University of Bern

Interfaculty Bioinformatics Unit

Advancing metal-binding protein predictions with deep learning

LAN Jingkai

Master of Science in Bioinformatics and Computational Biology

Supervisor: Prof. Thomas Lemmin

Co-supervisor: Giulia Peteani

August 2025

Abstract

While indispensable for diverse biological activities, metalloproteins are still under-annotated in large-scale protein databases. To improve sequence-based prediction, we integrated two public metalloprotein databases, MESPEUS and MbPA. Covering 44 metal types, and constructed a high-quality positive set after redundancy reduction and quality control. Two types of negative sets were also generated: (i) synthetic sequences obtained by redesigning residues within 5 Å of binding sites using ProteinMPNN, followed by verification and retention of successfully mutated sequences, and (ii) natural non-binding chains extracted from the same complexes. Based on these datasets, we developed three sequence-only models: MoN (distinguishing metalloproteins from non-metalloproteins), EoS (distinguishing enzymatic from structural metalloproteins), and ESNN (three-class classification). Using ProtBERT as the backbone, we explored different strategies for different tasks, including adapter fine-tuning, GANBERT semi-supervision, and embedding-based machine learning approaches (SVM, XGBoost) combined with lightweight ensembles. Results showed that ESNN achieved overall better performance than MoN and stand-alone binary classifiers on natural sequences, reaching 86.2% accuracy and 90.0% F1 score (vs. 78.6% and 82.2% for MoN), with particularly strong improvements for common ions such as Mg, Ca, and Na. The ensemble approach in EoS also outperformed the best single model (embedding + XGBoost), with 74.1% accuracy (vs. 73.2%) and 72.9% F1 score (vs. 72.8%). Overall, this study presents a sequence-only predictive framework that can support large-scale metalloprotein prediction and enzymatic annotation.

Contents

1 Introduction

Metalloproteins play crucial roles in biological processes, participating in catalysis, structural stability, and regulation, and are therefore essential for understanding cellular mechanisms [1]. Although metal-binding proteins account for approximately one-third of all proteins in nature, their annotation in large-scale protein databases remains incomplete due to the high cost and time required for experimental validation. As a result, the proportion of annotated metal-binding proteins in databases such as RCSB PDB (https://www.rcsb.org/) and UniProt (https://www.uniprot.org/) is significantly underrepresented [2–4]. Furthermore, about half of all enzymes are metalloenzymes, with their metal-binding sites contributing directly to catalytic activity [5]. To advance the study of metal-binding proteins and their catalytic functions, improving annotation efficiency in large protein databases and related high-throughput proteomic studies is of critical importance.

In the research on metalloprotein prediction, many mature models have already been developed. Typically, these models are trained by annotating the metal-binding sites of metalloproteins. For studies that employ three-dimensional structural information, the main approach is to annotate the spatial positions of metal ions within proteins. For example, the Metal3D method constructs a local cubic box of 16 Å in length centered on the Cα atom of a residue, voxelizes the atomic environment within the box, and thereby generates three-dimensional grid representations with annotated metal ion positions [6]. Similarly, Mohamadi et al. proposed an ensemble 3D-CNN model that also relies on voxelized local 3D environments, incorporating multiple biophysical features to predict a variety of metal ion binding sites [7].

In contrast, sequence-based prediction of metal-binding proteins currently lacks effective metal-binding site annotation strategies and instead mainly relies on feature engineering from the protein sequence for model training. For instance, MEBIPred predicts metal-binding proteins using 220 features, including amino acid composition and physicochemical properties. Kumar et al. proposed a method based on reduced amino acid alphabets combined with a random forest classifier, in which sequence patterns within the local neighborhood of a residue are extracted to predict metal-binding sites [8]. The M-Ionic model, in turn, leverages residue embeddings derived directly from sequences as predictive features for metal-binding site identification [9].

Overall, machine learning and deep learning models have demonstrated strong performance in the prediction of metal-binding proteins, ranging from approaches based on three-dimensional voxelized structural representations to those relying on sequence-derived features. For the prediction of catalytic activity in metalloproteins, research has mainly focused on the MAHOMES model, which predicts the catalytic potential of metal-binding sites by extracting and analyzing site-specific 3D structural features. However, sequence-only models for catalytic activity prediction are still lacking.

In this study, we constructed a high-quality dataset of metal-binding proteins for subsequent model training. By integrating sequence information from two large-scale metal-binding protein

databases, applying clustering and other redundancy-reduction strategies, and filtering out low-quality binding sites, we established a reliable positive set of metal-binding proteins [10, 11]. Unlike previous studies, which typically focused on only a limited number of common metal ions, our dataset encompasses nearly all available metal-binding proteins and covers 44 distinct metal ion types. Subsequently, based on the high-quality positive set of metal-binding proteins, this study innovatively created a synthetic negative set by redesigning the amino acids surrounding binding sites, thereby mutating the metal-binding residues so that metal ions could no longer bind to those sites. This strategy ensures that the only difference between positive and negative sequences lies at the binding site itself, strengthening the model's focus on metal-binding regions during training. In parallel, we constructed a natural negative set by extracting the non-binding regions of heteromeric metal-binding protein chains. This not only balanced the ratio of positive and negative samples in the dataset but also provided a reliable means for evaluating model performance on authentic protein sequences.

Building upon this dataset, we aimed to achieve accurate sequence-only predictions of both metal-binding capacity and enzymatic activity using deep learning approaches. This framework simplifies the prediction task, streamlines the workflow, and enhances the feasibility of large-scale functional annotation. Specifically, we first developed the MoN (Metalloprotein or Not) model for metal-binding protein prediction, which requires no feature engineering and relies solely on amino acid sequences as input. Next, leveraging enzymatic annotation data, we built the EoS (Enzymatic or Structural) model to perform binary classification of enzymatic versus structural metal-binding proteins. Finally, we trained a three-class model, ESNN (Enzymatic or Structural or Natural Negative), which simultaneously predicts both metal-binding and enzymatic properties.

For model training, we primarily used the pretrained language model (PLM) ProtBERT (https://huggingface.co/Rostlab/prot_bert), which was trained on large-scale protein data using masked language modeling (MLM) and has learned broad protein-related knowledge and features [12]. All of our models were built on this pretrained model and further trained using various approaches. For instance, the adapter method was applied, which enables PLMs to be adapted to new tasks by training only a small number of additional parameters, thereby improving training efficiency [13]. We also experimented with GANBERT, a semi-supervised approach that combines supervised fine-tuning of BERT modules with an unsupervised generator to introduce noise, which is then used in adversarial training via a classifier [14]. This method has the advantage of leveraging unlabeled data more effectively to improve performance. Given the high redundancy in metal-protein datasets, GANBERT was particularly suitable for exploiting redundant data to generate noise and augment the training set. Finally, we extracted embeddings from different ProtBERT layers and trained machine learning models on these representations, which in some cases outperformed deep learning models.

Based on these deep learning strategies, we ultimately trained three models to perform distinct classification tasks. By relying exclusively on sequence data as input, we simplified the

prediction workflow while expanding the range of metal ion types considered, yet still achieved excellent predictive performance.

2 Methods

2.1 Data Sources

All PDB IDs of metal-binding proteins were retrieved from the publicly available metalloprotein databases MESPEUS (https://mespeus.nchu.edu.tw/) and MbPA (http://bioinfor.imu.edu.cn/mbpa). Duplicate IDs were removed, and the corresponding PDB files were downloaded. Based on the "REMARK 620" section (information on metal coordination) in the PDB files, the specific metal-binding chain information, metal ion details, and sequences were obtained. Sequences containing more than 10% non-proteinogenic amino acids, as well as DNA/RNA sequences involved in metal binding from complex proteins, were excluded. Furthermore, entries with identical sequences were removed.

2.2 Construction of the Positive Set

During natural evolution, small-scale genetic variations (such as point mutations or amino acid substitutions) generate many protein sequences that differ only slightly but are overall highly similar [15]. To further remove redundancy in the dataset, we used the MMseqs2 tool to cluster all metal-binding proteins at 90% sequence similarity, retaining only the representative center sequence of each cluster and set aside the rest as redundant sequences [16].

Subsequently, we used ProDy to identify and extract spherical binding sites of all sequences at radii of 2.88 Å and 5 Å, generating a table of metal-binding site information (S1) [17]. Based on this information, we discarded sequences in which all metal-binding sites either had no residues within 2.88 Å of the metal ion or involved multi-chain coordination, thereby removing low-quality metal-binding sequences. In addition, to ensure that all metal-binding sites could be completely and accurately fed into the model for training, we excluded sequences longer than 2048 amino acids. After this processing, we obtained the positive set used in this study.

