Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research article

Predicting the location of coordinated metal ion-ligand binding sites using geometry-aware graph neural networks

Clement Essien [a,1], Ning Wang [b,1], Yang Yu [a], Salhuldin Alqarghuli [a], Yongfang Qin [a], Negin Manshour [a], Fei He [a,b,*], Dong Xu [a,**]

[a] *Department of Electrical Engineering and Computer Science, Bond Life Sciences Center, University of Missouri, Columbia, MO, USA*
[b] *School of Information Science and Technology, Northeast Normal University, Changchun, Jilin, China*

ARTICLE INFO

Keywords:
Metal ions
Protein structures
Binding sites
Graph neural network
Point cloud

ABSTRACT

More than 50 % of proteins bind to metal ions. Interactions between metal ions and proteins, especially coordinated interactions, are essential for biological functions, such as maintaining protein structure and signal transport. Physiological metal-ion binding prediction is pivotal for both elucidating the biological functions of proteins and for the design of new drugs. However, accurately predicting these interactions remains challenging. In this study, we proposed GPred, a novel structure-based method that transforms the 3-dimensional structure of a protein into a point cloud representation and then designs a geometry-aware graph neural network to learn the local structural properties of each amino acid residue under specific ligand-binding supervision. We trained our model to predict the location of coordinated binding sites for five essential metal ions: Zn^{2+}, Ca^{2+}, Mg^{2+}, Mn^{2+}, and Fe^{2+}. We further demonstrated the versatility of GPred by applying transfer learning to predict the binding sites of 2 heavy metal ions, that is, cadmium (Cd^{2+}) and mercury (Hg^{2+}). We achieved greater than 19.62 %, 14.32 %, 36.62 %, and 40.69 % improvement in the area under the precision-recall curve (AUPR) of Zn^{2+}, Ca^{2+}, Mg^{2+}, Mn^{2+}, and Fe^{2+}, respectively, when compared with 6 current accessible state-of-the-art sequence-based or structure-based tools. We also validated the proposed approach on protein structures predicted by AlphaFold2, and its performance was similar to experimental protein structures. In both cases, achieving a low false discovery rate for proteins without annotated ion-binding sites was demonstrated. © 2017 Elsevier Inc. All rights reserved.

1. Introduction

Proteins carry out their functions by interacting with other molecules [1]. Greater than 50 % of proteins bind to small metal ions to regulate their biological functions and stabilize their structures [2]. Among them, coordination bonds represent the case where an ion is directly coordinated by several nearby protein atoms, typically involving the electron pair donors such as the oxygen of carboxyl groups, or the sulphur of cysteine residues. The coordination bonds between ions and proteins often play more important functional roles than ionic bonds from electrostatic attractions between charged ions and oppositely charged amino acid residues. For example, Zn^{2+} ions are important for nucleases and transcription factors, enabling them to recognize DNA and RNA and regulate gene expression [3]. Hemoglobin requires Fe^{2+} to transport oxygen in the blood of vertebrates and some invertebrates [4]. Mn^{2+} is required by glycosyltransferases as a cofactor for synthesizing proteoglycans and mucopolysaccharides, which are necessary for bone and cartilage production [5].

Accurately identifying coordinated ion-binding sites is important because it helps to explain protein functions and is useful in developing novel drugs. Unfortunately, detecting coordinated ion-binding sites using experimental methods, such as nuclear magnetic resonance [6] and particle-induced X-ray emission (PIXE) [7] is both time-consuming and expensive. Consequently, there is a need to develop computational tools that can accurately predict coordinated protein-ion binding sites. The computational problem is formulated as follows: given an ion type and a protein, predict which residues, if any, can bind to the ion. Currently, there are 2 main computational approaches to addressing this problem: sequence-based and structure-based methods. Sequence-based methods analyze and learn patterns of residue conservation through

[*] Corresponding author at: Department of Electrical Engineering and Computer Science, Bond Life Sciences Center, University of Missouri, Columbia, MO, USA.
[**] Corresponding author.
E-mail addresses: hefe@missouri.edu (F. He), xudong@missouri.edu (D. Xu).
[1] Contributed to this work equally.

https://doi.org/10.1016/j.csbj.2024.12.016
Received 26 September 2024; Received in revised form 15 December 2024; Accepted 20 December 2024
Available online 21 December 2024

sequence-derived features. For example, IonSeq [8] and TargetS [9] use a sliding-window strategy to extract information on evolutionary conservation, predicted secondary structure, and ligand-specific binding propensity from the sequence context. These methods then utilize a support vector machine to learn local binding patterns for predicting metal and acidic radical ion-binding sites. TargetATPSite [10], NsitePred [11], and ZincBinder [12] make use of the same features. ZinCaps [13] implements a deep learning framework based on a Capsule Network to predict zinc-binding sites from quantitative descriptors obtained from the multidimensional scaling of the 237 physical-chemical properties of raw protein sequences. Because they are solely based on protein sequences, their performance may be limited due to the absence of structural information. Tertiary structures can provide more geometric information related to ion binding, for example, pockets or cavities in the protein structure. Furthermore, some sequence patterns are not directly related to ion binding; for example, many nonbinding residues are often conserved due to their role in maintaining protein structure or function unrelated to ion binding.

Structure-based methods use the 3-dimensional 3-D Cartesian coordinates of protein atoms as input, and they tend to be more accurate than their sequence counterparts. They can be classified into 3 categories: template-based, machine-learning-based, and hybrid. Template-based methods, like MIB [14], utilize alignment algorithms to transfer structural information from templates to infer binding sites. However, these methods can be severely limited when a high-quality template is unavailable. Machine-learning-based methods for protein structures extract geometric features and then train neural networks. For example, DELIA [15] treats protein structures as 2-D images and uses a convolutional neural network to extract features from protein distance matrices. Alternatively, methods such as GraphBind [16] encode protein structures as graphs and use graph neural networks (GNNs) to learn local tertiary patterns for binding site prediction. And GASS-Metal [17] presents a method using genetic algorithms to find candidate metal-binding sites structurally similar to curated templates from M-CSA and MetalPDB databases. However, GraphBind constructs its graphs using the centroids of residues instead of atomic coordinates. These pseudo-positions can distort learned structural features and lead to misrecognition of ion-binding sites. In our proposed method, GPred, protein structures are modeled directly with the real coordinates of the atoms, thus representing the structural information more accurately to obtain finer-grained features. In addition, we make use of the attention mechanism to better model the interactions between individual atoms in the protein. Hybrid methods, such as COACH [18], IonCom [8] and Master of Metals [19], integrate template-based and machine-learning-based methods simultaneously, and they are also limited by the availability of structure templates.

Although several existing structure-based methods tend to have better performance than sequence predictors, they are time-consuming due to the complexity of 3-D computations, the search for templates, energy evaluations, and molecular dynamics simulations. This could sometimes take up to an hour Central Processing Unit (CPU) based machine for a typical protein, and the time increases exponentially with the protein length. This makes such tools hard to use, especially in situations that require running the tool on many sequences in an input file. Meanwhile, most existing methods are designed for medium- to large-sized ligands and are not optimal for small ligand prediction, such as metals. Small ions have more versatile and flexible interactions with proteins than larger ligands [20]. Additionally, current methods have been developed using generic approaches based on geometry that do not discriminate against the different physicochemical characteristics of ligand types. Therefore, ligand-specific features should be used to improve small-ion binding recognition.

The 3-D structures of proteins play an essential role in regulating binding between proteins and their partners. For instance, the geometry of cavities on protein surfaces can facilitate the binding of specific drug targets [6]. Residues around an ion may form a specific local geometric pattern that is informative for binding residue recognition [16]. Such a pattern is a good indicator to recognize specific binding partners of the protein. In this work, we aim to learn the local structural properties under specific ligand-binding supervision, which leads to adaptive descriptors toward the ligand-binding pockets. In our proposed architecture, we transformed the 3-D coordinates of protein atoms into a point cloud and calculated spatial geometric features into the message-passing operation of a GNN to efficiently capture the structural patterns from proteins potentially interacting with metal ions. This approach enables the utilization of proteins' 3-D coordinates to extract intricate geometric properties, including angles and concavity. When metal ions bind to proteins, the surrounding amino acid residues typically adopt specific geometric configurations, represented by tetrahedral, octahedral, or hexahedral patterns. These patterns are crucial for the stability and activity of metal ions. They combine with ligand-specific physiochemical features to effectively reflect the local environment of proteins, thereby enhancing the accuracy of physiological coordinated metal ion binding site prediction.

2. Materials and methods

2.1. Data collection and preprocessing

In this study, we focus on predicting the binding sites of 7 small metal ion ligands, namely, Zn^{2+}, Ca^{2+}, Mg^{2+}, Mn^{2+}, Fe^{2+}, Cd^{2+}, and Hg^{2+}. Among these, 5 (Zn^{2+}, Ca^{2+}, Mg^{2+}, Mn^{2+}, Fe^{2+}) are frequently encountered in the literature and are relatively abundant in living organisms as trace elements. The remaining two, Cd^{2+} and Hg^{2+}, are nonessential metals for living organisms, with excessive levels posing potential toxicity [21]. To develop and evaluate our model, we sourced proteins with annotated binding sites for the 5 essential metal ions from the BioLiP database [22] and proteins associated with the 2 nonessential metal ions from the MetalPDB database [23,24]. To prevent overestimation of accuracy in our assessment, we employed the CD-HIT tool to reduce data redundancy between the training and test sets, ensuring that any 2 proteins across these sets shared less than 40 % sequence identity [25]. Furthermore, we excluded database entries shorter than 50 residues. We retained only those protein chains that featured at least 1 ion-binding site. After these steps, a total of 5393 protein chains were assembled for the training and testing phases of our model. During training, proteins in the training set were distributed into 5 groups, adopting a five-fold cross-validation strategy, where 4 groups served as the training set and the remaining group as the validation set for each fold. Detailed information on the training and testing data is summarized in Table 1.

