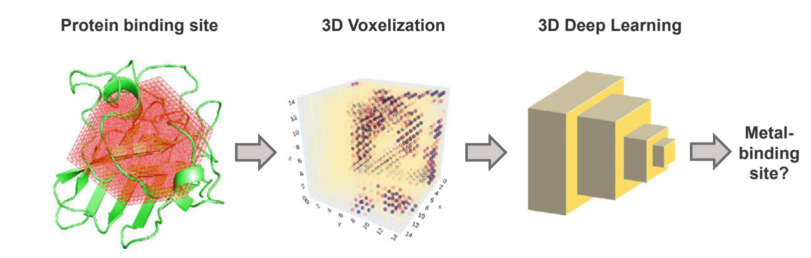
Cell Reports
Physical Science

CellPress
OPEN ACCESS

Article

An ensemble 3D deep-learning model to predict protein metal-binding site



Metal ions play pivotal roles either structurally or functionally in the (patho) physiology of human biological systems, and discovering metal-binding sites of the protein will help gain a better understanding of the proteins' biological functions. Here, Mohamadi et al. use a deep-learning approach to automatically identify the proteins' metal-binding sites.

Ahmad Mohamadi, Tianfan Cheng, Lijian Jin, Junwen Wang, Hongzhe Sun, Mohamad Koohi-Moghadam

hsun@hku.hk (H.S.)
koohi@hku.hk (M.K.-M.)

Highlights

An ensemble 3D deep-learning model has been developed to predict metal-binding sites

3D biophysical features of protein structures are used to train the model

The 3D-CNN method shows high performance in Zn, Fe, Ca, Na, and Mg metal binding

Mohamadi et al., Cell Reports Physical Science 3, 101046
September 21, 2022 © 2022 The Author(s).
https://doi.org/10.1016/j.xcrp.2022.101046

Cell Reports
Physical Science

CellPress
OPEN ACCESS

Article

An ensemble 3D deep-learning model to predict protein metal-binding site

Ahmad Mohamadi,[1] Tianfan Cheng,[2] Lijian Jin,[2] Junwen Wang,[1,3,4] Hongzhe Sun,[5,*] and Mohamad Koohi-Moghadam[1,6,*]

SUMMARY

Predicting metal-binding sites in proteins is critical for understanding the protein's biological function. Here, we develop an ensemble deep convolutional neural network (CNN) method for predicting metal-binding sites based on their three-dimensional (3D) structure. We build multi-channel 3D voxels based on biophysical characteristics obtained from raw atom coordinates of each protein-binding pocket. Then, we use these 3D voxels as the input of an ensemble 3D CNN model. We train and evaluate the model using a curated dataset of 3D protein structures. Our proposed model shows high performance in predicting metal-binding sites for Zn, Fe, Mg, Mn, Ca, and Na. Our approach offers a framework to use 3D spatial features to train 3D-CNN, which may be used to predict complicated metal-binding sites directly from their biophysical characteristics. The source code and webserver of the model are publicly available.

INTRODUCTION

Metal ions are involved in a variety of metabolic activities and cellular pathways, and they play important roles in a wide range of physiological, pathological, and clinical processes. Indeed, nearly 40% of all enzymes with known three-dimensional (3D) structures need a metal ion to catalyze properly.[1] Metal-binding proteins, also known as metalloproteins, are a type of protein family that serves as cofactors in a variety of functions such as metabolic management, signal transduction control, and metal homeostasis.[2] A protein may interact with many metal ions, and this interaction may occur with one or more residues of the protein. The most abundant metal-binding residues in amino acids are ASP, HIS, CYS, and GLU, whereas lesser-known metal-binding residues include SER, GLN, MET, ASN, THR, and TYR.[3–5] Different experimental procedures such as mass spectrometry, electrophoretic mobility shift assay, metal ion affinity column chromatography, gel electrophoresis, nuclear magnetic resonance spectroscopy, absorbance spectroscopy, X-ray crystallography, and electron microscopy can be used to discover the metal-binding sites.[6–8] These methods necessitate time-consuming steps and specialized equipment, making them costly and potentially ineffective for unknown targets. Therefore, providing a cheap and accessible approach to identify protein metal-binding sites will be encouraging.

In recent years, the advancement of computer approaches has allowed for the quick examination of protein metal-binding sites. Several algorithms have been developed to predict protein metal-binding sites.[9–17] These techniques either use explicit biophysical features in amino acid interactions of protein structures or extract features from 2D amino acid sequences to train the models. While these techniques

[1]Division of Applied Oral Sciences and Community Dental Care, Faculty of Dentistry, The University of Hong Kong, Hong Kong S.A.R., PRC

[2]Division of Periodontology and Implant Dentistry, Faculty of Dentistry, The University of Hong Kong, Hong Kong S.A.R., PRC

[3]Department of Quantitative Health Sciences and Center for Individualized Medicine, Mayo Clinic, Scottsdale, AZ 85259, USA

[4]College of Health Solutions, Arizona State University, Scottsdale, AZ 85259, USA

[5]Department of Chemistry and CAS-HKU Joint Laboratory of Metallomics on Health and Environment, The University of Hong Kong, Hong Kong S.A.R., PRC

[6]Lead contact

*Correspondence: hsun@hku.hk (H.S.), koohi@hku.hk (M.K.-M.)

