

PinMyMetal: a hybrid learning system to accurately model transition metal binding sites in macromolecules

Received: 13 February 2024

Accepted: 24 February 2025

Published online: 28 March 2025

 Check for updates

Huihui Zhang^{1,2,3}, Juanhong Zhong^{2,3}, Michal Gucwa^{4,5}, Yishuai Zhang^{2,3}, Haojie Ma², Lei Deng³, Longfei Mao², Wladek Minor⁴✉, Nasui Wang⁶✉ & Heping Zheng¹✉

Metal ions are vital components in many proteins for the inference and engineering of protein function, with coordination complexity linked to structural, catalytic, or regulatory roles. Modeling transition metal ions, especially in transient, reversible, and concentration-dependent regulatory sites, remains challenging. We present PinMyMetal (PMM), a hybrid machine learning system designed to accurately predict transition metal localization and environment in macromolecules, tailored to tetrahedral and octahedral geometries. PMM outperforms other predictors, achieving high accuracy in ligand and coordinate predictions. It excels in predicting regulatory sites (median deviation 0.36 Å), demonstrating superior accuracy in locating catalytic sites (0.33 Å) and structural sites (0.19 Å). Each predicted site is assigned a certainty score based on local structural and physicochemical features, independent of homologs. Interactive validation through our server, CheckMyMetal, expands PMM's scope, enabling it to pinpoint and validate diverse functional metal sites from different structure sources (predicted structures, cryo-EM, and crystallography). This facilitates residue-wise assessment and robust metal binding site design. The lightweight PMM system demands minimal computing resources and is available at <https://PMM.biocloud.top>. The PMM workflow can interrogate with protein sequence to characterize the localization of the most probable transition metals, which is often interchangeable and hard to differentiate by nature.

Metal ions play a crucial role in the structure and function of macromolecules¹, acting as essential cofactors for many enzymes and influencing various molecular and cellular processes^{2,3}. About one-third of proteins in known genomes require metal ions to maintain

their natural structure and function⁴. However, only a small fraction of metal binding proteins has been elucidated^{5,6}. Understanding the location of metals in proteins and their interactions is essential for designing new drug synthesis pathways and modifying biological

¹Department of Cardiology, First Affiliated Hospital of Shantou University Medical College, Shantou, Guangdong, People's Republic of China. ²Hunan University College of Biology, Bioinformatics Center, Changsha, Hunan, People's Republic of China. ³Hunan Provincial Key Laboratory of Medical Virology, Hunan University, Changsha, Hunan, People's Republic of China. ⁴Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA, USA. ⁵Department of Computational Biophysics and Bioinformatics, Jagiellonian University, Cracow, Poland. ⁶Division of Endocrinology and Metabolism, First Affiliated Hospital of Shantou University Medical College, Shantou, Guangdong, People's Republic of China.

✉ e-mail: wladek@iwonka.med.virginia.edu; nswang@stu.edu.cn; dust.zheng@hotmail.com

functions^{7–9}. For example, the most abundant metal ion in the Protein Data Bank (PDB) is zinc, crucial in diseases, drug targeting, stability, and regulation¹⁰.

Metal-protein complex studies benefit from experimental methods yet face artifacts like incorrect metal incorporation and ion removal during purification^{11,12}. In addition, experimental methods face resolution limitations when determining metal binding structures, particularly in cryo-electron microscopy (cryo-EM). Despite the success of cryo-EM in large and complex macromolecules, electron penetration depth and scattering effects hinder high-resolution imaging of metal ions¹³. Computational predictions offer advantages, including cost-effectiveness, scalability, and high throughput. Combining both approaches provides a more comprehensive understanding of metal sites in proteins. Metal sites typically comprise amino acids close in 3D structure but distant in sequence, posing a challenge to identify sites with short amino acid spacers between ligands, such as regulatory sites¹⁴. Hence, structure-based predictions are expected to outperform sequence-based methods^{15–17}. Advancements in protein structure prediction, exemplified by AlphaFold2, show promise for accurate predictions of protein structures, offering opportunities and challenges in annotating metal sites in computational models¹⁸.