Table 1
Brief description of the benchmarking datasets in this study.

Ligand type	Dataset	No. of proteins	No. of binding residues	No. of nonbinding residues
Zn^{2+}	Train	388	2329	11,030
	Test	175	741	6700
Ca^{2+}	Train	1048	4003	84,668
	Test	263	988	18,731
Mg^{2+}	Train	1009	2186	58,483
	Test	253	507	14,368
Mn^{2+}	Train	372	1194	19,297
	Test	94	255	4375
Fe^{2+}	Train	452	1495	21,070
	Test	85	256	4425
Cd^{2+}	Train	630	2480	20,884
	Test	160	613	5256
Hg^{2+}	Train	395	891	16,771
	Test	69	168	3327

2.2. Data encoding

Our method captures the intricate 3-D architectures of proteins by leveraging point cloud technology. We generate representations that comprise numerous points within a 3-D matrix, thus effectively illustrating an object's shape. Every point in this representation is loaded with comprehensive data, including spatial coordinates, which renders a point cloud to catalog the complex 3-D configurations of proteins. Our strategy involves constructing point clouds of protein structures at the atomic scale, given that data on this scale provides deeper structural insights than basic residue data, enabling the prediction models to become more sensitive and accurate.

To enhance our model's capability to intricately learn the nuanced interplay between the structures of proteins and their functional roles, we have not only integrated the 3-D spatial coordinates of the protein atoms but also enriched the point feature with biochemical attributes. Initially, leveraging RDKit [26], we distilled 5 fundamental physicochemical characteristics of atoms. These essential attributes of each atom were then encoded into a 39-dimensional 1-hot vector, describing its atomic properties across 23 types of atoms, 6 types of atomic connection degrees, 5 types of atomic charges, 4 types of atomic chirality, and 1 binary value for atomic aromaticity, providing a foundation for understanding the complex characteristics and molecular behaviour of proteins. Furthermore, we integrated evolutionary attributes of protein residues—namely, position-specific scoring matrices (PSSMs) [27]. PSSM enables the identification of pivotal residues instrumental to the structure and functionality within protein families. By executing PSI-BLAST searches against the Swiss-Prot database [28] and completing 3 iterations, we extracted a 20-dimensional evolutionary feature vector for each residue. These residue-specific properties were then assigned to their respective atoms, enriching them with comprehensive atomic features, thereby outfitting each atom within the point cloud with an elaborate 59-dimensional feature vector. Fig. 1 delineates the entire methodology, from harvesting atomic data and encoding features to incorporating evolutionary insights, thereby establishing a robust framework for a thorough analysis of protein structures and functionalities.

2.3. Model architecture

Our model architecture comprises 4 modules: point neighborhood grouper, Point Transformer, residual pooler, and classifier, as shown in Fig. 2.

In the analysis of point clouds of protein structures, the process of delineating each point's immediate surroundings, termed as **point neighborhood grouper**—equipped us with the precision needed to capture the nuances of local geometry and structure. By employing a ball query strategy [29], we constructed a local graph for each point, pinpointing neighboring points within a predetermined radius *r*. This method enhanced the simulation of interactions among adjacent atoms, thus accurately depicting the geometric and physicochemical dynamics at play between binding sites and their immediate structural context. For this study, we settled on a radius of 5 Å, reflecting the typical span of interactions observed among atoms or molecules [30], to ascertain adjacency relations. Upon detecting the local structural features, the Point Transformer layer, which harnesses the principles of self-attention alongside information propagation techniques [31], was adept at assimilating the embeddings that encapsulate atomic properties, geometric attributes, and evolutionary data.

The **Point Transformer** adds a position encoding δ to both the attention vector γ and the transformed features α, and is defined as Eq. (1) for each point *i* and its point feature X_i:

$$y_i = \sum\nolimits_{x_j \in \chi(i)}^{\rho} \rho\left(\gamma\left(\varphi(X_i) - \psi(X_j) + \delta\right)\right) \odot \left(\alpha(X_j) + \delta\right) \tag{1}$$

Here, the subset $X(i) \subseteq X$ is a set of local neighboring points of X_i. ϕ, ψ, andα are linear projectors to map raw point features into embedding. The mapping function γ is an MLP with 2 linear layers and 1 ReLU layer, which produces attention scores to differentiate the contributions from different neighboring points. ρ is a layer of normalization function, such as SoftMax [31]. $\odot$ is a dot-product operator used to apply the learned attention scores to integrate the embeddings of neighboring points for the center point. This message-passing mechanism allows the embedding of the center point to capture indirect effects by aggregating information from neighboring residues. This ensures that the residual embeddings account for indirect interactions within the context of the overall structure, including non-candidate amino acids. Furthermore, because the geometric properties of protein structures define their coordinated ion-binding sites [32], we leveraged the geometric encoding function δ as shown in Eq. (2) to encode local concavities, density, and torsion angles of proteins:

$$\mathrm{\delta_{ij}} = \mathrm{h_\theta}\left(\mathrm{P_i} - \mathrm{P_j}\right) + \mathrm{\lambda_{ij}} \tag{2}$$

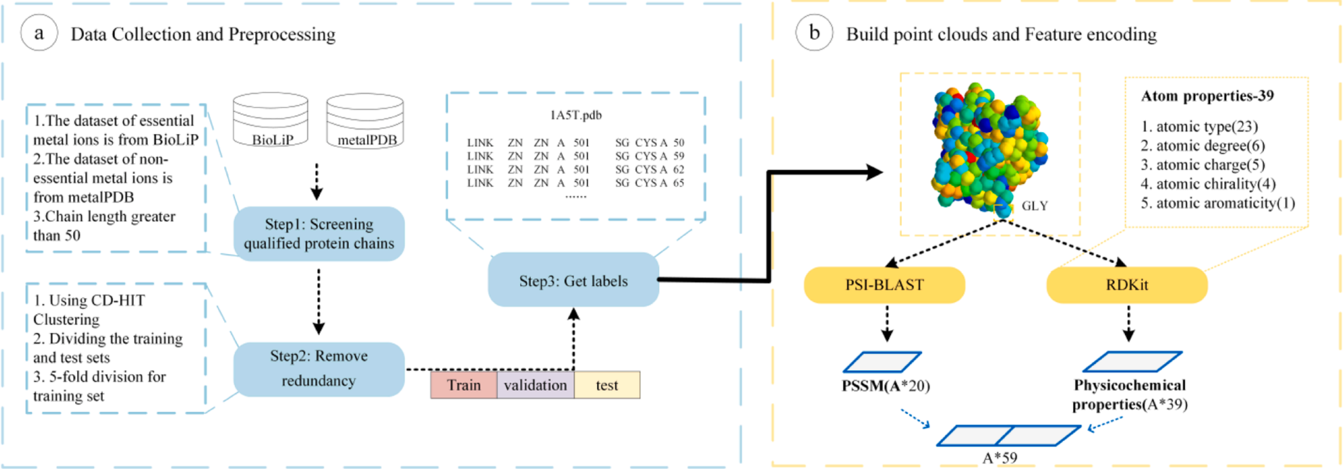

a
Data Collection and Preprocessing
1.The dataset of essential metal ions is from BioLiP
2.The dataset of non-essential metal ions is from metalPDB
3.Chain length greater than 50
BioLiP
metalPDB
Step1: Screening qualified protein chains
1. Using CD-HIT Clustering
2. Dividing the training and test sets
3. 5-fold division for training set
Step2: Remove redundancy
Train
validation
test
1A5T.pdb
LINK ZN ZN A 501 SG CYS A 50
LINK ZN ZN A 501 SG CYS A 59
LINK ZN ZN A 501 SG CYS A 62
LINK ZN ZN A 501 SG CYS A 65
......
Step3: Get labels
b
Build point clouds and Feature encoding
GLY
Atom properties-39
1. atomic type(23)
2. atomic degree(6)
3. atomic charge(5)
4. atomic chirality(4)
5. atomic aromaticity(1)
PSI-BLAST
RDKit
PSSM(A*20)
Physicochemical properties(A*39)
A*59

Fig. 1. Data preparation and feature encoding. (a) Initially, we collected datasets of critical and non-critical metal ions from protein sequences exceeding 50 amino acids in length from the BioLip and MetalPDB databases. In the second step, we used the CD-HIT tool to remove redundant entries and then divided the data into training and test sets under fivefold cross-validation. Additionally, we extracted related ion-binding annotations from the LINK field in PDB files. (b) We constructed 3-D point clouds to represent involved proteins at the atomic level using a rich 59-dimensional atomic feature derived from PSI-BLAST and RDKit, resulting in a 20-dimensional PSSM descriptor and a 39-dimensional atomic descriptor. "A" represents the total number of atoms.

