OXFORD

Briefings in Bioinformatics, 2022, **23(6)**, 1–10
https://doi.org/10.1093/bib/bbac444
Advance access publication date 23 October 2022
Problem Solving Protocol

Alignment-free metal ion-binding site prediction from protein sequence through pretrained language model and multi-task learning

Qianmu Yuan, Sheng Chen, Yu Wang, Huiying Zhao and Yuedong Yang

Corresponding authors: Yuedong Yang, School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510000, China. Tel.: +86 020-37106046. E-mail: yangyd25@mail.sysu.edu.cn; Huiying Zhao, Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou 510000, China. Tel.: +86 020-81332199. E-mail: zhaohy8@mail.sysu.edu.cn

Abstract

More than one-third of the proteins contain metal ions in the Protein Data Bank. Correct identification of metal ion-binding residues is important for understanding protein functions and designing novel drugs. Due to the small size and high versatility of metal ions, it remains challenging to computationally predict their binding sites from protein sequence. Existing sequence-based methods are of low accuracy due to the lack of structural information, and time-consuming owing to the usage of multi-sequence alignment. Here, we propose LMetalSite, an alignment-free sequence-based predictor for binding sites of the four most frequently seen metal ions in BioLiP (Zn^{2+}, Ca^{2+}, Mg^{2+} and Mn^{2+}). LMetalSite leverages the pretrained language model to rapidly generate informative sequence representations and employs transformer to capture long-range dependencies. Multi-task learning is adopted to compensate for the scarcity of training data and capture the intrinsic similarities between different metal ions. LMetalSite was shown to surpass state-of-the-art structure-based methods by more than 19.7, 14.4, 36.8 and 12.6% in area under the precision recall on the four independent tests, respectively. Further analyses indicated that the self-attention modules are effective to learn the structural contexts of residues from protein sequence. We provide the data sets, source codes and trained models of LMetalSite at https://github.com/biomed-AI/LMetalSite.

Keywords: metal ion-binding site, alignment-free, pretrained language model, multi-task learning

Introduction

Almost 40% of the proteins in the Protein Data Bank (PDB) [1] bind to metal ions [2], which are indispensable for the protein structural stability [3] and biological functions in cells such as enzyme catalysis and regulation of gene expression [4–6]. For example, Zn^{2+} ions can bind with specific nucleases and transcription factors to form Zn finger domains that recognize DNA and RNA for regulation of gene expression [6]. Hence, identifying amino acids involved in protein–metal–ion interactions helps to understand protein functions and design novel drugs [7]. Unfortunately, experimental methods for metal ion-binding site detection such as nuclear magnetic resonance [8] and absorption spectroscopy [9] are costly and time-consuming. Therefore, it is desirable to develop computational methods for making reliable metal ion-binding site prediction.

Many computational methods have been developed for predicting metal ion-binding sites, but the problem remains challenging due to the small size and high versatility of metal ions. Current methods can be classified into structure-based and sequence-based methods according to their used information. Structure-based approaches using experimental structures as input are often more accurate, which can be generally categorized into template-based methods, machine-learning-based methods and hybrid methods. Template-based methods such as MIB [10] employ alignment algorithms to transfer the structure information of templates for binding site inference. Nevertheless, these methods will be seriously restricted when no high-quality template can be found. Structure-based machine learning methods handle protein structures by extracting geometric features and then feeding to the neural networks, or explicitly taking the structural context topology into account and training in an end-to-end way. One example of the former is DELIA [11], which treats protein structures as 2D images and uses convolutional neural networks to extract characteristics from protein distance matrices. Alternately, an example of the latter is GraphBind [12], which encodes protein structures as graphs and adopts graph neural networks to learn the local tertiary patterns for binding site prediction. Hybrid methods such as COACH [7] and IonCom

Qianmu Yuan is a PhD student in the School of Computer Science and Engineering at Sun Yat-sen University. His research interests lie in deep learning, graph neural network, protein structure prediction and protein function prediction.
Sheng Chen is a PhD student in the School of Computer Science and Engineering at Sun Yat-sen University. His research interests include deep learning, protein design, protein structure prediction and graph neural network.
Yu Wang is a research professor in Peng Cheng National Laboratory at Shenzhen. His research interests include AI for systems biology, particularly foundation models in biomedicine research.
Huiying Zhao is an associate research fellow in the Sun Yat-sen Memorial Hospital at Sun Yat-sen University. Her research interests include pathogenic gene analysis, protein function and RNA function prediction.
Yuedong Yang is a professor in the School of Computer Science and Engineering at Sun Yat-sen University. Currently he focuses on developing AI algorithms and the HPC platform for biomedicine.
Received: May 21, 2022. **Revised:** September 2, 2022. **Accepted:** September 17, 2022

[13] integrate template-based methods and machine-learning-based methods simultaneously. Albeit powerful, the structure-based methods are not applicable to most proteins whose tertiary structures are unavailable due to the difficulties to determine protein structures experimentally [14].

By comparison, sequence-based methods learn local patterns of metal ion-binding characteristics through sequence-derived features. For example, IonSeq [13] and TargetS [15] extract evolutionary conservative information, predicted secondary structure and ligand-specific binding propensity from sequence context using sliding-window strategy, and then employ support vector machine to learn local binding patterns. Sequence-based approaches have a potentially wider range of applications since they require only readily available protein sequences, yet the lacks of tertiary structure information usually cause their limited performances.

In addition to the weaknesses described above, both structure-based and sequence-based approaches are mostly time-consuming owing to the usage of evolutionary sequence profile from multi-sequence alignment. For example, the profile generation using PSI-BLAST [16] requires about an hour for individual protein on single CPU, and the computational time is even growing due to the exponential growth of the sequence library. This has partly limited their large-scale applications in proteomes. Unsupervised pretraining with contextual language models has yielded ground-breaking improvements in natural language processing, which has recently been applied to protein sequence representation learning and has displayed highly promising results in downstream predictions including secondary structure, tertiary contact, mutational effect and ontology-based protein function [17–19]. Such breakthroughs inspire us to develop a fast and accurate sequence-based metal ion-binding site predictor. Besides, current methods in this field learn different ion-binding patterns separately, ignoring the underlying relations between these similar ligands. Multi-task learning aims to improve predictive performance for related tasks (e.g. binding sites of different metal ions) by exploiting shared networks [20], which has been shown to benefit bioinformatics problems including predictions of inter-residue distances [21], cleavage sites [22] and binding sites [23, 24]. Therefore, it is promising to further advance metal ion-binding site prediction by effectively modeling the intrinsic relations between different ions through multi-task learning technique.

