

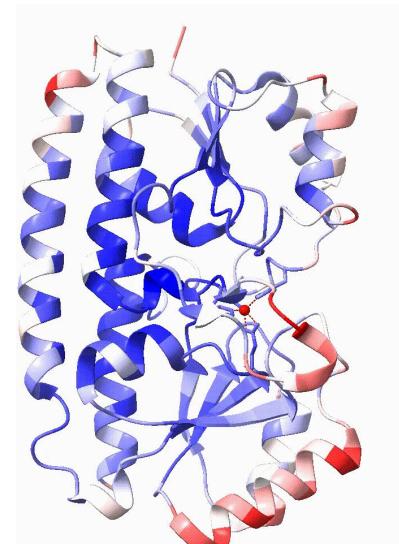
Advancing metal-binding protein predictions with deep learning

Jingkai LAN

Supervisor: Thomas Lemmin
Co-supervisor: Giulia Peteani

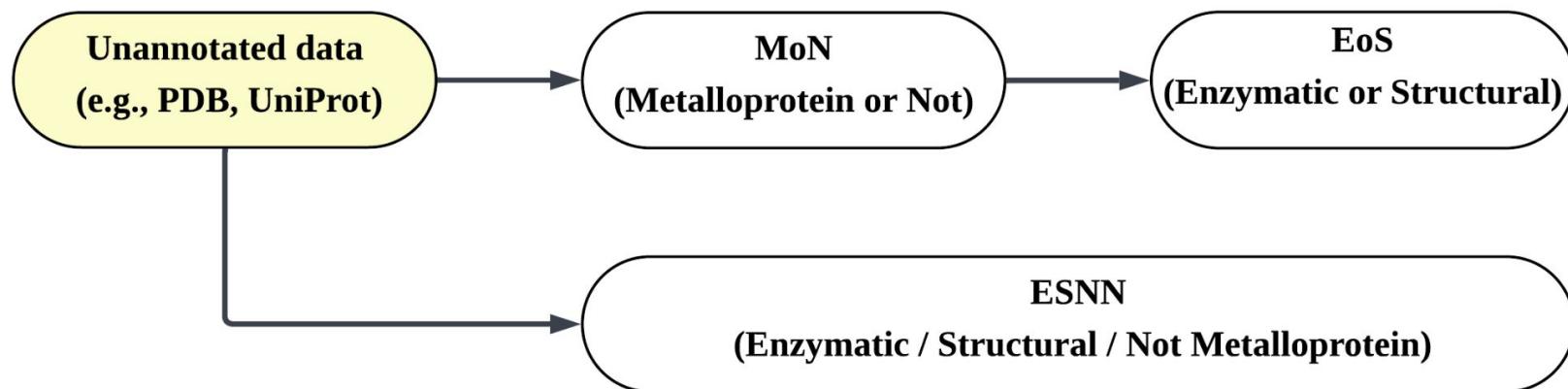
Introduction

- ❖ **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.