https://doi.org/10.1016/j.xcrp.2022.101046

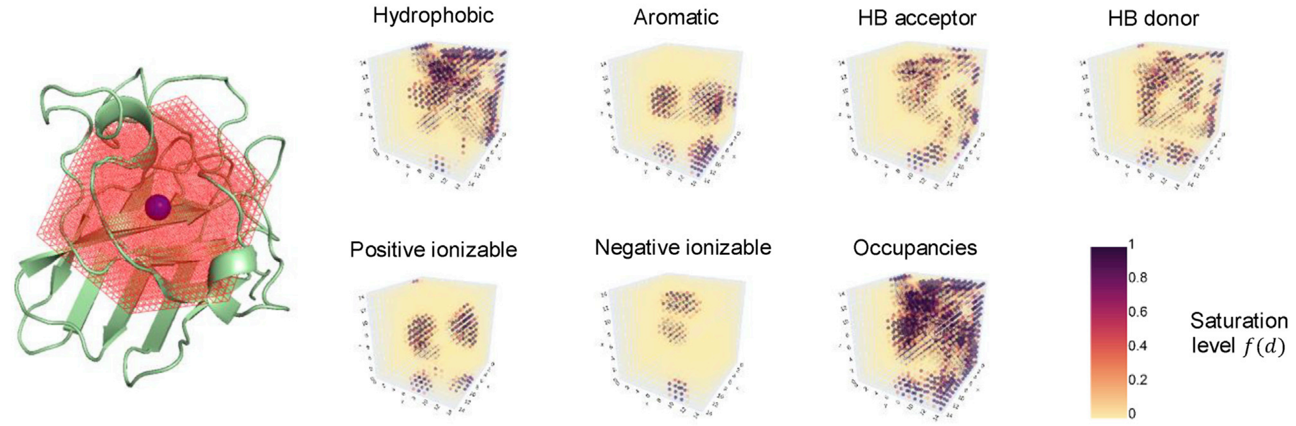


Figure 1. Voxelization of the protein binding pocket

Each PDB sample in the dataset converted to a 3D image that represents the protein structure in a 20 × 20 × 20 Å^3 cubic. Each voxel sample consists of seven channels, in which each channel illustrates a specific biophysical feature: hydrophobic, aromatic, hydrogen-bonding acceptor, hydrogen-bonding donor, positive ionizable, negative ionizable, and occupancies.

have proven to be beneficial in a variety of situations, they do have some drawbacks. For example, biophysics-based approaches are computationally intensive and low throughput, preventing them from being used for large-scale protein structures. On the other hand, studies have found that machine-learning-based approaches that only use 2D sequential features of the metal-binding sites have significant bias in their predictions.[18] The primary bias is due to the fact that training sets are dominated by predefined 2D features and machine-learning algorithms are prone to overfitting to this sequential template data. As a result, there is a need for novel approaches that can generate predictions based on the protein's 3D structure.

Here, we implement MetalSiteHunter, a computational framework based on deep 3D convolutional neural network (CNN) for predicting protein metal-binding sites. Our approach uses the 3D biophysical features derived from raw atom coordinates to train a 3D deep-learning model.[19] We use atom biophysical properties to generate 3D voxels from the protein structures. These 3D voxels can be considered as multi-channel 3D images to train an ensemble 3D deep-learning model to predict protein metal-binding sites. The advantage of this strategy is that it trains the model on the protein's spatial properties, which enables it to incorporate features from amino acids that are close in their 3D coordination but are far apart in the 2D sequence. We exploit the strength of 3D-CNN architecture to detect spatially proximate features. We demonstrate that MetalSiteHunter is highly accurate at predicting Fe, Zn, Mg, Mn, Ca, and Na metal-binding sites on our validation and test dataset. To make our 3D deep-learning method publicly accessible, we developed the MetalSiteHunter webserver. Our webserver will help scientists to gain a better knowledge of protein structures by predicting metal-binding sites.

RESULTS

Overview of MetalSiteHunter

We use recent advances in 3D deep CNN for computer vision to predict the metal-binding sites of proteins based on 3D biophysical properties. We treat protein structures as 3D images, with voxels parameterized according to their atomic biophysical properties (Figure 1). The model requires a 3D structure of the protein (either experimentally determined or predicted). Our pipeline begins by constructing a 3D feature map of the particular protein binding site using the HTMD Python package

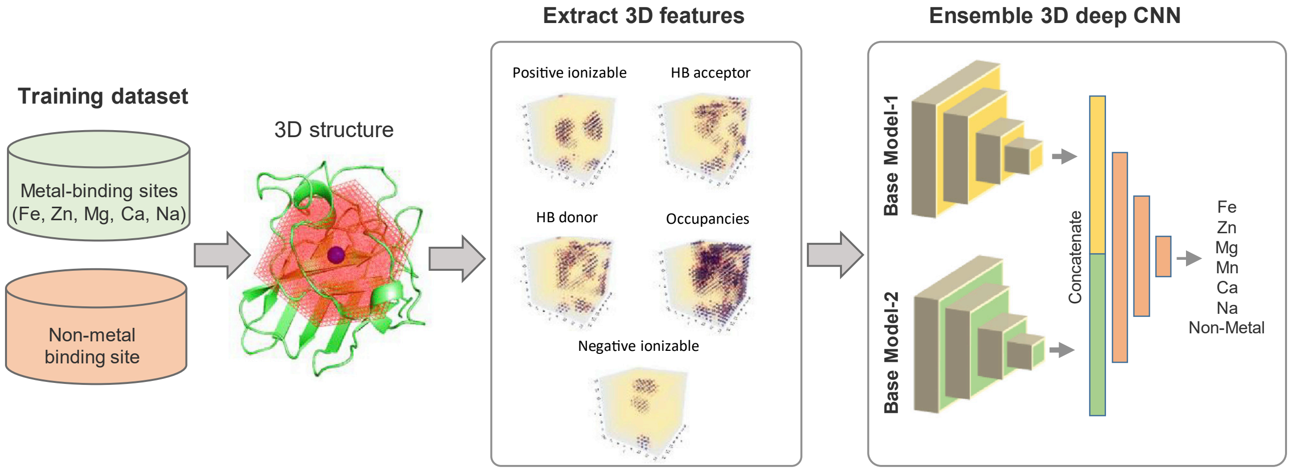


Figure 2. An overview of MetalSiteHunter pipeline

We started by collecting metal- and non-metal-binding site structures from public databases. Then, we built a 3D grid to voxelize the protein binding sites based on their biophysical properties. We used a feature selection approach to select the five top 3D features to feed into our ensemble 3D deep CNN model. The proposed ensemble model has two based models. We concatenated the base models' output, followed by a dense layer to predict the final class label.