In this study, we present LMetalSite, a novel alignment-free sequence-based method for binding site predictions of the four most frequently seen metal ions (Zn^{2+}, Ca^{2+}, Mg^{2+} and Mn^{2+}) in structures from BioLiP [25]. LMetalSite leverages the recently published pretrained language model (ProtTrans [18]) to bypass slow database searches and generate informative sequence representations within a short time. Multi-task learning is adopted to further improve the predictive quality by compensating for the scarcity of training data and better modeling the intrinsic similarities between different metal ions. Concretely, we employ the well-acknowledged transformer model [26, 27] as shared networks to capture common binding mechanisms such as long-range dependencies in protein sequence, followed by four ion-specific multilayer perceptrons (MLP) to learn the binding patterns of particular metal ions. LMetalSite was shown to surpass state-of-the-art methods MIB, TargetS, IonCom, DELIA and GraphBind by more than 19.7, 14.4, 36.8 and 12.6% in area under the precision-recall curve (AUPR) on the four independent tests, respectively. Further analyses indicated that the self-attention modules are effective to learn the structural contexts of residues from protein sequence representation, which partly explains the superior performance of LMetalSite. In the future, our framework can be easily extended to sequence-based predictions of other functional sites, such as nucleic-acid-binding sites.

Materials and methods

Data sets

To evaluate the performance of LMetalSite, we constructed four benchmark data sets for the four most frequently seen metal ion-binding proteins from the BioLiP database [25], including Zn^{2+}, Ca^{2+}, Mg^{2+} and Mn^{2+} ions. This database is a collection of biologically relevant protein-ligand complexes primarily from PDB. Concretely, we collected proteins that bind with these ions from BioLiP released on 29 December 2021. Only protein chains with resolutions of ≤3.0 Å and lengths of 50–1000 were kept. In these data sets, a binding site/residue was defined if the smallest atomic distance between the target residue and the ligand molecule is <0.5 Å plus the sum of the Van der Waal's radius of the two nearest atoms. Then, we removed redundant proteins sharing sequence identity >25% over 30% alignment coverage within each data set using CD-HIT [28]. Finally, each benchmark data set was further split into a training set that contains proteins released before 1 January 2020, as well as an independent test set that contains proteins released from 1 January 2020 to 29 December 2021. Specifically, the training sets of Zn^{2+}, Ca^{2+}, Mg^{2+} and Mn^{2+} ions contain 1647, 1554, 1730 and 547 chains, respectively, and the corresponding independent test sets contain 211, 183, 235 and 57 chains, respectively. Details of the statistics of these benchmark data sets are given in Table 1, where the data sets are named by the amounts of the included proteins.

Protein sequence representation

LMetalSite solely employs pretrained language models to effectively generate informative sequence representation as input feature. Moreover, we have also tested other widely used features including evolutionary information and structural properties to demonstrate the superiority of the representation produced by pretrained language models (Results are shown in section Features from pretrained language models are informative for binding site detection.)

Language model representation

LMetalSite leverages the recent language model ProtT5-XL-U50 [18] (denoted as ProtTrans) for feature extraction, which is a transformer-based auto-encoder named T5 [29] pretrained in a self-supervised manner, essentially learning to predict masked amino acids. Concretely, the ProtTrans model contains 24 layers and 32 heads with 3B parameters, which was first trained on BFD [30] and then fine-tuned on UniRef50 [31]. The BERT's denoising objective [32] was adopted to corrupt and reconstruct single tokens using a masking probability of 15% (details shown in Supplementary Note 1). We extracted the output from the last layer of the encoder part of ProtTrans as sequence representation, which is a 1024-dimensional per-residue feature matrix. We have also investigated another similar language model, ESM-1b [17] (denoted as ESM), which was also pretrained on UniRef50 using transformer. Sequence representation by ESM is a 1280-dimensional per-residue feature matrix. Note that the inference costs of ProtTrans and ESM are really low, and the feature extraction process of our whole benchmark data sets (~6000 sequences) using these pretrained models can be done within 10 min on an Nvidia GeForce RTX 3090 GPU.

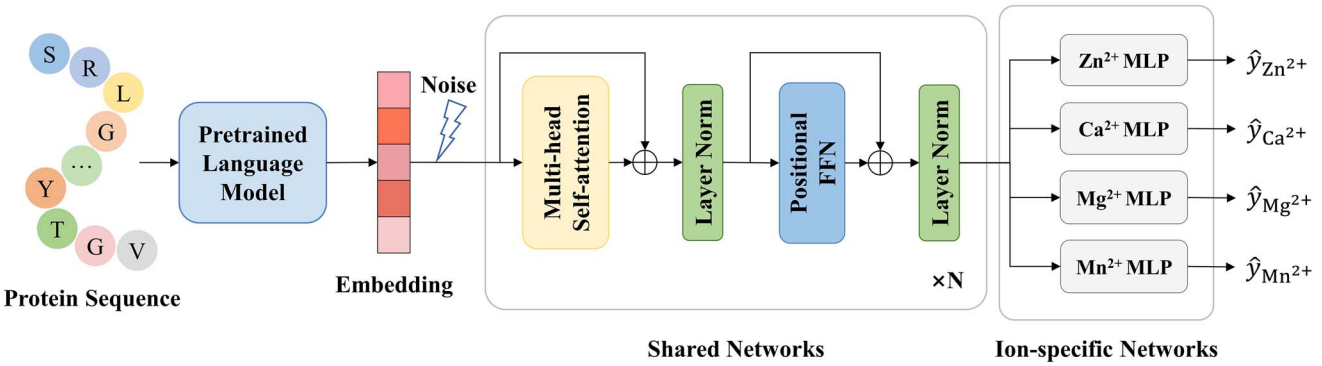

Figure 1. The overall architecture of the proposed framework LMetalSite. First, the protein sequence is input to the pretrained language model to produce the sequence embedding, which is augmented by Gaussian noise in the training steps. Then, the shared transformer networks are employed to capture the common binding-relevant characteristics such as long-range dependencies. Finally, four ion-specific MLPs are adopted to learn the binding patterns of particular metal ions.

Table 1. Statistics of the four metal ion benchmark data sets used in this study

Ligand type	Data set	Binding residues	Non-binding residues	% of binding residues
Zn^{2+}	ZN_Train_1647	7731	467 184	1.63
	ZN_Test_211	1039	54 981	1.85
Ca^{2+}	CA_Train_1554	8442	495 700	1.67
	CA_Test_183	1034	65 820	1.55
Mg^{2+}	MG_Train_1730	6321	569 572	1.10
	MG_Test_235	893	87 913	1.01
Mn^{2+}	MN_Train_547	2556	179 143	1.41
	MN_Test_57	225	20 194	1.10

Note: The columns give, in order, the ligand type, the data set name, the numbers of binding and non-binding residues and the percentage of the binding residues out of total in each data set.

Evolutionary information

Evolutionarily conserved residues may contain motifs related to important protein properties. Here, we additionally tested the widely used evolutionary features position-specific scoring matrix (PSSM) and hidden Markov models (HMM) profile. PSSM was produced by running PSI-BLAST [16] to search the query sequence against UniRef90 [31] with three iterations and an *E*-value of 0.001. HMM profile was generated by running HHblits [33] against UniClust30 [34] with default parameters. Each residue was encoded into a 20-dimensional vector in PSSM or HMM.

Structural properties

We also tested the structural properties extracted by DSSP [35] using native PDB structures: (1) eight-dimensional one-hot secondary structure profile. (2) Sine and cosine of the torsion angles PHI and PSI. (3) Relative solvent accessibility, which is the normalized solvent accessible surface area (ASA) by the maximal ASA of the corresponding amino acid. This 13-dimensional structural feature is named DSSP hereinafter.

