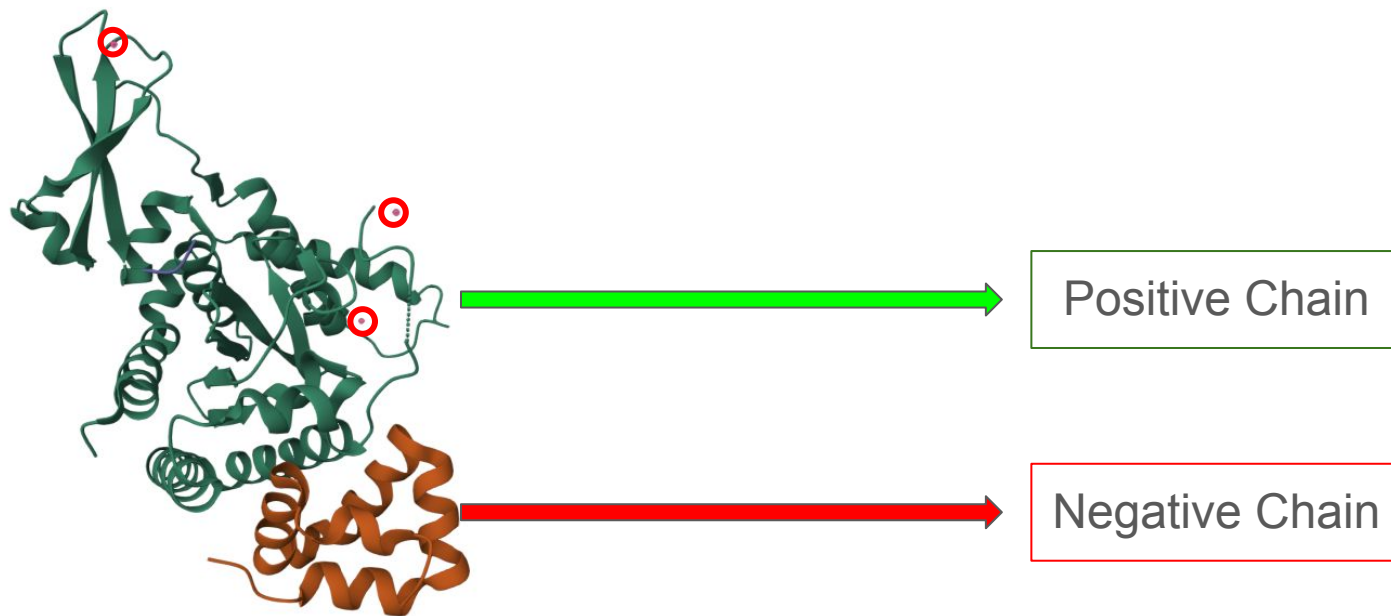


Dataset creation: Negative set

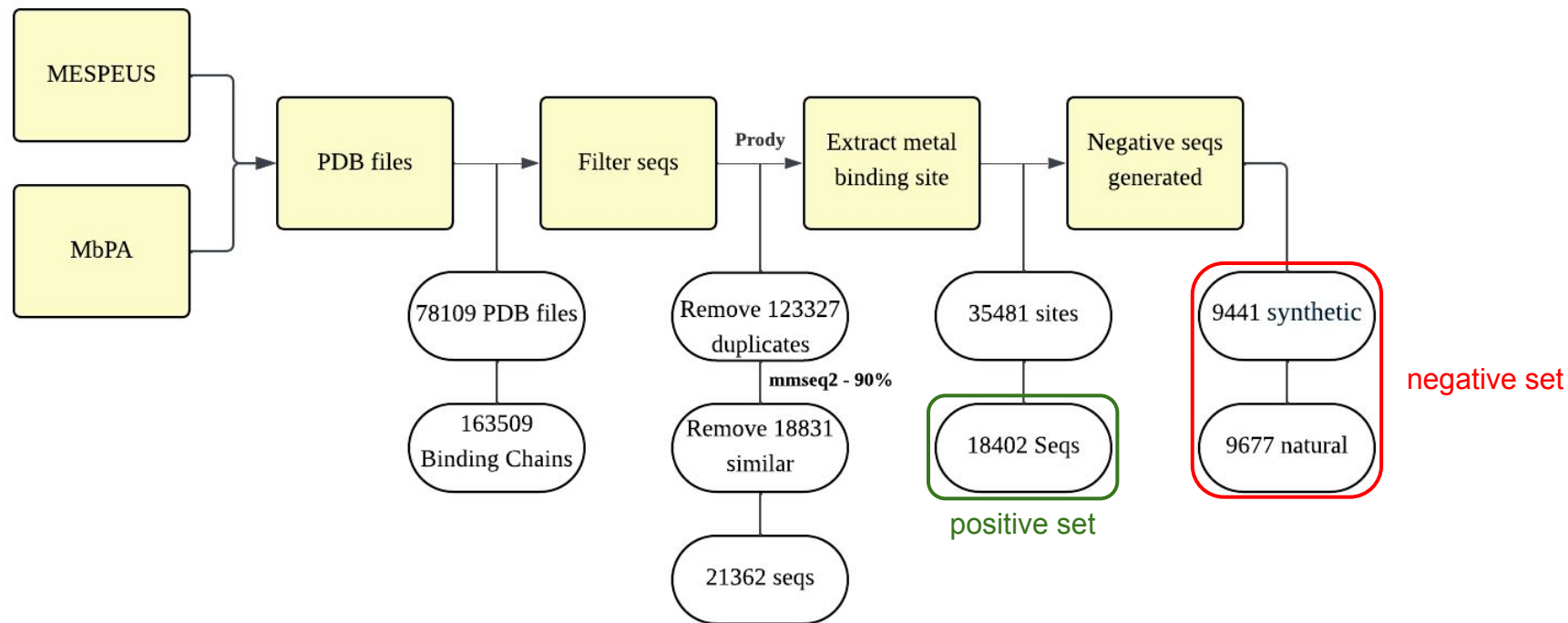


Second approach to get negative sequences