(v.1.23.5).[20] The HTMD python package has been widely used by researchers to generate 3D features from protein structures.[21–23] Each voxel is parameterized with seven predetermined criteria (Table S1) that describe the biophysical properties of the atoms that surround it. We further performed feature importance analysis using random forest and selected the five most important criteria based on our training dataset. The feature maps are then layered to build a tensor with the dimensions [20, 20, 20, 5], which is fed into an ensemble deep 3D-CNN with two base models (Figure 2). The based models are two 3D CNNs with different architectures. In one of them, we used global average pooling, and in another one, we used dense layer concatenation. Each 3D-CNN base model is composed of 3D convolutional and 3D pooling layers. The size of convolutional and pooling filter and the size of densely connected layers were tuned using a 5-fold cross-validation procedure. To perform a fair evaluation, we used an unseen test dataset to evaluate our model.

Dataset

We used the MetalPDB[24] database to collect the 3D structures of the protein metal-binding sites. MetalPDB contains around 300,000 sites from more than 50,000 structures. We downloaded 271,792 metal-binding sites for eleven different metals, namely Mg, Zn, Fe, Ca, Na, Mn, Cd, Cu, Sr, Ni, and Co, from the MetalPDB database (downloaded in January 2021). We cleaned the PDB structures by removing water, small-molecule ligands, and metal ions from them. We used the "filter" method of HTMD python package to extract the clean protein structure. The final cleaned PDB structures have been saved into the local machine for further analysis. As MetalPDB compiled all the binding sites of a single protein, some of the binding sites were identical structures that occurred in different domains of the same protein. We used TM-align[25] to remove redundant 3D structure binding sites with a TM score greater than 0.5. We found that 29,039 of these structures are unique among a total of 271,792. As training a deep-learning model requires a sufficient number of training samples, we selected metals with at least 2,000 unique structures (Fe, Zn, Mg, Mn, Ca, and Na) and excluded those with fewer than 2,000 unique structures for the next step (Cd, Cu, Sr, Ni, and Co) (Table S2). Moreover, to build a training dataset for the non-metal-binding site, we downloaded 3,000 random PDBs from the RCSB PDB[26] based on the list of all PDBs provided on the PDB website. We developed a Python code to parse the PDB file to verify that the selected file does not contain a metal. Then, we used Fpocket[27] to find pockets in the PDBs and rank them based on the Fpocket score. We added the pocket with the highest score to our non-metal-binding site database. We checked the similarity between these non-metal-binding sites using TM-align and finally selected 2,469 unique pockets for the next step (Table S2).

Voxelization to extract 3D features

Each protein structure was represented by a set of 3D volume cubes (voxels). We parameterized voxels in a metal-binding site of the protein structure using a set of k biophysical property channels, similar to how we parameterize images with color channels: $[V_1, V_2, \ldots, V_k]$, where the saturation level of the property i at this voxel is represented by the value V_i of each property channel (Figure 1). First, we filtered the PDB to only keep the protein structure, then we used the getVoxelDescriptors function of HTMD to extract a grid of [20 × 20 × 20] voxels from the structure for each particular protein binding site. Each voxel was parameterized with seven property channels, each of which corresponded to a separate biophysical class as defined by the HTMD python module[20] (Table S1). This produced a tensor representing the form of a specific structure, which could be used to create a 3D feature map of the protein binding sites. The final 3D grid was centered on the average location of all C_α atoms in the protein metal-binding site, with each voxel made up of a unit cube with a 1 Å long side. The van der Waals radius r_{vdw} of the atom with a given property and its distance d from the voxel center are used to calculate the saturation level $f(d)$ of each property channel using the formula

$$f(d) = 1 - \exp\left[-\left(\frac{r_{vdw}}{d}\right)^{12}\right]. \qquad \text{(Equation 1)}$$

Tensors were computed from protein structures using functions included in the HTMD Python package for molecular simulations (v.1.23.5). A Python program for constructing input tensors from the protein binding sites is available in the https://github.com/ClinicalAI/metal-site-hunter repository. The 3D orientation of protein structures was taken directly from the obtained PDB files. Although, we implemented the code to rotate input structures about all three cartesian axes to augment the data. We applied the code to the protein binding sites we collected to build the voxels. Finally, we generated 9,027, 3,647, 2,849, 3,315, 2,396, 1,800, and 2,187 voxels for the Mg-, Zn-, Fe-, Ca-, Na-, Mn-, and non-metal-binding sites respectively (Table S2).

Feature selection

We performed a feature selection step to select those feature channels that are more important in metal-binding prediction. To find the most important channels, we trained a random forest classifier over 5-fold cross-validation using 100 estimators. We then ranked the channels that made contributions in building the classifier base on their feature importance attribute.[28] The result shows that positive_ionizable, hbond_acceptor, occupancies, negative_ionizable, and hbond_donor are the five most important channels in our training dataset, with feature importance more than 0.1 (Figure S1). We used these five channels for further training of our model. Using this approach, our model will operate on only five channels, allowing for significant saving of resource during model training and testing.

Training models

We had the lowest number of voxels for Mn-binding sites in our training dataset with 1,800 3D structures, while other binding sites such as Mg had four times as many. We first split the Mn-binding site to 70% for train and 30% for test, which makes 1,260 training samples. For other metal types, we used a down-sampling approach to select 1,260 samples randomly to make a balanced training dataset. The remaining samples have been used as unseen test dataset for these metals. The details of partitioning can be found in Table S3.