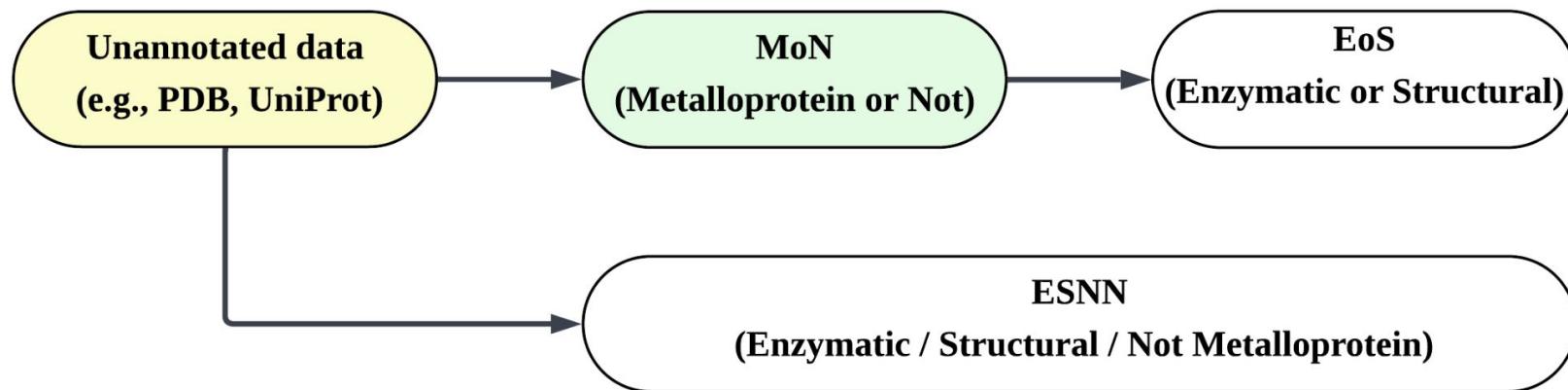
Project goal

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



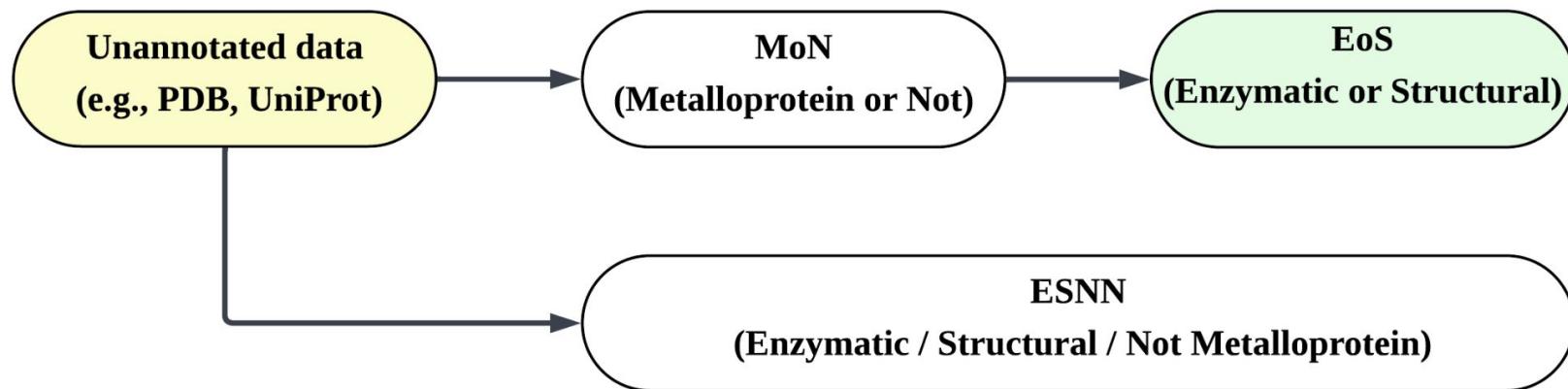
Project goal

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



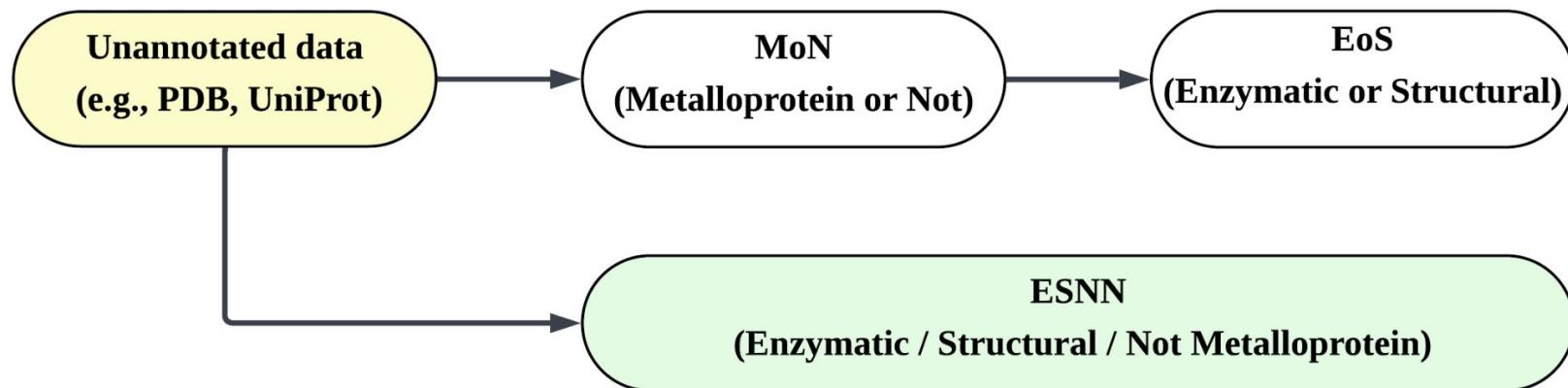
Project goal

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

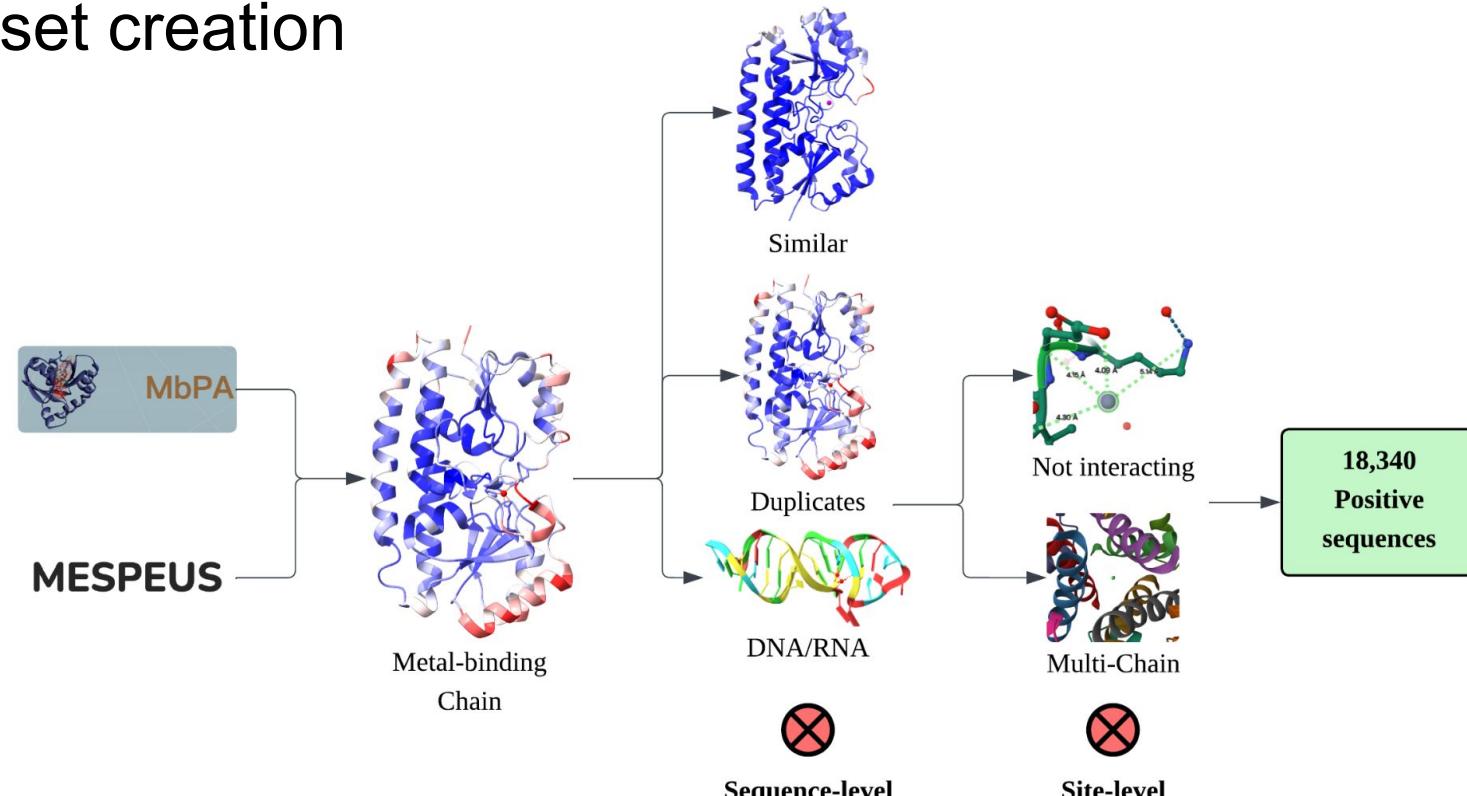


Project goal

