MetaLLM: Residue-wise Metal ion Prediction Using Deep Transformer Model

Fairuz Shadmani Shishir[1], Bishnu Sarker[2], Farzana Rahman[3], and Sumaiya Shomaji[1]

[1] Electrical Engineering and Computer Science, University of Kansas, Lawrence, USA.
shishir@ku.edu
[2] Computer Science and Data Science, Meharry Medical College, Nashville, USA.
[3] School of Computer Science and Mathematics, Kingston University London, London, UK.

Abstract. Proteins bind to metals such as copper, zinc, magnesium, etc., serving various purposes such as importing, exporting, or transporting metal in other parts of the cell as ligands and maintaining stable protein structure to function properly. A metal binding site indicates the single amino acid position where a protein binds a metal ion. Manually identifying metal binding sites is expensive, laborious, and time-consuming. A tiny fraction of the millions of proteins in UniProtKB – the most comprehensive protein database – are annotated with metal binding sites, leaving many millions of proteins waiting for metal binding site annotation. Developing a computational pipeline is thus essential to keep pace with the growing number of proteins. A significant shortcoming of the existing computational methods is the consideration of the long-term dependency of the residues. Other weaknesses include low accuracy, absence of positional information, hand-engineered features, and a pre-determined set of residues and metal ions. In this paper, we propose MetaLLM, a metal binding site prediction technique, by leveraging the recent progress in self-supervised attention-based (e.g. Transformer) large language models (LLMs) and a considerable amount of protein sequences publicly available. LLMs are capable of modelling long residual dependency in a sequence. The proposed MetaLLM uses a transformer pre-trained on an extensive database of protein sequences and later fine-tuned on metal-binding proteins for multi-label metal ions prediction. A stratified 10-fold cross-validation shows more than 90% precision for the most prevalent metal ions. Moreover, the comparative performance analysis confirms the superiority of the proposed MetaLLM over classical machine-learning techniques.

Keywords: metal binding-site prediction, deep learning, attention, self-supervised learning language model, transformers, bio-transformers.

1 Introduction

Proteins are biomolecules composed of amino acid chains that form the building blocks of life and play fundamental roles in the entire cell cycle. They perform

multitudes of functions including catalyzing reactions as enzymes, participating in the body's defense mechanism as antibodies, forming structures, and transporting important chemicals. In addition, they interact with other molecules including proteins, DNAs, RNAs, and drug molecules to act on metabolic and signaling pathways, cellular processes, and organismal systems. Protein structures and interactions describe the molecular mechanism of diseases and can convey important insights about disease prevention, diagnosis, and treatments. Likewise, proteins bind to different metal ions, such as zinc, iron, copper etc. to play necessary roles in many biological processes, including enzyme catalysis, regulation of gene expression, and oxygen transport. Metal ions are often bound to specific sites on proteins, known as metal binding sites, which play a key role in determining protein's structure and function. Identifying metal binding sites manually using various experimental procedures such as mass spectrometry, electrophoretic mobility shift assay, metal ion affinity column chromatography, gel electrophoresis, nuclear magnetic resonance spectroscopy, absorbance spectroscopy, X-ray crystallography, and electron microscopy is an expensive, laborious, and time-consuming process [23]. A very small fraction of the millions of proteins stored in UniProtKB [1] – the most comprehensive protein database – are annotated with metal binding sites. Millions of other proteins are awaiting for metal binding site annotation. To keep pace with the exponential increase of protein sequences in the public databases, it is essential to develop computational approaches for predicting metal binding sites in proteins. Considering the benefits it can provide in understanding function and structure of proteins as well as having practical implications in drug design and biotechnology, automatic prediction of metal binding sites in proteins is considered to be an important problem in computational biology.

Metal Binding Site Prediction: Predicting the binding sites for metals is a challenging problem in computational biology. Decades of research has been dedicated to discovering computational approaches that can accurately predict the metal ions as well as the positions where they bind to the proteins [3,19,9,5]. A comprehensive review of recent advances in computational approaches for predicting metal binding sites can be found in [34]. Broadly, these approaches can be categorized into following three groups based on the type of attributes they take into account: 1) structure-based methods that use three dimensional secondary structure of proteins as primary data; 2) sequence-based methods that use amino acid sequence as primary data; and 3) combined methods that leverage both structure and sequence attributes.