We first trained the two based models on the training dataset using 5-fold cross-validation approach. Here, we divided the training dataset into five parts. We trained the base models with four parts and used the one remaining part to evaluate the model. We repeated this procedure five times to cover all training datasets. Based on our cross-validation results, we optimized hyperparameters like the different number of convolutional filters and the size of the fully connected layers. We found that for model-1, 16, 32, and 64 convolutional filters followed by a 64-64-64 fully connected layer is the best configuration (Figure S2), while model-2 showed the best performance with 16 and 32 convolutional filters and a 256-256-256 fully connected layer (Figure S3). Model-1 has 84,966 trainable parameters, and the second model has 214,694 trainable parameters.

After tuning the parameters, we trained the base models for 150 epochs using the Adam optimizer with default hyperparameters (learning rate = 0.001, 1 = 0.9, 2 = 0.999). We then froze the weights of the based models to use them in our ensemble model. For the ensemble model, we concatenated a layer before the softmax layer of each base model and used it as the input of a fully connected layer. Using 5-fold cross-validation, we found that 160-80-40-20-10-6 is the best structure of the fully connected layer for our ensemble model (Figure S4). The final ensemble model has 68,565 trainable parameters. We used again the training dataset to fine-tune our final ensemble model for 500 epochs and Adam optimizer with learning rate = 0.00005. During training, the cross-entropy was employed as the loss function.

Evaluation of the model using 5-fold cross-validation

We first trained and evaluated the base models using our training dataset. We used the 5-fold cross-validation method to train and evaluate each model. We used confusion matrix to evaluate our model. The results of confusion matrix for model-1 showed that this model can predict Ca-, Fe-, Mg-, Mn-, Na-, Zn-, and non-metal-binding sites with a sensitivity of 0.8, 0.9, 0.79, 0.54, 0.86, 0.78, and 0.89. Also, the result for model-2 showed a sensitivity of 0.79, 0.91, 0.82, 0.59, 0.85, 0.86, and 0.88 for these classes (Figure 3). This result shows that model-1 performs better in predicting Ca- and non-metal-binding sites, while model-2 performs better in predicting Fe-, Mg-, Na-, and Zn-binding sites. Both models showed a sensitivity of less than 0.6 for Mn-binding sites, as we had the minimum number of training samples for this metal-binding site.

Model-1 and model-2 were ensembled with parameters and weights that performed well in our previous cross-validation results. We concatenated the layer before softmax layer of model-1 and model-2 and fed it to a fully connected layer to predict the class of binding site label in an ensemble mode. The confusion matrix and accuracy results show that our ensemble model outperformed each of the base models in predicting the metal-binding sites (Figure 4; Table 1). The detailed results of the base models and ensemble models can be found in Tables S4–S6.

Evaluation of the model using unseen data

To provide a fair evaluation, we used the unseen test data to evaluate our model. For each class, we used the remaining samples, except for 1,260, which we used in

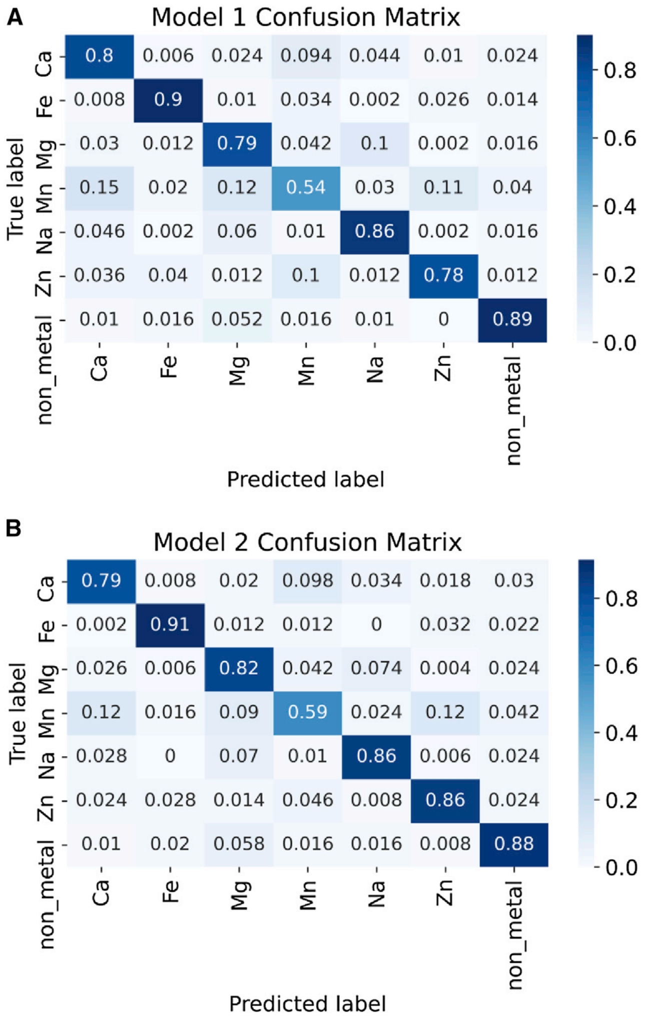


Figure 3. The confusion matrix of base models

The confusion matrix results for 5-fold cross-validation for model-1 (A) and model-2 (B).

training phase as a completely unseen test dataset. The result shows a sensitivity of 0.77, 0.92, 0.89, 0.63, 0.82, 0.84, and 0.92 for Ca, Fe, Mg, Mn, Na, Zn, and non-metal, respectively, of our unseen dataset (Figure 5).