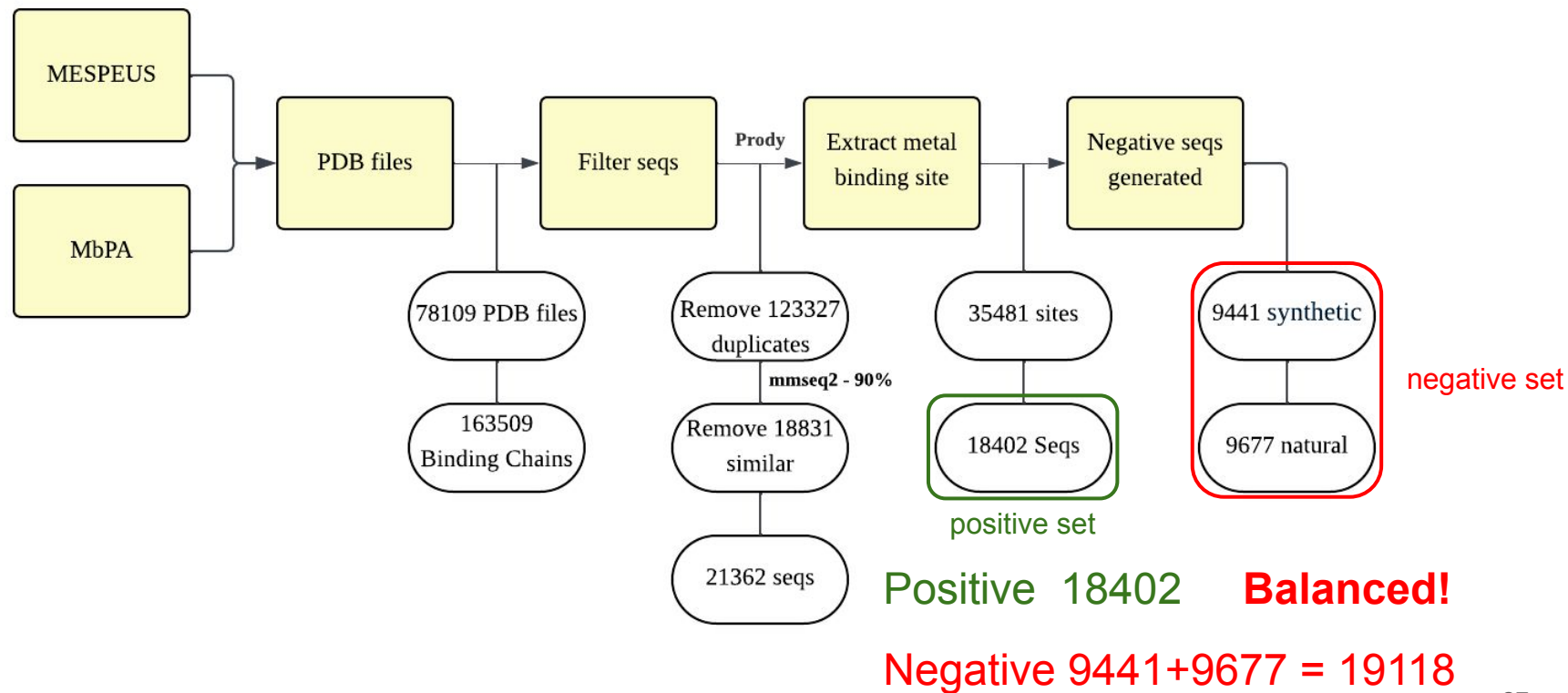
Identify heteromeric protein complexes within the positive set and use their non-binding chains to expand negative set.