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

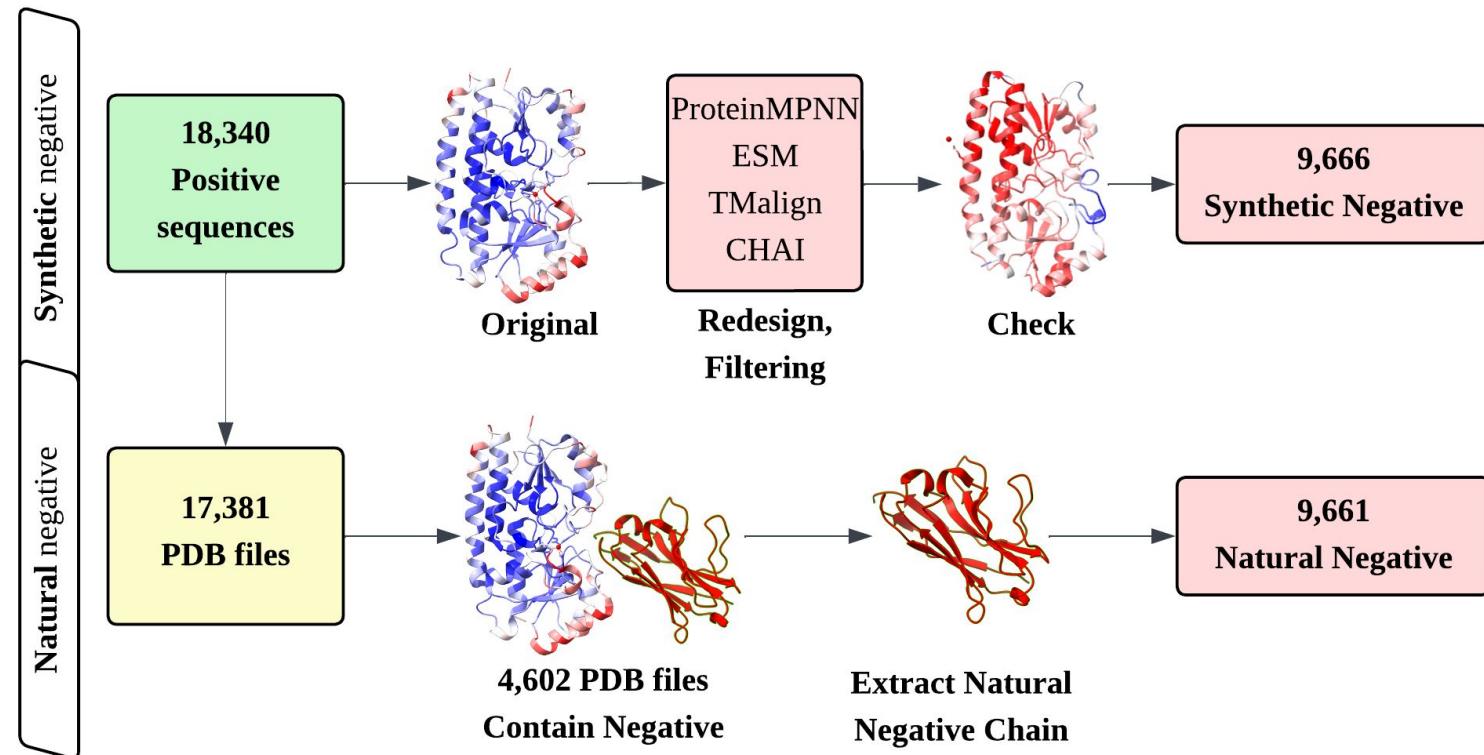


Dataset creation

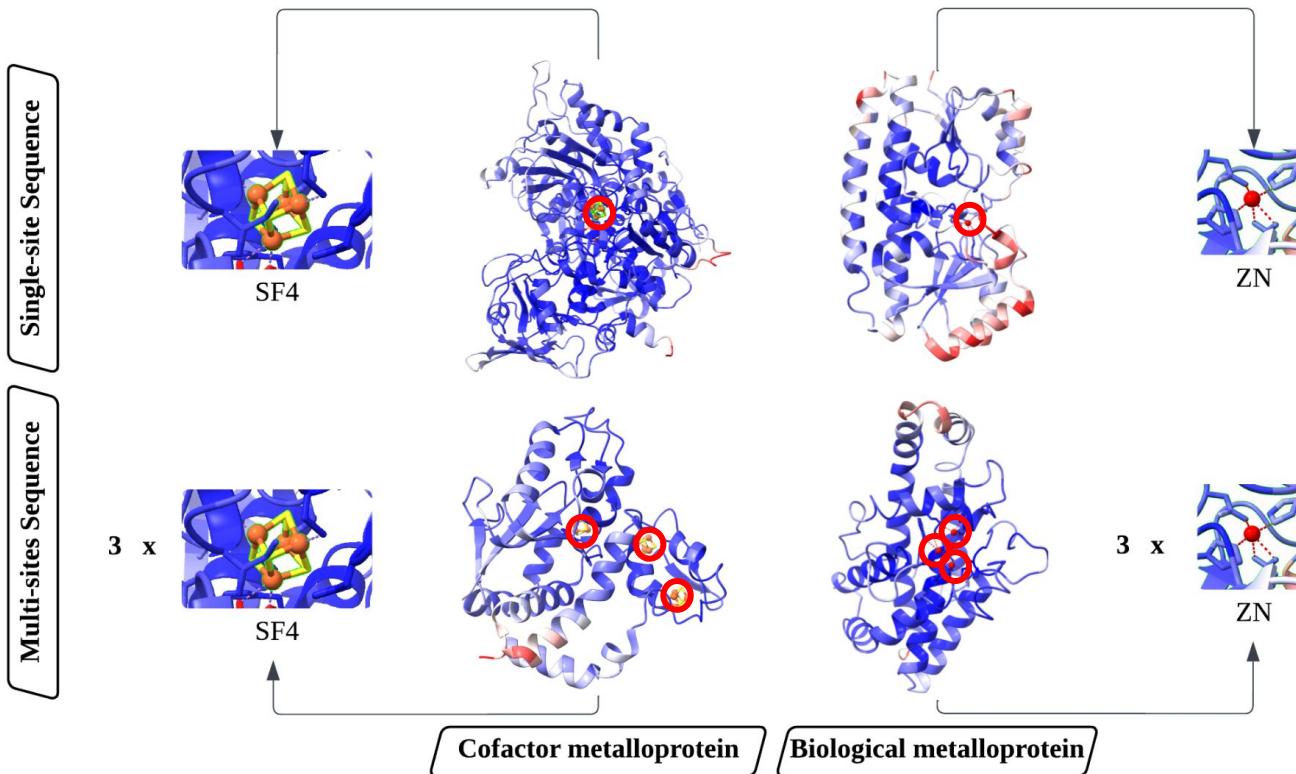


Database	Extract target PDB Chain	Quality Control	Positive set
----------	--------------------------	-----------------	--------------

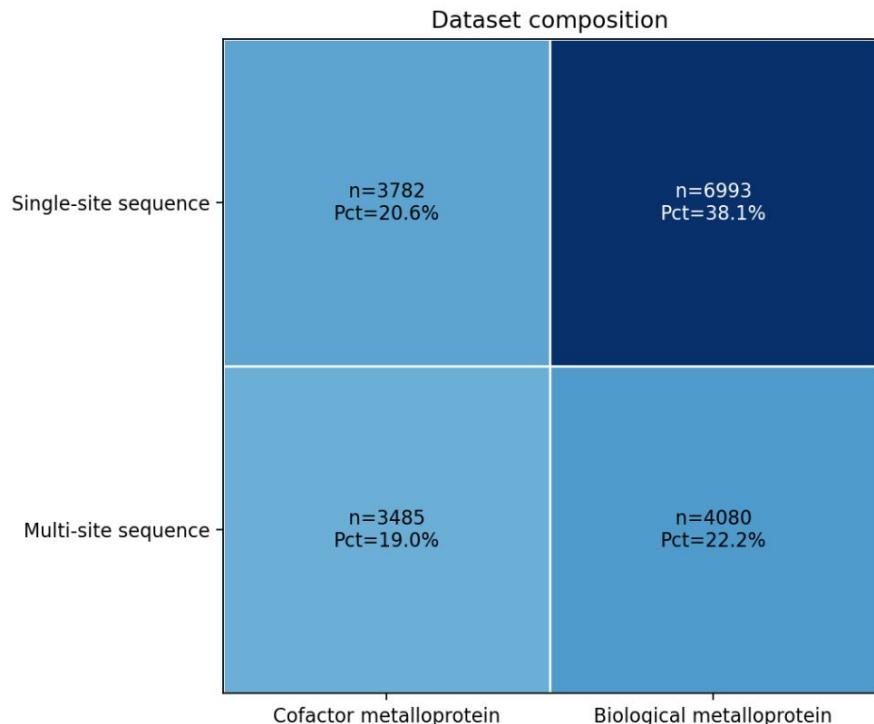
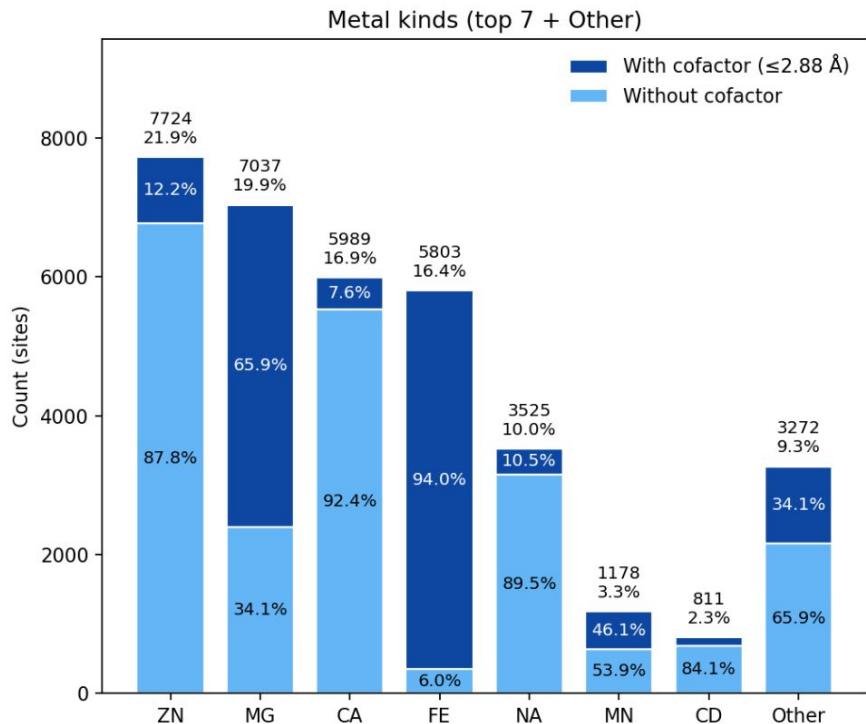
Dataset creation