Comparison study

We compared the performance of our webserver with MIB,[14] IonCom,[17] GASS-Metal,[29] and BioMetAll.[30] As there was not a standalone tool for the MIB webserver, we could not compare performances for the entire unseen dataset. We carried out the comparison on 300 samples that were randomly chosen from our unseen dataset. The results show that our webserver's performance is comparable to the other methods in predicting Ca-, Fe-, Zn-, Mg-, and Na-binding sites, but it performs slightly worse in predicting Mn-binding sites (Figure 6).

DISCUSSION

Accurate modeling of the prediction of metal-binding sites is a complex task due to the complexity of the 3D protein structures. The major objective of this work is to demonstrate a novel 3D deep-learning approach to predict metal-binding sites

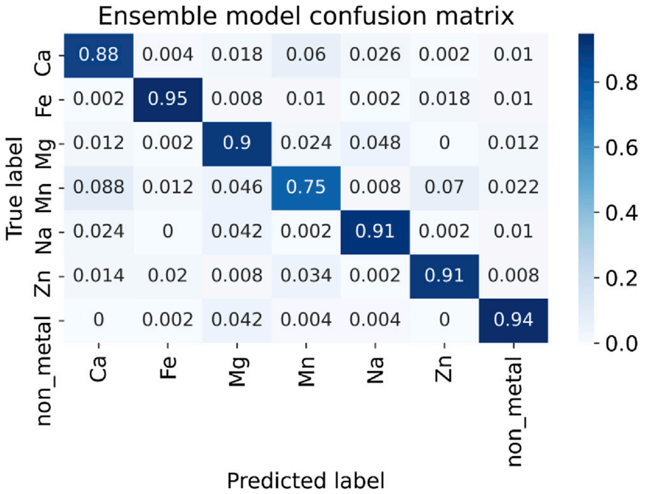


Figure 4. The confusion matrix of ensemble model

The confusion matrix results for 5-fold cross-validation for the ensemble model shows that using both model-1 and model-2 in an ensemble approach improves the final performance of the model.

based on their 3D structures. We addressed the challenge of predicting metal-binding sites from the standpoint of 3D computer vision, utilizing the capability of CNNs to learn the spatial features available in the protein binding sites. Our 3D-CNN method shows high performance in Zn, Fe, Ca, Na, and Mg metal-binding site prediction on an unseen dataset. However, our webserver showed slightly lower performance in predicting the Mn-binding site compared with other tools. We believe that the model's performance can be improved once sufficient samples from Mn-binding sites are available in the future to enrich the training dataset. Besides, our approach may be extended to predict other metal types upon having enough training samples for them.

The advantage of our approach is that it builds 3D biophysical channels of the protein binding sites automatically. The 3D feature extraction in our approach apparently captures more details of biophysical non-linearity. We believe that the 3D-CNN model can predict biophysically relevant information from protein structures, showing potential for protein engineering. These 3D features could potentially be incorporated with other features to enhance the metal-binding site prediction performance in the future. However, the fact that our approach relies on the availability of 3D experimental structures of the protein is the limitation of our model. It is estimated that around 25% of the human proteome has an experimental 3D structure,[31] although the new AlphaFold[32,33] deep-learning model has already predicted almost all the 3D structures of human proteins with high accuracy. To overcome this limitation, the predicted structure of the proteins by Alpha-Fold can be used as the input of our model to predict their metal-binding sites. In addition, the current version of our model is incapable of predicting the coordination number for the predicted metal type due to a lack of relevant training data. We

Table 1. The accuracy results for 5-fold cross-validation of the base and ensemble models

Models	Accuracy
Model-1	0.79 ± 0.003
Model-2	0.81 ± 0.001
Ensemble model	0.89 ± 0.014

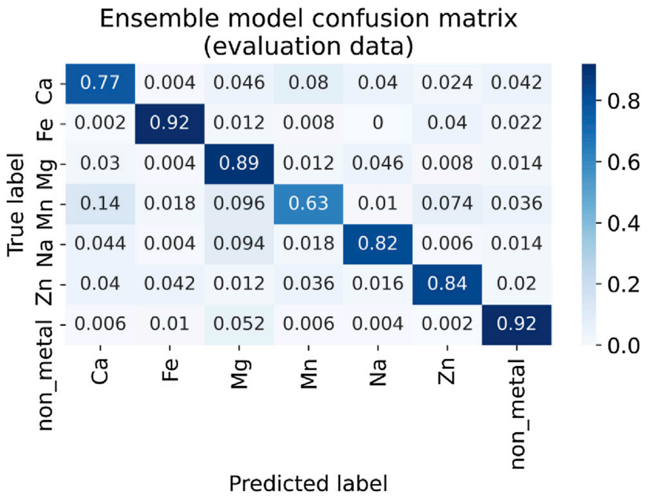


Figure 5. Evaluation of the model using unseen data

The confusion matrix of ensemble model for the unseen data shows that the model is robust enough to predict unseen dataset.

are working in our laboratory to extend the model to support coordination number prediction. We anticipate that more functions of metalloproteins can be uncovered based on the prediction of metal-binding sites in proteins, opening a new avenue of metalloproteomics[34] and their application in life processes.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Further information and requests should be directed to and will be fulfilled by the lead contact, Mohamad Koohi-Moghadam (koohi@hku.hk).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The code and database are publicly available in https://github.com/ClinicalAI/metal-site-hunter The webserver is accessible from https://mohamad-lab.ai/metalsitehunter/

Deep-learning model architecture

CNNs are a deep-learning model type that are frequently utilized in image processing applications. Here, we considered the protein structure as a multi-channel 3D image and used these data as inputs for an ensemble of 3D CNNs to predict metal-binding sites. Our ensemble model has two base models with different architectures. We used global average pooling in model-1, while in model-2, we used a simple flattening layer. Both models use 3D convolutional (Conv3D) layers and a 3D max pooling (MaxPool3D) layer to extract 3D features and reduce the feature map dimension.