Dataset creation: Negative set



Dataset creation: Negative set

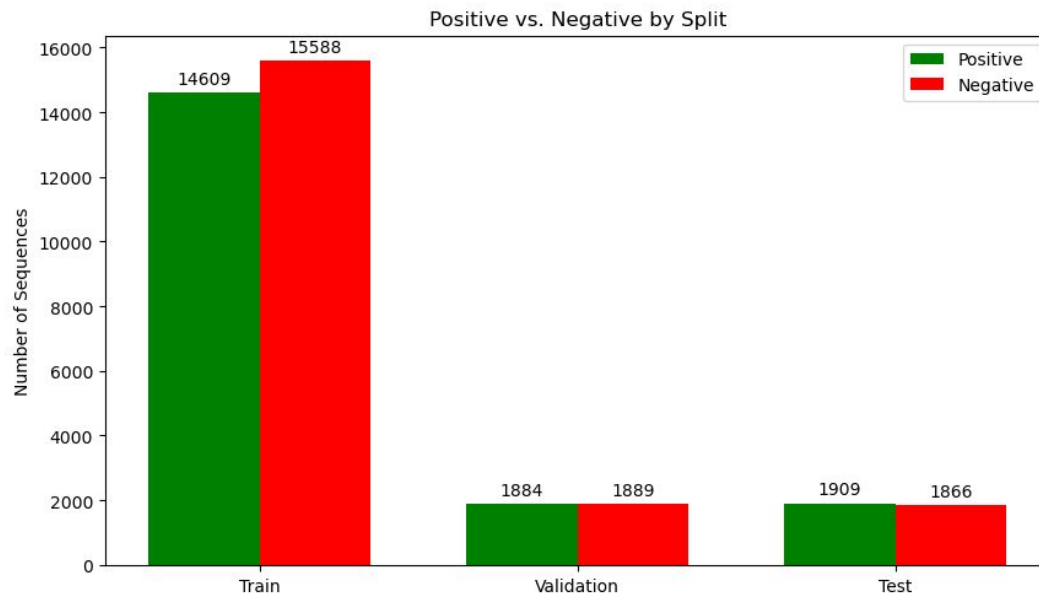


Dataset split

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.