2.3 Enzyme Information Annotation

Based on the sequence information in our final positive set, we retrieved and annotated enzyme information using the BRENDA (`https://www.brenda-enzymes.org/`) and PDB enzyme databases (`https://www.rcsb.org/stats/explore/enzyme_classification_name`) [18]. If both BRENDA and PDB provided enzyme information for a positive sequence, the annotation from BRENDA was used as the representative. If only PDB contained enzyme information, the PDB annotation was adopted. If neither database contained enzyme information, the sequence was labeled as Structural.

2.4 Construction of the Negative Set

2.4.1 Construction of the Synthetic Negative Set

To enable the model to focus on learning the core feature of whether a protein binds metal ions and thereby enhance its discriminative ability, we employed an artificial synthesis strategy to construct part of the negative set. ProteinMPNN was used to redesign the amino acids located within 5 Å of the metal-binding sites in all sequences [19]. The parameters were set as follows: (1) Temperature = 1.0; (2) to balance the redesign rate across different amino acid types, the *bias* parameter "-2 -1 -1 -1" was applied to residues "C D E H". For each positive sequence, 10 redesigned sequences were generated, and only the sequence with the highest degree of redesign was retained as the representative. To maximize the redesign success rate for metal-binding sites, we used a relatively high temperature for sequence synthesis. During this process, many redesigned sequences showed substantial deviations from their original positive counterparts, so further filtering was applied.

First, we generated structural models of all synthetic sequences using ESM and calculated their average pLDDT scores [20]. Sequences with pLDDT < 60, indicating low confidence, were removed. Then, TMalign was used to compare the ESM-predicted structures with their corresponding original positive structures, and RMSD as well as TM-score were computed [21]. Sequences with RMSD < 3.0 and TM-score > 0.5 were retained, as they remained structurally similar to the original after redesign at the binding sites. Finally, all remaining sequences, together with their corresponding metal ions, were input into CHAI1 simulations to assess whether the positions of the metal ions shifted compared with the original structures, or whether the ions no longer interacted with the binding chains (distance > 2.88 Å) [22]. If all metal-binding sites in a redesigned sequence were rendered non-functional, the redesign was considered successful. These sequences were retained as our final synthetic negative set.

2.4.2 Construction of the Natural Negative Set

Since the synthetic negative set underwent multiple filtering steps, its final size did not match that of the positive set. To address this, we constructed a natural negative set following the methodology of MEBIPred. Specifically, we identified all heteromeric proteins present in the positive set and selected the non-positive chains as natural negative sequences. After removing duplicates, DNA/RNA sequences, and sequences composed entirely of non-proteinogenic amino acids, we obtained our natural negative set.

2.5 MoN Setting: Metalloprotein or Not

To train a model for classifying sequences as either metal-binding or non-metal-binding, we performed a specific partitioning of the datasets described above. Since the sequences in the synthetic negative dataset were obtained by redesigning those from the positive dataset, they

are highly similar to their positive counterparts, with only minor differences. Therefore, we first merged the positive set with the synthetic negative set and clustered them using MMseqs2 at 30% sequence similarity. The clusters were then ordered by size (i.e., number of sequences per cluster), and the dataset was then split at the cluster level into overall training and testing sets at a 9:1 ratio, in descending order of cluster size. This ensured that the test set contained greater diversity, thereby improving the reliability of evaluation. The same procedure was applied to the natural negative set, and the partition results were then merged. The overall training set was further divided into 5 folds, with 20% of randomly selected sequences in each fold used as the validation set.

For each fold, we used ProtBERT as the backbone model and trained an Adapter module to perform classification. The maximum input length was set to 2048 amino acids, with a dropout probability of 0. A parameter search was conducted over the following combinations: learning rate = [1e-5, 5e-6] and batch size = [128, 512]. During training, early stopping was applied when no new minimum validation loss was observed within 10 epochs. After training, each fold was independently evaluated on the test set, and finally, the results from all five folds were combined using a majority voting strategy for comprehensive evaluation.

2.6 EoS Setting: Enzymatic or Structural

2.6.1 Dataset Split for EoS

In addition to the MoN model for distinguishing metalloproteins from non-metalloproteins, we further trained a binary classification model to predict whether a metalloprotein is enzymatic or structural. To construct this binary dataset, the positive set was clustered using MMseqs2. The clusters were ordered by size, and the dataset was split into training, validation, and test sets at an 8:1:1 ratio. Multiple approaches were explored for training this model.

2.6.2 GANBERT

First, we fine-tuned a GANBERT model with ProtBERT as the backbone. The maximum input length was set to 2048 amino acids, with a dropout probability of 0. The generator and discriminator shared the same learning rate, searched within [1e-6, 5e-6], and the batch size was searched within [256, 512]. Since GANBERT requires unlabeled data for generator training, We used the redundant sequences that were removed from the positive dataset by MMseqs2 clustering at 90% similarity as unlabeled data. These sequences are known metalloprotein sequences and are highly similar to our training data, making them particularly suitable for the generator to learn meaningful representations.

In GANBERT, the CLS embedding from the BERT module is used as input to the discriminator. We further attempted to use the 23rd layer instead of the default final layer (30th) for training on the same dataset. Due to limitations of time and computational resources, only one

exploratory trial was performed with maximum input length = 2048 aa, dropout probability = 0, learning rate = 1e-6, and batch size = 128.

We also tested replacing the BERT backbone with esm2_t33_650M_UR50D [23]. Owing to the input length limitation of ESM, sequences longer than 1024 aa were removed under the current split, and training was conducted with 1024 aa as the maximum input length. Again, only one parameter setting was tested: dropout probability = 0, learning rate = 1e-6, and batch size = 128.

2.6.3 Embedding + Machine Learning

In addition, we attempted an embedding + machine learning approach. For embeddings, we tested two sources: (1) ProtBERT and (2) the BERT module from the best-performing GANBERT-based binary classification model (also ProtBERT but fine-tuned). For embedding extraction, we compared two methods: mean pooling and direct CLS extraction. Each of these four combinations was evaluated by extracting embeddings from four different layers [7, 15, 23, 30]. These embeddings were then used to train SVM and XGBoost models, with parameter searches performed as follows:

SVM: 51 parameter combinations were tested: tol=1e-3; C=[0.1, 1, 10]; kernel=["linear", "rbf", "poly"]; gamma=["scale", 0.01, 0.1, 1.0]; degree=2; coef0=[0.0, 0.5, 1.0].

XGBoost: 720 parameter combinations were tested: learning rate=[0.001, 0.005, 0.01, 0.05, 0.1]; max depth=[3, 5, 7]; subsample=[0.3, 0.5, 0.7]; colsample bytree=[0.2, 0.4, 0.6, 0.8]; reg lambda=[0.0, 1.0, 2.0, 5.0]. Early stopping was applied if the cross-entropy loss on the validation set did not improve for 1000 rounds, and the best model was saved.

Since the numbers of enzymatic and structural sequences in our training set were imbalanced, class weighting was applied in both SVM and XGBoost. Class weights were computed as:

$$w_c = \frac{N}{K \times n_c} \tag{1}$$

where N is the total number of training samples, K is the number of classes (=2), and n_c is the sample count of class c, ensuring balanced contributions of each class to the objective function.

2.6.4 Ensemble Learning

Finally, we performed ensemble learning by using the output probabilities from the three best-performing models (GANBERT, SVM, and XGBoost) as inputs to a simple MLP. Using the same dataset split, we applied random upsampling of the minority structural class to balance the training set. A parameter search was conducted over the following combinations: hidden dim=[8, 16, 32]; dropout=[0.0, 0.2]; learning rate=[0.001, 0.0005, 0.0001, 0.00005]; batch size=[16, 32, 64, 128]; and hidden layers=[1, 2, 3].

2.7 ESNN Setting: Enzymatic, Structural, or Natural Negative

Both MoN and EoS are binary classification models. Here, we further developed a three-class model to simultaneously classify proteins into three categories: enzymatic, structural, and non-metalloprotein. To maintain a more balanced class distribution, only the positive set and the natural negative set were used, thereby avoiding excessive non-metalloprotein samples.