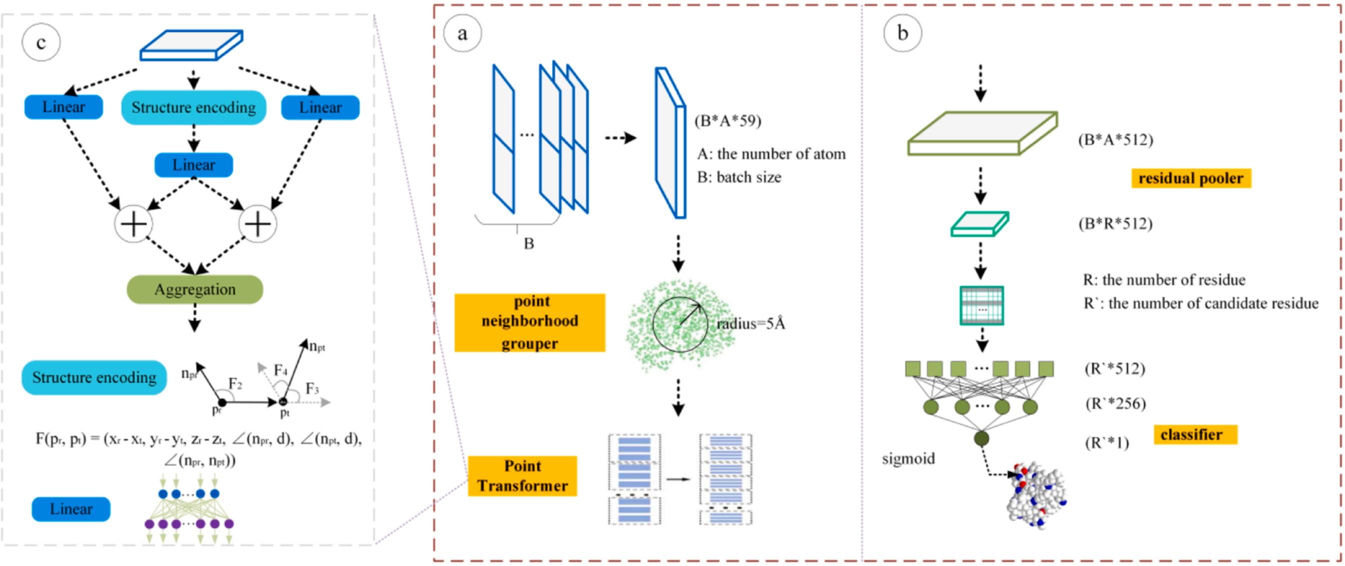


Fig. 2. Model Architecture. (a) To accurately represent the micro-environment for each point (atom) in our constructed point cloud, a ball query with a radius of 5 Å was conducted to channel neighboring points into a Point Transformer layer, which learns both the physical and geometric nuances of the structure. (b) The learned atomic embeddings were pooled into residual representations for further binding-site prediction. To exclude nonbinding-site candidates, a masking strategy was applied, followed by a sigmoid layer to classify each candidate amino acid. (c) The illustration shows how the Point Transformer learns structural representations through the coordinates of indexing points and their neighboring points.

$$\lambda_{ij} = \left(\angle(n_i, d), \angle(n_j, d), \angle(n_i, n_j)\right) \tag{3}$$

where $P \in R^{N*3}$ denotes the coordinates of each point and function h_θ generates the distance encoding between any neighboring point pair i and j. $\angle$ is their torsion angles. d denotes their Euclidean distance, and n_i and n_j are their normal of surface, respectively.

Hence, a **residual pooler** using the MaxPooling layer aims to bridge this granularity gap by upgrading the embeddings from atomic to residue level. This strategic elevation not only deepens our insight into molecular interactions, but the **classifier** distinguishes residues as either coordinated binding sites or nonbinding sites. Specific coordinated metal ions exhibit clear affinity for certain types of amino acids. The specific information is shown in Table 2. For instance, zinc ions tend to bind with amino acids C, H, E, and D but not with other types of amino acids. Leveraging this characteristic, we defined specific amino acids as candidate residues [33] for each metal ion while treating all other amino acids as non-candidates. In the GPred model, we masked non-candidate amino acids in the loss function. This masking operation is designed to minimize the risk of irrelevant information interfering with the model's ability to distinguish between positive and negative candidate pairs. By doing so, the approach enables the model to focus on the key candidate residues that directly bind metal ions, while still allowing for the incorporation of indirect effects from other residues through the message-passing mechanism. Using a fully connected layer with a Sigmoid activation function, we calculated prediction scores for all candidate residues and classified them based on these scores. We used a classification threshold (0.5 by default). If no candidate residues in a particular protein achieved a higher predicted probability than the threshold, GPred would yield no binding site outputs. The architecture of the model is presented in Fig. 2.

Table 2
Candidate residues for each ion.

Ligand type	Candidate residues
Zn^{2+}	C\H\E\D
Ca^{2+}	D\E\G\N
Mg^{2+}	D\E\N
Mn^{2+}	D\E\H
Fe^{2+}	D\E\H
Cd^{2+}	C\H\E\D
Hg^{2+}	C\H\E\D

2.4. Model training

In the training phase of the model, acknowledging the predominance of negative samples over positive ones within our dataset, we opted for a weighted binary cross-entropy (BCE) loss function to mitigate this imbalance. By fine-tuning the weight accorded to positive samples within this loss function, we aimed to equilibrate the skewed distribution of positive and negative samples. BCE loss is represented in Eq. (4):

$$l_2(x, y) = L_2 = \{l_{1,2}, \ldots, l_{N,2}\}^T \tag{4}$$

$$l_{n,2} = -w_{n,2}\left[p_2 y_{n,2}\cdot \log\sigma(x_{n,2}) + \left(1 - y_{n_2}\right)\cdot \log\left(1 - \sigma(x_{n,2})\right)\right] \tag{5}$$

$$\sigma(x) = \frac{1}{1 + e^{-x}}, \tag{6}$$

where x is the predicted value, y is the actual value, N is the batch size, n is the number of samples in the batch, and p_2 is the weight of the positive samples. Adam optimizer [34] was used here to update the model parameters, whereas the Graphics Processing Unit (GPU) memory was utilized to accelerate the computation. After each training iteration, cross-entropy loss and the ROC-AUC (area under the curve) scores [12] were calculated on the validation set. If the loss and ROC-AUC scores on the validation set were not both improving during 20 consecutive epochs, we triggered an early stop to prevent model overfitting to the training data.

3. Results

In this segment of our study, we have detailed the performance metrics employed to evaluate our models, which include precision, recall, F1-Score, ROC-AUC, Matthew's correlation coefficient (MCC), and the area under the precision-recall curve (PR-AUC) [12,13]. These metrics provide a holistic view of our model's effectiveness, particularly in scenarios marked by class imbalance. We conducted a comprehensive evaluation by training 5 distinct models for the coordinated metal ions

Zn^{2+}, Ca^{2+}, Mg^{2+}, Fe^{2+}, and Mn^{2+} using a fivefold cross-validation approach. The results are expressed as the mean and variance of the performance metrics across all validation folds. For the calculation of precision, recall, F1 score, and MCC, we adopted a threshold that optimizes the F1 score for each validation fold, ensuring that our model's performance is robustly gauged.

3.1. Performance of GPred on 5 essential metal ions

Within the framework of our implemented fivefold cross-validation study, the performance of GPred on 5 essential metal ions (Zn^{2+}, Fe^{2+}, Ca^{2+}, Mg^{2+}, Mn^{2+}) is described as shown in Table 3. This dataset reveals a comprehensive and promising array of performance indicators for these ion types. Notably, the model demonstrates exceptional performance in the classification of Zn^{2+} and Fe^{2+} ions. It achieves high recall, precision, and F1 score, which reflects its robust ability to accurately identify the coordinated binding sites of these ions while maintaining a low false positive rate. Additionally, for these 2 categories, the model performs remarkably well in terms of MCC, ROC-AUC, and PR-AUC, underlining its excellent robustness. However, it is important to note that the performance of Ca^{2+} and Mg^{2+} ions appear relatively lower than other ions. This is potentially due to that the preference of these cations for coordination by backbone oxygen atoms, which are not sensitive to amino acid variations, thus lowering the information content of PSSMs.

It is worth mentioning that, for each ion, the variability among the 5 models in various metrics is extremely low. These metrics, including recall and precision, changed by less than 0.1 in most cases, whereas changes in the F1 score and MCC were all less than 0.06. As for the composite metrics such as ROC-AUC and PR-AUC, which did not vary with the threshold, their range of variation was between 0.005 and 0.07. These data adequately demonstrate the high robustness and stability of our algorithm.

Fig. 3 shows the zinc-coordinated binding sites identified by GPred, located at the interior, edge, and exterior of the protein. This visualization demonstrates GPred's discerning power across scenarios.

3.2. Hyperparameter optimization

In the point cloud representation of GPred, the local environment of a point is defined by the ball area around it, determined by radius size. To identify the most effective ball query radius for generating edges, we conducted a series of experiments with datasets related to a representative ion Zn^{2+}. We tested 3 different radii—3 Å, 5 Å, and 7 Å—to explore the effect of aggregating neighboring information across different ranges. According to the results in Table 4, the performance metrics AUC and AUPR on the validation set are optimal at a 5 Å radius. It was observed that performance metrics decrease when the radius is less than or greater than 5 Å, revealing a critical insight: the optimal radius of 5 Å is essential for obtaining the local environment information necessary for efficient model performance. As shown in Fig. 4, we also experimented with multiple grouping using 2 different radii to explore the potential benefits of aggregating atomic information from different ranges. Interestingly, this multi-grouping approach did not result in better performance. Potentially such aggregation of multiple grouping neighboring points (atoms) would over-smooth the discriminations among points (atoms). Based on these observations, we confirmed 5 Å as the optimal ball query radius and kept the single grouping strategy in our experiments.

Table 3
Performance and variance of GPred on 5 essential metal ions from fivefold cross validation. There are five models for each of the five ions, totaling twenty-five models. The table presents the average performance and variance of GPred.