Structured-Based: Structure-based approaches for predicting metal binding sites use a combination of geometric, chemical, and electrostatic criteria to explain the metal-protein interaction that eventually works to identify possible binding sites in a protein structure. A relatively early work described in [29] uses electrostatic energy computation [22,10] to find binding affinity of metal ion to a site in a protein structure. In [4], the proposed method learns geometric constraints to differentiate binding sites for different metals based on the statis-

tical analysis of structures of metal binding proteins. Another structured-based method is proposed in [38] for predicting only the zinc binding sites. A template-based method is proposed in [21] where a database of pre-computed structural templates for metal binding sites is searched against each residue in a query protein to find which metal binds to it. mFASD [13] is a structure based model to predict metal binding sites. From the structures of metal binding proteins, mFASD computes the functional atom set (FAS) - the set of the atoms that are in contact with the metal - for each metal, and store it as reference. The distance between FASs of different metals is used to distinguish binding sites for different metals. Given a query protein structure, mFASD scans the database reference FASs against FAS of each sites, and computes the distance. The decision is made based on how many reference FASs matched for a given metal. One of the shortcomings of the structure based models is that they are dependent on structural databases such as Protein Data Bank (PDB) which is very limited in terms of amount of protein structures in the database.

Sequence-Based: On the other hand, sequence-based methods for predicting metal binding sites use sequence conservation, alignment, and similarity to identify metal binding sites. For example, [2] is a sequence-based method that find the patterns of binding sites from the metal binding proteins. Sliding window-based feature extraction and biological feature encoding techniques are proposed to predict the protein metal-binding amino acid residues from its sequence information using neural network in [17] and using support vector in machine [18]. MetalDetector [20] is a sequence-based technique that uses decision tree to classify histidine residues in proteins into one of two states as 1) free, or 2) metal bound; and cysteine residues into one of three states as 1) free, 2) metal bound, or 3) disulfide bridged. A two stage machine learning model is proposed in [24] that includes support vector machine as local classifier in the first stage, and a recurrent neural network (RNN) [14] in second stage to refine the classification based on dependencies among residues. A combined approach proposed in [30] where support vector machine, sequence homology, and position specific scoring matrix (PSSM) are put together to predict zinc-binding Cys, His, Asp and Glu residues.

Additionally, there are combined methods for predicting metal binding sites that use an ensemble of sequence, structural, and physicochemical features. For example, MetSite [32] is a method that uses both the sequence profile information and approximate structural data - PSSM scores together with secondary structure, site residue distances, and solvent accessibility - are fed into neural network machine learning technique.

Deep Learning and Transformers: Classical machine learning techniques such as decision tree, random forests, support vector machine etc. have been widely applied to the problem of metal binding site prediction. These model face several challenges, such as 1) they mostly depends on hand-engineered features e.g. K-mers, 2) they can not model variable length sequences data, and 3) they fail to comprehend long-distance residual dependency. Moreover, scaling these models to work on millions of variable length sequences is another big challenge.

Over the last decade, Deep Learning based techniques such as convolutional neural network (CNN) [16], long short term memory (LSTM) [14], transformers [33] etc. have grown to be extensively powerful. Generally, Deep Learning is a type of machine learning built using neural networks inspired by the structure, function and deep interconnection of the brain neurons [16]. In a typical deep neural network, many layers of interconnected nodes process vast amounts of complex data for learning from the experience. The strength of the deep learning lies in its ability to automatic feature engineering by learning low rank vector representation of data points. These advanced models are very efficient in handling large datasets. For example, [11,12] applied different deep learning architectures namely 2D CNN, LSTM, RNN coupled with various feature extraction techniques for predicting metal-binding sites of Histidines (HIS) and Cysteines (CYS) amino acids.

Very recently Transformer - a deep learning model for natural language processing tasks such as language translation, text summarization, question answering etc - is introduced by historic paper "Attention is All You Need" [33]. Generally, a transformer comprises of an encoder and a decoder neural network. Encoder and decoder composed of multiple layers of self-attention layers. The input of the encoder consists of queries as well as keys of dimension d_k, and values of dimension d_v. There are dot products of the query with all keys, which will divide each by $\sqrt{d_k}$, and apply a softmax function to obtain the attention weights which is shown in **equation 1**. The attention function on a set of queries was done simultaneously, packed together into a matrix Q. The keys and values are also packed together into matrices K and V. Transformer is capable of looking into the data from multiple perspectives following a multi-head mechanism shown in **equation 3**. Each head computes an attention weights that provides a new perspective that is shown in **equation 2**.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{2}$$

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^O \tag{3}$$

Transformer-based language models overcome the challenge of modeling long-distance dependency in natural text by introducing attention weights. This property ideally make it suitable for using in protein sequence modeling that long chain of amino acids.