The merged dataset was clustered using MMseqs2 at 30% sequence similarity, and clusters were ordered by size. The dataset was then split into training, validation, and test sets at an 8:1:1 ratio. Based on prior training experience, we extracted mean-pooled embeddings from ProtBERT at layers 7, 15, 23, and 30. Using these embeddings, we performed XGBoost training with parameter searches under the same settings as before, applying class weights to address label imbalance.

3 Results

3.1 Metal Distribution in the Positive Dataset

After processing two public metalloprotein databases, a total of 40,182 unique metal-binding protein sequences were obtained. After filtering redundant and low-quality data, 35,339 valid metal-binding sites were extracted, resulting in a final set of 18,340 positive metalloprotein sequences (Figure 1). Among these sites, 44 distinct metal types were identified (Figure 2), namely: ['AG', 'AL', 'AS', 'AU', 'BA', 'BE', 'CA', 'CD', 'CE', 'CO', 'CS', 'CU', 'EU', 'FE', 'GA', 'GD', 'HG', 'HO', 'IN', 'IR', 'K', 'LA', 'LI', 'MG', 'MN', 'MO', 'NA', 'NI', 'PB', 'PD', 'PR', 'PT', 'RB', 'RU', 'SM', 'SR', 'TB', 'TL', 'U', 'V', 'W', 'Y', 'YB', 'ZN'].

However, the distribution is dominated by five metals—Zn, Mg, Ca, Fe, and Na—which together account for 85.11% of all binding sites. Among them, Zn, Ca, and Na ions are predominantly coordinated solely by amino acids (without cofactors within 2.88 Å), representing 87.8%, 92.4%, and 89.5% of their respective categories. In contrast, Mg and Fe ions mostly involve cofactors within 2.88 Å, accounting for 65.9% and 94.0% of their categories, respectively. Overall, across all metal-binding sites, the average proportion of amino-acid-based coordination atoms is 63.37%.

In our final positive dataset, each sequence interacts with an average of 1.9 metal ions. Specifically, 10,775 sequences (58.75%) contain a single metal ion, while 7,565 sequences (41.25%) bind multiple metal ions. Sequences containing at least one metal ion coordinated by a cofactor (within 2.88 Å) were categorized as cofactor metalloproteins (7,267 sequences, 39.62%). Conversely, sequences in which all metal ions are coordinated exclusively by amino acid residues were categorized as biological metalloproteins (11,073 sequences, 60.38%).

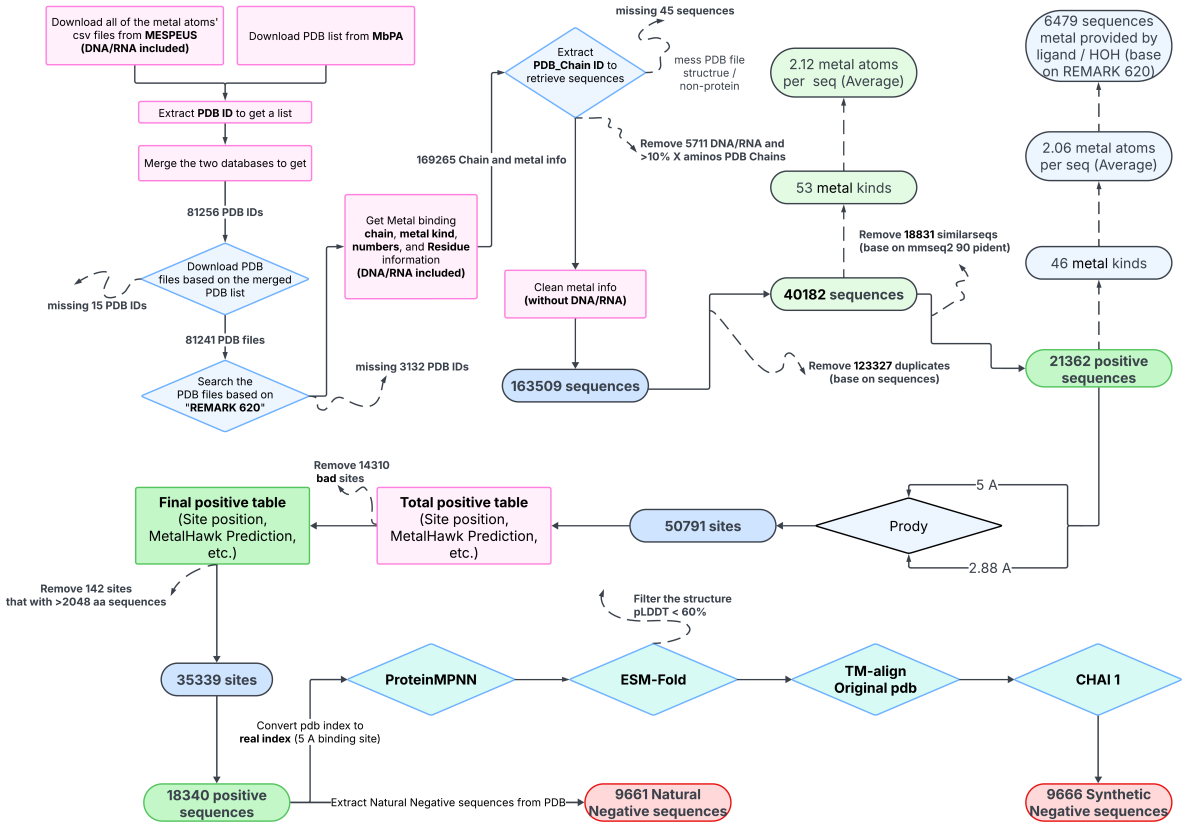

Figure 1: Metalloprotein Dataset Construction and Negative Set Generation Pipeline

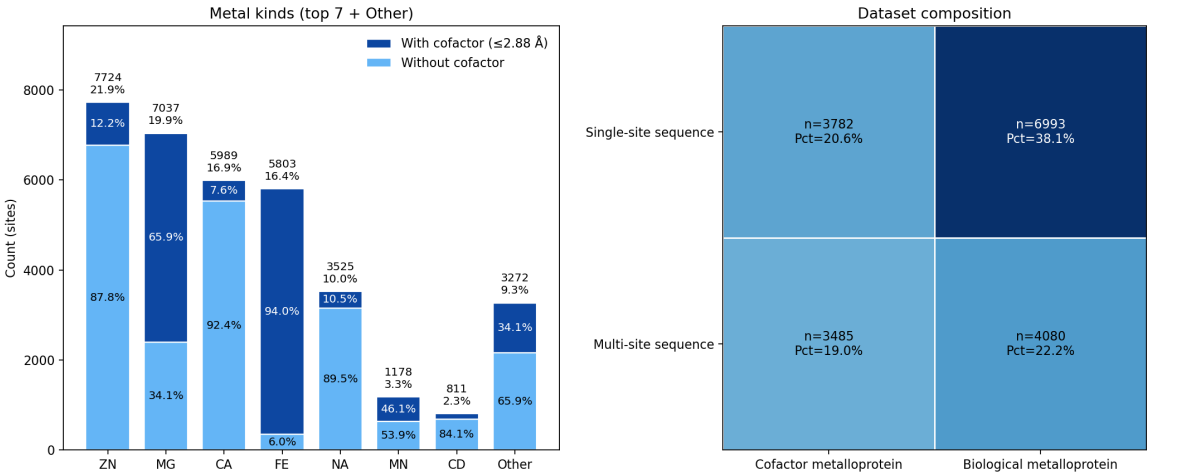

Figure 2: Metal ion distribution and dataset composition in the positive set

3.2 Enzyme Distribution in the Positive Dataset

Based on enzyme information databases, we successfully annotated 10,447 metalloprotein sequences as enzymatic metal-binding proteins and 7,893 metalloprotein sequences as structural, and summarized the counts and proportions of each category (Figure 3). Among them, 9,870 metalloprotein chains were associated with a single enzymatic class, while 577 metalloprotein chains exhibited multiple enzymatic functions. In our dataset, the three most represented classes were hydrolases, transferases, and oxidoreductases, which together accounted for 80.3% of all enzymes (counting multiple enzymatic functions separately).