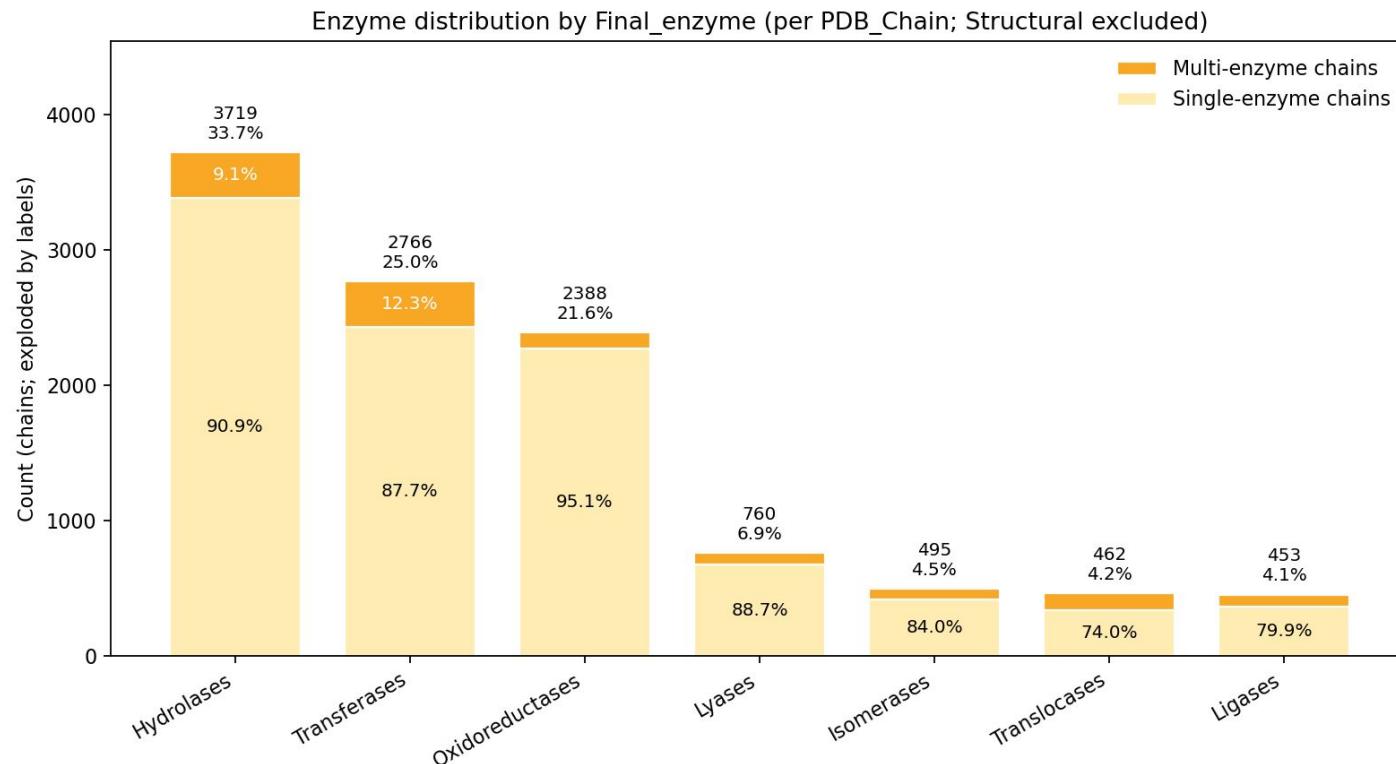
Dataset distribution



Dataset distribution

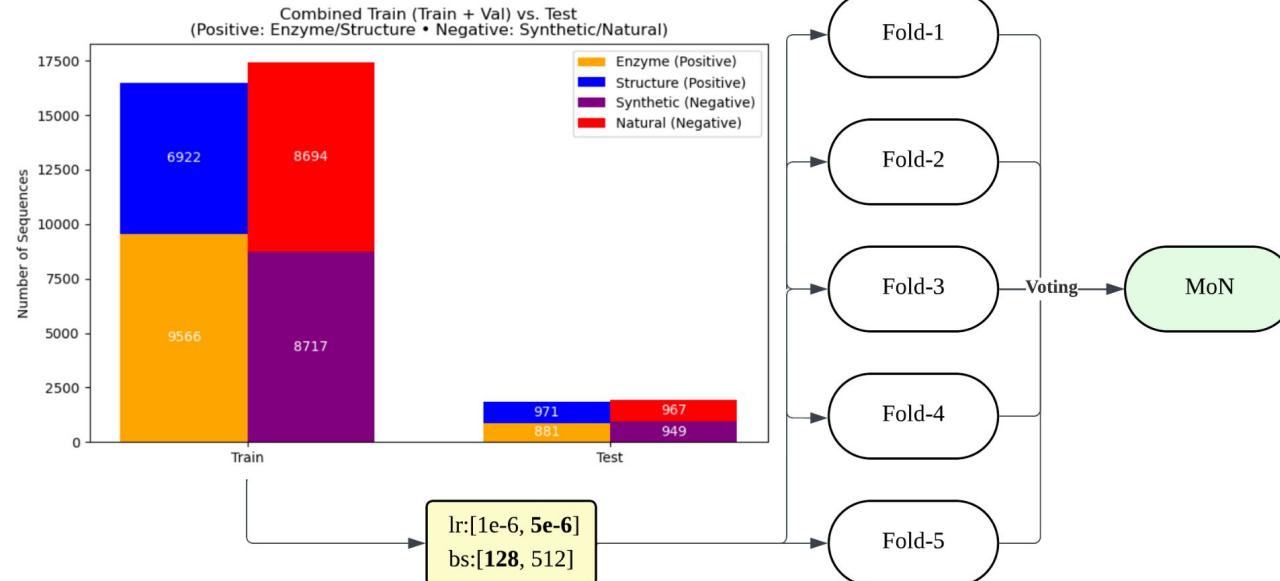


Dataset distribution



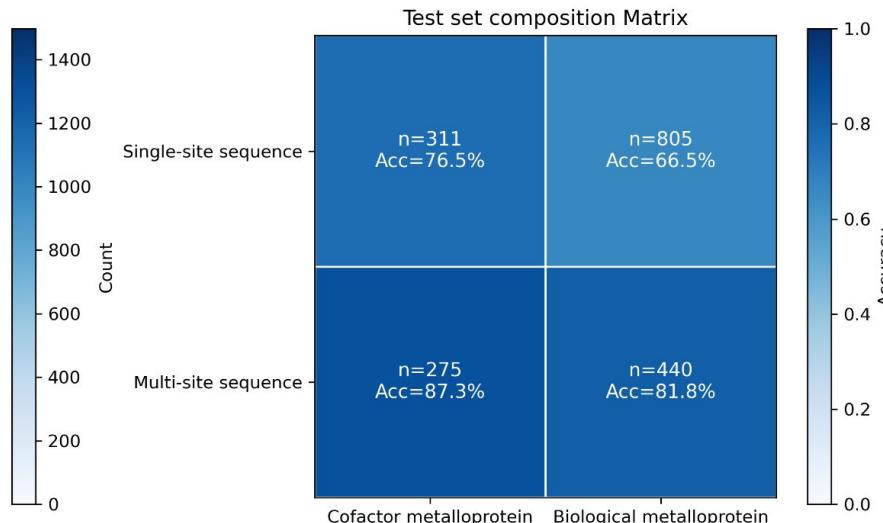
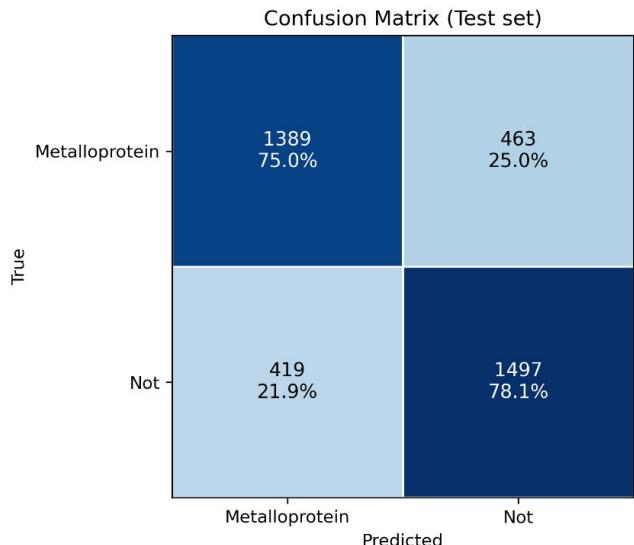
MoN - Metalloprotein or Not

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



MoN - Metalloprotein 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%

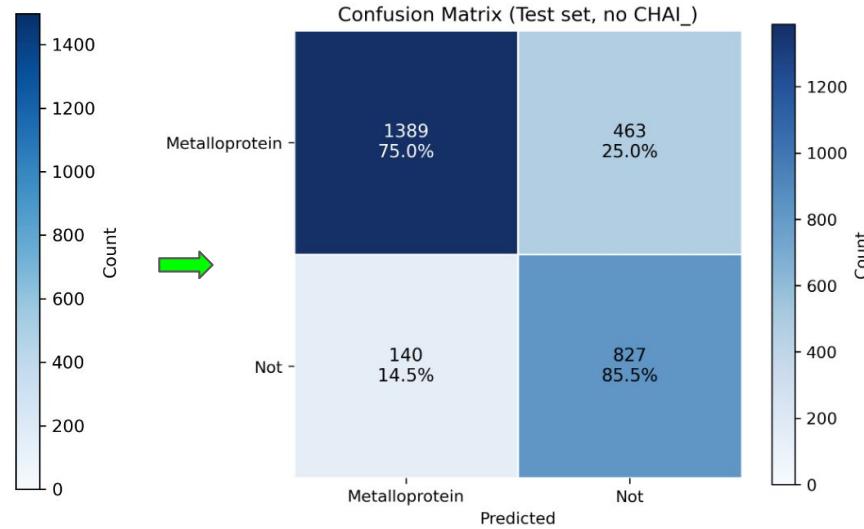
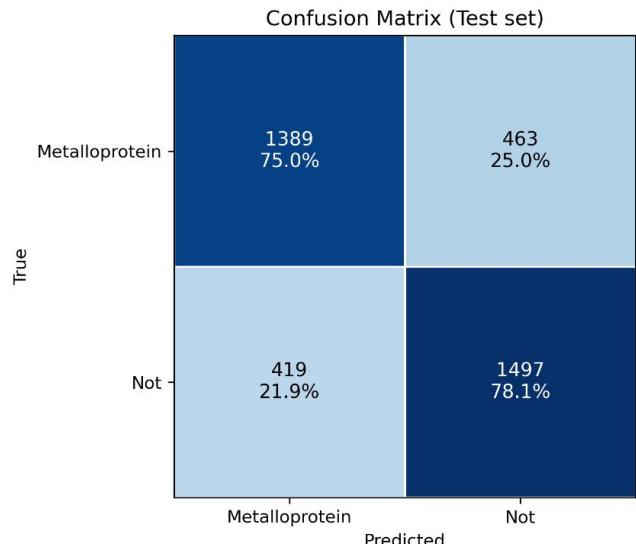