While many of the existing methods are performing well, predicting metal binding sites is still a challenging problem as well as an open problem in computational biology. Partly because metals exhibit similar chemical properties making it hard for machine learning models to differentiate. A major shortcoming of the existing computational methods is in taking into account the long-distance dependency of the residues to distinguish the presence of distinct metal ions. There are other shortcomings such as low accuracy, absence of positional information, hand-engineered features, and pre-determined set of residues and metal

ions. Building high performing prediction model would require a comprehensive understanding of the structural, chemical, and biological factors that influence metal binding. Therefore, ongoing research effort is important to improve our understanding of metal binding in proteins and to develop more accurate and efficient computational pipeline to keep pace with the growing number of proteins. Considering the challenges and the availability of protein data, in this paper, we propose MetaLLM, a metal and binding site prediction technique by leveraging the recent progress in self-supervised attention-based (e.g. Transformer) large language models (LLMs) and huge amount of protein sequences publicly available. LLMs are capable of modeling long residual dependency in a sequence. The proposed MetaLLM uses a transformer pre-trained on large database of protein sequences, and later fine-tuned on metal binding proteins for multi-label metal ions prediction. In the fine-tuning step, the low rank sequence embeddings are concatenated with positional one-hot embeddings and fed into the fully connected neural network layer for metal ions prediction. A stratified 10-fold cross-validation shows more than 90% precision for the most prevalent metal ions.

2 Methodology

In this section, the proposed methodology and overall workflow have been discussed thoroughly. The workflow starts by computing the sequence embeddings for training and testing datasets using pre-trained protein language model provided by the Bio-Transformers[4] python wrapper of ProtTrans [8]. After that, each sequence embedding is combined with the positional one-hot encoding of metal binding sites. Finally, this combined vector is fed to a fully connected multi-layer deep neural network to predict the metal ion that finds the one-hot encoded position. The performance is validated on a set of text examples following the same workflow. The working principle is depicted in **Figure 1**. In the following subsections, we detail the workflow of model development and evaluation.

2.1 MetaLLM: Residue-wise Metal ion Prediction

MetalLLM - the proposed metal ion prediction model - is built on Large Language Model (LLM) pretrained on millions of protein sequences to provide sequence embeddings. Powerful LLMs are built using encoder-decoder based deep learning technique called Transformer [26,33]. Thanks to the capability of transformers in ingesting large amount of data and learning meaningful representation, it has already found application in various life science and biomedicine tasks such as protein folding, drug discovery, and gene expression prediction [35] in academia and industry alike. ProtTrans [8] is a such model using transformer architecture that includes self-attention layers in both encoder and decoder and is

[4] https://github.com/DeepChainBio/bio-transformers

trained on large amount of sequence data. Pretrained ProtTrans provides fixed dimensional embeddings for given sequence. We used ProtTrans as the background model to transform the variable length protein sequence into a fixed length numerical vector for the purpose of predicting metal ion that binds to the protein. To include the binding site in the pipeline, we have combined the positional one-hot encoding with sequence embedding.

Problem Statement: Therefore, the problem that we solve in this paper is defined as predicting metal ion given an input protein sequence and the residue position we are interested to look at as possible binding site. We formulate this as a machine learning problem where the sequence is represented by latent features computed using ProtTrans combined with positional enconding vector constitute the features, and finally we train a fully connected deep neural network for predicting the probability of metal ions as a downstream task.

Using transformer-based LLM has many benefits; 1) pretrained model reduce the computation cost significantly, 2) can be easily fine-tuned towards transfer learning for new data as well as new tasks, 3) they are capable to capturing long distance dependancy among the residues and tokens, 4) easily scalable.

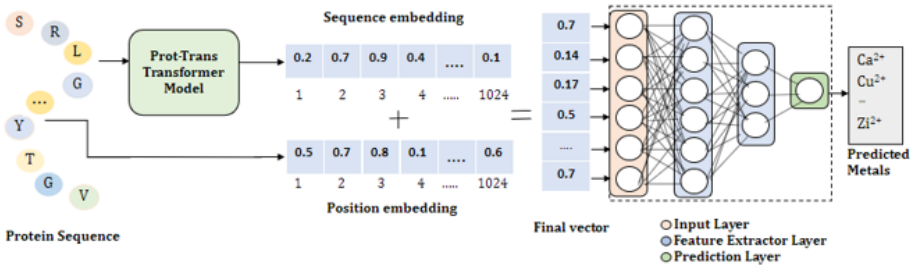

Fig. 1. The overall architecture of the proposed framework of MetaLLM: (i) the protein sequences are delivered as an input to the pre-trained language model to produce the sequence embedding, (ii) the embedding for identifying the binding positions of the amino acids are extracted (iii) both the embedding are concatenated to obtain a final vector, and (iv) the final vectors are fed into a deep neural network to predict the metal ions.