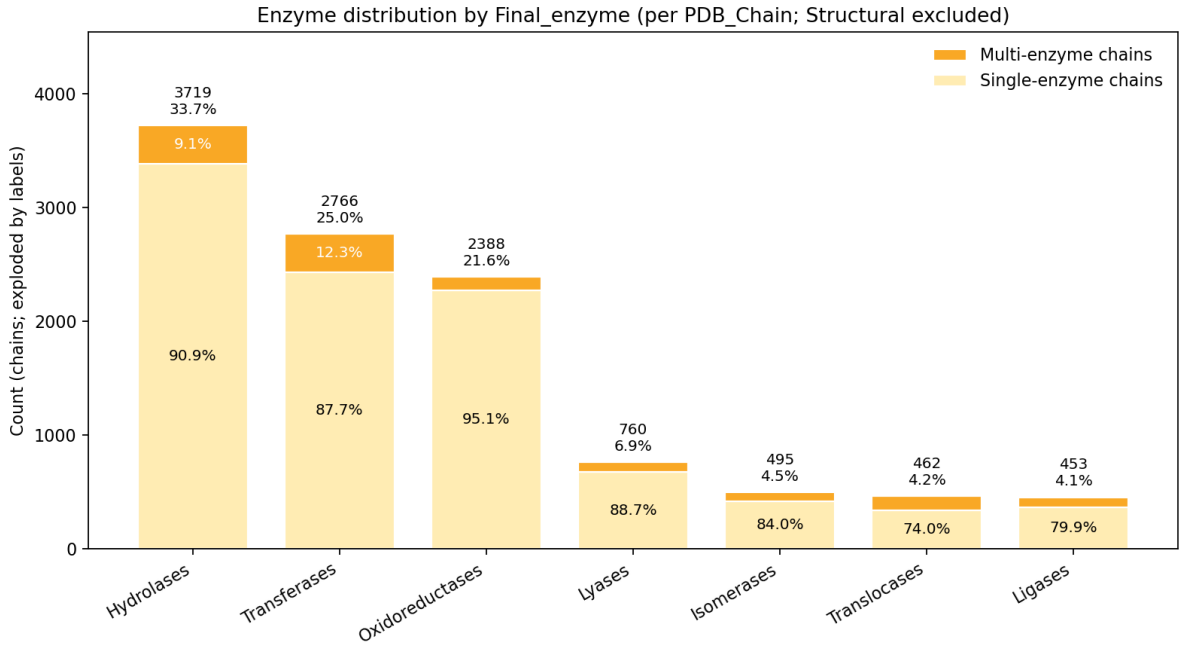

Figure 3: Distribution of enzymatic classes in the positive dataset

3.3 Processing of the Synthetic Negative Set

After redesigning the metal-binding sites of all sequences in the positive set, we analyzed the results (Figure 4). The most frequent amino acids at metal-binding sites were C, D, E, and H, which together accounted for 72.68% of all binding-site residues. Notably, by tuning the ProteinMPNN settings, we reduced the non-mutated rate of these amino acids to below 20%. For the four major amino acids, only 9.5%–17.8% of residues remained unchanged after redesign.

Following quality control of the redesigned sequences, a total of 9,666 synthetic negative sequences were retained, forming the synthetic negative set (Figure 5).

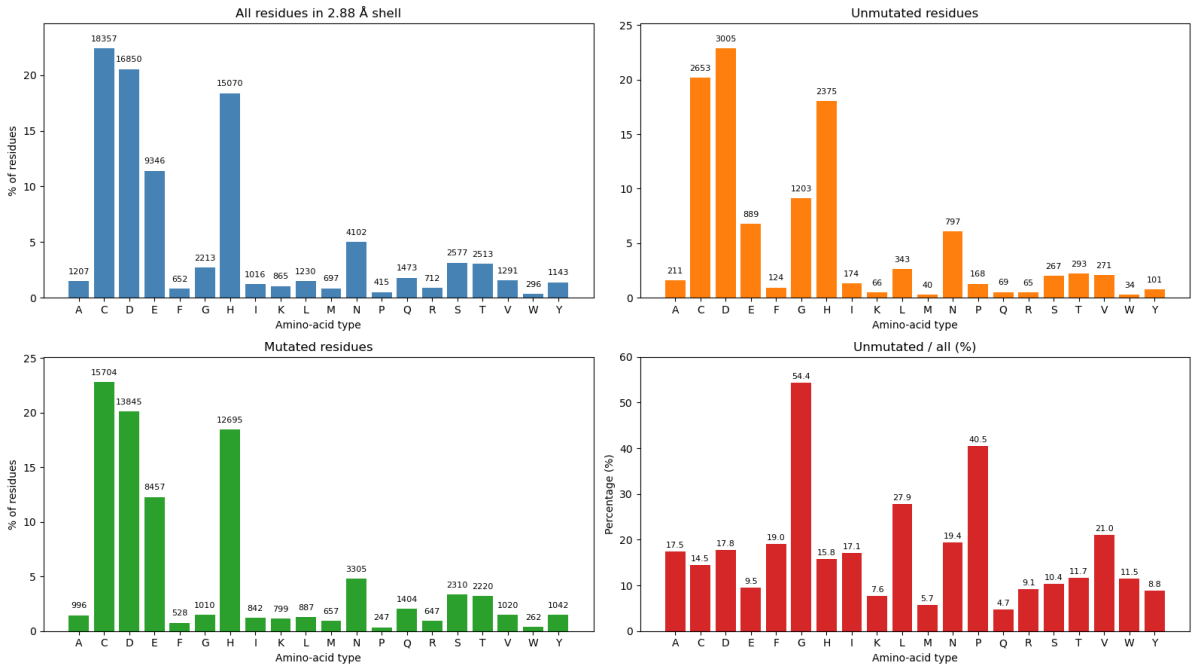

Figure 4: Amino acid composition and redesign outcomes in synthetic negative sequences

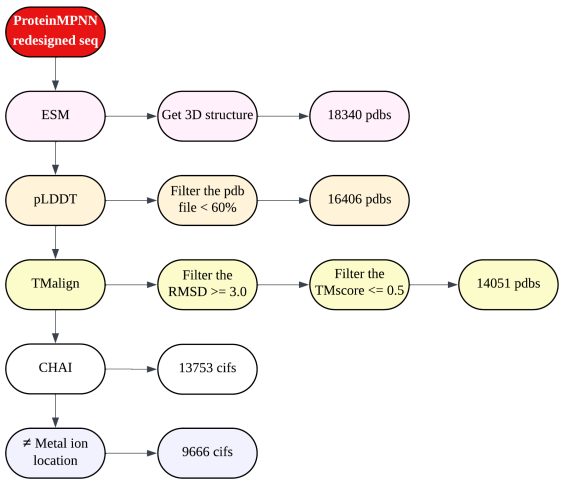

Figure 5: Workflow for generating the synthetic negative dataset

3.4 Processing of the Natural Negative Set

Due to the imbalance between the positive set and the synthetic negative set, we supplemented the negatives by extracting non-metal-binding chains from heteromeric proteins in the PDB entries of the positive set. Among all metalloprotein chains in the positive set, a total of 17,381 PDB structures were identified, of which 4,602 contained non-metal-binding protein chains within the same protein complex. After extraction and cleaning, we ultimately obtained 9,661 natural non-metal-binding protein chains, which together constituted the natural negative set.

3.5 MoN: Metalloprotein or Not

By merging, clustering, and partitioning the positive set, synthetic negative set, and natural negative set, we obtained a nearly balanced training and test set (Figure 6).

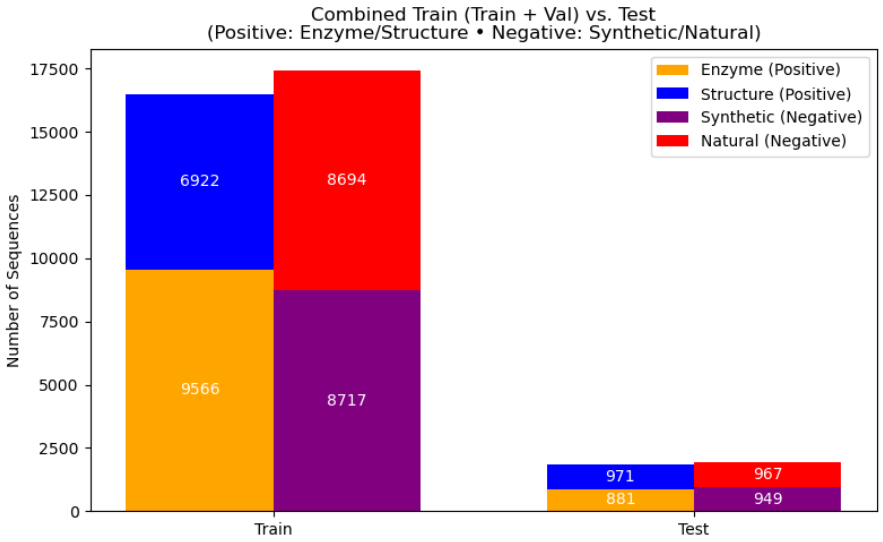

Figure 6: Composition of training and test sets for the MoN model

Through parameter search, the optimal hyperparameters were determined as: learning rate = 5e-6, batch size = 128, dropout probability = 0. On average, early stopping occurred after 36 epochs across the five folds, with Fold-3 training the longest (43 epochs). Under these settings, we evaluated the five folds individually as well as the final ensemble model (via majority voting) on the test set using four standard metrics (Table 1). The results showed that applying the voting strategy improved performance across all four metrics to varying degrees.