MoN - Metalloprotein or Not

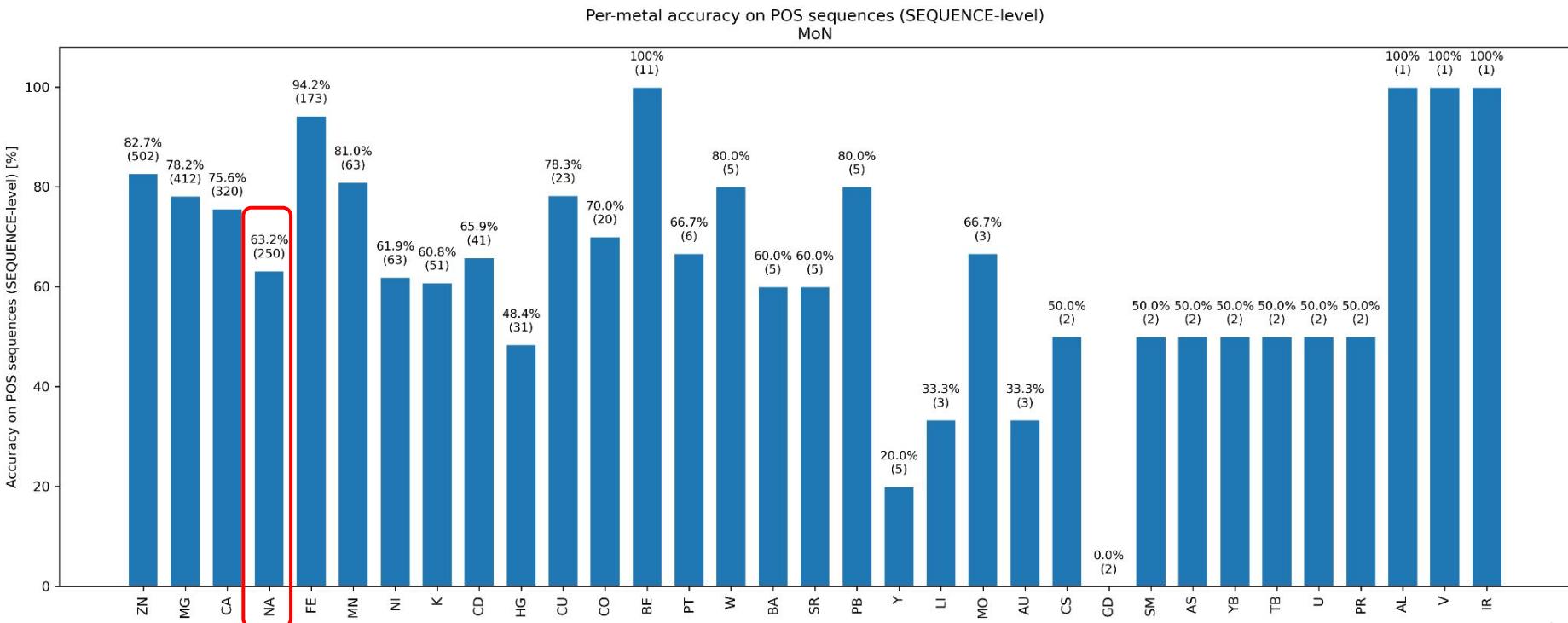
If we evaluate the model only on the test set that contains **Natural Negatives**:

Model	Accuracy	Precision	Recall	F1 Score
MoN (Voting)	76.59%	76.83%	75.00%	75.90%

MoN (Natural test) $+ 2.02\% = 78.61\%$ $+14.01\% = 90.84\%$ $+0\% = 75.00\%$ $+6.27\% = 82.17\%$

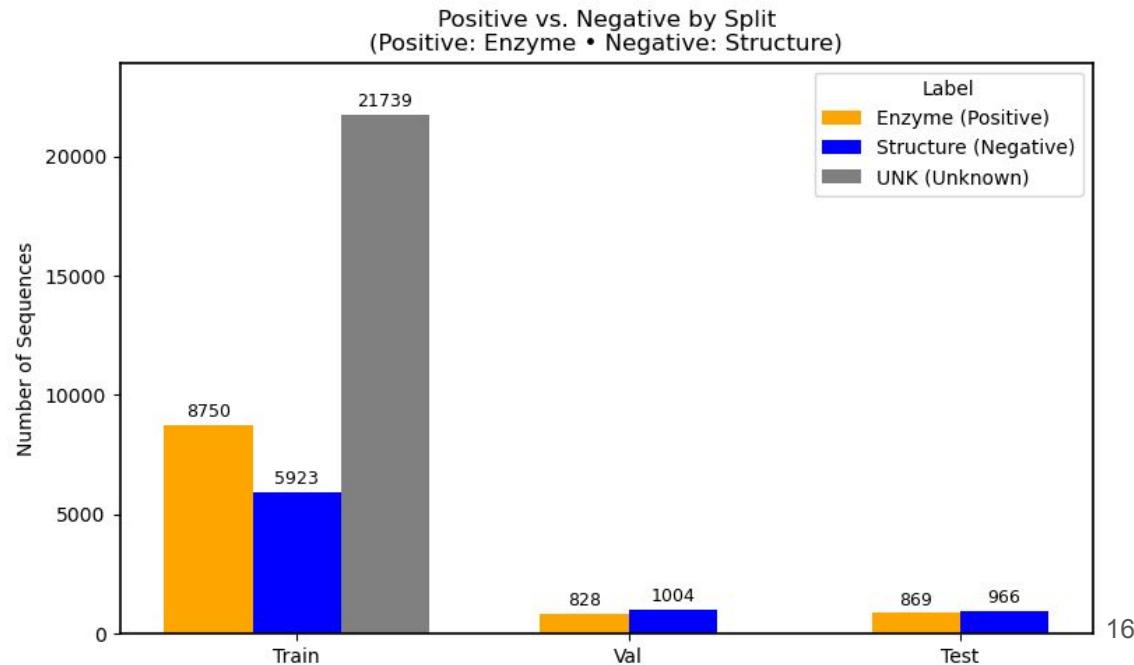
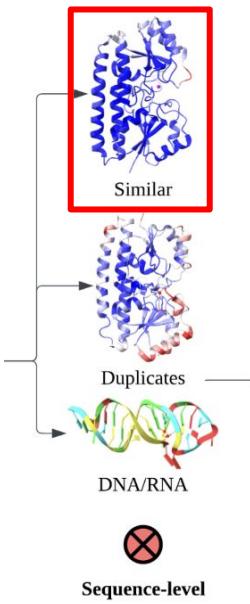