The feature values in the sequence representations from pretrained language models, PSSM and HMM were all normalized to scores between 0 and 1 using Equation (1), where υ is the original feature value, and Min and Max are the smallest and biggest values of this feature type observed in the training set.

$$\upsilon_{\text{norm}} = \frac{\upsilon - \text{Min}}{\text{Max} - \text{Min}}. \tag{1}$$

The architecture of LMetalSite

The overall architecture of the proposed framework LMetalSite is shown in Figure 1. First, the protein sequence (FASTA format) is input to the pretrained language model to produce the sequence embedding, which is augmented by Gaussian noise to avoid overfitting in the training steps. Then, the shared networks consisting of N transformer blocks are employed to capture the common binding-relevant characteristics such as long-range dependencies of the residues. Finally, four ion-specific MLPs are adopted to learn the binding patterns of particular metal ions.

Shared transformer networks

We stack multiple standard transformer encoder layers as shared networks to capture common binding characteristics of different metal ions. Each transformer layer consists of a multi-head self-attention module and a positional fully connected feed-forward network. A residual connection [36] is employed around each of the two sub-layers, followed by layer normalization [37]. Let $H = [h_1^T, \ldots, h_n^T]^T \in \mathbb{R}^{n \times d}$ denote the input of the self-attention module, where n is the sequence length, d is the hidden dimension and $h_i \in \mathbb{R}^{1 \times d}$ is the hidden representation of the ith amino acid. The input of the lth layer $H^{(l)}$ is projected by three matrices $W_Q \in \mathbb{R}^{d \times d_K}$, $W_K \in \mathbb{R}^{d \times d_K}$ and $W_V \in \mathbb{R}^{d \times d_V}$ to the corresponding query, key and value representations Q, K, V:

$$Q = H^{(l)}W_Q, K = H^{(l)}W_K, V = H^{(l)}W_V. \tag{2}$$

The self-attention is then calculated as:

$$A = \frac{QK^T}{\sqrt{d_K}}, \tag{3}$$

$$H^{(l+1)} = \text{Attn}\left(H^{(l)}\right) = \text{softmax}(A)V, \tag{4}$$

where A is a matrix capturing the similarities between queries and keys. To jointly attend to information from different representation subspaces at different positions, multi-head attention is used to linearly project the queries, keys and values h times, perform

the attention function in parallel and finally concatenate them together. In this study, $d_K = d_V = d/h$.

Ion-specific MLP

The output of the last transformer layer is input to the ion-specific MLPs to predict the binding probabilities of particular metal ions for all n amino acid residues:

$$Y'_m = \text{sigmoid}\left(H^{(N+1)}W_m + b_m\right), \tag{5}$$

where $H^{(N+1)} \in \mathbb{R}^{n\times d}$ is the output of the Nth transformer layer; $W_m \in \mathbb{R}^{d\times 1}$ is the weight matrix for the specific metal ion m; $b_m \in \mathbb{R}$ is the bias term for metal ion m; $Y'_m \in \mathbb{R}^{n\times 1}$ is the predictions of metal ion m for the n residues. The sigmoid function normalizes the output of the network into binding probabilities ranging from 0 to 1. These four ion-specific MLPs adopt the same hyperparameters but different network weights in order to further mine specific binding patterns of Zn^{2+}, Ca^{2+}, Mg^{2+} and Mn^{2+} ions from the common binding characteristics captured by the shared transformer networks.

Noise augmentation and multi-task training

Due to the limited training data and high dimension of the feature vectors, we employ feature augmentation by Gaussian noise as suggested in References [38, 39], to avoid overfitting and enhance model robustness. In the training steps, we add a matrix of random values drawn from a Gaussian distribution to the sequence representation from the pretrained language model before feeding it to the shared networks:

$$H^{(1)} = H^{(0)} + \varepsilon \times X, X \sim \mathcal{N}(0,1), \tag{6}$$

where $H^{(0)}$ is the sequence representation from the pretrained language model, X is a matrix with the same size as $H^{(0)}$ which is filled with random values from the standard normal distribution and ε is a hyperparameter. During the validation and test phases, this technique is turned off.

LMetalSite utilizes multi-task learning for simultaneous predictions for four types of metal ion-binding sites, which means that all proteins of different ion-binding types are input to the same network, and the predictions of the four types of ions can be obtained. Nevertheless, only the predictions of the corresponding known ion-binding types are used to calculate loss and perform backpropagation, while the predictions of other ions without ground truth are masked in the training steps. That is, each protein is used to train the shared networks and the corresponding ion-specific network(s) of its known ion-binding type(s) without affecting other irrelevant MLPs.

Implementation details

We performed the 5-fold cross-validation (CV) on the training data, where the four training sets were mixed and split into 5-folds randomly, and then each time a model was trained on 4-folds and evaluated on the remaining fold. This process was repeated for five times and the performances on the 5-folds were averaged as the overall validation performance, which was used to choose the best feature combination and optimize all hyperparameters through grid search (Supplementary Table S1). In the test phase, all five trained models from the CV were used to make predictions, which were averaged as the final predictions of LMetalSite.

Specifically, we employed a two-layer shared transformer network with 64 hidden units and the following set of hyperparameters: $h = 4$, $\varepsilon = 0.05$ and batch size of 32. We utilized the Adam optimizer [40] with $\beta_1 = 0.9$, $\beta_2 = 0.99$, weight decay of 10^{-5} and learning rate of 3×10^{-4} for model optimization on the binary cross-entropy loss. The dropout rate was set to 0.2 to avoid overfitting. We implemented the proposed model with Pytorch 1.7.1 [41]. Within each epoch, we randomly drew 30 000 samples from the training data with replacement to train our model. The training process lasted at most 30 epochs and we performed early stopping with patience of 6 epochs based on the validation performance, which took about 45 min on an Nvidia GeForce RTX 3090 GPU. During the test phase, it took ~10 seconds to make predictions for one batch of proteins.

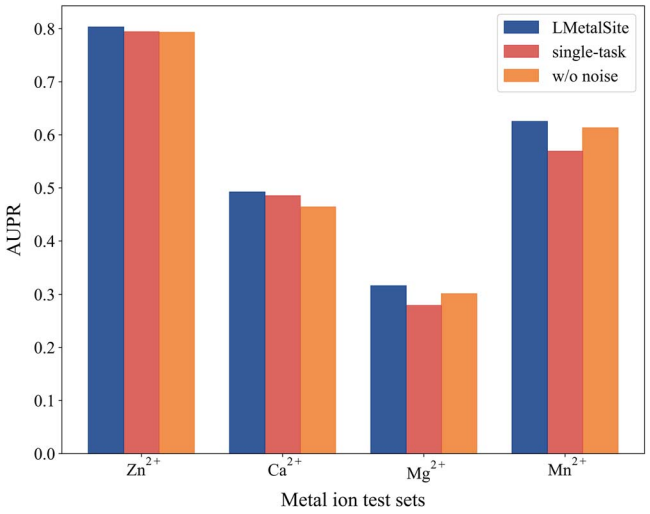

Figure 2. Ablation study on multi-task learning and noise augmentation in the four metal ion test sets.