Existing structure-based metal site predictors employ diverse approaches. BioMetAll¹⁹, TEMSP²⁰, and GRE4Zn¹⁵ use geometric features, such as metal-ligand distances. CHED²¹ focuses on triads of metal-coordinating ligand residues in apoprotein structures. Zinc-BindDB classifies zinc sites into ten classes, employing machine learning models based on structural characteristics²². MIB^{23,24}, and AlphaFill²⁵ infer the presence of metal ions based on homology to known metal binding structures. Metal3D employs a deep learning algorithm with a voxelized protein environment representation²⁶.

These predictors can be divided into three categories: (I) binding site predictors for metal binding residues (CHED²¹, ZincBindDB²²); (II) binding position predictors for metal ion coordinates (Metal3D²⁶, BioMetAll¹⁹, AlphaFill²⁵); (III) predictors that identify both residues and coordinates (TEMSP²⁰, GRE4Zn¹⁵, MIB^{23,24}). However, these methods have significant drawbacks. BioMetAll lacks templates and a confidence metric but provides many potential binding site locations on a grid, whose strategy finds the site at the cost of increasing site uncertainty¹⁹. CHED, TEMSP, GRE4Zn, and MIB exclude metal sites with two or fewer coordinating ligands. Metal3D can only predict the coordinates of metal ions and has a long prediction time, unsuitable for large-scale predictions²⁶. Homology-based predictors like MIB, AlphaFill, and ZincBindDB can successfully find sites that match known metal site patterns, while identifying metal binding sites (MBSs) in proteins lacking sufficient homologous structural domains or motifs remains challenging. The structure-based hybrid machine learning system developed herein, named PinMyMetal (PMM), overcomes these drawbacks to predict both metal location and coordinating ligands.

Metal ions in proteins are typically coordinated by Cysteine(C), Histidine(H), Glutamate(E), and Aspartate(D)²⁷. The ligands are mainly oxygen, nitrogen, and sulfur-containing groups^{28,29}, with coordination numbers typically ranging from four to six, forming tetrahedral or octahedral geometries³⁰. Due to its weaker ligand field, C commonly forms tetrahedral structures with metal ions, while E, with stronger ligand fields, favors octahedral coordination³¹. The varied functions of transition metal binding sites exhibit distinct structures^{16,32}. Taking tetrahedral zinc binding sites as an example, these sites are functionally divided into structural, catalytic, and regulatory (inhibitory) sites, predominantly coordinated by four, three, and two residues, respectively^{33–35}.

The PMM system employs a hybrid learning approach to identify MBSs based on different geometries. For tetrahedral coordination, the algorithm uses a CH-based approach, focusing on C and H residues,

while for octahedral coordination, it applies the EDH-based approach, considering combinations of E, D, and H residues. We have applied CH as the primary measure for copper and zinc, while used EDH as the primary measure for all transition metals, formulating a distinctive tool that can characterize the localization and environment of transition metal ions by certainty scores. This approach improves the accuracy of identifying potential MBSs by providing adequate training data for each geometric category^{16,32}. Trained on 4984 non-redundant high-quality MBSs validated by CheckMyMetal (CMM)^{36,37}, the PMM system incorporates predicted sites into protein structures and further validates them using CMM. It efficiently screens and validates both metal ion locations and coordinating ligands throughout the protein based on amino acid type, location coordinates, structural characteristics, and surrounding hydrophilic profile.

In this work, we present PMM, a hybrid machine learning system that accurately predicts transition metal binding sites and their coordination environments. Unlike transition metals involved in redox reactions, electron transport, and catalysis³⁸, alkali and alkaline earth metals (e.g., Na, K, Ca, Mg) primarily maintain electrolyte balance, nerve transmission, and muscle contraction^{39,40}. The CH and EDH geometry-based models in PMM require substantial modification to accommodate the distinct chemical properties, weaker ligand binding, and more flexible coordination of these metals⁴¹.

Results

PinMyMetal (PMM) employs a hybrid learning system to predict MBSs by integrating geometric and chemical features. First, geometric constraints are applied to identify candidate sites, focusing on ligand pairs suitable for tetrahedral (CH sites) and octahedral (EDH sites) geometries based on specific amino acid compositions. Next, candidate sites are validated with certainty scores assessed using an ensemble learning model for low-coordination sites (LCS, \leq half of the full coordination number) and the Pearson correlation coefficient for high-coordination sites (HCS, $>$ half). Finally, for validated sites with certainty scores above 0.5, the system predicts the most probable metal type using another specialized ensemble learning model. PMM is trained on a benchmark dataset of CMM-validated, non-redundant MBSs, comprising 4984 transition MBSs from 2778 structures, including metal site information at crystal interfaces (Supplementary Table 1).