MoN - Metalloprotein or Not

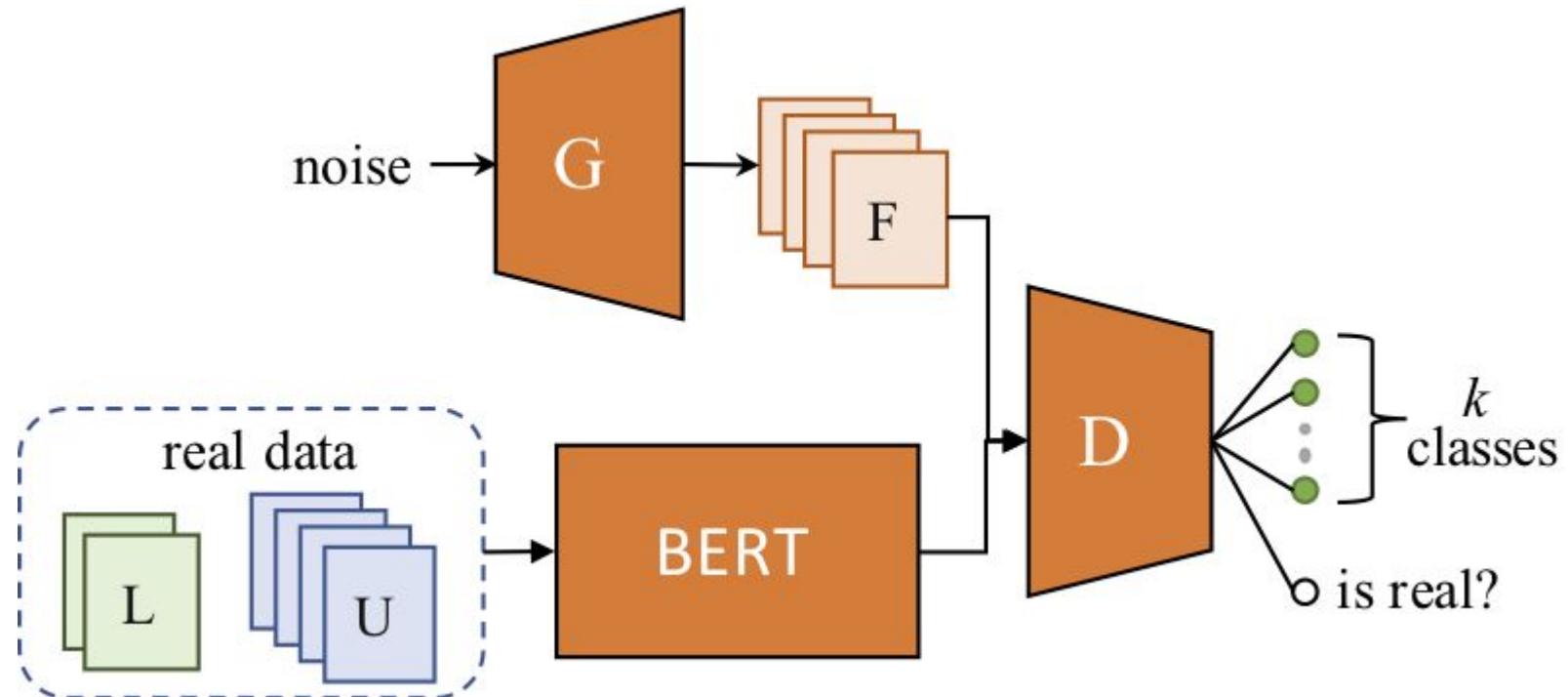


EoS - Enzymatic or Structural

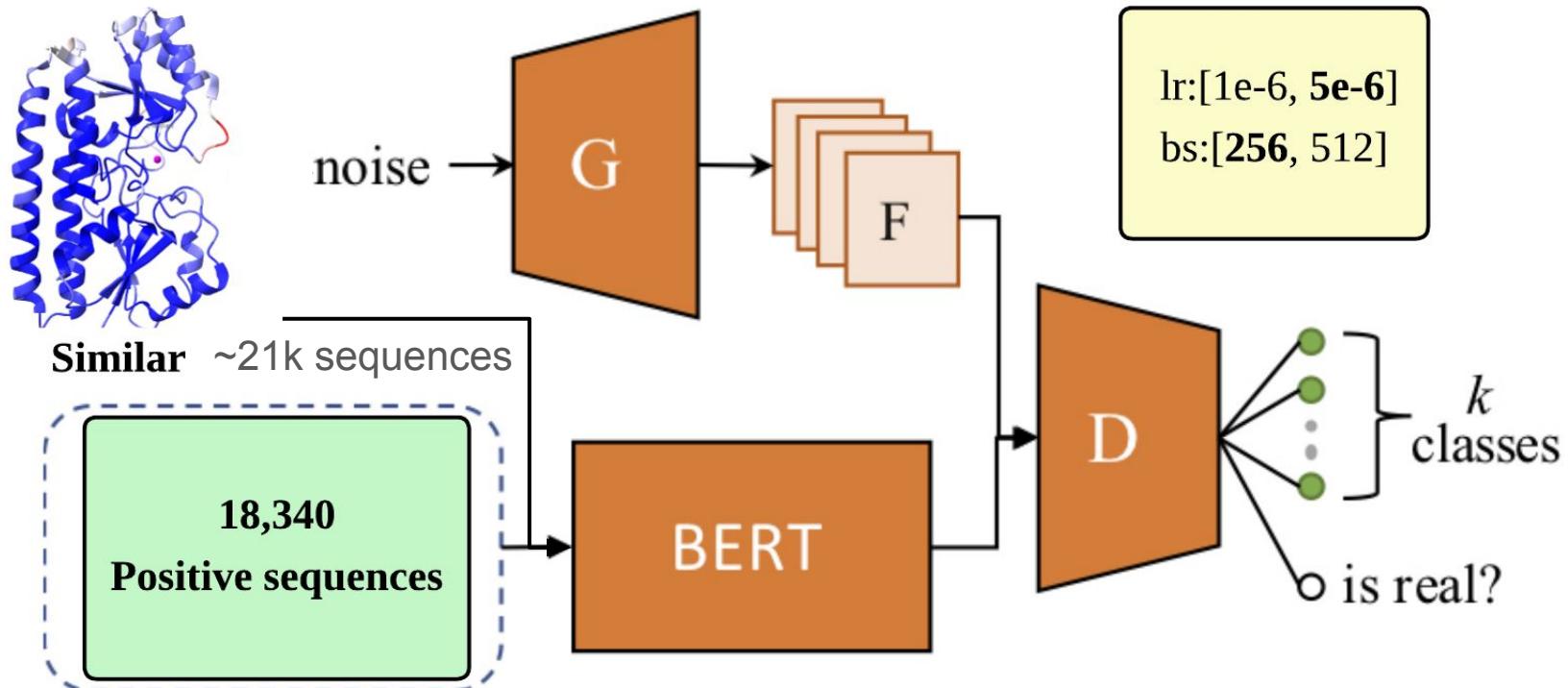
- ❖ MMseq2 Positive set at 30 pident (8:1:1)
- ❖ Using the previously discarded similar sequences as unlabeled data.



EoS - Enzymatic or Structural



EoS - Enzymatic or Structural



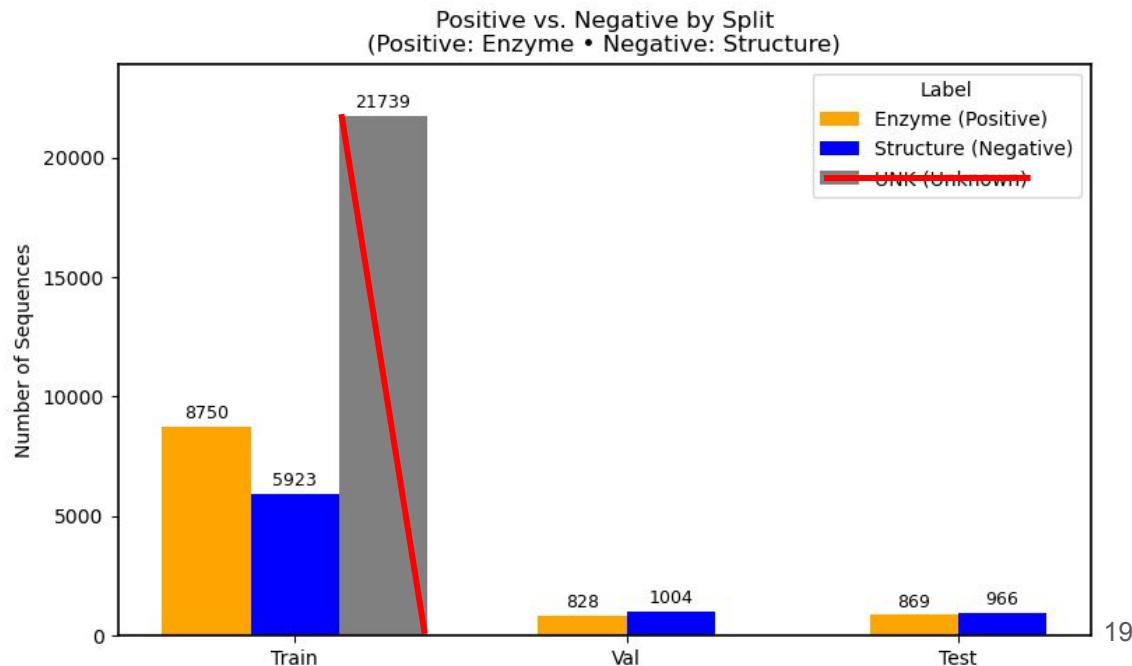
EoS - Enzymatic or Structural

- ❖ MMseq2 Positive set at 30 pident (8:1:1) - same as before
- ❖ Using **class weights** to handle **class imbalance**.
- ❖ Remove unlabeled data.