Model Description: The proposed model is shown in **Figure 1** initially employs a pre-trained language model called ProtTrans which effectively generate embedded and informative protein sequence representation from the raw protein sequence. ProtTrans, a large protein language model, was used as the back-end of our proposed architecture. Moreover, we have also tested our representation for the classification of various metals to justify the novelty of the language model. Our proposed architecture leverages ProtTrans model for the feature extraction, which was trained on computationally expensive hardwares. [8]. Then,

the embedded 1024-dimensional features were extracted from the last layer of ProtTrans model and classification was done integrating the sites information with the protein sequence embedding. The fixed 1024 dimension was chosen through the investigating in our experiment and 1024 dimentional embedding has given most discriminant features set which performed good for our methods. To achieve better performance from our model, we also concatenated the positional information by one-hot encoding since the specific binding location of a the metal ion is very important in a protein sequence. Our novel contribution on this paper is to carefully design a neural network which is more robust and cost effective with limited number of parameters without dropping the accuracy of the prediction.

The metal prediction was done by the classification model on top of the language model. The input of the classification layer was the 1024 size embedding from the language model and then concatenated with the same dimensional sites information. The weights of the four layers of classification model was initialized randomly and for the non-linearity, relu activation function was used. In order to get the final prediction, sigmoid activation function was considered. We also used stratified 10-fold cross validation to evaluate the whole dataset better and maximize the performance of our model. The mathematical representation are given in equations respectively.

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{4}$$

$$Relu(z) = max(0, z) \tag{5}$$

$$BCE = -(y \log(p) + (1 - y) \log(1 - p)) \tag{6}$$

$$multiclass\ classification = -\sum_{c=1}^{M} y_{o,c} \log(p_{o,c}) \tag{7}$$

3 Experiments and Results

3.1 Experimental Details

Dataset Description: To conduct the experiment, we have prepared a dataset with sequences retrieved from UniProt Knowledge Base (UniProtKB)[1]. UniProtKB is the largest protein database divided into two parts; 1) UniProtKB/SwissProt contains manually reviewed protein data and 2) UniProtKB/TrEMBL contains unreviewed protein data. We have accessed UniProtKB/SwissProt in March 2022 to retrieve protein sequences, binding sites and metal ions. Therefore, all of the protein data selected for this experiment were manually reviewed. To eliminate the possible fragmented protein sequences, the sequences smaller than 50 amino acids long are removed. We have also removed the sequences longer than 1000 amino acids long. Moreover, metals that does not have a sufficient number of instances are removed during further pre-processing. Finally, it contains 18,348 protein sequences holding 6 different metals in total. The distribution of metals is shown in the **Figure 2**.

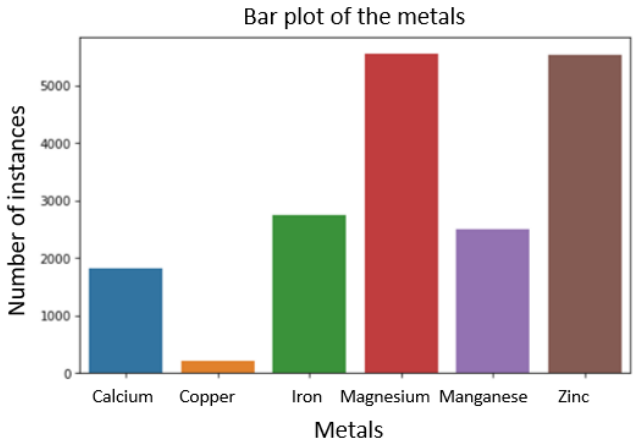

Fig. 2. This figure shows the number of samples of our dataset.

Implementation Details: MetaLLM is implemented using Bio-Transformers[5] python package that wraps the implementation of ProtTrans (Protein Transformer) [8] and ESM (Evolutionary Scale Modeling) [27,26,25] - two of the large language models trained on million of protein sequences to predict embeddings. MetaLLM gets the sequence embeddings using Bio-Transformer with ProtTrans backend. Sequence embeddings are added to multi-hot position encodings and the resultant vector is fed into the fully connected neural network. MetaLLM proposes a four-layer fully connected neural network with 500 hidden neurons in the first layer, 300 neurons in the second and third layer, and the final layer is with 100 hidden neurons before it goes to the soft-max layer for predicting the likelihood of different metals. To prevent the over-fitting of the network, we carefully used a dropout with a rate of 0.02. Batch size of 64 and learning rate in the range of 0.0001 to 0.1 was found to be optimal. The model is trained for 200 epochs with Adam optimizer to optimize binary cross-entropy loss function. The training takes 45 minutes on Nvidia GeForce RTX 3090 GPU-enabled computer. Once the model is trained, it takes 5 sec to make a prediction for one batch of protein sequences.