Evaluation metrics

Similar to the previous works [42, 43], we used recall (Rec), precision (Pre), F1-score (F1), Matthews correlation coefficient (MCC), area under the receiver operating characteristic curve (AUC) and AUPR to evaluate the predictive performance:

$$\text{Rec} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{7}$$

$$\text{Pre} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{8}$$

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{9}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FN} \times \text{FP}}{\sqrt{(\text{TP} + \text{FP}) \times (\text{TP} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TN} + \text{FN})}}, \tag{10}$$

where true positives (TP) and true negatives (TN) denote the number of correctly predicted binding and non-binding residues, and false positives (FP) and false negatives (FN) denote the number of incorrectly predicted binding and non-binding residues, respectively. AUC and AUPR are independent of thresholds, thus reflecting the overall performance of a model. The other metrics were calculated using a threshold to convert the predicted binding probabilities to binary predictions, which was determined by maximizing MCC for the model. We adopted AUPR for hyperparameter selections as it is more sensitive and informative than AUC in imbalanced two-class classification tasks [44].

Table 2. The predictive performance on the 5-fold CV and the four metal ion test sets using different features

Feature	CV AUC*	CV AUPR*	Test AUC*	Test AUPR*	Zn^{2+} AUPR	Ca^{2+} AUPR	Mg^{2+} AUPR	Mn^{2+} AUPR
Evo	0.853	0.374	0.870	0.392	0.694	0.239	0.173	0.462
Evo + DSSP	0.871	0.413	0.884	0.443	0.734	0.312	0.213	0.511
ESM	0.900	0.521	0.921	0.537	0.780	0.463	0.310	0.596
ProtTrans (LMetalSite)	**0.913**	**0.540**	**0.928**	**0.559**	**0.803**	**0.492**	**0.316**	**0.625**

Note: * denotes the average metrics of the four metal ions. Evo denotes the evolutionary features PSSM and HMM. Bold fonts indicate the best results.

Results

Features from pretrained language models are informative for binding site detection

We evaluated LMetalSite by AUC and AUPR using the 5-fold CV and independent test sets of Zn^{2+}, Ca^{2+}, Mg^{2+} and Mn^{2+} ions. As shown in Table 2, the LMetalSite model obtained average AUC values over the four metal ions of 0.913 and 0.928 on the 5-fold CV and independent tests, respectively; as well as average AUPR values of 0.540 and 0.559, respectively. The consistent performances on the CV and independent tests indicated the robustness of our model. Specifically, LMetalSite achieved AUPR of 0.803, 0.492, 0.316 and 0.625 on the four independent test sets.

To demonstrate the effect of sequence representation by the ProtTrans language model, we conducted feature ablation experiments to compare ProtTrans with another language model ESM and other widely used handcrafted features in this field. As shown in Table 2, when using the computationally efficient pretrained model ProtTrans or ESM to extract sequence representation, the model gained an average AUPR of 0.559 or 0.537 on the independent tests, respectively, higher than the one (0.443) by using evolutionary information (PSSM and HMM) and native structural features (DSSP). To further understand the advantages of ProtTrans over evolutionary information, we plotted their AUC values against the number of effective homologous sequences (Neff) in Supplementary Figure S1. Neff is an HHblits parameter measuring the effective size of homologous sequence cluster. The figure shows that ProtTrans consistently outperformed PSSM and HMM, especially for the target proteins with more homologous sequences (higher Neff), which may be due to the better information extraction ability of language models for larger amount of data points. Note that ProtTrans performed slightly better than ESM, which might be ascribed to the different network architectures they adopted for pretraining. Moreover, Supplementary Table S2 shows that combining ProtTrans and ESM is redundant and could not attain any further improvement, suggesting that these two language models are similar. Additionally, further integrating evolutionary and structural features to ProtTrans only brought minor improvements on the test sets (<0.01 of AUPR on average). These results indicated that the ProtTrans language model may potentially capture the evolutionary and structural information of the protein.

The impact of multi-task learning

LMetalSite employs multi-task learning to capture the intrinsic similarities between different metal ions. To investigate the impact of multi-task technique, we changed the four ion-specific MLPs in LMetalSite into a single MLP, and then trained and tested on the four metal ion benchmark data sets separately with the same input sequence features. As shown in Figure 2 and Supplementary Table S3, the removal of the multi-task strategy caused AUPR drops of 0.009, 0.007, 0.037 and 0.056 on the Zn^{2+}, Ca^{2+}, Mg^{2+} and Mn^{2+} test sets, respectively. As expected, the ion with the smallest training set (Mn^{2+}) benefited the most from multi-task learning, since the transformer networks could be better trained with a larger data set containing other types of metal ion-binding proteins. This suggested that different types of metal ions might potentially share common chemical mechanisms and binding patterns, and the predictions for one metal ion type could benefit from the binding information of other ion types. Supplementary Table S3 also shows the performance of using a fully shared architecture, in which all proteins of different ion-binding types were input to the same transformer and MLP networks, while the information of different tasks was represented as a four-dimensional vector and concatenated to the input features. We also found that the Gaussian noise feature augmentation is effective in preventing overfitting, since its removal caused small but consistent performance drops in AUPR of 0.010, 0.028, 0.015 and 0.012 on the four test sets, respectively.

Comparison with state-of-the-art methods

We compared LMetalSite with one sequence-based (TargetS) and four structure-based (MIB, IonCom, DELIA and GraphBind) predictors on the Ca^{2+}, Mg^{2+} and Mn^{2+} test sets describe in Table 1. Since binding site prediction of Zn^{2+} ion is not supported by DELIA and GraphBind, we only compared with the remaining three methods on the Zn^{2+} test set. As reported in Table 3, binding sites of Ca^{2+} and Mg^{2+} ions seem harder to distinguish, which may be due to the small differences between frequencies of the 20 native amino acids among binding and non-binding residues in Ca^{2+} and Mg^{2+} ion-binding proteins (Supplementary Note 4 and Supplementary Figure S2). Howsoever, LMetalSite outperformed all other sequence-based and even structure-based methods in F1, MCC, AUC and AUPR. Undoubtedly, LMetalSite substantially surpassed the sequence-based method TargetS by 35.4, 201.8, 113.5 and 94.1% in AUPR on Zn^{2+}, Ca^{2+}, Mg^{2+} and Mn^{2+} test sets, respectively. Interestingly, though our method is a sequence-based predictor, it outperformed the state-of-the-art structure-based method GraphBind by 14.4, 36.8 and 12.6% in AUPR on the Ca^{2+}, Mg^{2+} and Mn^{2+} test sets, respectively. This is expected because the sequence representation from the pretrained language model used by LMetalSite is more informative and powerful than the handcrafted evolutionary and structural features employed by GraphBind. Another reason may be that the experimental structures also brought noises because the protein structures are flexible. In addition, the multi-task learning adopted by LMetalSite could further boost the predictive quality through better trained shared networks that capture common binding patterns among different ions. Our method also outperformed IonCom by 19.7% on the Zn^{2+} test set, likely because IonCom is a hybrid method and the newly resolved proteins did not always have high-quality templates. The poor performance of MIB may