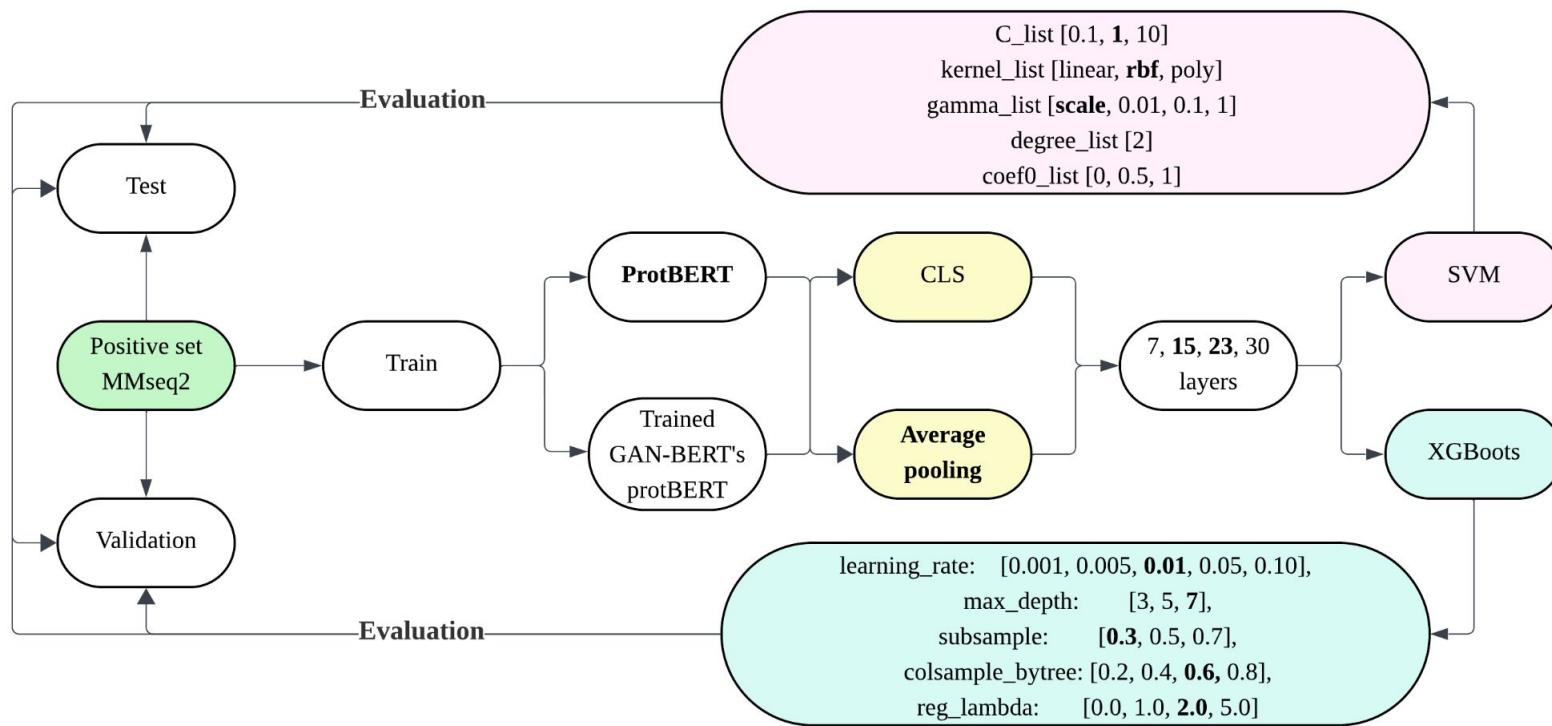
- ❖ $w_c = \frac{N}{K \times n_c}$

Enzymatic weight = 0.8384

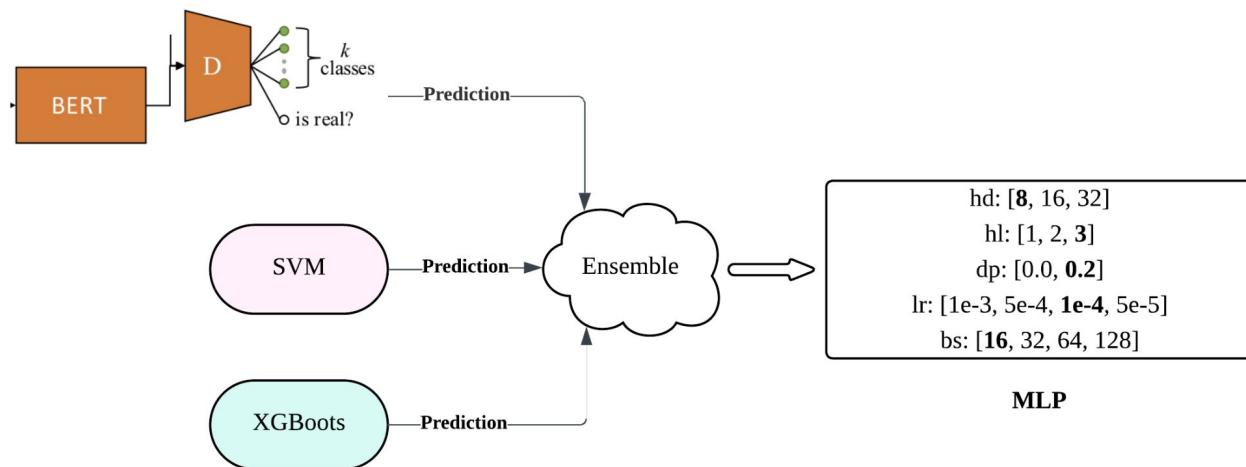
Structural weight = 1.2386



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

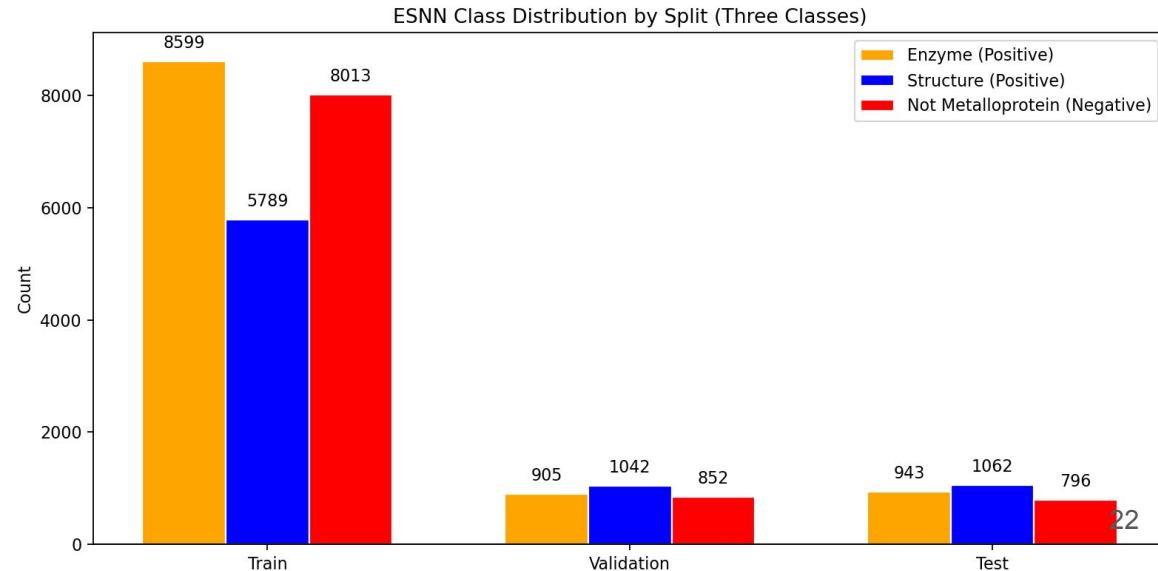
- ❖ Merge Positive and Natural Negative → MMseq2 at 30 pident
- ❖ Using class weights to handle class imbalance.