Evaluation Metrics: To evaluate the performance of our proposed model, we have considered the following evaluation metrics; 1) $Accuracy = (\frac{TP+TN}{TP+TN+FP+FN})$, 2) $Precision = \frac{TP}{TP+FP}$, 3) $Recall = \frac{TP}{TP+FN}$, and 4) $F1\text{-}max = \frac{2*Precision*Recall}{Precision+Recall}$. F1-max score is also computed as $\frac{2*TP}{2*TP+FP+FN}$. Here, TP = true positives and TN = true negatives denote correctly predicted positive and negative examples respectively, whereas FP = false positives and FN = false negatives denote the number of incorrectly predicted metals respectively.

[5] https://github.com/DeepChainBio/bio-transformers

3.2 Result Analysis

Model Performance: We evaluated MetaLLM by Precision, recall, and F1-score using the stratified 10-fold cross-validation (CV). Since the dataset is class imbalanced, we considered stratified cross-validation which ensures that the proportion of instances is conserved across the metals per fold. **Table 1** shows the model performance across various metals for with and without positional information. Precision, recall, and F1-scores are computed as average over the 10-folds. Overall, MetaLLM has performed very well across all the metals. However, the superior performance is observed for ion like magnesium **(Precision=0.97, Recall=0.94)** and iron **(Precision=0.96, Recall=0.94)** considering positional information. **Table 1** also lists precision, recall, and F1-scores without considering positional information - the model is fed with a protein sequence embedding and asked to predict the likelihood of metal ions binding to it. It is evident from the **Table 1** that performance dropped significantly for copper, iron, and magnesium. MetaLLM achieved an average CV accuracy of 90% when positional encoding is used. The box plots shown in **Figure 3** depict the spread of performance scores along the folds and across the metals. Finally, **Figure 4** shows the training and validation accuracy of our proposed model for a single fold across the epochs.

Table 1. Performance metrics of the proposed model

Metal	**With Position**			**Without Position**		
Name	Precision	Recall	F1	Precision	Recall	F1
Calcium	0.83	0.84	0.84	0.94	0.71	0.81
Copper	0.94	0.67	0.78	0.67	0.93	0.78
Iron	0.96	0.94	0.95	0.89	0.90	0.90
Magnesium	0.97	0.94	0.95	0.72	0.94	0.82
Manganese	0.92	0.93	0.92	0.71	0.47	0.56
Zinc	0.91	0.89	0.90	0.84	0.95	0.89

Comparison with Classical Machine Learning Methods: To briefly discuss our proposed method, we first consider the Bio-Transformer model for generating the embedding from the protein sequences, next we add the positional information for metal binding to the embeddings to create new representations, and finally, we train a fully connected deep neural network considering the new representation to classify the metal ions. By leveraging the promising feature representation techniques from the individual steps, we maximize the prediction of unknown metal ions. However, unlike ours, there exist several studies which rely on the traditional classifiers for predicting the binding sites and metal ions [37]. To compare the performance of our work with the existing classical ones,