For the first step, we applied Conv3D filter with the size of 3 × 3 × 3 on the input voxels. The number of filters for the model-1 is 16, 32, and 64, while for model-2, we used 16 and 32 filters. For each filter *L* we have

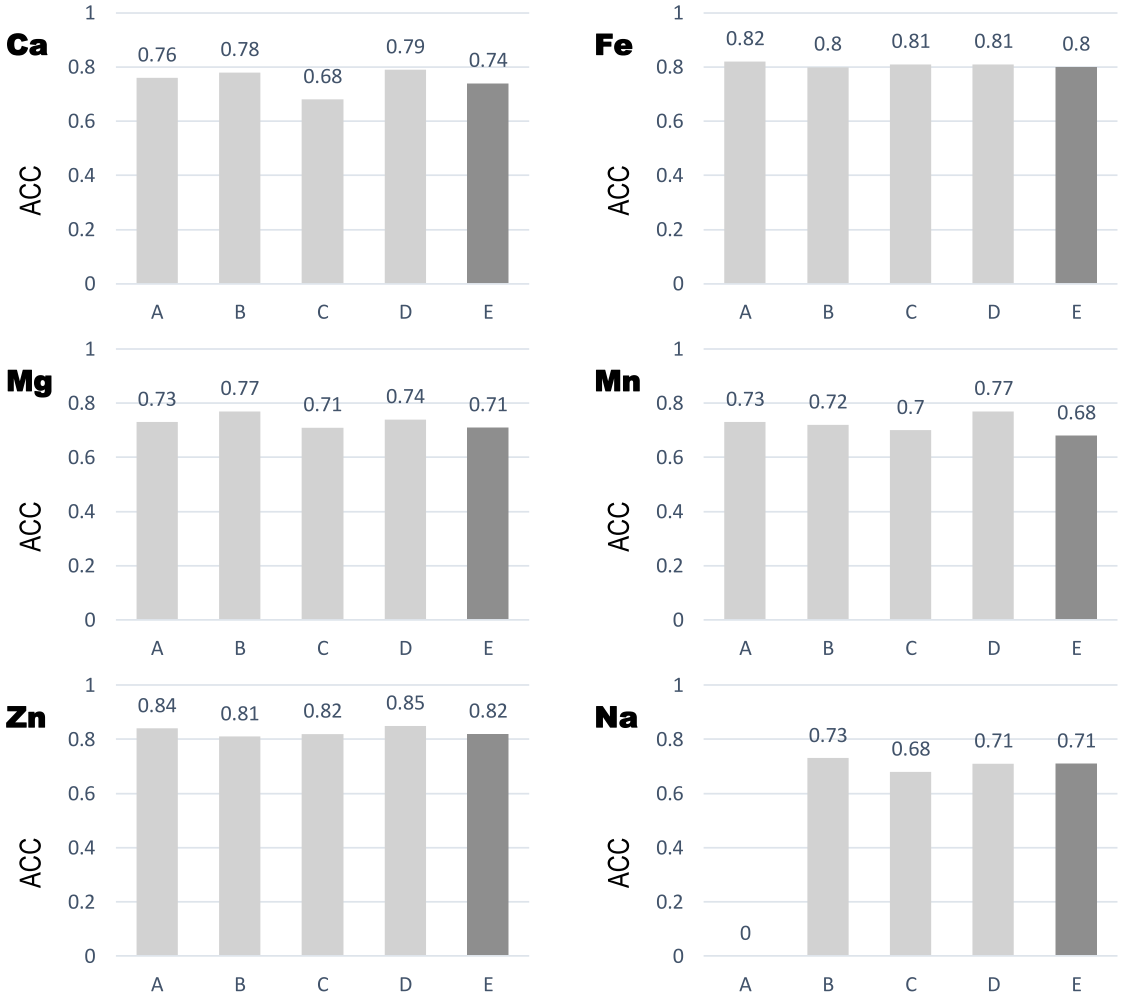


Figure 6. Comparison of the accuracy results of our method with other metal site prediction methods

The accuracy results of our method compared with (A) MIB, (B) IonCom, (C) GASS-Metal, (D) BioMetAll, and (E) MetalSiteHunter (MIB does not support Na-binding site prediction).

$$u^{L}_{i,j,k} = ReLU\left[\sum_{p=0}^{l-1}\sum_{q=0}^{w-1}\sum_{m=0}^{h-0} W^{L}_{i+p,j+q,k+m}X_{i+p,j+q,k+m} + b^{L}\right], \qquad \text{(Equation 2)}$$

$$ReLU = \begin{cases} x, & if\ x \geq 0 \\ 0, & if\ x < 0 \end{cases}$$

where *I*, *j*, and *k* are the indices of the output matrix *u* and *p*, *q*, and *m* denote the filter's indices; *l*, *w*, and *h* denote the filter's size (length, width, and height, respectively); *W* denotes the weight matrix; *X* is the input; and *b* denotes the bias value. The activation function for the convolution layer was a rectified linear unit (ReLU), and the output of the convolution layer was saved as a 3D matrix. Additionally, a MaxPool3D filter was utilized to minimize the dimension of the output of the convolution layer followed by a dropout layer (with a dropout rate of 0.3 for model-1 and 0.4 for model-2). The following equation represents the max-pooling layer function, where *i*, *j*, and *k* are the output cube's indices and *p*, *q*, and *m* are the input cube's indices. $P_{i,j,k}$ denotes the subset of the input matrix's indices that overlap with the max-pooling filter.

$$u_{i,j,k} = \max_{p,q,m \in P_{i,j,k}} X_{p,q,m} \tag{Equation 3}$$

We applied global average pooling on the output of final MaxPool3D layer in model-1. The output of the global average pooling is f_{gap} vector with a size of 64:

$$f_{gap} = \frac{1}{\left|u_{i,j,k}\right|} \sum_{p,q,m \in u_{i,j,k}} u_{p,q,m}. \tag{Equation 4}$$

Model-2 uses a simple flattening layer by concatenating the output of five channels to generate the final vector f_{flat} with a size of 864. These vectors used as input of a fully connected layer:

$$f_n = \sum_{i=0}^{N-1} w_{i,n} x_i + b_n. \tag{Equation 5}$$

Here, f_n is the n^{th} neuron's activation value; w is the weight array with size N and x is the input; and b is the n^{th} node's bias value. The details of the architecture of based model-1 and model-2 can be found in Figures S5 and S6.