Ion	Recall	Precision	F1	MCC	ROC-AUC	PR-AUC
Zn^{2+}	0.771 ± 0.025	0.803 ± 0.004	0.786 ± 0.014	0.763 ± 0.015	0.957 ± 0.005	0.823 ± 0.024
Fe^{2+}	0.852 ± 0.057	0.861 ± 0.050	0.835 ± 0.048	0.826 ± 0.051	0.987 ± 0.006	0.861 ± 0.051
Mn^{2+}	0.605 ± 0.070	0.529 ± 0.037	0.560 ± 0.013	0.536 ± 0.013	0.892 ± 0.005	0.518 ± 0.019
Ca^{2+}	0.528 ± 0.136	0.529 ± 0.135	0.501 ± 0.055	0.489 ± 0.043	0.901 ± 0.010	0.511 ± 0.070
Mg^{2+}	0.538 ± 0.033	0.504 ± 0.016	0.520 ± 0.015	0.503 ± 0.016	0.887 ± 0.006	0.444 ± 0.026

3.3. Comparison with state-of-the-art tools

We compared GPred with 2 sequence-based methods (TargetS and ZinCaps) and 4 structure-based methods (IonCom, DELIA, MIB, and GraphBind), which are all the current runnable related tools to our best knowledge. In this section, to enhance the practicality of GPred, we merged the models for each ion from fivefold cross-validation into a comprehensive predictive model by a major voting strategy, facilitating further research. Under this voting strategy, for each candidate residue, its final classification is determined by the majorly predicted category from the 5 models. This integration approach aims to leverage the predictive strengths from fivefold cross-validation models. ZinCaps supports only Zn^{2+}, whereas TargetS, DELIA, and GraphBind do not support Fe^{2+}. Table 5 presents the performance of GPred, now integrated into a single model, compared with 6 competitors, along with the types they support. For Zn^{2+}, GPred has surpassed ZinCaps by 17 % and 49.02 % in F1 and MCC, respectively, and surpassed IonCom by 32.72 %, 31.55 % in F1 and MCC, respectively. As such they are both the next top performers for Zn^{2+} ion. For Ca^{2+}, we see a 13.55 % and 13.99 % improvement in the F1 score when compared to GraphBind, which is the next best performer in the calcium ion category. For Mg^{2+}, we see a 49.43 % and 26.70 % improvement in F1 score and MCC, respectively, compared with GraphBind, which is the next top performer for magnesium ions. The limited formation of protein-ion binding complexes involving s-block metals like Mg and Ca could be a contributing factor to GPred's poorer performance with these ions. The relatively smaller dataset for these metals may have resulted in insufficient training for GPred. Although for Mn^{2+}, there is a clear outperformance in the F1 score and MCC by 21.29 % and 21.27 %, respectively, when compared to TargetS. Finally, our method significantly outperforms IonCom and MIB by 57.35 % and 84 %, respectively, in the F1 score for Fe^{2+}. The superior performance of GPred is further represented in the ROC curve in Fig. 5, where the ROC curves for GPred are consistently located at the upper part of the figures. Such improvements are beneficial; GPred not only incorporates sequence features but also places a heightened emphasis on structural features to capture the interactions among individual atoms within proteins.

GPred currently runs on the command line and is computationally more efficient compared to other methods; on average it takes about 6 seconds to make a prediction for each protein chain, whereas GraphBind and IonCom take 12 seconds and more than 252 seconds, respectively, to accomplish the same task.

3.4. GPred Scalability on New Ion-binding Site Prediction

Next, we extended GPred to other types of coordinated ion-binding site prediction. Retraining a well-performing model for a new task hypothesizes that the new task shares some features with pretrained tasks. Considering elements in the same group of the periodic table have the same valence electrons, leading to reacting similarly with other elements [35], we treated the model for Zn^{2+} as a base model and fine-tuned it on the data of Cd^{2+} and Hg^{2+}, which are in the same family group of Zn^{2+}. Due to relatively smaller training datasets with Cd^{2+} and Hg^{2+}, we froze Point Transformer layers and tuned the model parameters in the classifier. The learning rates for Cd^{2+} and Hg^{2+} were reduced to 0.0001 for guaranteeing the model convergence. As shown in Table 6, the average ROC-AUC of the model after transfer learning on Cd^{2+} and

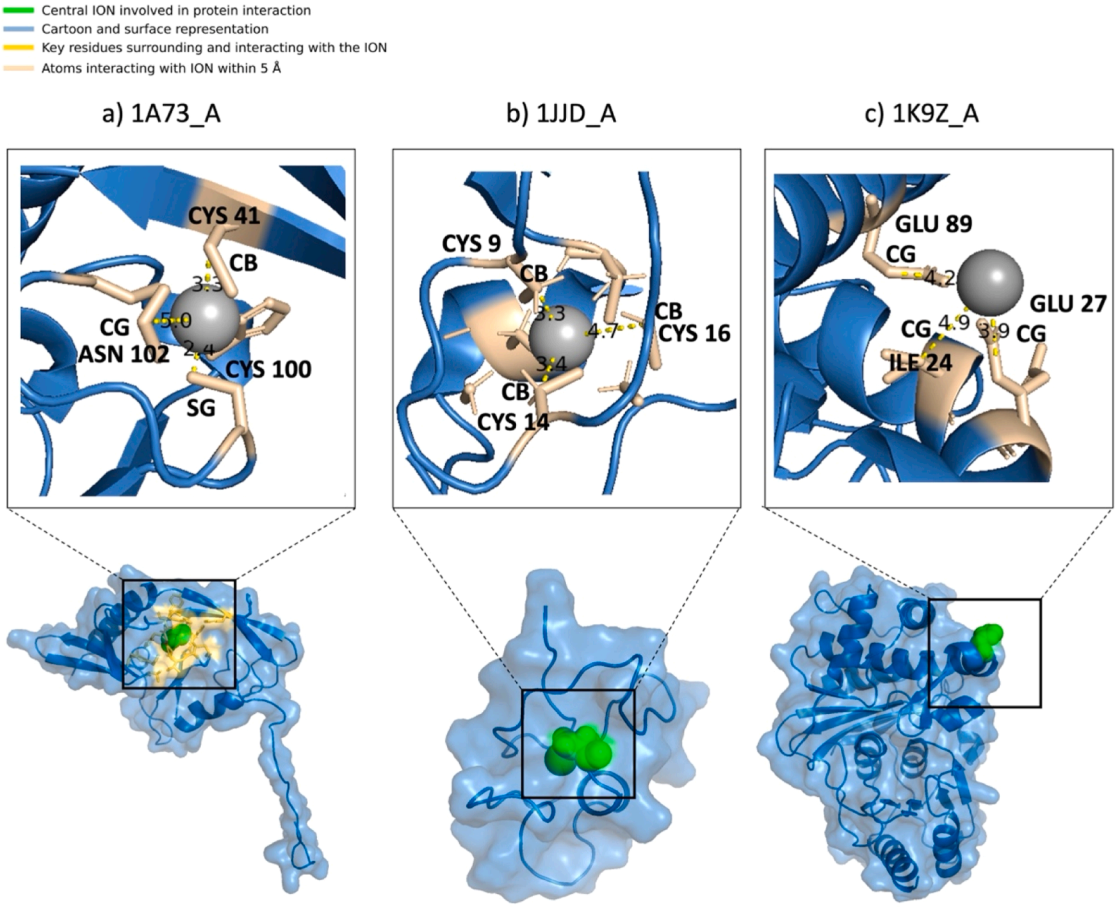

Central ION involved in protein interaction
Cartoon and surface representation
Key residues surrounding and interacting with the ION
Atoms interacting with ION within 5 Å
a) 1A73_A
CYS 41
CB
3.3
CG
5.0
ASN 102
2.4
CYS 100
SG
b) 1JJD_A
CYS 9
CB
3.3
4.7
CB
CYS 16
3.4
CB
CYS 14
c) 1K9Z_A
GLU 89
CG
4.2
GLU 27
CG
4.9
3.9
CG
ILE 24

Fig. 3. Examples of GPred predictions of coordinated zinc-binding sites at different locations of proteins. Our tool, GPred, can accurately predict coordinated zinc-binding sites located at the interior residue C of protein 1A73-A (a), the edge residue C of protein 1JJD-A (b), and the exterior residue H of protein 1K9Z-A (c), where the predicted binding sites were marked in green, their neighboring residues at surface were colored in yellow and other residues at surface were plotted in blue. The Zn^{2+} ions, depicted as grey spheres, are positioned based on the annotations from the PDB file 1A73 for visualization purposes. Their locations were not predicted by GPred.

Table 4
The results from different grouping radius settings.

Radius			ROC-AUC	PR-AUC
3A	5A	7A		
✓			0.935	0.770
	✓		0.957	0.823
		✓	0.930	0.774
✓	✓		0.953	0.810
	✓	✓	0.940	0.793

Hg^{2+} binding sites reached 0.805 and 0.919, whereas the MCC reached 0.351 and 0.546, respectively. The promise indicates the scalability of GPred to other relevant ion-binding site prediction problems. Hg^{2+} has a coordination chemistry radically different from the other six metal ions. Transfer learning is not a good choice since Hg(II) has unique coordinated chemistry properties than other mental ions. On the other hand, although cadmium shares chemical similarities with zinc, its performance in GPred is not comparable to that of zinc. We hypothesize this is because cadmium binds to proteins in a less specific manner than zinc, often causing structural instability. This binding diversity and the formation of unstable complexes make it more challenging for GPred to learn robust features.

3.5. Analysis of feature importance

In our study, to elucidate how features affect coordinated protein ion-binding site prediction, we utilized GNNExplainer [36], an innovative, model-agnostic technique that provides interpretability analysis of GNN models through feature perturbation analysis. Specifically, GNNExplainer identifies information-dense subgraphs that are highly relevant to predictions. The feature importance derived from GNNExplainer is shown in Fig. 6, illustrating the significance of PSSM and atomic type features as well as the contributions of structural features, atom degree, charge, chirality, and aromaticity to GPred’s performance. PSSM, which reveals the evolutionary conservatism toward coordinated metal ion-binding residues, contributes the most to prediction accuracy. And the atoms forming each amino acid determine its physio-chemical properties and its interaction capabilities with metal ions. The extensive atom types than other studies potentially enhanced the semantic capacity of representations of amino acids leading to outperformance of GPred. Moreover, the geometric features learned from our point cloud architecture contributed extra fine-grained information to improve the discerning power of GPred, compared to other sequence-based and structure-based competitive tools.