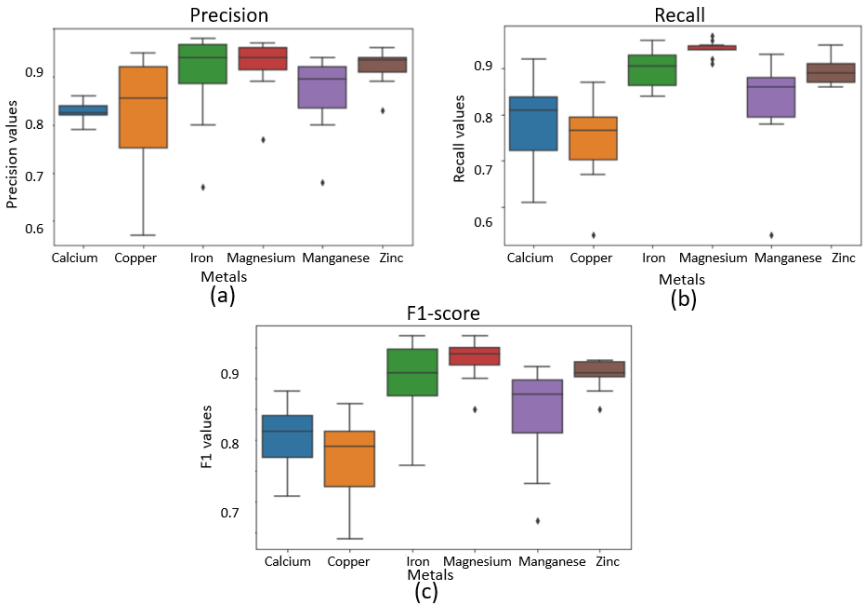

Fig. 3. The box-plot (a) shows the precision of 10-fold CV results for 6 metals. A comparison of the metals clearly indicates that zinc and magnesium have stable precision values of $93 \pm 3\%$ and $91 \pm 3\%$ respectively. The overall distribution of these two metals is not very scattered and also has low variance. On the contrary, copper, iron, and manganese have larger percentile values with some outliers also. In box-plot (b) zinc has outperformed compared to other metals with a recall value of $95 \pm 2\%$. However, there are some outliers in that distribution due to the limitation of the samples. The calcium and copper, although having limited number of samples, performed well with a recall value of $82 \pm 5\%$ and $78 \pm 4\%$. The box plot (c) shows the F1-score of CV. Iron, magnesium, and zinc have performed well in terms of F1-score and variation is also low compared to other metals.

we determined five widely used methods: K-Nearest Neighbor [36], Logistic Regression [7], Xgb classifier [31], Gaussian Naïve Bayes [15], and Support Vector Machine [28]. To train the classical models with our dataset, we have considered the embeddings extracted from Bio-Transformer and added positional information to create the final feature set for training. Once the models are trained, we compared the findings from these five methods with that of our proposed study. As shown in **Table 2**, the classical methods offer $42\% \sim 76\%$ of F1- score for Calcium. Similarly, it offers $19\% \sim 62\%$ for Copper, $73\% \sim 90\%$ for Iron, $53\% \sim 90\%$ for Magnesium, $59\% \sim 85\%$ for Manganese, and $65\% \sim 86\%$ for Zinc respectively. Whereas our proposed model achieved F1- score of 84% for Calcium, 78% for Copper, 95% for Iron, 95% for Magnesium, 92% for Manganese, and 90% for Zinc. As deliberated earlier, our proposed method is established on state-of-the-art deep learning models which provide promising distinctive features, therefore it can offer better prediction F1-score than the classical methods.

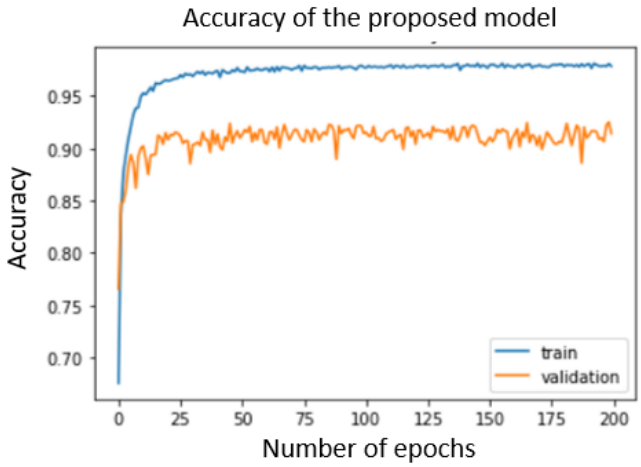

Fig. 4. Train and validation accuracy score of the proposed model. The overall trend of validation curve remained stable over the epochs during training.

Table 2. Metal wise F1-scores of the traditional classifiers compare against MetaLLM

Classifier	Calcium	Copper	Iron	Magnesium	Manganese	Zinc
Logistic Regression	0.72	0.59	0.88	0.89	0.81	0.84
Naive Bayes	0.42	0.19	0.73	0.53	0.59	0.65
SVM	0.75	0.51	0.89	0.90	0.84	0.86
KNN	0.76	0.38	0.90	0.84	0.85	0.86
XGB	0.63	0.62	0.84	0.80	0.70	0.74
MetaLLM	**0.84**	**0.78**	**0.95**	**0.95**	**0.92**	**0.90**

The Impact of Positional Information in Proposed Model: It has been estimated that more than one-third of the entire proteomes are metal-binding proteins [6]. Therefore, metal position in a protein plays a significant role in protein binding. Metal ions, such as zinc and copper, can act as co-factors in enzyme reactions and can also play a role in stabilizing protein-protein interactions. In some cases, metal ions can act as a "switch" in protein activity, allowing or preventing binding to other molecules. Considering the aforementioned reasons, our experiment was designed with and without concatenating sites information with the sequences. We also investigated further to understand whether site of a metal can contribute to increase accuracy while making prediction or not. The network architecture was developed based-on that workflow. It was found in the experiment that the model performed better with the positional information than without providing it while making the prediction of the metals, which is indeed one of the major findings in our work.