To further assess model performance across the two classes, and to analyze accuracy differences among different types of metalloproteins in the positive set, we generated a confusion matrix and a per-class accuracy matrix (Figure 7). The results indicated that classification accuracy for non-metalloproteins was slightly higher than for metalloproteins (by 3.1%), although overall predictions were well balanced between the two categories. Moreover, the per-class accuracy analysis revealed considerable variability across metalloprotein subtypes. In general,

Table 1: Performance of the model across 5 folds and MoN(majority voting).

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 performed 8% better on proteins with multiple metal-binding sites compared to single-site proteins, and 7% better on cofactor metalloproteins than on biological metalloproteins.

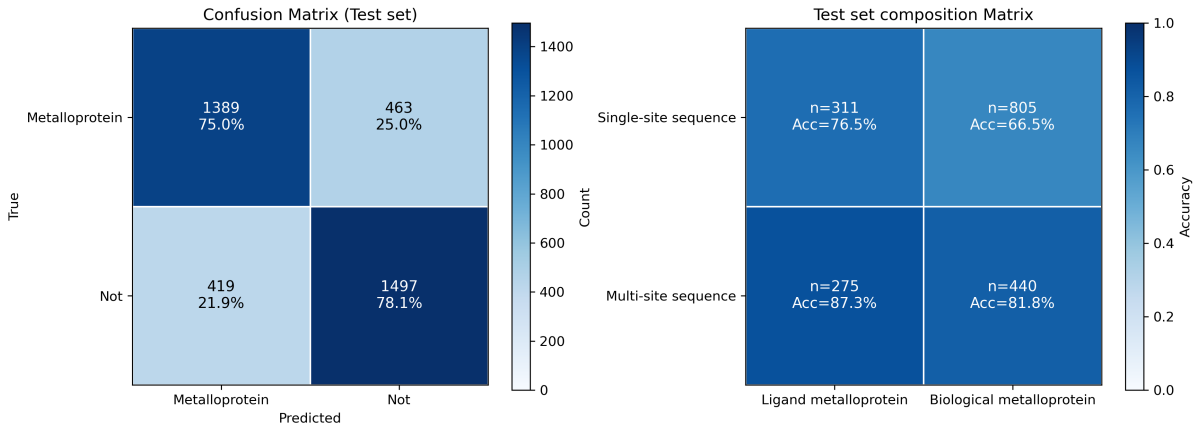

Figure 7: Confusion matrix and subclass accuracy of the MoN model on the test set

We further evaluated the model's predictive ability across different metal ion types by analyzing the relationship between ion type and classification accuracy in the test set (Figure 8). For the major ions present in the training set (Zn, Ca, Mg, Fe), accuracies all exceeded 75%, with Fe reaching as high as 94.22%. In contrast, Na had lower accuracy (63.20%), making it the primary factor limiting further improvements in overall model performance.

Additionally, since half of the negatives in the test set were synthetic sequences, we re-evaluated MoN after removing these synthetic sequences to assess its performance on real sequences. On this refined test set, all metrics improved significantly: accuracy increased by 2% (to 78.61%), precision by 14% (to 90.84%), recall by 0% (to 75.00%), and F1 score by 6% (to 82.17%).

3.6 EoS: Enzymatic or Structural

Based on clustering and partitioning of the metalloprotein positive set, combined with unlabeled data, we constructed the dataset for training the GANBERT binary classifier. In the EoS model, enzymatic proteins were treated as the positive class and structural proteins as the negative class (Figure 9).

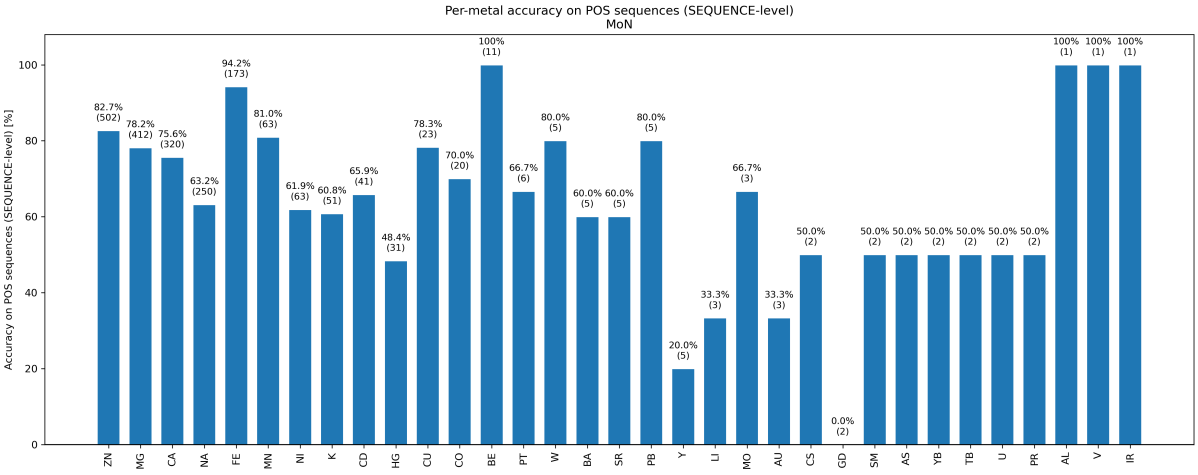

Figure 8: Per-metal classification accuracy of the MoN model on the test set

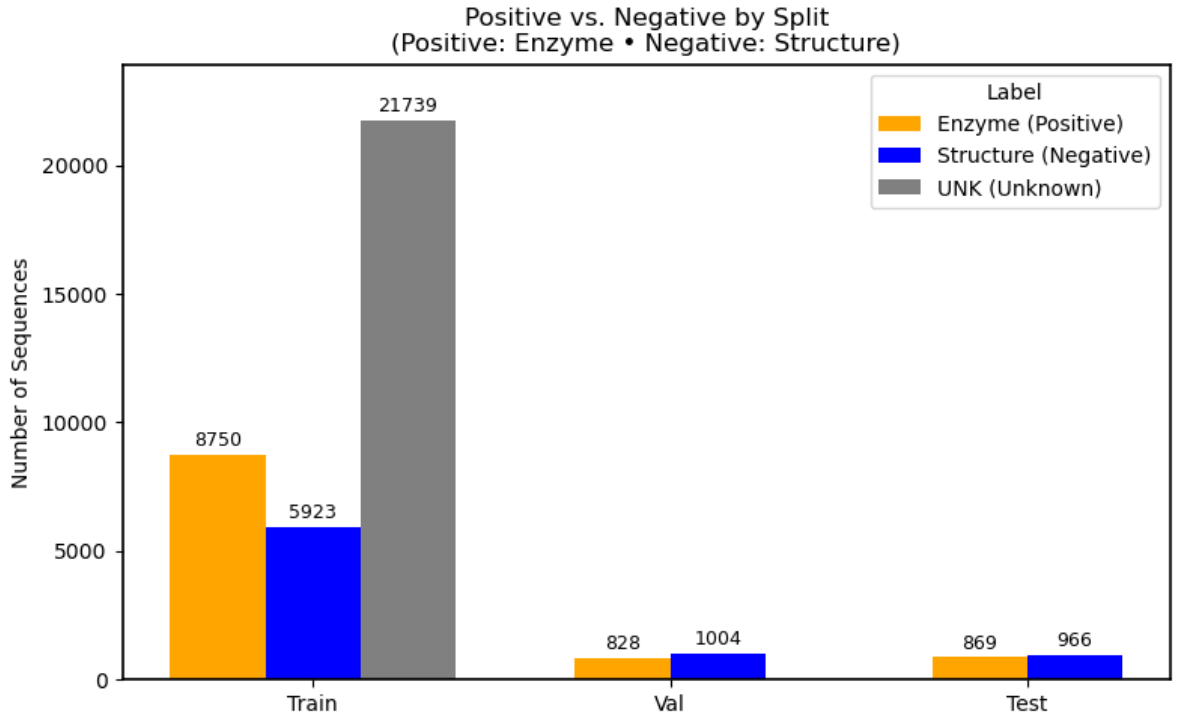

Figure 9: Dataset composition and split for the EoS model

For the GANBERT model with ProtBERT as the backbone, the optimal parameters were: learning rate = 5e-6, batch size = 256, and dropout probability = 0. Early stopping occurred after 10 epochs. The evaluation metrics on the test set were: accuracy = 72.75%, precision = 71.23%, recall = 71.23%, and F1 score = 71.23%. Despite the notable class imbalance between enzymatic and structural proteins, the optimal ProtBERT-based GANBERT showed balanced predictive ability across both classes (Figure 10).