Prediction of candidate MBSs and ion coordinates with PMM

Using the CMM-validated benchmark dataset (Supplementary Table 1), we assess the accuracy of predictions for candidate MBS ligands by calculating the overlap between predicted and actual ligand sets at MBSs, evaluated by IoUR defined in formula (6). An IoUR threshold of ≥ 0.5 indicates that the predicted ligands overlap with at least 50% of the actual ligands and is marked as positive prediction. The results show that for the transition metals, including Mn, Fe2, Fe, Co, Ni, Cu, and Zn, the recall rates at IoUR ≥ 0.5 exceed 90%. Specifically, PMM demonstrate the highest recall for Fe2, followed by Zn, Fe, Ni and other metals. An IoUR value of 1 indicates that the predicted ligands exactly match the actual ligands, the recall rates are: Mn 60.76%, Fe2 69.09%, Fe 69.38%, Co 78.46%, Ni 75.10%, Cu 58.74%, and Zn 83.77% (Supplementary Table 2). The procedure may exclude some experimental sites from consideration due to certain complications, e.g., for sites with distance between ligands exceeding 4.5 Å or for sites with coordinated atoms being N or O of the backbone peptide bond.

Based on the predicted candidate MBS ligands, PMM employs an innovative algorithm that integrates factors such as ligand composition, atomic positions, bond angles, and distances to deduce the coordinates of the metal ion. Different coordination environments are handled with specific strategies to predict the optimal ion position based on the spatial arrangement of all ligands. The accuracy is evaluated using the benchmark dataset by measuring the average metal