3.6. Performance comparison on experimental and AlphaFold2-predicted protein structures

Determining the experimental structure of a protein is a complex and

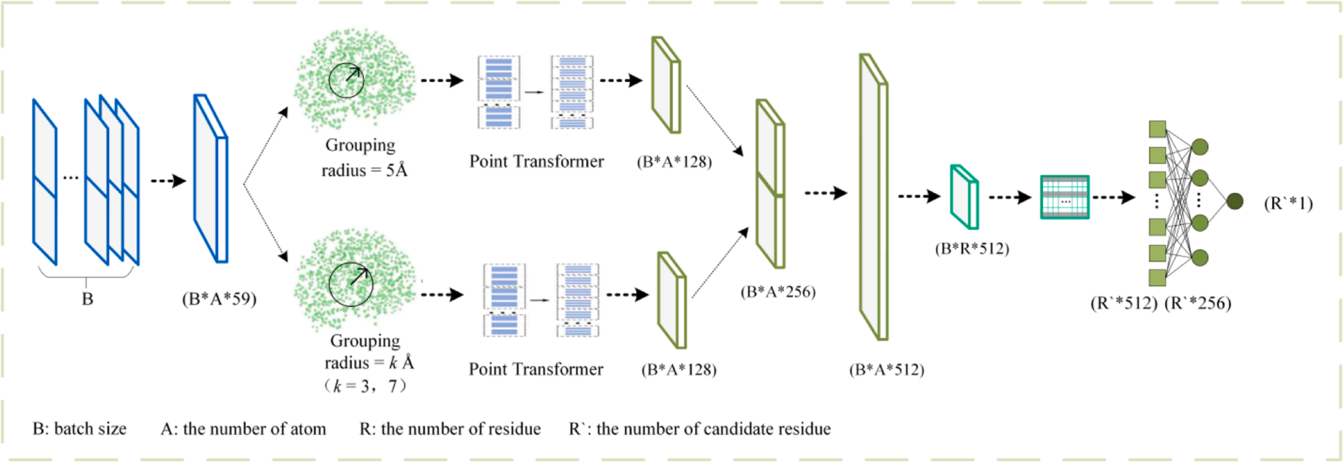


Fig. 4. Different hyperparameter settings in grouping radii. To explore the optimal definition of neighboring points, we designed a series of comparative experiments using different grouping radii and combining 2 options of grouping radii in 2 Point Transformers.

Table 5
Performance comparison of GPred with state-of-the-art methods. The bold font indicates the highest performance in the category.

Ion	Method	Recall	Precision	F1	MCC	ROC-AUC	PR-AUC
Zn^{2+}	MIB	0.794	0.205	0.326	0.386	0.913	0.404
	TargetS	0.514	0.700	0.593	0.580	0.856	0.608
	ZinCaps	0.725	0.620	0.668	0.512	0.919	0.669
	IonCom	**0.819**	0.128	0.222	0.318	0.919	0.688
	GPred	0.778	**0.846**	**0.810**	**0.797**	**0.957**	**0.823**
Ca^{2+}	MIB	0.432	0.057	0.101	0.122	0.765	0.107
	TargetS	0.155	0.358	0.216	0.215	0.768	0.169
	IonCom	0.380	0.180	0.245	0.233	0.694	0.172
	DELIA	0.220	0.463	0.298	0.293	0.781	0.258
	GraphBind	0.474	0.455	0.465	0.429	0.883	0.447
	GPred	**0.553**	**0.539**	**0.546**	**0.501**	**0.901**	**0.511**
Mg^{2+}	MIB	0.540	0.030	0.056	0.098	0.690	0.074
	TargetS	0.259	0.340	0.294	0.285	0.714	0.208
	IonCom	0.527	0.173	0.261	0.285	0.700	0.259
	DELIA	0.283	0.450	0.348	0.345	0.752	0.278
	GraphBind	**0.599**	0.287	0.388	0.397	0.795	0.325
	GPred	0.545	**0.564**	**0.525**	**0.533**	**0.887**	**0.444**
Mn^{2+}	MIB	0.456	0.071	0.122	0.157	0.778	0.139
	TargetS	0.267	0.365	0.309	0.294	0.796	0.267
	lonCom	0.504	0.180	0.266	0.279	0.768	0.252
	DELIA	0.495	0.489	0.492	0.465	0.831	0.405
	GraphBind	0.421	0.519	0.465	0.442	0.858	0.460
	GPred	**0.615**	**0.533**	**0.575**	**0.544**	**0.892**	**0.518**
Fe^{2+}	IonCom	0.601	0.379	0.465	0.581	0.868	0.612
	MIB	0.781	0.417	0.544	0.426	0.826	0.528
	GPred	**0.856**	**0.878**	**0.867**	**0.833**	**0.985**	**0.861**

time-consuming process, typically using experimental methods such as X-ray crystallography [37], Nuclear Magnetic Resonance (NMR) spectroscopy [6], and cryo-electron microscopy [38]. Additionally, some proteins are difficult to crystallize due to their size, complexity, flexibility, or membrane-bound nature. Hence, despite the vast number of protein structures determined, there remains a significant number of protein structures not fully solved. Although AlphaFold2 [39] provides predicted protein structures with high quality, its predicted structures might not always provide atomic-level accuracies, particularly for highly dynamic or disordered regions of a protein [40]. Considering that the training data for GPred were all from crystal protein structures, we compared the performance of GPred on crystal structures and AlphaFold2-predicted structures from the same protein set, validating the adaptation of GPred to AlphaFold2-predicted protein structures. Specifically, we collected newly released zinc-binding proteins ions in PDB [41] after 2021 to guarantee that they are independent of the previous training data and testing data of GPred. We used the advanced search functionality of the RCSB database. We included 'zn' as an additional structural keyword in the structure attributes, selected 'protein' for the Polymer Entity Type, and set the release date range from 2021 to 2024. Similarly, we also applied the redundancy removal process described in Section 2.1 to minimize the sequence similarity between these data and our training data. As a result, we collected a total of 671 crystal protein structures for this comparison. Subsequently, we fed the sequences of these 671 proteins to AlphaFold2 and selected the conformations with the highest pLDDT from all outputs from AlphaFold2 for each input protein. Thereby, we built a dataset with crystal structures and AlphaFold2-predicted structures for zinc-binding proteins.

We assessed the GPred model on zinc-binding site prediction on these protein structures. Fig. 7 provides the F1 scores from GPred on crystal structures versus AlphaFold2-predicted structures, depicting that most proteins in this set displayed consistent F1 scores using their crystal structures and predicted structures. However, there is still a small number of samples getting inconsistent F1 scores from GPred between their crystal structures and predicted structures. We investigated a case

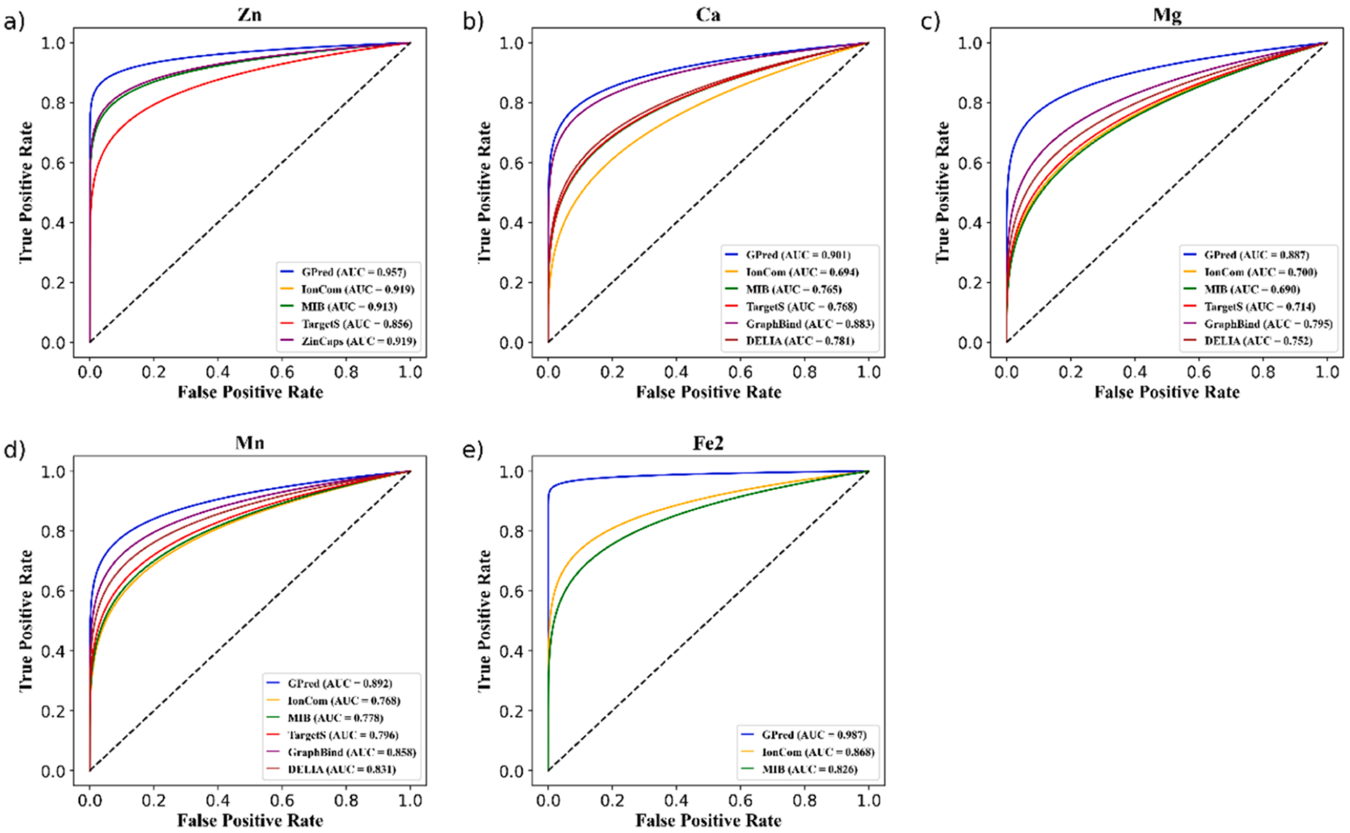


Fig. 5. Receiver operating characteristic curves for GPred and competitive tools.