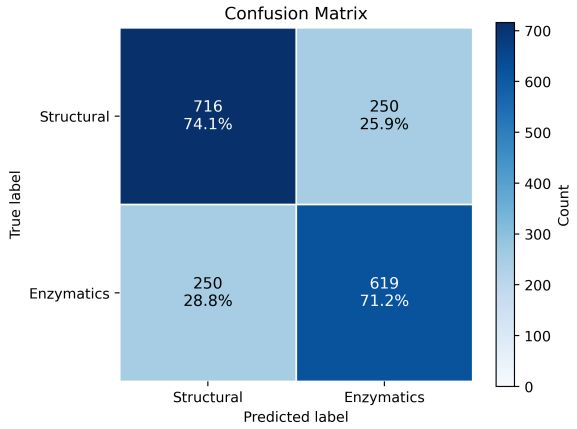

Figure 10: Confusion matrix of the EoS model on the test set

We further tested using the CLS embedding from the 23rd layer of ProtBERT instead of the default last layer (30th) as input to GANBERT. On the test set, the evaluation metrics were accuracy = 69.05% and F1 score = 64.37%. While accuracy was close to the result from the 30th layer, the F1 score dropped substantially. Finally, we replaced the backbone with esm2_t33_650M_UR50D, extracting the final-layer CLS embedding to construct a GANESM model. With only one tested parameter setting, all metrics fluctuated around 60%, about 10% lower than the optimal GANBERT, showing weaker performance in distinguishing enzymatic from structural metalloproteins.

Next, using only the clustered and partitioned metalloprotein positive set, we extracted embeddings from different backbone models, layers, and extraction types, and trained SVM and XGBoost classifiers (Figure 11).

The best-performing models from parameter searches were selected as representatives (Table 2). Results indicated that ProtBERT embeddings performed better than embeddings extracted from the fine-tuned GANBERT ProtBERT. Mean-pooled embeddings outperformed CLS embeddings.

In Table 2, the best-performing method for each of the two machine learning approaches is marked with a smiley face on the left side of the table. For the SVM classifier, the best test-set performance was obtained using the mean-pooled embedding from ProtBERT layer 15, with kernel = RBF, C = 1.0, and gamma = scale. For the XGBoost classifier, the best result was achieved using mean-pooled ProtBERT layer 23 embeddings, with parameters: colsample

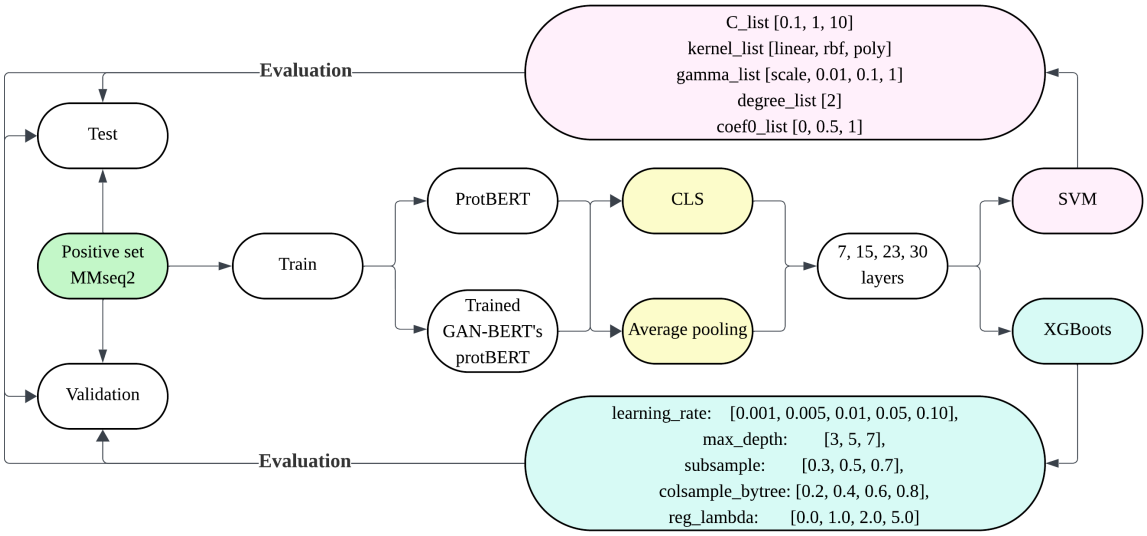

Figure 11: Workflow of embedding extraction and machine learning for the EoS model

Table 2: Performance of embedding-based SVM and XGBoost models for the EoS task

	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

bytree = 0.6, learning rate = 0.01, max depth = 7, reg lambda = 2.0, and subsample = 0.3, with training stopping after 2423 rounds.

Based on the same task and dataset, we further built a simple MLP ensemble using the output probabilities of the three best-performing models (GANBERT, SVM, and XGBoost). Through parameter search, the best setting was found with three hidden layers, hidden dimension = 8, dropout probability = 0.2, learning rate = 1e-4, and batch size = 16, with early stopping after only four epochs. Comparison of results showed that the ensemble outperformed each of the three individual models on the same test set (Table 3). Therefore, we selected the ensemble as the final EoS model.

Table 3: Performance of individual models and the ensemble EoS model

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%**

3.7 ESNN: Enzymatic, Structural, or Natural Negative

By merging the metalloprotein positive set with the natural negative set and applying clustering and partitioning, we constructed a three-class dataset (Enzymatic, Structural, Not Metalloprotein) for training the ESNN model (Figures 12). Mean-pooled embeddings were computed from ProtBERT at four different layers and subsequently trained using an XGBoost model. In addition, we performed PCA analysis on the embeddings of the three protein classes to visualize their distribution in the reduced-dimensional space (Figures 13). In the figures, the centroid of each class is highlighted with a larger marker in the same color as the corresponding class.

Through parameter search, the best performance in terms of both accuracy and F1 score was achieved using embeddings from the 15th layer. The optimal parameter combination was: colsample bytree = 0.2, learning rate = 0.005, max depth = 5, reg lambda = 2.0, subsample = 0.3, with early stopping after 6,232 rounds. On the test set, the three-class performance was: accuracy = 68.87%, precision = 68.92%, recall = 70.15%, and F1 score = 69.20%. We further evaluated the three-class results using a confusion matrix to explore model performance across different categories (Figure 14).

Because the ESNN model incorporates three classification labels, we were also able to evaluate its performance on the two previous binary tasks.

When only metalloproteins and non-metalloproteins(natural negative) were considered in the test set (equivalent to the MoN task), performance was very high: accuracy = 86.15% (+7.54%), precision = 93.07% (+2.23%), recall = 87.13% (+12.50%), and F1 score = 90.01% (+7.84%). Compared with MoN tested only on natural sequences, all metrics were substantially improved.