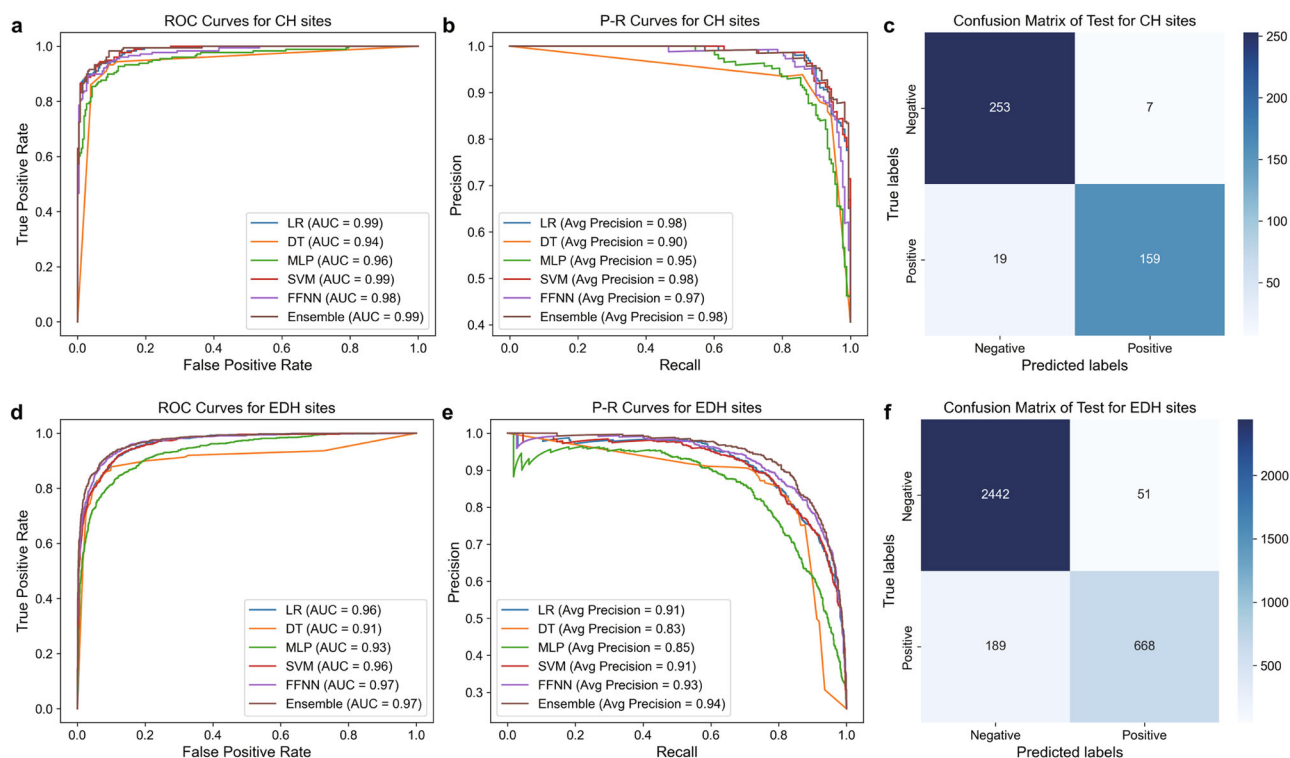


Fig. 1 | Performance metrics for predicting low-coordination CH and EDH sites. Performance metrics for predicting low-coordination CH (a–c) and EDH (d–f) sites. **a, d** Receiver operating characteristic (ROC) curves show strong model performance, with AUC values of 0.99 for CH and 0.97 for EDH. **b, e** Precision–

recall (P-R) curves highlight the ensemble model's superior balance, achieving 0.98 and 0.94 average precision for CH and EDH, respectively. **c, f** Confusion matrices show higher accuracy for CH sites, with EDH predictions proving more challenging.

deviation between predicted and experimental positions for correctly identified sites (certainty scores > 0.5).

The overall average deviation and its dispersion among different metal ions at CH and EDH sites remain relatively small (Supplementary Fig. 1). For CH sites, Zn demonstrates higher prediction accuracy with a concentrated distribution (0.279 ± 0.29 Å) than Cu that shows a larger deviation with a wider range (0.558 ± 0.48 Å) (Supplementary Fig. 1a). For EDH sites, the deviation distributions for Ni, divalent iron Fe2, and Fe suggest relatively accurate predictions with average deviations below 0.41 Å, while Mn and Cu have slightly higher deviations (Supplementary Fig. 1b). This indicates that despite the differences in average metal deviations, the PMM models achieve consistently high accuracy in predicting metal ion positions at both CH and EDH sites. PMM demonstrates high accuracy in predicting candidate MBSs by identifying the ligands at these sites and subsequently inferring the metal positions, providing a strong foundation for further identification and classification of MBSs.

Handling LCS: certainty score by ensemble learning

In a hybrid learning system, we adopt different strategies to assign certainty score to candidate sites for LCS and HCS. For LCS prediction, two independent ensemble learning models are applied: one for predicting CH sites and another for predicting EDH sites. Each model computes certainty score by integrating five classifiers, including Logistic Regression (LR), Decision Tree (DT), MLPClassifier (MLP), Support Vector Machine (SVM), and Feedforward Neural Network architecture (FFNN). The models' performance is assessed using receiver operating characteristic (ROC) curves, precision–recall (P-R) curves, and confusion matrices.

For CH sites, the ROC curves (Fig. 1a) indicate that most models perform well, with SVM, LR, and the ensemble model achieving an area under the curve (AUC) of 0.99, followed by FFNN with an AUC of 0.98 and the decision tree with a lower AUC of 0.94. The precision–recall

curves (Fig. 1b) further highlight that SVM, LR, and the ensemble method achieve an average precision of 0.98, demonstrating high precision and robustness. The confusion matrix of the ensemble model (Fig. 1c) shows that 159 out of 178 positive samples are correctly predicted, and 253 out of 260 negative samples are correctly identified, with only 19 false negatives (FN) and 7 false positives (FP). Based on the confusion matrix, the ensemble model achieves a recall of 0.893, precision of 0.958, F1-score of 0.924, and accuracy of 0.940, demonstrating excellent performance with minimal misclassification in predicting CH sites.

For EDH sites, the ROC curves (Fig. 1d) reveal that both the ensemble model and FFNN achieve the highest AUC of 0.97, followed by SVM and LR (AUC = 0.96). The decision tree exhibits the lowest AUC (0.91). The precision–recall curves (Fig. 1e) show that the ensemble model has the highest average precision (0.94), while FFNN (0.93) and SVM (0.91) also perform well, with the decision tree lagging behind at 0.83. The confusion matrix of the ensemble model (Fig. 1f) indicates that 668 out of 857 positive samples are correctly predicted, and 2442 out of 2493 negative samples are correctly identified, with 189 false negatives and 51 false positives. The recall is 0.779, precision is 0.929, F1-score is 0.847, and accuracy is 0.928, showing that EDH site prediction present higher rates of misclassification compared to CH site prediction. Predicting EDH sites is more challenging due to the complexity of the E, D, and H ligand arrangements.