Table 6
Performance comparison of GPred on the test sets.

Ion	Recall	Precision	F1	MCC	ROC-AUC	PR-AUC
Cd^{2+}	0.478	0.380	0.422	0.351	0.805	0.370
Hg^{2+}	0.696	0.451	0.554	0.546	0.919	0.435

study of 8AMY-A with an apparent gap between F1 scores from its crystal structure and predicted structure. In Fig. 7b, we first emphasize the significant overall differences between the crystal structure and the predicted structure. Following this, to delve deeper into the geometric differences at the binding sites of these 2 structures, we have extracted and presented the geometric features h_θ of the binding site (i.e., the distance encoding between neighboring atom pairs) within each structure. The geometric features exhibit notable differences as shown in Fig. 7b.

Section 3.5 discusses the importance of geometric features extracted from point cloud data. By analyzing this case, we can infer that the differences between crystal and predicted structures, especially the differences in geometric features at the candidate residues, might be the main reason for the inconsistency in F1 scores. Although there is inconsistency in F1 scores for a few samples, the consistency of F1 scores in the majority of samples validates the accuracy of the AlphaFold2 tool. This is remarkable because, unlike the native PDB structures, AlphaFold2-predicted structures do not contain ions. It demonstrates the effectiveness of geometric features in representing the local structure environment of an ion-binding site and the applicability of our GPred tool to AlphaFold2-predicted protein structures.

3.7. False discovery performance of GPred

Given that our training is anchored in annotated data, it becomes critical to assess the GPred tool's efficacy on unannotated datasets to unearth potential predictive inaccuracies. To this end, we curated a new dataset devoid of annotations for evaluation. Utilizing the advanced search capabilities offered by the RCSB official website, we filtered and downloaded protein structure files based on their publication date and the absence of any ion annotations. Following the procedures outlined in Sections 2.1 and 2.2, we ultimately compiled a dataset comprising 100 proteins for testing the false discovery rate (FDR) [42] of GPred. It is calculated at the protein level, with the formula as follows:

$$FDR_p = \frac{FP_p}{P_p} \tag{7}$$

where FP_p is the number of proteins with false predictions, and P_p is the number of all involved proteins. We treated the integrated Zn^{2+} model from fivefold cross-validation models as representative and observed an FDR_p of 0.07. This finding implies that out of 100 testing proteins, only 7 proteins had incorrectly predicted binding sites, and the remaining 93 proteins were accurately predicted without any zinc-binding sites.

We mixed this dataset with original Zn^{2+} data to form a testing dataset mimicking the realistic condition, and benchmarked on Zn^{2+} predictor of Gred, TargetS and ZinCaps. For this independent test dataset, GPred outperformed both TargetS and ZinCaps. Fig. 8 shows that GPred, exhibiting the highest AUC in both the ROC and Precision-Recall curves, suggests that it is not only accurate in ideal conditions—where annotations are known—but also maintains robust performance in less-defined scenarios.

4. Discussion and conclusion

The identification of coordinated metal ion-binding sites is essential for understanding protein functions and the design of novel drugs. Current state-of-the-art structure-based methods are inefficient because most proteins do not have reliable tertiary structures. Although

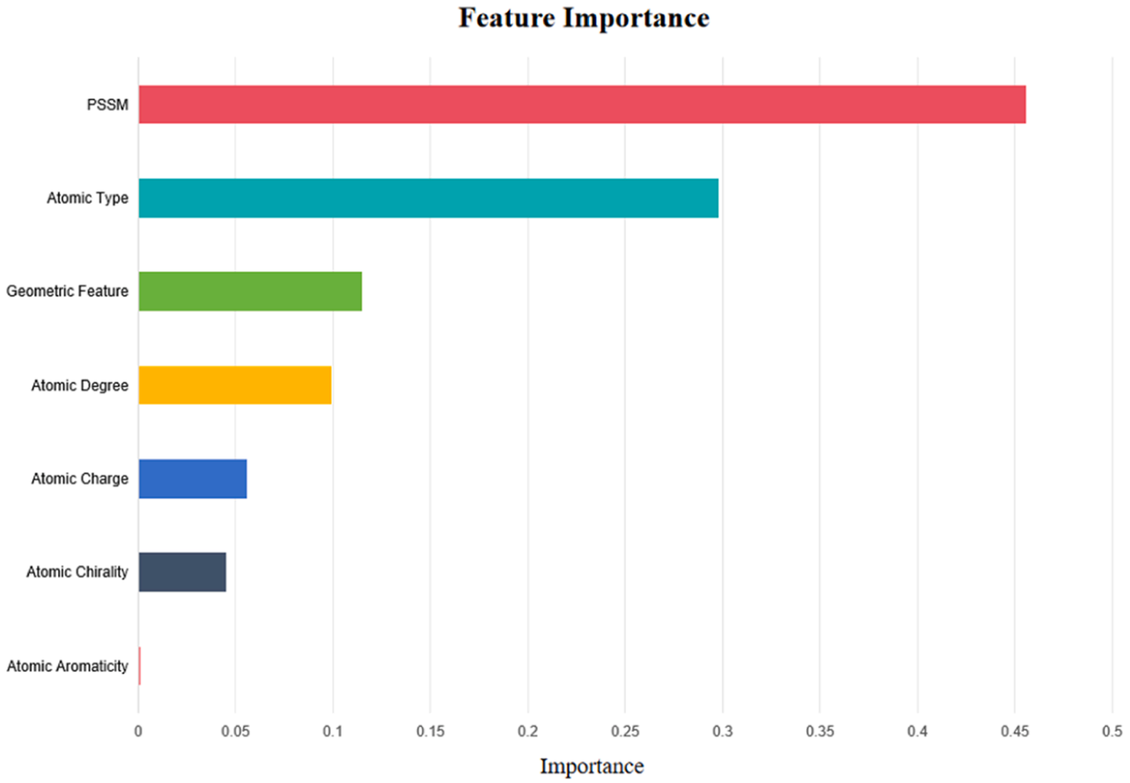
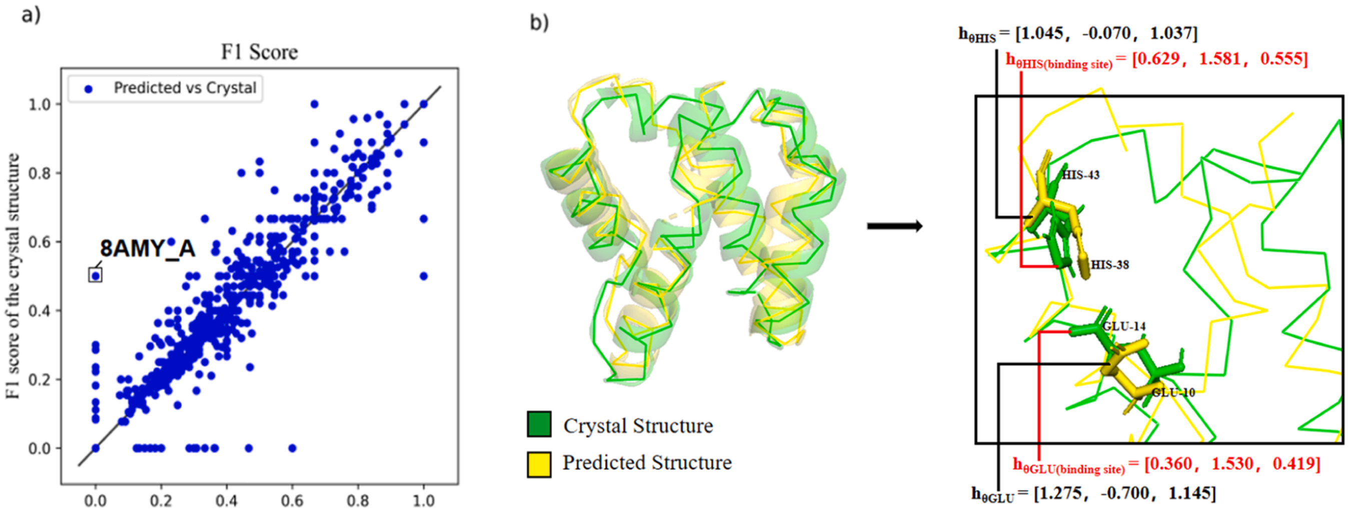


Fig. 6. Importance of each feature to the predictive performance.



Crystal Structure vs. Predicted Structure(8AMY_A)

Fig. 7. Difference in F1-Scores between the crystal structure and the AlphaFold2-predicted structures. (a): In most cases, the F1 scores between crystal structure and predicted structure are close, but there are also inconsistent samples, such as 8AMY_A. (b): In the case of 8AMY_A, the crystal structure is presented in green, whereas the predicted structure is presented in yellow. The geometric configuration of the binding sites is shown through a detailed local graph. There are significant differences in geometric features between the crystal structure and the predicted structure at the binding site. h_θ is the geometric function described in Section 2.3.

sequence-based are more practical for use, they have low precision. Most existing tools that adopt structure-based and sequence-based approaches are time-consuming because they rely on generating multi-sequence alignments.