We finally combined the output of model-1 and model-2 to build a robust ensemble model (Figure S7). To overcome the vulnerability of the model to the rotation and transition of the protein structure during training and prediction, we used a data augmentation approach. We rotated the 3D voxel structures around the xy, xz, yz plane with 15°, 30°, 45°, 60°, 90°, 120°, and 180° randomly. We implemented the code of the models in Keras python package, with TensorFlow serving as the backend.

Webserver architecture

We developed a web server for the MetalSiteHunter to make the model publicly available (Figure S8). The webserver is accessible from https://mohamad-lab.ai/metalsitehunter/. We used Vue.js as the frontend framework to create a single-page app that communicates with our server through a REST API. Also, for the visualization of protein structures and selecting the location for the metal-binding site prediction, we used the NGL library.[35] In this implementation, after selecting the desired location to search for a metal-binding site, the request with the protein file will be sent to the server. After receiving a request from the client, the server will start the processing procedures that filter the protein structure to ensure it does not have any unnecessary structures. In the next step, the voxels, based on the location information from the client, will be created. After all, the prepared voxels will be fed to the pretrained model, and the result of the probability of each class will be sent back to the client.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xcrp.2022.101046.

ACKNOWLEDGMENTS

The authors gratefully acknowledge support of the University of Hong Kong for seed funding for new staff to M.K.-M. and Research Grants Council of Hong Kong (17308921, 2122-7S04, and 17318322) to H.S.

AUTHOR CONTRIBUTIONS

For the work described herein, M.K.-M., H.S., and A.M. conceived the idea. A.M. implemented the codes and analyzed the results. A.M. and T.C. prepared and validated the dataset. A.M. and M.K.-M. performed result validation. A.M., M.K.-M., and T.C. wrote the paper. L.J. and J.W. commented on and edited the manuscript. M.K.-M. and H.S. provided overall project leadership.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: May 17, 2022
Revised: August 2, 2022
Accepted: August 19, 2022
Published: September 9, 2022