In both tasks, the ensemble model consistently outperforms the other base learners, achieving higher AUC and precision scores. However, the prediction of CH sites is more accurate, with fewer misclassifications, while the prediction of EDH sites proves more challenging due to the increased complexity of the ligand arrangements. The model is tuned to prioritize precision over recall to achieve optimal performance. Overall, the ensemble model performs robustly predicting LCS, with higher precision and F1-score, especially for CH sites.

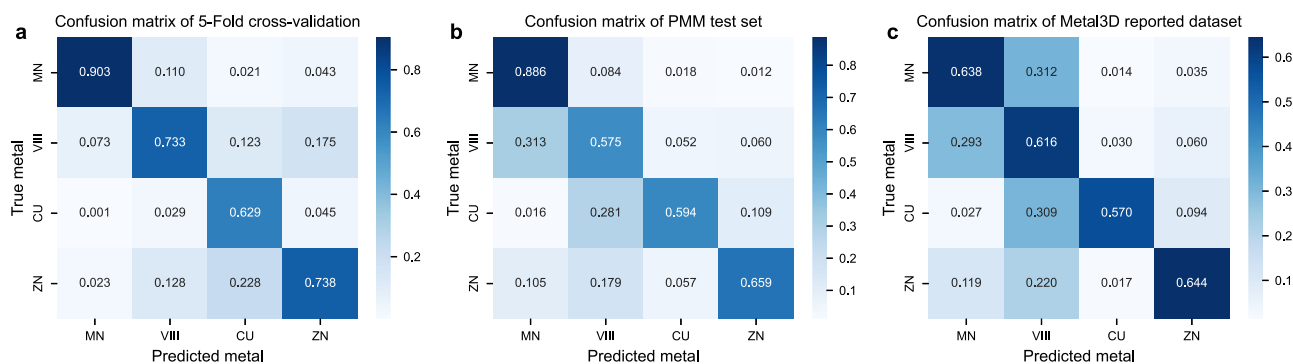


Fig. 2 | Confusion matrices displaying the classification performance of the ensemble model for determining metal types. a Fivefold cross-validation on training set. **b** Results on the PMM test set. **c** Results on the unseen data

from Metal3D, which was used in the Metal3D paper to test its selectivity for other transition metals. This is the same dataset referenced in Fig. 4b of Section 2.6.

The certainty score ranges from 0 to 1. By evaluating the model's performance metrics (Recall, Precision, F1-Score, and Accuracy) on the test dataset across different thresholds, we identify 0.5 as the optimal probability threshold for confirming MBSSs. Candidates with a score greater than 0.5 are considered verified sites. At this threshold, the F1-Score for CH sites is 0.924, with high precision (0.958) and recall (0.893), and an overall accuracy of 0.941. The F1-Score for EDH sites at 0.5 is 0.848, with no improvement at 0.55. Thus, 0.5 is selected as the best threshold for optimal model performance (Supplementary Table 3).

Handling HCS: certainty score by Pearson correlation

The certainty score for HCS is derived from hydrophobicity contrast function values (C) and mean atomic solvation parameters ($\Delta\sigma$) obtained from 21 radii around the metal coordinates within 2 Å to 7 Å. For predicted sites classified using the CH-based or EDH-based algorithm, the score is calculated as the average of the Pearson correlation coefficient (r) values between the predicted site curves and the corresponding standard curves of the same algorithm for both C (Supplementary Fig. 2a, b) and $\Delta\sigma$ (Supplementary Fig. 2c, d). A certainty score greater than 0.5 indicates that the predicted site is classified as a metal site.

To validate the accuracy of the method, the similarity between each experimental site curve and the standard curve is evaluated using box and violin plots (Supplementary Fig. 3). The experimental data is derived from the benchmark dataset of CMM-validated non-redundant HCS. The results show an average r value of 0.91 for CH sites, with 1904 of 1905 sites having values greater than 0.5. For EDH sites, the r value is 0.98, with all 745 sites exhibiting values above 0.5. These findings demonstrate that the method effectively captures the relationship between individual site characteristics and the standard curve for both CH and EDH sites.