In contrast, GPred is not only a computationally efficient tool but also has a strong predictive power. GPred utilizes a structure-based approach, representing protein 3-D structures in a point cloud format, enabling the matching of each point with precise and multifaceted feature information. Its utilization of a Point Transformer layer equipped with an attention mechanism facilitates the learning of comprehensive embeddings, encompassing crucial information such as atomic properties, geometric nuances, and evolutionary profiles—factors that significantly influence protein function. In GPred, each candidate residue aggregates its subordinate atoms to make separate predictions for a specific binding ion, enabling the model to identify multiple ion-binding sites within a single query protein.

Furthermore, GPred showcases remarkable performance even with small datasets, particularly targeting ions with limited labels.

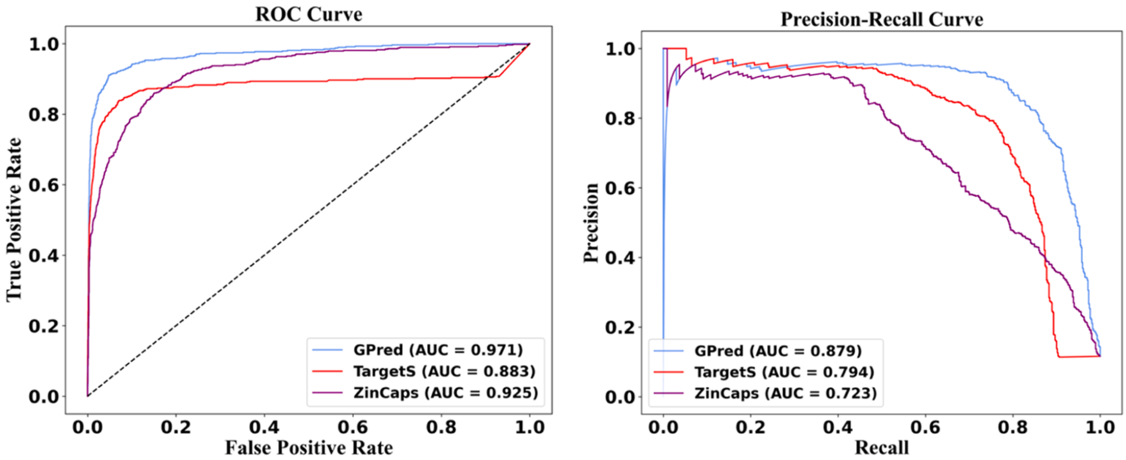


Fig. 8. Receiver operating characteristic and precision-recall curves for the GPred and competitive tools on the dataset mixing with non-binding site data.

Leveraging transfer learning, we achieved commendable results when training the model on such datasets. In addition, GPred can also be applicable in scenarios where experimental protein structure data is unavailable by using predicted structures from AlphaFold.

The definition of “True Negative” is debatable to build an ion-binding site predictor. The negatives in this study were defined by the candidate amino acids, which technically were unknown binding sites rather than actual non-binding sites. An interesting idea to address this issue is treating the annotated binding sites for other metals as negatives. We combined Fe, Mn, Ca and Mg binding sites as Zn predictor’s negatives to form a new independent testing set. This new dataset posed a larger challenge to GPred since such negatives hold some common features across various binding ligands, which may confuse GPred in prediction. In this independent test dataset, GPred demonstrated superior performance compared to both TargetS and ZinCaps as shown in Fig. 9. For ROC analysis, the AUC values were 0.945, 0.938, and 0.859 for GPred, TargetS, and ZinCaps, respectively. Similarly, GPred outperformed the other tools in the precision curve, with AUCs of 0.754 for GPred, 0.731 for TargetS, and 0.424 for ZinCaps. Through the results, we observed a performance drop of GPred, indicating that the different negative definitions revealed a limitation of GPred in learning ligand-specific features but GPred still outperformed other competitors with such rigorous negatives.

Controlling structural similarity between training data and testing data is also utilized to investigate possible overevaluation in this study. We applied TM-score to measure the structural similarity between any two proteins from training and testing sets. And then we removed the testing data with a higher TM-score than 0.5 compared to any training sample [43]. This process constructed a rigorous testing subset to evaluate GPred. We benchmarked GPred, TargetS, and ZinCap using this rigorous testing subset, as shown in Fig. 10. The AUC values for the ROC curves were 0.934, 0.807, and 0.827 for GPred, TargetS, and ZinCap, respectively. However, in the Precision-Recall curves, the AUCs for GPred, TargetS, and ZinCap decreased to 0.670, 0.552, and 0.405, respectively. These results suggest that GPred implicitly relies on structural similarity in its decision-making, indicating that there is still room for improvement in learning ion-binding site-specific features. Nevertheless, GPred maintains a clear advantage over the other two available ion-binding prediction tools.

In our previous experiments, the testing data excluded peptides, However, they usually exhibit biological activity in complexes with metal ions and frequently have stable structures. To explore the generalizability of GPred to peptides, we retrieved peptides with ion-binding annotations using the same data retrieval protocol we employed for proteins, but with the additional constraint that the peptides have sequence lengths of less than 50 amino acids. We then evaluated the five GPred models, each trained on a 5-fold ion-binding protein dataset, to predict potential ion-binding sites in the retrieved peptides. The performance of GPred on these peptides is summarized in Table 7: Upon analysing the performance on 15 peptides with 60 Zn^{2+}-binding sites, the results were consistent with the evaluations conducted on proteins. These findings suggest two key points: First, GPred is extensible to

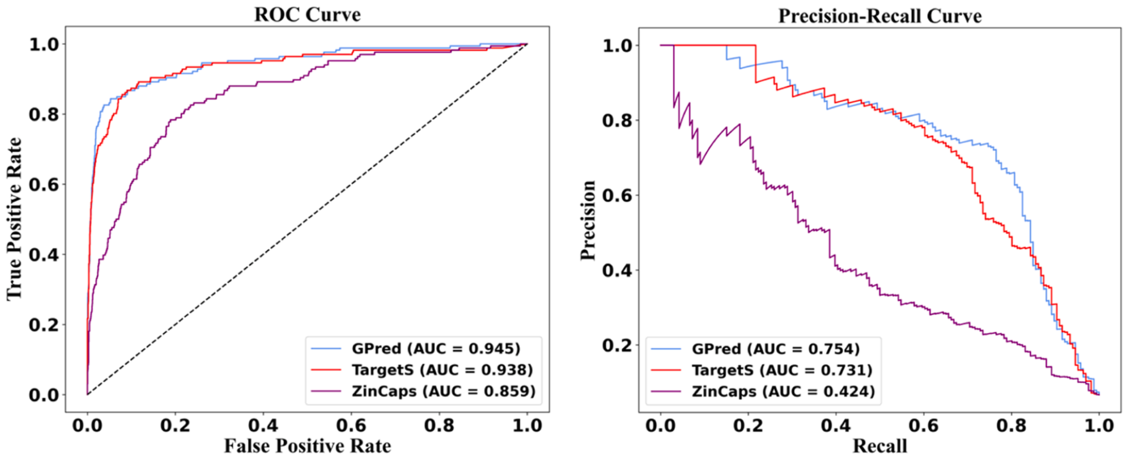


Fig. 9. Receiver operating characteristic and precision-recall curves for the GPred and competitive tools on dataset defining Fe, Mn, Ca and Mg binding sites as Zn’s negatives.

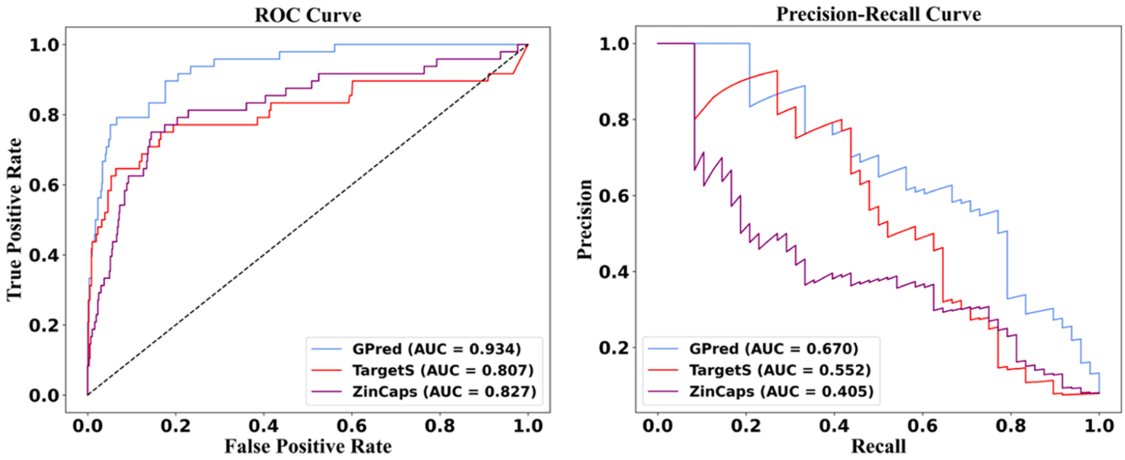


Fig. 10. Receiver operating characteristic and precision-recall curves for the GPred and competitive tools on dataset with lower structural similarity than TM-score of 0.5 compared to training data.

Table 7
Performance of GPred on Zn^{2+}-binding peptides.

Metric	1-fold	2-fold	3-fold	4-fold	5-fold	Mean
Precision	1.000	0.978	1.000	1.000	1.000	0.996
Recall	0.767	0.750	0.750	0.750	0.750	0.753
F1	0.870	0.849	0.857	0.857	0.857	0.858
MCC	0.803	0.771	0.790	0.790	0.790	0.789
ROC-AUC	0.945	0.943	0.940	0.938	0.947	0.943
PR-AUC	0.954	0.939	0.945	0.945	0.951	0.947

peptides with stable structures. Second, GPred appears to learn generalizable local features around ion-binding sites rather than being overly dependent on specific protein features. We believe these results further confirm the broad applicability of GPred in the ion-binding field.