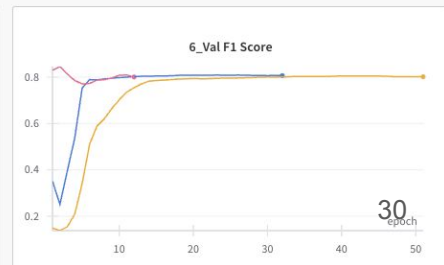
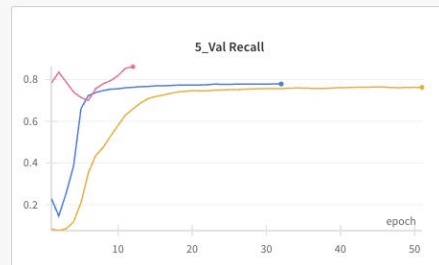
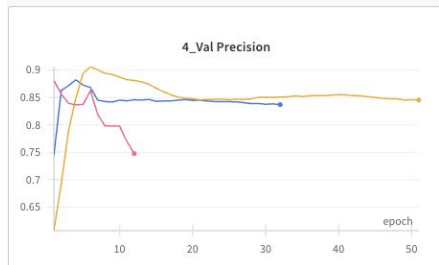
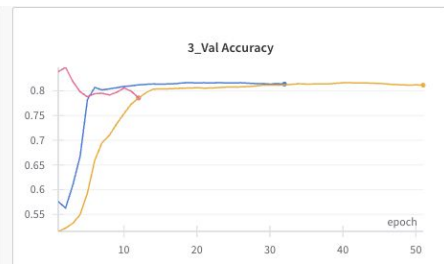
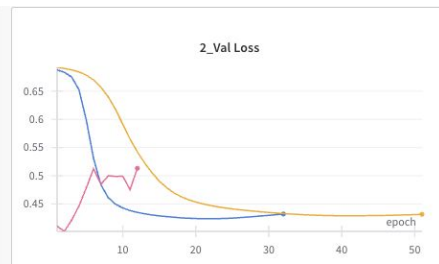
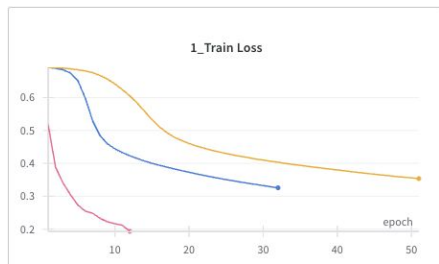
Training - Parameter search

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

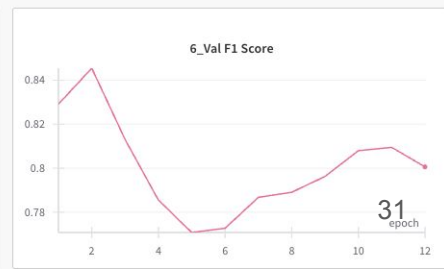
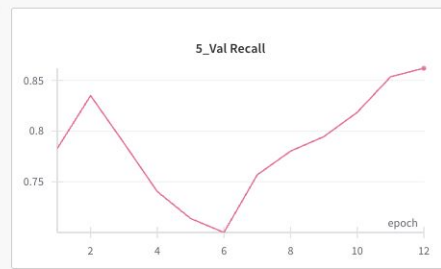
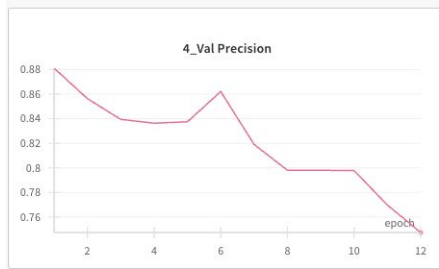
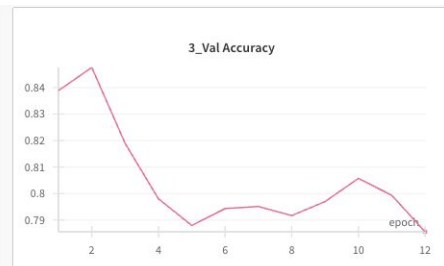
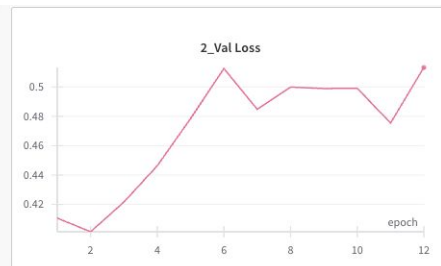
Model: Training ProtBERT model with adapters.

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

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

Best performing model

dropout_p: [0.3, **0.5**]



Training - Parameter search

Model: Training ProtBERT model with adapters.