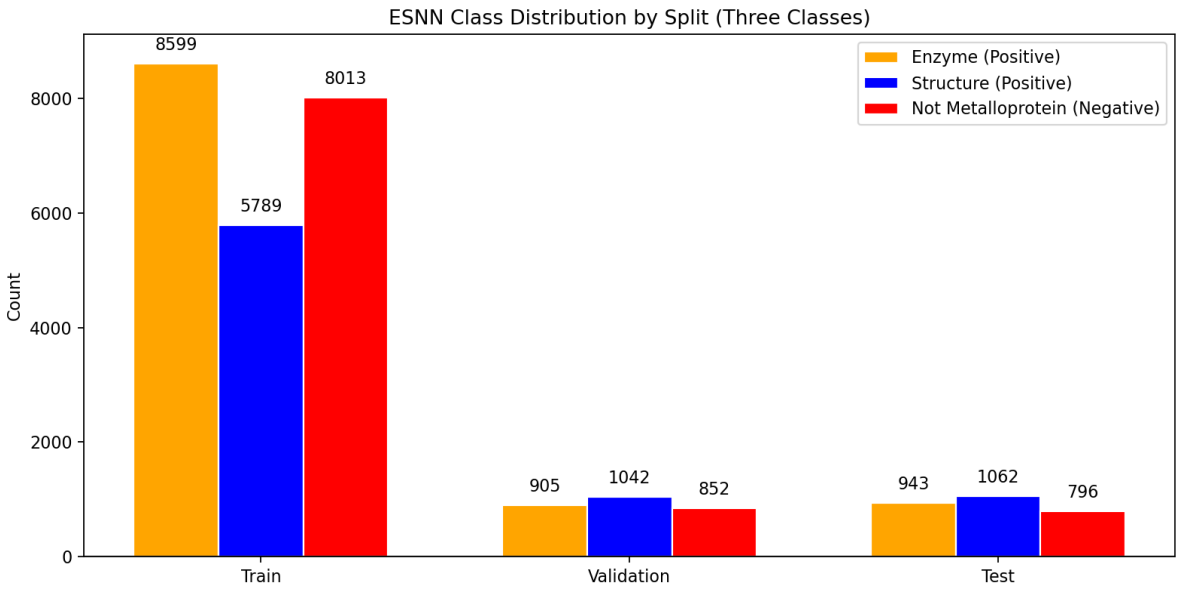

Figure 12: Class distribution of the ESNN dataset across training, validation, and test sets

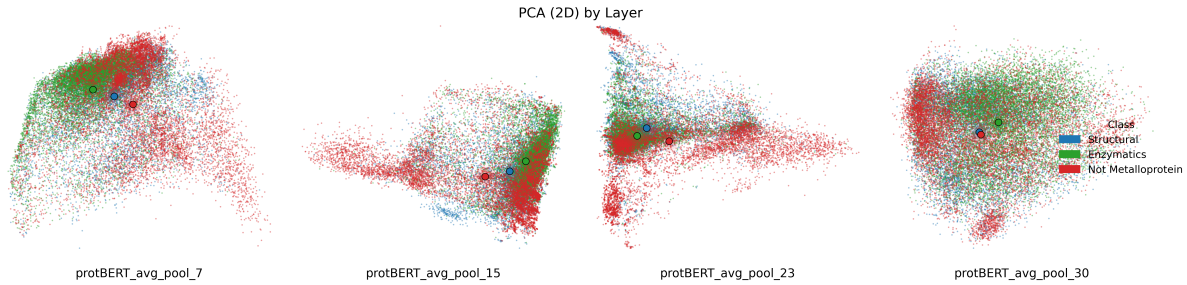

Figure 13: PCA of ProtBERT embeddings for the ESNN dataset across different layers

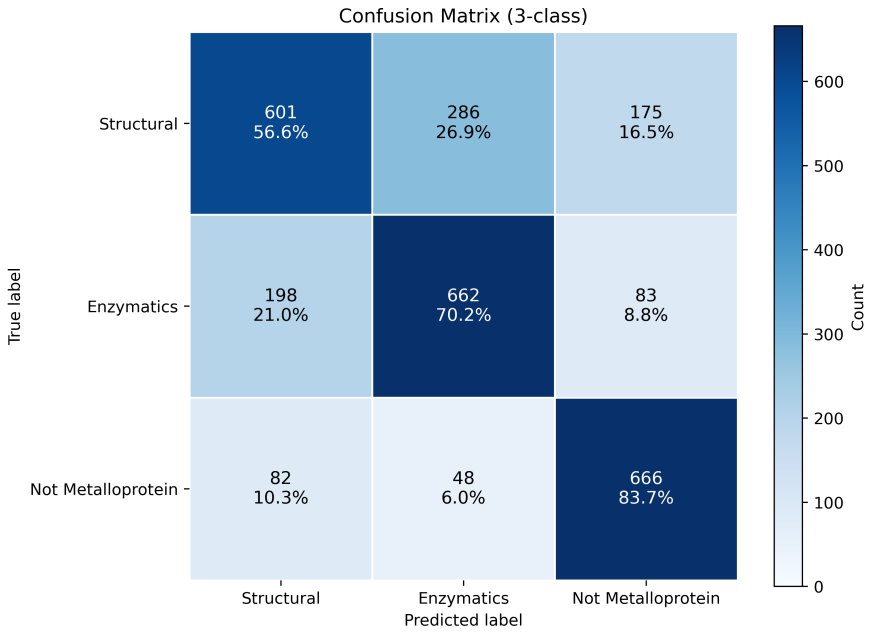

Figure 14: Confusion matrix of the ESNN model on the test set (three-class classification)

When the Not Metalloprotein class was excluded from the test set and only enzymatic vs. structural proteins were considered (equivalent to the EoS task), the metrics were: accuracy = 71.72%, precision = 69.83%, recall = 70.20%, and F1 score = 70.02%. These results were slightly lower than those of the EoS model.

In addition, we further evaluated the performance of the ESNN model in binary classification (metalloprotein vs. not) by comparing the prediction accuracy for different metal-binding ions relative to the MoN model (Figure 15). Analysis showed that, for the 33 metal ions present in the test sets of both models, ESNN achieved accuracy that was either higher than or equal to that of MoN across all ions. In particular, the accuracy for major ions such as Mg, Ca, and Na was substantially improved in ESNN compared to MoN. For certain ions, such as Mn, the accuracy reached as high as 98.81%.

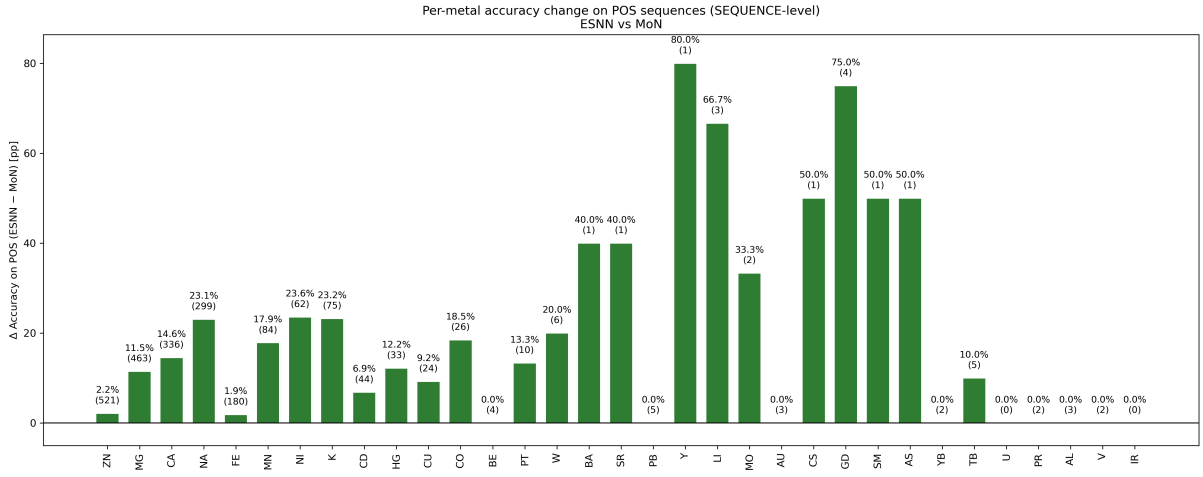

Figure 15: Per-metal accuracy improvement of the ESNN model compared with the MoN model

4 Discussion

This study performed redundancy reduction and quality control on a large collection of metalloprotein data and further annotated their enzymatic properties. A distinctive strategy was applied by mutating metal-binding sites in metalloprotein sequences to generate part of the negative set, while the dataset also incorporated natural negative sequences derived from non-binding regions of heteromeric metalloproteins within the same structure, thereby successfully constructing a curated metalloprotein dataset. Based on this dataset, and relying solely on protein primary sequences, we developed three models: MoN for identifying metalloproteins, EoS for predicting whether metalloproteins are enzymatic or structural, and ESNN for simultaneous three-class prediction (enzymatic, structural, and non-metalloprotein).