Identification of metal types

A classifier is trained to differentiate the type of transition metals for the metal binding sites that pass the certainty score test. Six commonly encountered transition metals in the fourth period (Mn, Fe, Co, Ni, Cu, Zn) are taken into consideration. While Mn, Cu, Zn each represent a separate class, Fe, Co, Ni are combined to form a congregate class VIII due to the high similarity in physicochemical properties, local environments, and binding geometries, along with the scarcity of structural data for Co and Ni. This 4-classes model reduces classification ambiguity and improves model accuracy compared to the model that considers each of the six transition metals in the fourth period as a separate class. Using 5-fold cross-validation on the training set, PMM achieves the best performance for Mn with an accuracy of 90.3% and the worst performance for Cu with an accuracy of 62.9%. The classification accuracies for Zn and VIII are 73.8% and 73.3%, indicating a

moderate level of misclassifications. The most notable misclassifications highlight Zn misclassified as Cu (22.8%) or VIII (12.8%) (Fig. 2a).

We also evaluate the model using the test set and an unseen dataset from Metal3D. The confusion matrices for both datasets show that PMM performs robustly in predicting Mn and Zn binding sites. Specifically, Mn has accuracies of 88.6% in the test set and 63.8% in the Metal3D dataset, while Zn achieves accuracies of 65.9% in the test set and 64.4% in Metal3D. However, the misclassification rates between VIII and Mn or Cu are higher. The prediction accuracies for VIII are 57.5% in the test set and 61.6% in Metal3D. These results suggest that while PMM can effectively distinguish Mn from Zn, challenges remain in differentiating VIII from either Mn or Cu (Fig. 2b, c).

Prediction of unknown functional sites supported by experimental data

In addition to accurately predicting known experimental binding sites, PMM identifies numerous unknown, putative MBSSs, including both LCS and HCS that are not determined in experimental structures. Some of these predicted sites are found in cryo-EM structures. LCS often serve as regulatory sites that can reversibly bind metals depending on factors such as metal concentration or the presence of chaperone proteins in the environment. Therefore, the absence of a predicted LCS under certain experimental conditions does not necessarily exclude its potential to bind metals. HCS, on the other hand, may be skipped due to oversight during model building or uncertainty in metal ion modeling from low-resolution structures.

For instance, in the low-resolution X-ray structure of wild-type RNA polymerase II (1nik)⁴² determined to a resolution of 4.1 Å, a CH4 site with four cysteine residues does not have the zinc ion modeled despite the presence of electron density (Fig. 3a). Another example is the TRAP-Anti-TRAP complex structure with a resolution of 3.2 Å (2zp9)⁴³, PMM predicted a CH4 site with four cysteine residues. While electron density is observed at this site, it is not modeled in the experimental structure (Fig. 3b).

To assess the efficiency of metal annotation, the number of transition metal ions per 100 amino acids serves as a useful metric, particularly given the association between lower resolution and increased uncertainty in metal ion modeling. Structures with resolutions better than 2.5 Å are excluded due to the scarcity of atomic-resolution cryo-EM structures (41 structures). The cryo-EM method is commonly used for determining large, complex, or challenging-to-crystallize structures. However, the annotation efficiency for transition metal ions is lower in cryo-EM structures compared to X-ray structures of the same resolution range, consistently decreasing from 0.25 metal ions per 100 amino acids at 3 Å to 0.05 metal ions per 100 amino acids at 5 Å (Supplementary Fig. 4). PMM is well suited