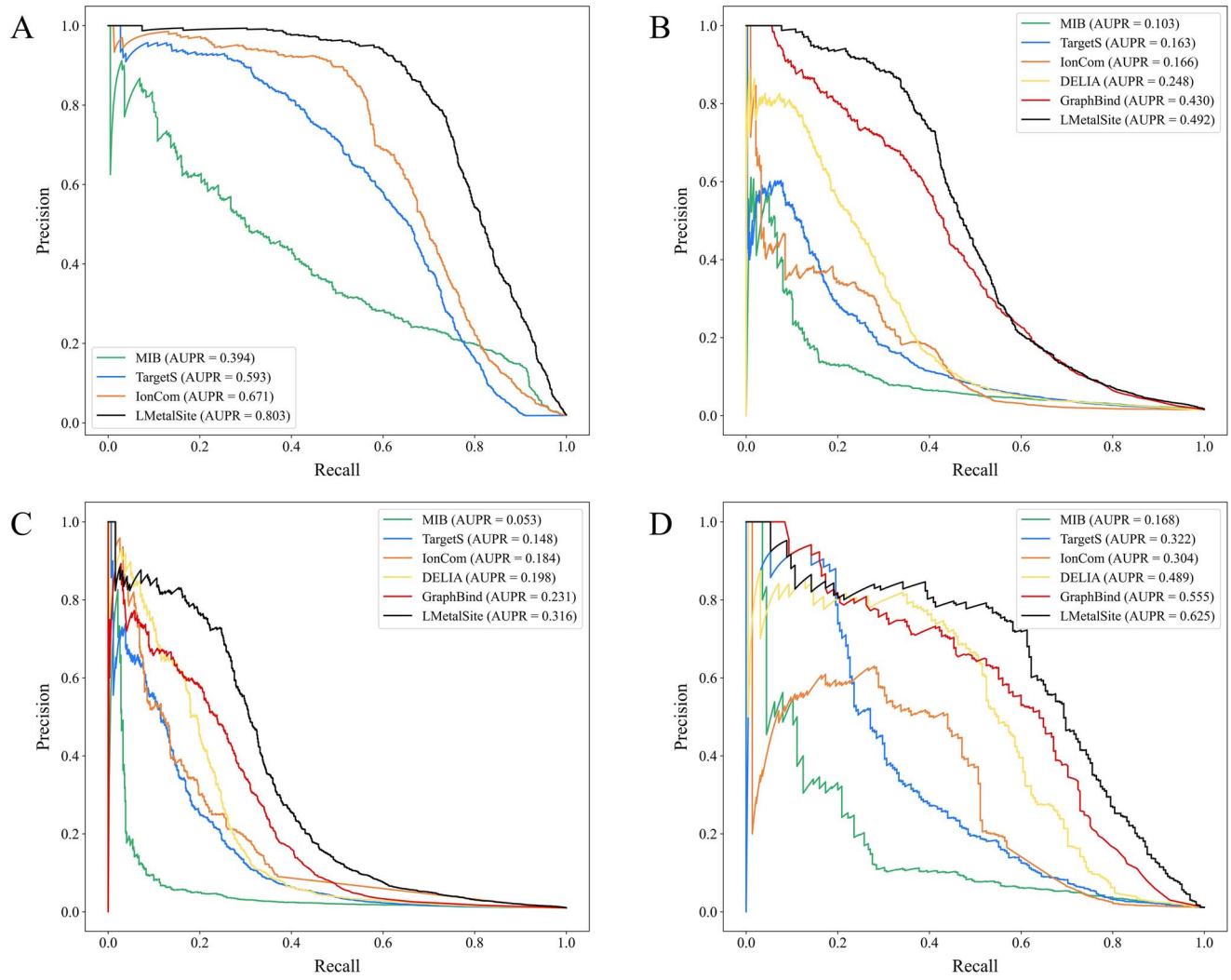

Figure 3. Precision-recall curves of LMetalSite and other methods on the Zn^{2+} (**A**), Ca^{2+} (**B**), Mg^{2+} (**C**) and Mn^{2+} (**D**) ion test sets.

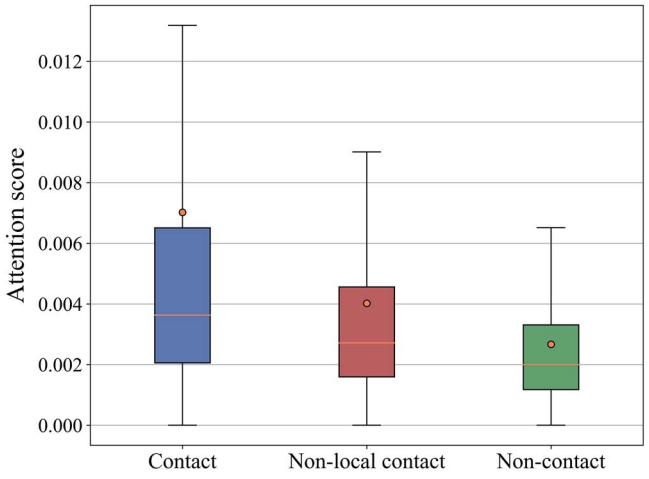

Figure 4. The distributions of the attention scores for the residue pairs in contact, in non-local contact, and not in contact. Each boxplot indicates the mean (orange dot), median (orange line), and quartiles with whiskers reaching up to 1.5 times the interquartile range. The outliers are not shown.

be partly ascribed to the slightly different definition of binding sites (residues partially within 3.5 Å of a metal ion) in their template database. The improved performance of LMetalSite over other methods on the four metal ion test sets can be further illustrated by the precision-recall curves (Figure 3) and the ROC curves (Supplementary Figure S3), where the curves of LMetalSite are mostly located on the upper portion of the figures. Note that our method did not always have the highest recall and precision, because they are unbalanced measures strongly depending on thresholds. On the other hand, our method is also computationally efficient. Empirically, it takes <1 min to make predictions for a protein with 200 residues using LMetalSite, but about 30 min using GraphBind or 6 h using IonCom.