Poor model performance due to small sample data and an unbalanced ratio of positive and negative samples is often an unsolved problem in this field. Although the use of transfer learning and the addition of positive sample weights to the loss function in our model are helpful in solving these problems, they are still problems that we need to overcome. Future endeavors could explore employing meta-learning strategies, enabling models to quickly adapt and predict ions with minimal data instead of tailoring a model specifically for each ion. This approach could significantly enhance our ability to address the challenges posed by limited data and imbalanced sample ratios in coordinated metal ion-binding site prediction. GPred was trained and tested exclusively on intact and static protein structures, which limits its ability to handle more complex, dynamic structures. GPred only takes protein structures and annotated binding sites to learn specific binding features at the protein structures for any molecular binding site prediction. This work represents an initial attempt to develop an ion-binding site predictor using a point cloud and geometry-aware graph neural network. Therefore, we selected only a few representative ions with sufficient binding site annotations to demonstrate the effectiveness of GPred. In future work, we aim to extend its application to a wider variety of ion types. We also plan to upgrade GPred to atom-level binding site prediction. Consequently, GPred will be adapted for diverse specific scenarios, such as identifying varying binding geometries, polymorphic binding states, and the pH of the environment in which a given protein exists, once sufficient specific datasets from these scenarios are available.

Code availability

The source code of GPred is available at https://github.com/wn1225/GPred. A webserver to run GPred with no installation, and no account required is available at https://www.musite.net.

Funding

This work was partially supported by the Paul K. and Diane Shumaker Endowment Fund. Ning Wang was supported by the Science and Technology Development Plan Project of Jilin Province, China (No. 20210101174JC).

Author statement

D.X. conceived and supervised the study, C.E and N.W. developed the algorithms, Y.Y. investigated the performance, S.A. and N.M. involved in data analysis, Y.D. developed the webserver, C.E, N.W. and F.H. wrote the manuscript and D.X. revised the manuscript. All authors have read and agreed to the published version of the manuscript.

CRediT authorship contribution statement

Fei He: Writing – review & editing, Visualization, Validation, Methodology, Conceptualization. **Clement Essien:** Visualization, Methodology, Data curation. **Dong Xu:** Writing – review & editing, Supervision, Conceptualization. **Yang Yu:** Validation, Methodology. **Ning Wang:** Writing – original draft, Visualization, Methodology. **Yongfang Qin:** Software. **Salhuldin Alqarghuli:** Visualization, Validation. **Negin Manshour:** Visualization.

Declaration of Competing Interest

The authors declare that they have no conflict of interest.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.12.016.

References

[1] Raff M, Alberts B, Lewis J, et al. Mol Biol Cell 4th Ed 2002.
[2] Von Hippel PH, Schleich T. Ion effects on the solution structure of biological macromolecules. Acc Chem Res 1969;2(9):257–65.
[3] Urnov FD, Rebar EJ, Holmes MC, et al. Genome editing with engineered zinc finger nucleases. Nat Rev Genet 2010;11(9):636–46.
[4] Hardison RC. A brief history of hemoglobins: plant, animal, protist, and bacteria. Proc Natl Acad Sci 1996;93(12):5675–9.
[5] Ram BP, Munjal DD, Fraser IH. Galactosyltransferases: physical, chemical, and biological aspect. Crit Rev Biochem 1985;17(3):257–311.

[6] Jensen MR, Petersen G, Lauritzen C, et al. Metal binding sites in proteins: identification and characterization by paramagnetic NMR relaxation. Biochemistry 2005;44(33):11014–23.

[7] Mandò PA, Przybyłowicz WJ. Particle-induced X-ray emission (PIXE). Encycl Anal Chem: Appl Theory Instrum 2006.

[8] Hu X, Dong Q, Yang J, et al. Recognizing metal and acid radical ion-binding sites by integrating ab initio modeling with template-based transferals. Bioinformatics 2016;32(21):3260–9.

[9] Yu DJ, Hu J, Yang J, et al. Designing template-free predictor for targeting protein-ligand binding sites with classifier ensemble and spatial clustering. IEEE/ACM Trans Comput Biol Bioinforma 2013;10(4):994–1008.

[10] Yu DJ, Hu J, Huang Y, et al. TargetATPsite: a template-free method for ATP-binding sites prediction with residue evolution image sparse representation and classifier ensemble. J Comput Chem 2013;34(11):974–85.

[11] Chen K, Mizianty MJ, Kurgan L. Prediction and analysis of nucleotide-binding residues using sequence and sequence-derived structural descriptors. Bioinformatics 2012;28(3):331–41.

[12] Srivastava A, Kumar M. Prediction of zinc binding sites in proteins using sequence derived information. J Biomol Struct Dyn 2018;36(16):4413–23.

[13] Essien C., Wang D., Xu D. Capsule network for predicting zinc binding sites in metalloproteins[C]//2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2019: 2337-2341.

[14] Lin YF, Cheng CW, Shih CS, et al. MIB: metal ion-binding site prediction and docking server[J]. J Chem Inf Model 2016;56(12):2287–91.

[15] Xia CQ, Pan X, Shen HB. Protein–ligand binding residue prediction enhancement through hybrid deep heterogeneous learning of sequence and structure data. Bioinformatics 2020;36(10):3018–27.

[16] Xia Y, Xia CQ, Pan X, et al. GraphBind: protein structural context embedded rules learned by hierarchical graph neural networks for recognizing nucleic-acid-binding residues. Nucleic Acids Res 2021;49(9):e51. e51.

[17] Paiva Vinícius A, Mendonça Murillo V, Silveira Sabrina A, Ascher David B, Pires Douglas EV, Izidoro Sandro C. GASS-Metal: identifying metal-binding sites on protein structures using genetic algorithms. Brief Bioinforma 2022;23(5):bbac178.

[18] Yang J, Roy A, Zhang Y. Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. Bioinformatics 2013;29(20):2588–95.

[19] Chakrabarti P. Anion binding sites in protein structures. J Mol Biol 1993;234(2):463–82.

[20] Laveglia Vincenzo, Bazayeva Milana, Andreini Claudia, Rosato Antonio. Hunting down zinc(II)-binding sites in proteins with distance matrices. Bioinformatics 2023;39(11):btad653.

[21] Ali S, Awan Z, Mumtaz S, et al. Cardiac toxicity of heavy metals (cadmium and mercury) and pharmacological intervention by vitamin C in rabbits. Environ Sci Pollut Res 2020;27:29266–79.

[22] Yang J, Roy A, Zhang Y. BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions. Nucleic Acids Res 2012;41(D1):D1096–103.

[23] Putignano V, Rosato A, Banci L, et al. MetalPDB in 2018: a database of metal sites in biological macromolecular structures. Nucleic Acids Res 2018;46(D1):D459–64.

[24] Andreini C, Cavallaro G, Lorenzini S, et al. MetalPDB: a database of metal sites in biological macromolecular structures. Nucleic Acids Res 2012;41(D1):D312–9.

[25] Fu L, Niu B, Zhu Z, et al. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 2012;28(23):3150–2.

[26] Landrum G. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling[J]. Greg Landrum, 2013, 8(31.10): 5281.

[27] cheol Jeong J, Lin X, Chen XW. On position-specific scoring matrix for protein function prediction. IEEE/ACM Trans Comput Biol Bioinforma 2010;8(2):308–15.

[28] Bairoch A. Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 2000;28(1):45–8.

[29] Qi CR, Yi L, Su H, et al. Pointnet+ +: deep hierarchical feature learning on point sets in a metric space. Adv Neural Inf Process Syst 2017:30.

[30] Krapp LF, Abriata LA, Cortés Rodriguez F, et al. PeSTo: parameter-free geometric deep learning for accurate prediction of protein binding interfaces. Nat Commun 2023;14(1):2175.

[31] Zhao H, Jiang L, Jia J, et al. Point Transform[C]//Proc IEEE/CVF Int Conf Comput Vis 2021:16259–68.

[32] Chen J, Xie ZR, Wu Y. Understand protein functions by comparing the similarity of local structural environments. Biochim Et Biophys Acta (BBA)-Proteins Proteom 2017;1865(2):142–52.

[33] Babor M, Gerzon S, Raveh B, et al. Prediction of transition metal binding sites from apo protein structures. Proteins Struct Funct Bioinf 2008;70:208. 17.

[34] Kingma D.P., Ba J. Adam: a method for stochastic optimization. CoRR[J]. arXiv preprint arXiv:1412.6980, 2014.

[35] Bhagi-Damodaran A, Lu Y. The periodic table's impact on bioinorganic chemistry and biology's selective use of metal ions. Period Table II: Catal, Mater, Biol Med Appl 2019:153–73.

[36] Ying Z, Bourgeois D, You J, et al. Gnnexplainer: generating explanations for graph neural networks. Adv Neural Inf Process Syst 2019:32.

[37] Kermani AA. A guide to membrane protein X-ray crystallography. FEBS J 2021;288(20):5788–804.

[38] Xie Q, Cao J, Zhang H, et al. Structural determination of glucosyltransferase C by cryo-electron microscopy[M]//the bacterial cell wall: methods and protocols. New York, NY: Springer US; 2023. p. 227–37.

[39] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596(7873):583–9.

[40] Peng CX, Liang F, Xia YH, et al. Recent advances and challenges in protein structure prediction. J Chem Inf Model 2023;64(1):76–95.

[41] Berman HM, Westbrook J, Feng Z, et al. The protein data bank. Nucleic Acids Res 2000;28(1):235–42.

[42] Hastie T, Tibshirani R, Friedman JH, et al. The elements of statistical learning: data mining, inference, and prediction. New York: Springer; 2009.

[43] Xu J, Zhang Y. How significant is a protein structure similarity with TM-score= 0.5? Bioinformatics 2010;26(7):889–95.