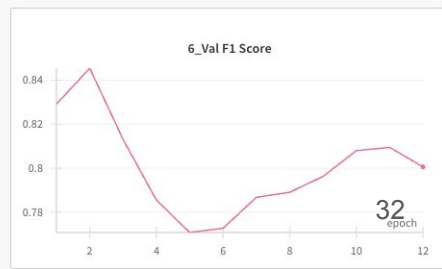
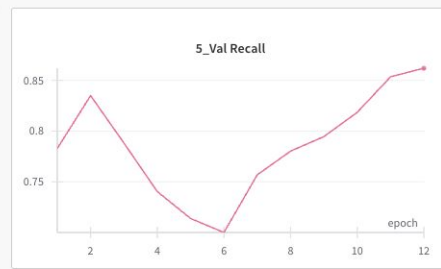
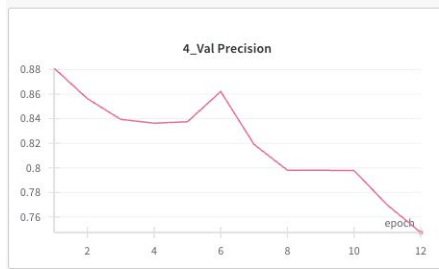
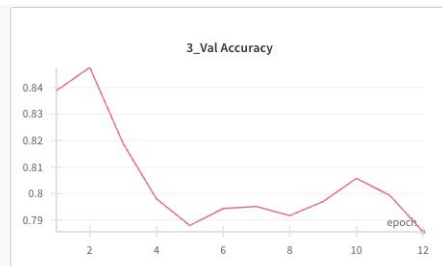
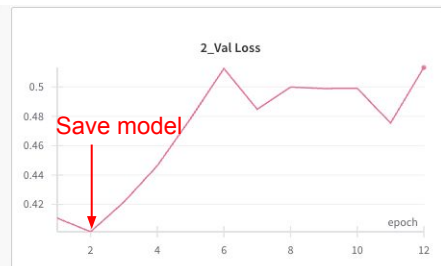
learning_rate: [1e-4, **5e-4**, 1e-5, 5e-6, 1e-6]

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

dropout_p: [0.3, **0.5**]

Best performing model

After 2 epochs, model overfits!



Evaluation

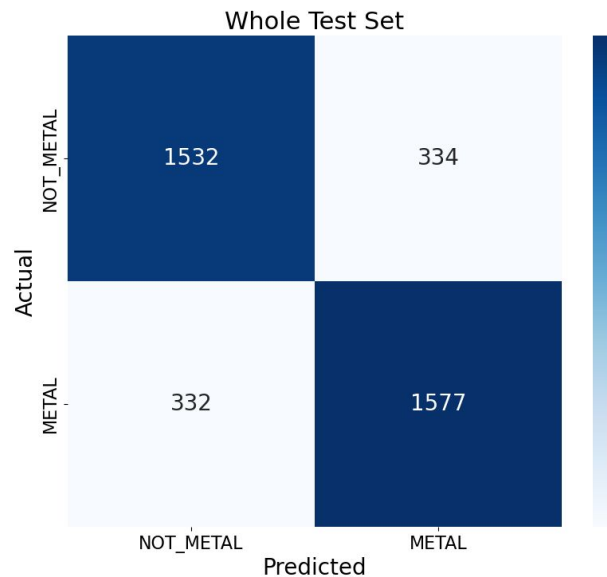
Metrics:

Accuracy: 82.36%

Precision: 82.52%

Recall: 82.61%

F1 Score: 82.57%



Evaluation - only on real sequences

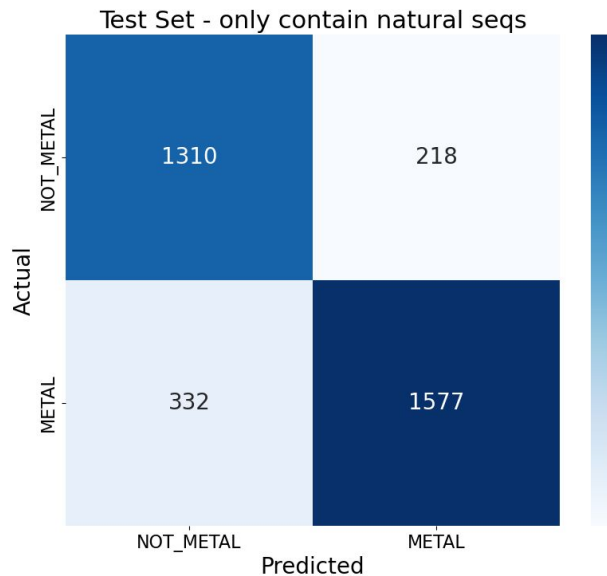
Metrics (excluding ProteinMPNN redesigned sequences):

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!