To further validate the adaptability of the LMetalSite framework to other data sets and metal ions, we adopted another benchmark data set used by MIonSite [45] and GASS-Metal [46] to retrain and evaluate our method. MIonSite is an evolutionary profile-based metal ion-binding residue predictor, while GASS-Metal is a recently published template-based metal ion-binding site (set of binding residues) predictor. The original data set contains 12 types of metal ion-binding proteins, but here we only considered the nine ions (Zn^{2+}, Ca^{2+}, Mg^{2+}, Mn^{2+}, Fe^{3+}, Cu^{2+}, Fe^{2+}, Co^{2+}, K^{+}) with more than two sequences in the test data sets to avoid potential bias (details shown in Supplementary Note 2 and Supplementary Table S4). We retrained LMetalSite with simply the same hyperparameters, and the performance comparison of LMetalSite with other state-of-the-art methods on the nine metal

Table 3. Performance comparison of LMetalSite with state-of-the-art methods on the four metal ion test sets

Data set	Method	Rec	Pre	F1	MCC	AUC	AUPR
Zn^{2+}	MIB	0.744	0.219	0.339	0.385	0.935	0.394
	TargetS	0.454	0.749	0.566	0.578	0.874	0.593
	IonCom	0.852	0.137	0.236	0.317	0.937	0.671
	LMetalSite	0.681	0.859	**0.760**	**0.761**	**0.976**	**0.803**
Ca^{2+}	MIB	0.338	0.078	0.126	0.135	0.775	0.103
	TargetS	0.121	0.490	0.194	0.238	0.776	0.163
	IonCom	0.297	0.247	0.269	0.258	0.698	0.166
	DELIA	0.172	0.633	0.271	0.325	0.785	0.248
	GraphBind	0.371	0.623	0.465	0.475	0.888	0.430
	LMetalSite	0.413	0.724	**0.526**	**0.542**	**0.905**	**0.492**
Mg^{2+}	MIB	0.246	0.043	0.074	0.082	0.675	0.053
	TargetS	0.118	0.491	0.190	0.237	0.724	0.148
	IonCom	0.240	0.250	0.245	0.237	0.688	0.184
	DELIA	0.129	0.650	0.215	0.287	0.744	0.198
	GraphBind	0.273	0.414	0.329	0.331	0.776	0.231
	LMetalSite	0.245	0.728	**0.367**	**0.419**	**0.865**	**0.316**
Mn^{2+}	MIB	0.462	0.096	0.159	0.193	0.856	0.168
	TargetS	0.271	0.496	0.351	0.362	0.864	0.322
	IonCom	0.511	0.245	0.331	0.344	0.833	0.304
	DELIA	0.502	0.665	0.572	0.574	0.902	0.489
	GraphBind	0.427	0.706	0.532	0.545	0.930	0.555
	LMetalSite	0.613	0.719	**0.662**	**0.661**	**0.966**	**0.625**

Note: The results of IonCom and GraphBind were obtained from their standalone programs, while the predictions by other competitive methods were generated from their web servers. The inputs of TargetS were protein sequences, while the inputs for other methods were native protein structures. Bold fonts indicate the best results.

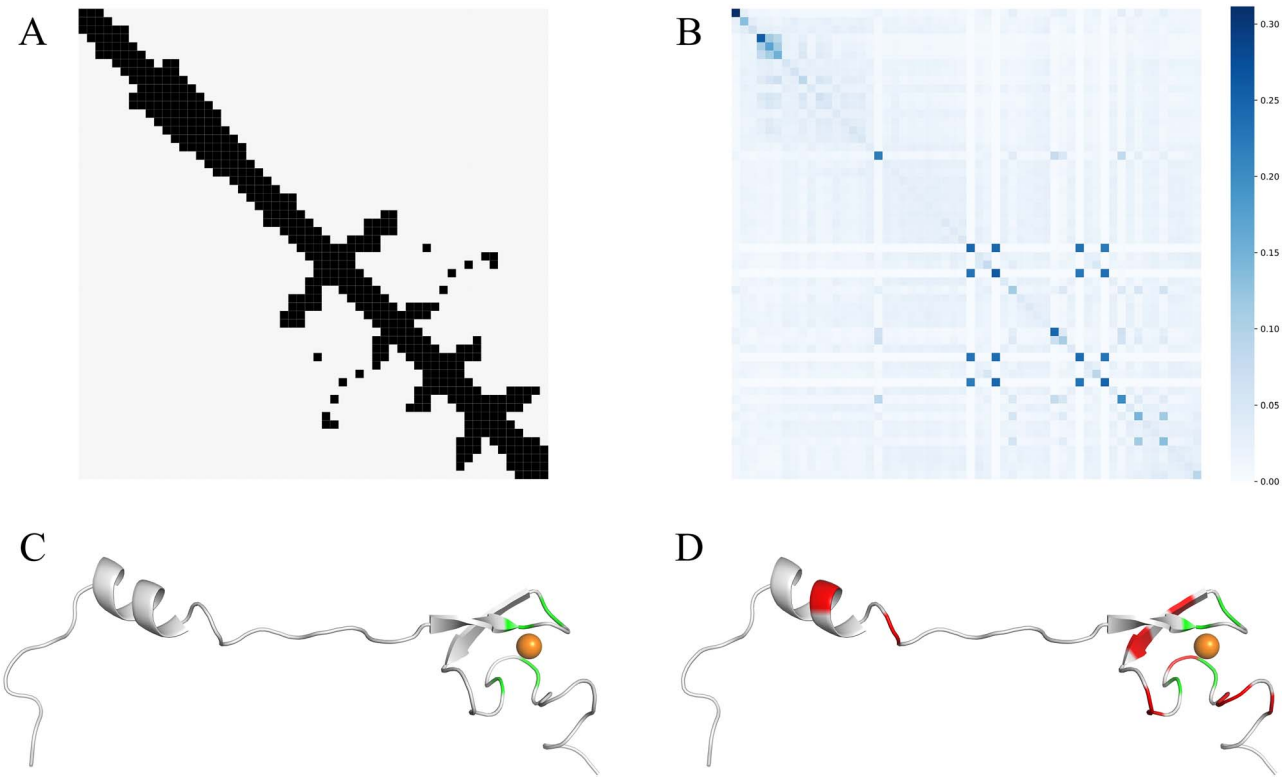

Figure 5. Visualization of one example (PDB ID: 6WU9, chain 2) from ZN_Test_211. **A**. The experimental protein contact map. **B**. The attention map from LMetalSite. **C**. Predictions by LMetalSite. **D**. Predictions by IonCom. The orange spheres denote Zn^{2+} ions, and true positives and false positives are colored in green and red, respectively.

ion independent test sets is shown in Supplementary Note 3 and Supplementary Table S5. LMetalSite obtained an average MCC of 0.698, surpassing GASS-Metal (0.617), MIonSite (0.547) and other methods by more than 13.1%, which further demonstrated the adaptability and robustness of our proposed framework.

Attention mechanism effectively captures protein structural information

To validate if our method can implicitly capture the structural information, we extracted the attention maps from the final layer of the shared transformer networks for all proteins in the

four metal ion test sets, which were then averaged among the attention heads to obtain the final maps. We also collected the native PDB structures for these proteins to acquire the coordinate of the Cα atom in each residue, and then calculated the inter-residual Euclidean distance maps. Comparative study found that the average attention score of the residue pairs in contact (<8 Å) is higher than those not in contact (7.02×10^{-3} versus 2.07×10^{-3}). However, residues adjacent in sequence are inherently in contact spatially, so the non-local contacts are more meaningful to investigate, which are defined as the contacts between two residues that are ≥20 residues away in sequence positions, but <8 Å in terms of Cα atomic distances. In addition, since the average sequence interval between residues not in contact is larger than those in non-local contact (194 and 93, respectively), we randomly sampled uncontacted residues to ensure the same sequence intervals as those in non-local contact to eliminate bias. Still, the average attention score of the residue pairs in non-local contact is significantly higher than the residue pairs not in contact after sampling (4.02×10^{-3} versus 2.67×10^{-3}, P-value $<2.225 \times 10^{-308}$) according to Mann–Whitney U-test [47] (since the scores are non-normal tested by D'Agostino's K-squared test [48]). The distributions of the attention scores for the residue pairs in contact, in non-local contact and not in contact are shown in Figure 4. These results suggested that the self-attention mechanism in LMetalSite effectively captures the protein structural information by spontaneously paying more attention to the spatial neighbors, which could facilitate better representation learning for the target residues as shown in our previous study [43].