4 Conclusion

In conclusion, predicting metal binding sites in proteins is a complex and multifaceted problem that requires integrating multiple types of information and developing sophisticated computational methods. While significant progress has been made in this area, there is still much to be learned about the roles of metal ions in biological systems and the computational approaches appropriate to predict where they bind to proteins. To this end, in this paper, we have described a transformer-based model called MetaLLM to incorporate the benefits of highly sophisticated self-supervised deep learning technique to propose an end-to-end computational pipeline that is capable of predicting the presence of metal ions as well as the binding sites with state-of-art performance in terms of precision, recall and F1-score. The performance of MetaLLM is validated on a dataset of protein sequences extracted from UniPortKB/SwissProt. Being a transformer-based model, MetaLLM leverage the attention mechanism for capturing long-distance residual dependency which is crucial for identifying and distinguishing the binding sites. MetaLLM is limited to sequences in the range of 50 to 1000 amino acids. In the future, we envision to extend the model for longer sequences and compare our results with some of the existing tools for metal binding site prediction. Furthermore, we aim to include structural pipeline as a separate stream to validate the prediction from structural context.

References

1. Uniprot: the universal protein knowledgebase in 2023. Nucleic Acids Research **51**(D1), D523–D531 (2023)
2. Andreini, C., Bertini, I., Rosato, A.: A hint to search for metalloproteins in gene banks. Bioinformatics **20**(9), 1373–1380 (2004)
3. Aptekmann, A.A., Buongiorno, J., Giovannelli, D., Glamoclija, M., Ferreiro, D.U., Bromberg, Y.: mebipred: identifying metal-binding potential in protein sequence. Bioinformatics **38**(14), 3532–3540 (05 2022). https://doi.org/10.1093/bioinformatics/btac358, `https://doi.org/10.1093/bioinformatics/btac358`
4. Babor, M., Gerzon, S., Raveh, B., Sobolev, V., Edelman, M.: Prediction of transition metal-binding sites from apo protein structures. Proteins: Structure, Function, and Bioinformatics **70**(1), 208–217 (2008)
5. Bromberg, Y., Aptekmann, A.A., Mahlich, Y., Cook, L., Senn, S., Miller, M., Nanda, V., Ferreiro, D.U., Falkowski, P.G.: Quantifying structural relationships of metal-binding sites suggests origins of biological electron transfer. Science advances **8**(2), eabj3984 (2022)
6. Cheng, Y., Wang, H., Xu, H., Liu, Y., Ma, B., Chen, X., Zeng, X., Wang, X., Wang, B., Shiau, C., et al.: Co-evolution-based prediction of metal-binding sites in proteomes by machine learning. Nature Chemical Biology pp. 1–8 (2023)
7. Dreiseitl, S., Ohno-Machado, L.: Logistic regression and artificial neural network classification models: a methodology review. Journal of biomedical informatics **35**(5-6), 352–359 (2002)

8. Elnaggar, A., Heinzinger, M., Dallago, C., Rihawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., et al.: Prottrans: towards cracking the language of life's code through self-supervised deep learning and high performance computing. arXiv preprint arXiv:2007.06225 (2020)
9. Gucwa, M., Lenkiewicz, J., Zheng, H., Cymborowski, M., Cooper, D.R., Murzyn, K., Minor, W.: Cmm–an enhanced platform for interactive validation of metal binding sites. Protein Science p. e4525 (2022)
10. Guerois, R., Serrano, L.: The sh3-fold family: experimental evidence and prediction of variations in the folding pathways. Journal of molecular biology **304**(5), 967–982 (2000)
11. Haberal, İ., Oğul, H.: Deepmbs: Prediction of protein metal binding-site using deep learning networks. In: 2017 Fourth International Conference on Mathematics and Computers in Sciences and in Industry (MCSI). pp. 21–25. IEEE (2017)
12. Haberal, İ., Oğul, H.: Prediction of protein metal binding sites using deep neural networks. Molecular informatics **38**(7), 1800169 (2019)
13. He, W., Liang, Z., Teng, M., Niu, L.: mfasd: a structure-based algorithm for discriminating different types of metal-binding sites. Bioinformatics **31**(12), 1938–1944 (2015)
14. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
15. Jahromi, A.H., Taheri, M.: A non-parametric mixture of gaussian naive bayes classifiers based on local independent features. In: 2017 Artificial Intelligence and Signal Processing Conference (AISP). pp. 209–212 (2017). https://doi.org/10.1109/AISP.2017.8324083
16. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436–444 (2015)
17. Lin, C.T., Lin, K.L., Yang, C.H., Chung, I.F., Huang, C.D., Yang, Y.S.: Protein metal binding residue prediction based on neural networks. International journal of neural systems **15**(01n02), 71–84 (2005)
18. Lin, H., Han, L., Zhang, H., Zheng, C., Xie, B., Cao, Z.W., Chen, Y.Z.: Prediction of the functional class of metal-binding proteins from sequence derived physicochemical properties by support vector machine approach. In: BMC bioinformatics. vol. 7, pp. 1–10. BioMed Central (2006)
19. Lin, Y.F., Cheng, C.W., Shih, C.S., Hwang, J.K., Yu, C.S., Lu, C.H.: Mib: metal ion-binding site prediction and docking server. Journal of chemical information and modeling **56**(12), 2287–2291 (2016)
20. Lippi, M., Passerini, A., Punta, M., Rost, B., Frasconi, P.: Metaldetector: a web server for predicting metal-binding sites and disulfide bridges in proteins from sequence. Bioinformatics **24**(18), 2094–2095 (2008)
21. Lu, C.H., Lin, Y.F., Lin, J.J., Yu, C.S.: Prediction of metal ion–binding sites in proteins using the fragment transformation method. PloS one **7**(6), e39252 (2012)
22. Mendes, J., Guerois, R., Serrano, L.: Energy estimation in protein design. Current opinion in structural biology **12**(4), 441–446 (2002)
23. Mohamadi, A., Cheng, T., Jin, L., Wang, J., Sun, H., Koohi-Moghadam, M.: An ensemble 3d deep-learning model to predict protein metal-binding site. Cell Reports Physical Science **3**(9), 101046 (2022)
24. Passerini, A., Punta, M., Ceroni, A., Rost, B., Frasconi, P.: Identifying cysteines and histidines in transition-metal-binding sites using support vector machines and neural networks. Proteins: Structure, Function, and Bioinformatics **65**(2), 305–316 (2006)