$$w_c = \frac{N}{K \times n_c}$$

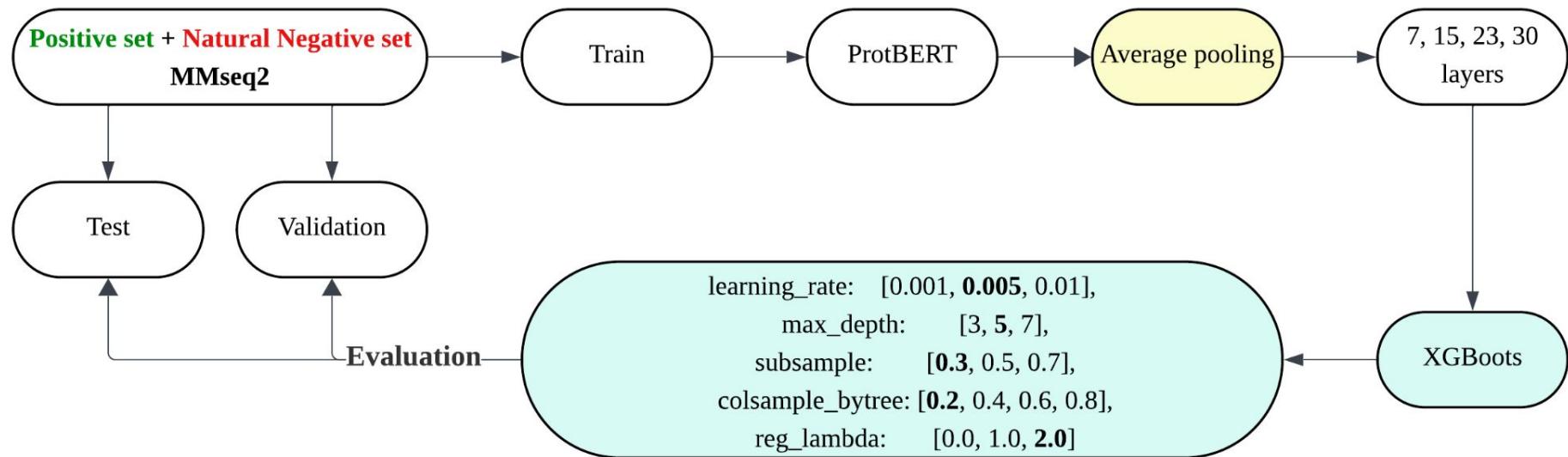
Enzymatic weight = 0.8683

Structural weight = 1.2899

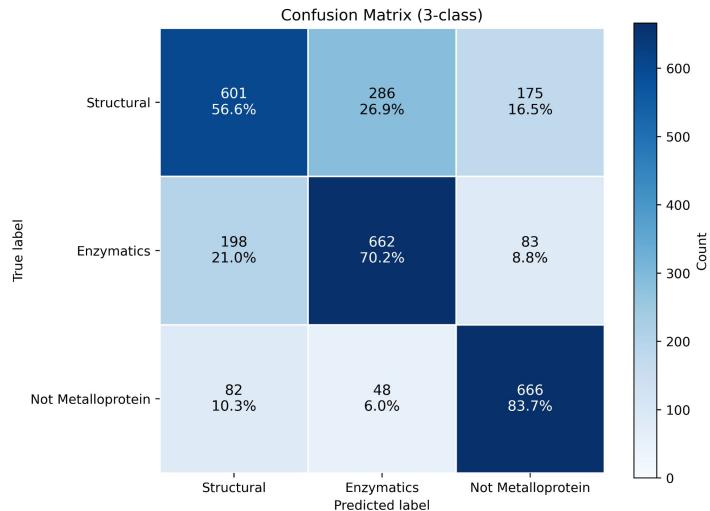
Natural Neg weight = 0.9319



ESNN - Enzymatic, Structural or Natural Negative



ESNN - Enzymatic, Structural or Natural Negative



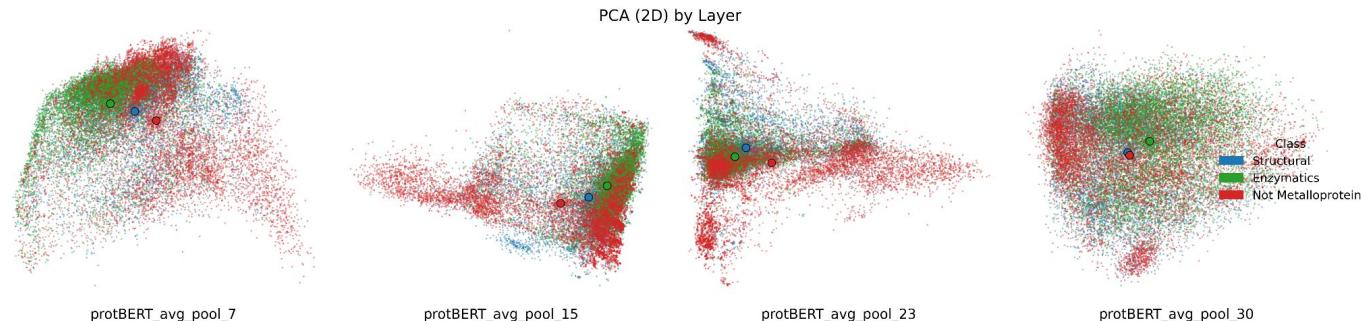
3-class performance

Accuracy: 68.87%

Precision: 68.92%

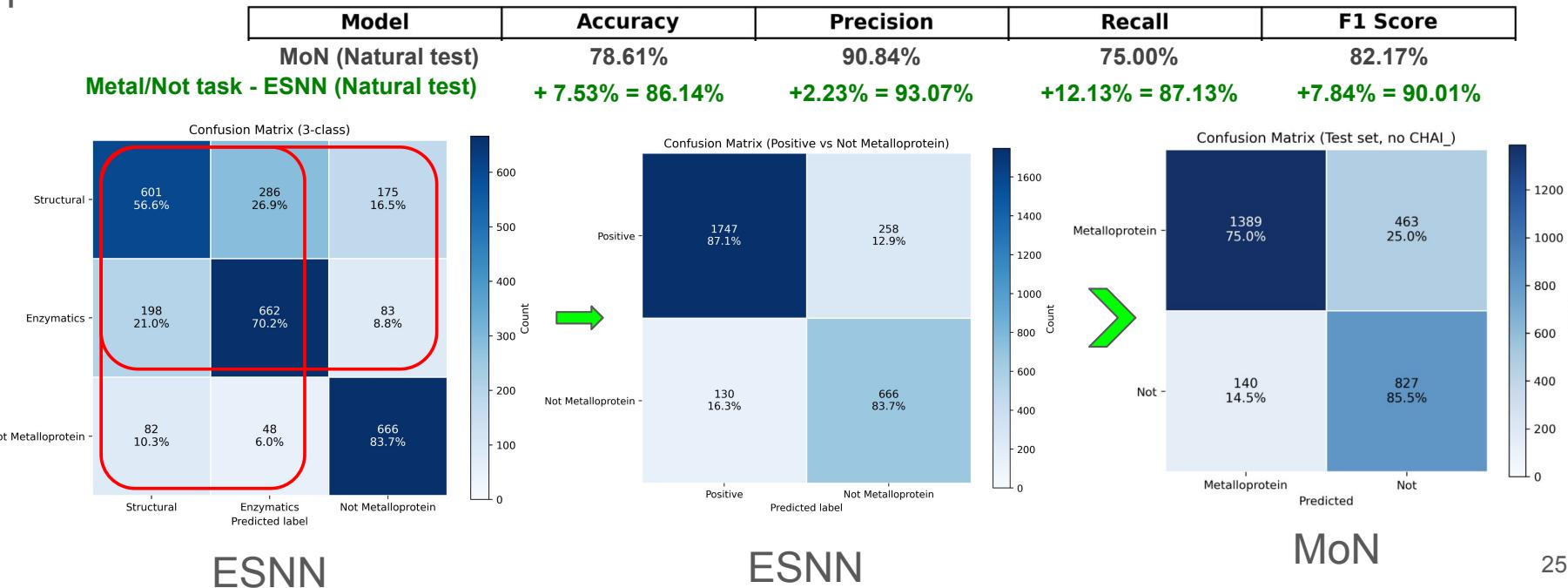
Recall: 70.15%

F1 score: 69.20%

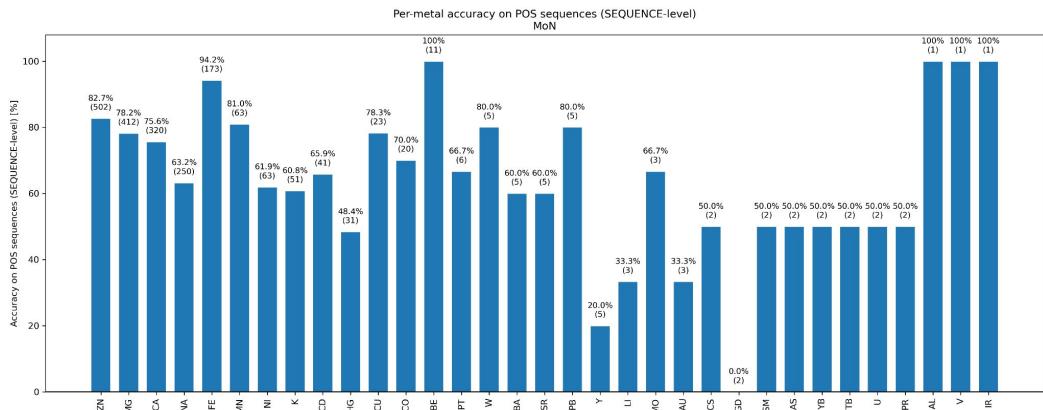
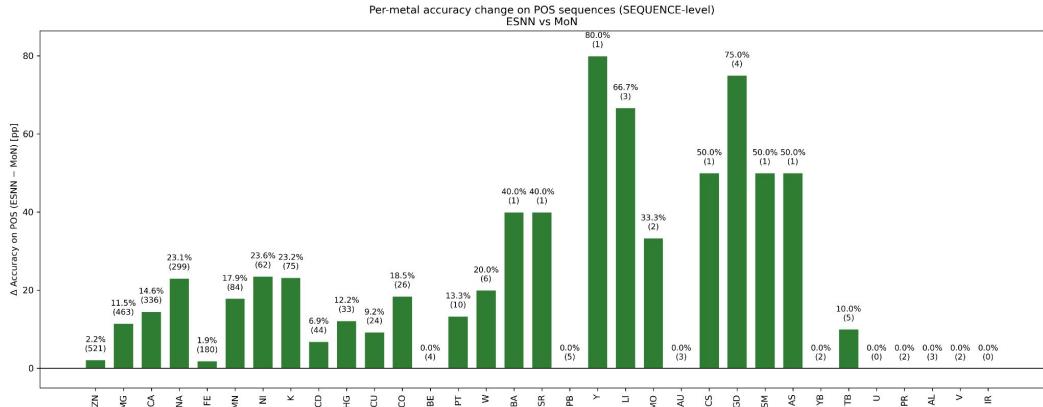


ESNN - Enzymatic, Structural or Natural Negative

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



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!