Case study

To visualize the superiorities of the attention module and our predictive model, we selected the 50S ribosomal protein L32 (PDB ID: 6WU9, chain 2) from ZN_Test_211 for illustration. Figure 5A shows its experimental contact map, where contacts were defined as the residue pairs with distances <8 Å. We found that the non-local contacts generally match the residue pairs with high attention scores in the attention map from LMetalSite (Figure 5B), suggesting that the attention module can partly capture the protein contact information. In this example, there are four Zn^{2+}-binding residues over a total of 56 residues, and the prediction results of our method and IonCom are shown in Figure 5C and D, respectively. LMetalSite correctly predicted all binding residues with high confidences (probability scores >0.99), and assigned non-binding residues with relatively low scores (<0.20), leading to an MCC of 1.000. By comparison, IonCom predicted 15 binding residues in which 11 are FP scattered spatially, leading to a lower MCC of 0.459. The results of MIB, TargetS and GASS-Metal for this case are also shown in Supplementary Note 5. Visualization of another case (PDB ID: 7PPP, chain A) can be found in Supplementary Note 6 and Supplementary Figure S4.

Discussion and conclusion

Identifying metal ion-binding sites is crucial for understanding protein functions and designing novel drugs. Existing structure-based methods are not applicable to most proteins with unknown tertiary structures, while sequence-based methods are limited in predictive performance. Moreover, both structure-based and sequence-based machine learning approaches are mostly time-consuming owing to the usage of multi-sequence alignment. Trained with the computationally efficient and informative sequence representations from pretrained language model, LMetalSite achieved great performance (surpassing the best structure-based methods) using only protein sequences, which is promising for simultaneously solving the limitations of current structure-based and sequence-based methods. The multi-task learning technique adopted by LMetalSite is able to further improve the predictive quality, while all other competitive methods ignore the underlying relations between similar ions. In summary, the superiority of LMetalSite mainly benefits from two aspects: (i) the informative sequence representation from the pretrained language model potentially captures evolutionary and structural information; (ii) multi-task learning is an effective algorithm to compensate for the scarcity of training data and better model the common binding mechanisms between different metal ions.

However, there is still room for further improvements on LMetalSite. First, directly fine-tuning the pretrained language model on the binding site tasks may yield better performance than using it for feature exaction. Second, meta-learning [49, 50] could also be explored in the multi-task problems, which allows fast adaptation to unseen tasks with limited labels. Third, although the pretrained language model may partly capture structural information such as secondary structure and solvent accessibility, and LMetalSite already exceeded the best structure-based methods, the binding site prediction could still benefit from known protein structures or high-quality predicted structures from AlphaFold2 [51] or RoseTTAFold [52]. For example, using distance maps to mask out spatially remote residues when calculating attention scores [43] or integrating pairwise geometric features between residues [38] may further enhance the predictive performance. Besides, spatial clustering of binding sites or constrained docking can be performed on the protein structures to additionally determine the coordinates of the binding locations. Fourth, inter-domain metal ion-binding sites could be considered as suggested in [46].

In conclusion, this study proposed an alignment-free sequence-based framework LMetalSite for metal ion-binding site prediction, which employed the pretrained language model for sequence representation and multi-task learning to capture common binding patterns between different ions. LMetalSite showed preferable performance than other methods in comprehensive evaluations. We suggest that our method may provide useful information for biologists studying metal ion-binding patterns or pathogenic mechanisms of mutations, and chemists interested in targeted drug design. For example, the prediction can help narrow down potential binding sites for further wet experiment validation [53] or provide hypotheses and insights for the mechanisms of disease-causing gene mutations [54] as shown in other similar fields. The binding site prediction can also be used for druggability prediction [55] or de novo drug design [56–58]. In the future, we would further extend our method to predict various functional sites, including binding sites with proteins [42] and nucleic acids [43].

Key Points

- Existing structure-based methods for identifying metal ion-binding sites are not applicable to most proteins with unknown tertiary structures, while sequence-based methods are limited in predictive performance.
- LMetalSite is an alignment-free sequence-based method for metal ion-binding site prediction trained with informative sequence representation from the pretrained language model, which potentially captures evolutionary and structural information.

- LMetalSite employs multi-task learning to further improve the predictive quality by compensating for the scarcity of training data and better modeling the common binding patterns between different metal ions.
- LMetalSite showed preferable performance than state-of-the-art sequence-based and even structure-based methods in the four independent data sets of Zn^{2+}, Ca^{2+}, Mg^{2+} and Mn^{2+} ion-binding proteins.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Data availability

We provide the data sets, source codes and trained models of LMetalSite at https://github.com/biomed-AI/LMetalSite.

Funding

This study has been supported by the National Key R&D Program of China (2020YFB0204803), National Natural Science Foundation of China (61772566 and 62041209), Guangdong Key Field R&D Plan (2019B020228001 and 2018B010109006), Introducing Innovative and Entrepreneurial Teams (2016ZT06D211) and Guangzhou S&T Research Plan (202007030010).