25. Rao, R., Liu, J., Verkuil, R., Meier, J., Canny, J.F., Abbeel, P., Sercu, T., Rives, A.: Msa transformer. bioRxiv (2021). https://doi.org/10.1101/2021.02.12.430858, https://www.biorxiv.org/content/10.1101/2021.02.12.430858v1
26. Rao, R.M., Meier, J., Sercu, T., Ovchinnikov, S., Rives, A.: Transformer protein language models are unsupervised structure learners. bioRxiv (2020). https://doi.org/10.1101/2020.12.15.422761, https://www.biorxiv.org/content/10.1101/2020.12.15.422761v1
27. Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C.L., Ma, J., Fergus, R.: Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. bioRxiv (2019). https://doi.org/10.1101/622803, https://www.biorxiv.org/content/10.1101/622803v4
28. Rossi, F., Villa, N.: Support vector machine for functional data classification. Neurocomputing **69**(7-9), 730–742 (2006)
29. Schymkowitz, J.W., Rousseau, F., Martins, I.C., Ferkinghoff-Borg, J., Stricher, F., Serrano, L.: Prediction of water and metal binding sites and their affinities by using the fold-x force field. Proceedings of the National Academy of Sciences **102**(29), 10147–10152 (2005)
30. Shu, N., Zhou, T., Hovmöller, S.: Prediction of zinc-binding sites in proteins from sequence. Bioinformatics **24**(6), 775–782 (2008)
31. Shwartz-Ziv, R., Armon, A.: Tabular data: Deep learning is not all you need. Information Fusion **81**, 84–90 (2022)
32. Sodhi, J.S., Bryson, K., McGuffin, L.J., Ward, J.J., Wernisch, L., Jones, D.T.: Predicting metal-binding site residues in low-resolution structural models. Journal of molecular biology **342**(1), 307–320 (2004)
33. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
34. Ye, N., Zhou, F., Liang, X., Chai, H., Fan, J., Li, B., Zhang, J.: A comprehensive review of computation-based metal-binding prediction approaches at the residue level. BioMed research international **2022** (2022)
35. Yuan, Q., Chen, S., Wang, Y., Zhao, H., Yang, Y.: Alignment-free metal ion-binding site prediction from protein sequence through pretrained language model and multi-task learning. bioRxiv (2022)
36. Zhang, S., Li, X., Zong, M., Zhu, X., Wang, R.: Efficient knn classification with different numbers of nearest neighbors. IEEE Transactions on Neural Networks and Learning Systems **29**(5), 1774–1785 (2018). https://doi.org/10.1109/TNNLS.2017.2673241
37. Zhao, J., Cao, Y., Zhang, L.: Exploring the computational methods for protein-ligand binding site prediction. Computational and structural biotechnology journal **18**, 417–426 (2020)
38. Zhao, W., Xu, M., Liang, Z., Ding, B., Niu, L., Liu, H., Teng, M.: Structure-based de novo prediction of zinc-binding sites in proteins of unknown function. Bioinformatics **27**(9), 1262–1268 (2011)