Compared with MoN, the ESNN model demonstrated clear improvements across all four standard metrics on test sets containing only natural sequences. Trained with embedding + machine learning, ESNN not only maintained broad applicability across metal ion types but also

exhibited higher predictive accuracy for metalloprotein identification. Particularly, prediction performance for certain ions such as Mg, Ca, and Na showed more than an order-of-magnitude improvement, even with over 200 test samples per ion. Similarly, during the training of the EoS model, a large number of unlabeled metalloprotein-like sequences (approximately 21k) were provided to the generator in the GANBERT framework to create fake samples and enhance training. In contrast, the embedding + machine learning approach did not make use of any unlabeled data. Nevertheless, we observed that models trained with embeddings and machine learning consistently achieved higher accuracy and F1 scores than GANBERT under the same labeled dataset splits. This may suggest that the embedding + machine learning approach has certain advantages for sequence-only classification tasks, although further studies are needed to confirm this.

In our experiments, artificially redesigned non-metalloprotein sequences were included in MoN training. The initial idea was that, by providing sequences differing only at metal-binding residues, the model could leverage deep learning to extract more informative features and improve recognition of metalloproteins. However, results showed that the model struggled to distinguish redesigned sequences, and models trained with them performed worse on natural sequences than the ESNN model trained only on natural data. One possible explanation is that the synthetic negative sequences differed from their positive counterparts only at the binding site. In principle, this design should help the model learn the key features that distinguish metal-binding from non-binding. However, the total number of synthetic negatives was only 9,666, and the differences were too subtle. As a result, the model may have struggled to learn effectively from them. In contrast, the natural negative sequences were too different from metalloproteins, which made them easy for the model to classify.

In addition, the model's discriminative ability varied considerably across different types of metalloproteins. Analysis of MoN's per-class accuracy matrix showed that single-metal proteins had the lowest prediction accuracy: only 66.5% for sequences without cofactors and 76.5% for those with cofactors, both far below the performance observed for multi-metal proteins. Overall, sequences with cofactors consistently achieved higher prediction accuracy, likely because their binding-site patterns are more uniform. For instance, Fe ions are commonly found in SF4 clusters, typically coordinated with a single Cys residue. This trend was also reflected in the per-ion accuracy: metal ions with higher cofactor proportions generally yielded better predictive results. For example, Fe achieved prediction accuracies of 94.22% in MoN and 96.11% in ESNN.

In the binary classification task of enzymatic versus structural metalloproteins, our final EoS model was built by ensembling one deep learning model with two machine learning models. Few studies have addressed this task, with the main related work being MAHOMES, which predicts whether a given metal-binding site is enzymatic using site-level input and extensive feature engineering, which limits its scalability. In contrast, our EoS model predicts enzymatic activity at the sequence level, making it more suitable for large-scale database annotation,

although its accuracy is slightly lower than that of MAHOMES.

In the three-class ESNN task, PCA analysis of embeddings showed that samples of the three labels were more distinctly separated at layer 15 than at other layers, which helps explain why embeddings from this layer yielded the best classification performance. At the same time, it is worth noting that across the four different layers, the Structural class (blue) consistently displayed a dispersed distribution. This observation partly explains why, in the evaluation of the three-class confusion matrix, the structural class exhibited a lower accuracy compared to the other two classes (only 56.6%).

In MoN and ESNN, compared with existing metalloprotein predictors such as MEBIPred (covering 9 metal ions), MIBPred (6 metal ions), and the Kumar model (8 metal ions), our models incorporated nearly all known metal ions capable of binding proteins (44 in total). Although this coverage substantially exceeds that of existing classification models, the top five metal ion types still account for 85% of the data, which limits the universality of metalloprotein prediction.

Beyond these findings, this study also has several limitations. When selecting sequences from the two metalloprotein databases, we did not filter based on experimental method or resolution, which may have limited further improvements in model performance. At the same time, the number of synthetic negative sequences was limited, making it necessary to include natural negative sequences for training. However, because natural negatives differ substantially from positive sequences, this may have reduced the model's ability to effectively learn from the synthetic negatives.

Looking ahead, with the continuous accumulation of proteomics and structural biology data, more high-quality data will become available. Incorporating higher-resolution structural information and experimental validation data may further enhance the reliability of model predictions. In addition, as AI tools continue to advance, it may become possible to design more efficient and biologically realistic methods for mutating metal-binding sites, thereby reducing the likelihood that synthetic sequences are filtered out for being overly artificial.

Overall, this study constructed a high-quality dataset of metalloproteins through rigorous data filtering and a distinctive approach to negative set generation. Building on this dataset, we developed several sequence-based prediction models that address both metalloprotein identification and enzymatic activity prediction, thereby establishing a more comprehensive predictive framework.

Data availability The data, code, and models of this study are available at: https://github.com/ibmm-unibe-ch/metal_binding_predictions

References

[1] H. H. Lin, L. Y. Han, H. L. Zhang, C. J. Zheng, B. Xie, Z. W. Cao, and Y. Z. Chen. Prediction of the functional class of metal-binding proteins from sequence derived physicochemical properties by support vector machine approach. *BMC Bioinformatics*, 7(Suppl 5):S13, 2006. Proceedings of APBioNet – Fifth International Conference on Bioinformatics (InCoB2006).

[2] A A Aptekmann, J Buongiorno, D Giovannelli, M Glamoclija, D U Ferreiro, and Y Bromberg. mebipred: identifying metal-binding potential in protein sequence. *Bioinformatics*, 38(14):3532–3540, 05 2022.

[3] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 01 2000.

[4] The UniProt Consortium. Uniprot: the universal protein knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617, 11 2024.

[5] Ryan Feehan, Lily Orellana, Diego Ochoa, Bo Zhang, Jens Kleinjung, Robert D. Finn, Christine A. Orengo, and Benoît H. Dessailly. Machine learning differentiates enzymatic and non-enzymatic metals in proteins. *Nature Communications*, 12(1):3980, 2021.

[6] Simon L. Dürr, Andrea Levy, and Ursula Rothlisberger. Metal3d: a general deep learning framework for accurate metal ion location prediction in proteins. *Nature Communications*, 14(1):2713, 2023.

[7] Farzaneh Mohamadi, Herman W. T. Van Vlijmen, and André H. Juffer. An ensemble 3d deep learning model to predict protein–metal binding sites. *Cell Reports Physical Science*, 3(12):101046, 2022.

[8] Kumar Suresh. Prediction of metal ion binding sites in proteins from amino acid sequences by using simplified amino acid alphabets and random forest model. *Genomics Inform*, 15(4):162–169, 2017.

[9] Aditi Shenoy, Yogesh Kalakoti, Durai Sundar, and Arne Elofsson. M-ionic: prediction of metal-ion-binding sites from sequence using residue embeddings. *Bioinformatics*, 40(1):btad782, 01 2024.

[10] Geng-Yu Lin, Yu-Cheng Su, Yen Lin Huang, and Kun-Yi Hsin. Mespeus: a database of metal coordination groups in proteins. *Nucleic Acids Research*, 52(D1):D483–D493, 11 2023.

[11] Jinzhao Li, Xiang He, Shuang Gao, Yuchao Liang, Zhi Qi, Qilemuge Xi, Yongchun Zuo, and Yongqiang Xing. The metal-binding protein atlas (mbpa): An integrated database for curating metalloproteins in all aspects. *Journal of Molecular Biology*, 435(14):168117, 2023. Computation Resources for Molecular Biology.

[12] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, DEBSINDHU BHOWMIK, and Burkhard Rost. Prottrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing. *bioRxiv*, 2020.

[13] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR, 09–15 Jun 2019.

[14] Danilo Croce, Giuseppe Castellucci, and Roberto Basili. 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, July 2020. Association for Computational Linguistics.

[15] Bruno Goeta. Bioinformatics-sequence and genome analysis. *Briefings in Bioinformatics*, 3(1):101–103, 03 2002.

[16] Martin Steinegger and Johannes Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, 2017.

[17] Ahmet Bakan, Lidio M. Meireles, and Ivet Bahar. Prody: Protein dynamics inferred from theory and experiments. *Bioinformatics*, 27(11):1575–1577, 04 2011.

[18] Antje Chang, Lisa Jeske, Sandra Ulbrich, Julia Hofmann, Julia Koblitz, Ida Schomburg, Meina Neumann-Schaal, Dieter Jahn, and Dietmar Schomburg. Brenda, the elixir core data resource in 2021: new developments and updates. *Nucleic Acids Research*, 49(D1):D498–D508, 11 2020.

[19] Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022.

[20] Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alexander Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *bioRxiv*, 2021.

[21] Yang Zhang and Jeffrey Skolnick. Tm-align: a protein structure alignment algorithm based on the tm-score. *Nucleic Acids Research*, 33(7):2302–2309, 2005.

[22] Chai Discovery. Chai-1: Decoding the molecular interactions of life. *bioRxiv*, 2024.

[23] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.