References

1. Berman HM, Westbrook J, Feng Z, *et al.* The Protein Data Bank. *Nucleic Acids Res* 2000;**28**:235–42.
2. 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**:D459–64.
3. Tainer JA, Roberts VA, Getzoff ED. Metal-binding sites in proteins. *Curr Opin Biotechnol* 1991;**2**:582–91.
4. Andreini C, Bertini I, Rosato A. Metalloproteomes: a bioinformatic approach. *Acc Chem Res* 2009;**42**:1471–9.
5. Andreini C, Bertini I, Cavallaro G, *et al.* Metal ions in biological catalysis: from enzyme databases to general principles. *J Biol Inorg Chem* 2008;**13**:1205–18.
6. Berg JM. Zinc finger domains: hypotheses and current knowledge. *Annu Rev Biophys Biophys Chem* 1990;**19**:405–21.
7. Yang J, Roy A, Zhang Y. Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics* 2013;**29**:2588–95.
8. Jensen MR, Petersen G, Lauritzen C, *et al.* Metal binding sites in proteins: identification and characterization by paramagnetic NMR relaxation. *Biochemistry* 2005;**44**:11014–23.
9. Reed GH, Poyner RR. Mn2+ as a probe of divalent metal ion binding and function in enzymes and other proteins. *Met Ions Biol Syst* 2000;**37**:231–56.
10. Lin Y-F, Cheng C-W, Shih C-S, *et al.* MIB: metal ion-binding site prediction and docking server. *J Chem Inf Model* 2016;**56**:2287–91.
11. Xia C-Q, Pan X, Shen H-B. Protein–ligand binding residue prediction enhancement through hybrid deep heterogeneous learning of sequence and structure data. *Bioinformatics* 2020;**36**:3018–27.
12. Xia Y, Xia C-Q, 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**:e51–1.
13. 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**:3260–9.
14. Nagarajan R, Ahmad S, Michael GM. Novel approach for selecting the best predictor for identifying the binding sites in DNA binding proteins. *Nucleic Acids Res* 2013;**41**:7606–14.
15. Yu D-J, 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 Bioinform* 2013;**10**:994–1008.
16. Altschul SF, Madden TL, Schäffer AA, *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–402.
17. Rives A, Meier J, Sercu T, *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci* 2021;**118**:e2016239118.
18. Elnaggar A, Heinzinger M, Dallago C, *et al.* ProtTrans: towards cracking the language of lifes code through self-supervised deep learning and high performance computing. *IEEE Trans Pattern Anal Mach Intell* 2021;**44**:7112–27.
19. Unsal S, Atas H, Albayrak M, *et al.* Learning functional properties of proteins with language models. *Nat Mach Intell* 2022;**4**:227–45.
20. Zhang Y, Yang Q. An overview of multi-task learning. *Natl Sci Rev* 2018;**5**:30–43.
21. Wu T, Guo Z, Hou J, *et al.* DeepDist: real-value inter-residue distance prediction with deep residual convolutional network. *BMC Bioinform* 2021;**22**:1–17.
22. Singh D, Sisodia DS, Singh P. Compositional framework for multitask learning in the identification of cleavage sites of HIV-1 protease. *J Biomed Inform* 2020;**102**:103376.
23. Sun Z, Zheng S, Zhao H, *et al. To improve the predictions of binding residues with DNA, RNA, carbohydrate, and peptide via multi-task deep neural networks, IEEE/ACM Trans Comput Biol Bioinform*, 2021.
24. Zhang F, Zhao B, Shi W, *et al.* DeepDISOBind: accurate prediction of RNA-, DNA-and protein-binding intrinsically disordered residues with deep multi-task learning. *Brief Bioinform* 2022;**23**:bbab521.
25. Yang J, Roy A, Zhang Y. BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic Acids Res* 2012;**41**:D1096–103.
26. Vaswani A, Shazeer N, Parmar N *et al.* Attention is all you need. *In: Advances in Neural Information Processing Systems*. Curran Associates Inc., USA. 2017, 5998–6008.
27. Zheng S, Rao J, Zhang Z, *et al.* Predicting retrosynthetic reactions using self-corrected transformer neural networks. *J Chem Inf Model* 2019;**60**:47–55.
28. Fu L, Niu B, Zhu Z, *et al.* CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;**28**:3150–2.
29. Raffel C, Shazeer N, Roberts A, *et al.* Exploring the limits of transfer learning with a unified text-to-text transformer. *J Mach Learn Res* 2020;**21**:1–67.
30. Steinegger M, Mirdita M, Söding J. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nat Methods* 2019;**16**:603–6.
31. Suzek BE, Huang H, McGarvey P, *et al.* UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* 2007;**23**:1282–8.
32. Kenton JDM-WC, Toutanova LK. BERT: Pre-training of deep bidirectional transformers for language understanding. *In: Proceedings of NAACL-HLT*. Association for Computational Linguistics, USA, 2019;4171–86.

33. Remmert M, Biegert A, Hauser A, *et al.* HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* 2012;**9**:173–5.
34. Mirdita M, von den Driesch L, Galiez C, *et al.* Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res* 2017;**45**:D170–6.
35. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules* 1983;**22**: 2577–637.
36. He K, Zhang X, Ren S *et al.* Deep residual learning for image recognition. *In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Institute of Electrical and Electronics Engineers, USA, 2016, 770–8.
37. Ba JL, Kiros JR, Hinton GE. Layer normalization. *Stat* 2016; **1050**:21.
38. Ingraham J, Garg V, Barzilay R, *et al.* Generative models for graph-based protein design. *Adv Neural Inf Process Syst* 2019;**32**: 15820–31.
39. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. *Journal of big data* 2019;**6**:1–48.
40. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. *In: 3rd International Conference on Learning Representations (Poster)*. 2015.
41. Paszke A, Gross S, Massa F, *et al.* Pytorch: An imperative style, high-performance deep learning library. *Adv Neural Inf Process Syst* 2019;**32**:8026–37.
42. Yuan Q, Chen J, Zhao H, *et al.* Structure-aware protein–protein interaction site prediction using deep graph convolutional network. *Bioinformatics* 2022;**38**:125–32.
43. Yuan Q, Chen S, Rao J, *et al.* AlphaFold2-aware protein–DNA binding site prediction using graph transformer. *Brief Bioinform* 2022;**23**:bbab564.
44. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 2015;**10**:e0118432.
45. Qiao L, Xie D. MIonSite: ligand-specific prediction of metal ion-binding sites via enhanced AdaBoost algorithm with protein sequence information. *Anal Biochem* 2019;**566**:75–88.
46. Paiva VA, Mendonça MV, Silveira SA, *et al.* GASS-Metal: identifying metal-binding sites on protein structures using genetic algorithms. *Brief Bioinform* 2022;**23**:bbac178.
47. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. *Ann Math Stat* 1947;**18**:50–60.
48. D'agostino R, Pearson ES. Tests for departure from normality. Empirical results for the distributions of b 2 and $\sqrt{}$ b. *Biometrika* 1973;**60**:613–22.
49. Finn C, Abbeel P, Levine S. *Model-agnostic meta-learning for fast adaptation of deep networks*. *In: International Conference on Machine Learning*. The Journal of Machine Learning Research, Australia, 2017, 1126–35. PMLR.
50. Wang J, Zheng S, Chen J, *et al.* Meta learning for low-resource molecular optimization. *J Chem Inf Model* 2021;**61**: 1627–36.
51. Jumper J, Evans R, Pritzel A, *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;1–11.
52. Baek M, DiMaio F, Anishchenko I, *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 2021;**373**:871–6.
53. Wang S, Liang K, Hu Q, *et al.* JAK2-binding long noncoding RNA promotes breast cancer brain metastasis. *J Clin Invest* 2017;**127**: 4498–515.
54. Kumar R, Corbett MA, Van Bon BW, *et al.* THOC2 mutations implicate mRNA-export pathway in X-linked intellectual disability. *Am J Hum Genet* 2015;**97**:302–10.
55. Schmidtke P, Barril X. Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. *J Med Chem* 2010;**53**:5858–67.
56. Xu M, Ran T, Chen H. De novo molecule design through the molecular generative model conditioned by 3D information of protein binding sites. *J Chem Inf Model* 2021;**61**:3240–54.
57. Zheng S, Li Y, Chen S, *et al.* Predicting drug–protein interaction using quasi-visual question answering system. *Nat Mach Intell* 2020;**2**:134–40.
58. Wang P, Zheng S, Jiang Y, *et al.* Structure-aware multimodal deep learning for drug–protein interaction prediction. *J Chem Inf Model* 2022;**62**:1308–17.