REFERENCES

1. Waldron, K.J., Rutherford, J.C., Ford, D., and Robinson, N.J. (2009). Metalloproteins and metal sensing. Nature *460*, 823–830. https://doi.org/10.1038/nature08300.
2. Messerschmidt, A., Huber, R., Poulos, T., Cygler, M., and Bode, W. (2001). Handbook of Metalloproteins (John Wiley & Sons). https://doi.org/10.1002/0470028637.
3. Dokmanić, I., Sikić, M., and Tomić, S. (2008). Metals in proteins: correlation between the metal-ion type, coordination number and the amino-acid residues involved in the coordination. Acta Crystallogr. D Biol. Crystallogr. *64*, 257–263. https://doi.org/10.1107/S090744490706595X.
4. Cun, S., Lai, Y.-T., Chang, Y.-Y., and Sun, H. (2013). Structure-oriented bioinformatic approach exploring histidine-rich clusters in proteins. Metallomics *5*, 904–912. https://doi.org/10.1039/c3mt00026e.
5. Cheng, T., Xia, W., Wang, P., Huang, F., Wang, J., and Sun, H. (2013). Histidine-rich proteins in prokaryotes: metal homeostasis and environmental habitat-related occurrence. Metallomics *5*, 1423–1429. https://doi.org/10.1039/c3mt00059a.
6. Li, H., and Sun, H. (2012). NMR studies of metalloproteins. Top. Curr. Chem. *326*, 69–98. https://doi.org/10.1007/128_2011_214.
7. She, Y.-M., Narindrasorasak, S., Yang, S., Spitale, N., Roberts, E.A., and Sarkar, B. (2003). Identification of metal-binding proteins in human hepatoma lines by immobilized metal affinity chromatography and mass spectrometry. Mol. Cell. Proteomics *2*, 1306–1318. https://doi.org/10.1074/mcp.M300080-MCP200.
8. Yan, M., Ma, J., and Ji, G. (2016). Examination of effects of Cu (II) and Cr (III) on Al (III) binding by dissolved organic matter using absorbance spectroscopy. Water Res. *93*, 84–90. https://doi.org/10.1016/j.watres.2016.02.017.
9. Lin, C.-T., Lin, K.-L., Yang, C.-H., Chung, I.-F., Huang, C.-D., and Yang, Y.-S. (2005). Protein metal binding residue prediction based on neural networks. Int. J. Neural Syst. *15*, 71–84. https://doi.org/10.1142/S0129065705000116.
10. Shu, N., Zhou, T., and Hovmöller, S. (2008). Prediction of zinc-binding sites in proteins from sequence. Bioinformatics *24*, 775–782. https://doi.org/10.1093/bioinformatics/btm618.
11. Lippi, M., Passerini, A., Punta, M., Rost, B., and Frasconi, P. (2008). MetalDetector: a web server for predicting metal-binding sites and disulfide bridges in proteins from sequence. Bioinformatics *24*, 2094–2095. https://doi.org/10.1093/bioinformatics/btn371.
12. Passerini, A., Lippi, M., and Frasconi, P. (2012 Jan-Feb). Predicting metal-binding sites from protein sequence. IEEE ACM Trans. Comput. Biol. Bioinf *9*, 203–213. https://doi.org/10.1109/TCBB.2011.94.
13. Zheng, H., Cooper, D.R., Porebski, P.J., Shabalin, I.G., Handing, K.B., and Minor, W. (2017). CheckMyMetal: a macromolecular metal-binding validation tool. Acta Crystallogr. D Struct. Biol. *73*, 223–233. https://doi.org/10.1107/S2059798317001061.
14. Lin, Y.-F., Cheng, C.-W., Shih, C.-S., Hwang, J.-K., Yu, C.-S., and Lu, C.-H. (2016). MIB: metal ion-binding site prediction and docking server. J. Chem. Inf. Model. *56*, 2287–2291. https://doi.org/10.1021/acs.jcim.6b00407.
15. Sobolev, V., and Edelman, M. (2013). Web tools for predicting metal binding sites in proteins. Isr. J. Chem. *53*, 166–172. https://doi.org/10.1002/ijch.201200084.
16. Passerini, A., Lippi, M., and Frasconi, P. (2011). MetalDetector v2. 0: predicting the geometry of metal binding sites from protein sequence. Nucleic Acids Res. *39*, W288–W292. https://doi.org/10.1093/nar/gkr365.
17. Hu, X., Dong, Q., Yang, J., and Zhang, Y. (2016). Recognizing metal and acid radical ion-binding sites by integrating ab initio modeling with template-based transferals. Bioinformatics *32*, 3260–3269. https://doi.org/10.1093/bioinformatics/btw396.
18. Walsh, I., Pollastri, G., and Tosatto, S.C.E. (2016). Correct machine learning on protein sequences: a peer-reviewing perspective. Briefings Bioinf. *17*, 831–840. https://doi.org/10.1093/bib/bbv082sl.
19. Koohi-Moghadam, M., Wang, H., Wang, Y., Yang, X., Li, H., Wang, J., and Sun, H. (2019). Predicting disease-associated mutation of metal-binding sites in proteins using a deep learning approach. Nat. Mach. Intell. *1*, 561–567. https://doi.org/10.1038/s42256-019-0119-z.
20. Doerr, S., Harvey, M.J., Noé, F., and De Fabritiis, G. (2016). HTMD: high-throughput molecular dynamics for molecular discovery. J. Chem. Theor. Comput. *12*, 1845–1852. https://doi.org/10.1021/acs.jctc.6b00049.
21. Jiménez, J., Škalič, M., Martínez-Rosell, G., and De Fabritiis, G. (2018). K deep: protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. J. Chem. Inf. Model. *58*, 287–296. https://doi.org/10.1021/acs.jcim.7b00650.
22. Skalic, M., Jiménez, J., Sabbadin, D., and De Fabritiis, G. (2019). Shape-based generative modeling for de novo drug design. J. Chem. Inf. Model. *59*, 1205–1214. https://doi.org/10.1021/acs.jcim.8b00706.
23. Wehmeyer, C., and Noé, F. (2018). Time-lagged autoencoders: deep learning of slow collective variables for molecular kinetics. J. Chem. Phys. *148*, 241703. https://doi.org/10.1063/1.5011399.
24. Putignano, V., Rosato, A., Banci, L., and Andreini, C. (2018). MetalPDB in 2018: a database of metal sites in biological macromolecular structures. Nucleic Acids Res. *46*, D459–D464. https://doi.org/10.1093/nar/gkx989.
25. Zhang, Y., and Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. *33*, 2302–2309. https://doi.org/10.1093/nar/gki524.
26. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The protein data bank. Nucleic Acids Res. *28*, 235–242. https://doi.org/10.1093/nar/28.1.235.
27. Le Guilloux, V., Schmidtke, P., and Tuffery, P. (2009). Fpocket: an open source platform for ligand pocket detection. BMC Bioinf. *10*, 168–211. https://doi.org/10.1186/1471-2105-10-168.
28. Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional variable importance for random forests. BMC Bioinf. *9*, 307–311. https://doi.org/10.1186/1471-2105-9-307.
29. Paiva, V.A., Mendonça, M.V., Silveira, S.A., Ascher, D.B., Pires, D.E., and Izidoro, S.C. (2022). GASS-Metal: identifying metal-binding sites on protein structures using genetic algorithms. Brief Bioinform. https://doi.org/10.1093/bib/bbac178.

30. Sánchez-Aparicio, J.E., Tiessler-Sala, L., Velasco-Carneros, L., Roldán-Martín, L., Sciortino, G., Maréchal, J.D., and Modeling. (2020). BioMetAll: identifying metal-binding sites in proteins from backbone preorganization. J. Chem. Inf. Model. *61*, 311–323. https://doi.org/10.1021/acs.jcim.0c00827.

31. Porta-Pardo, E., Ruiz-Serra, V., Valentini, S., and Valencia, A. (2022). The structural coverage of the human proteome before and after AlphaFold. PLoS Comput. Biol. *18*, e1009818. https://doi.org/10.1371/journal.pcbi.1009818.

32. Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., et al. (2021). Highly accurate protein structure prediction for the human proteome. Nature *596*, 590–596. https://doi.org/10.1038/s41586-021-03828-1.

33. Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A.W.R., Bridgland, A., et al. (2020). Improved protein structure prediction using potentials from deep learning. Nature *577*, 706–710. https://doi.org/10.1038/s41586-019-1923-7.

34. Zhou, Y., Li, H., and Sun, H. (2022). Metalloproteomics for biomedical Research: methodology and applications. Annu. Rev. Biochem. *91*, 449–473. https://doi.org/10.1146/annurev-biochem-040320-104628.

35. Rose, A.S., Bradley, A.R., Valasatava, Y., Duarte, J.M., Prlić, A., and Rose, P.W. (2018). NGL viewer: web-based molecular graphics for large complexes. Bioinformatics *34*, 3755–3758. https://doi.org/10.1093/bioinformatics/bty419.