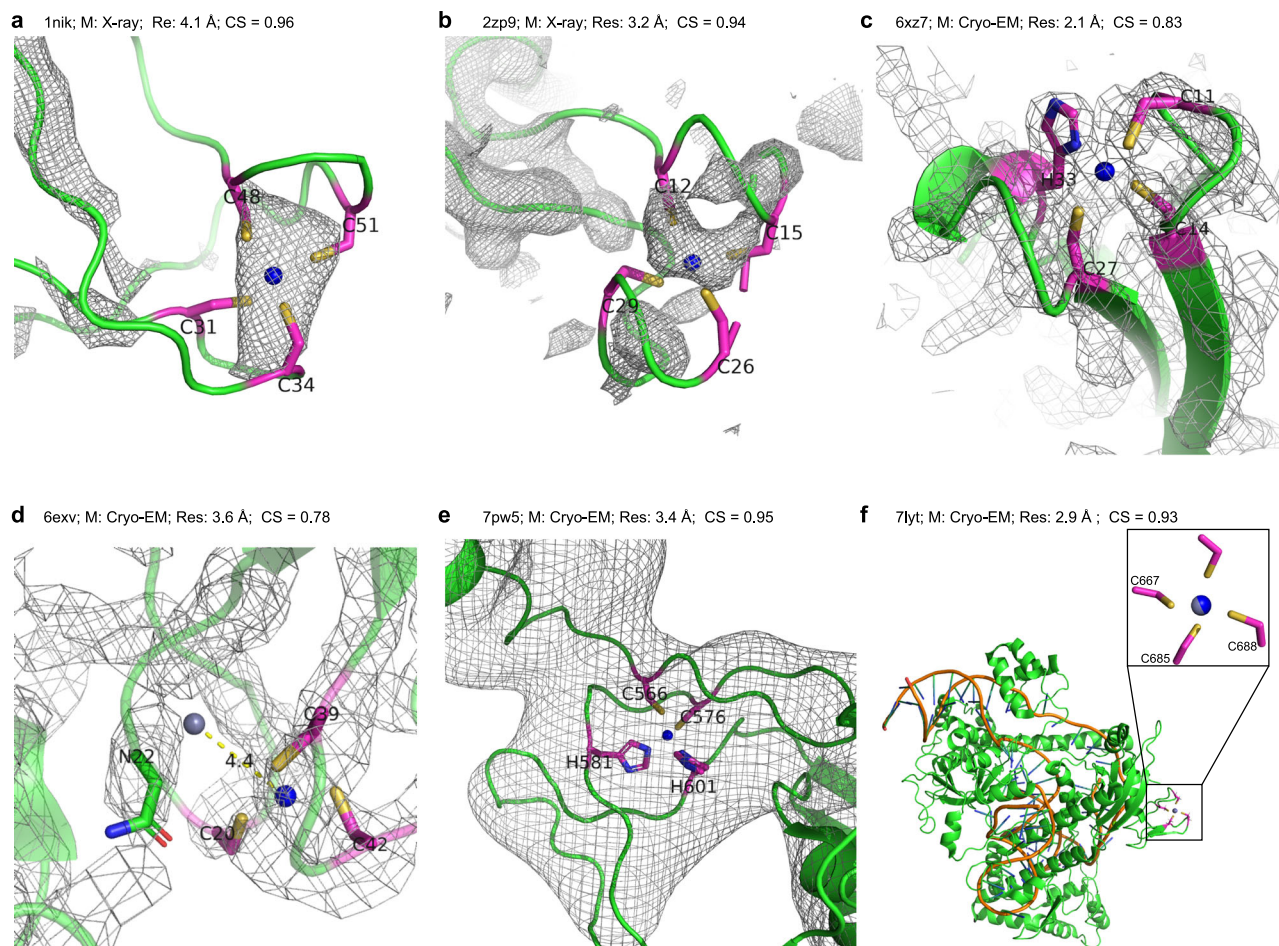


Fig. 3 | Zinc binding sites predicted by PMM. a 1nik, wild-type RNA polymerase II, chain L, 4-residue zinc site, $2F_o - F_c$ map with 3.0σ cutoff. **b** 2zp9, TRAP-Anti-TRAP complex, chain H, 4-residue zinc sites, $2F_o - F_c$ map with 1.0σ cutoff. **c** 6xz7, *E. coli* 50S ribosomal subunit complex, chain e, 4-residue zinc site, EM map with 5.0σ cutoff. **d** 6exv, mammalian RNA polymerase II subunit RPB7, chain I, 3-residue zinc site, EM map with 5.0σ cutoff. **e** 7pw5, human SMG1-8-9 kinase complex, chain B,

4-residue zinc site, EM map with 5.0σ cutoff. **f** PMM predicts a single zinc site in the cryo-EM structure 7lyt (CasPhi-2 (Cas12j) complex with crRNA and Phosphorothioate-DNA). EM map with 5.0σ cutoff. CS: certainty score; Blue spheres represent predicted zinc sites, while gray spheres depict experimentally determined zinc sites; Electron density maps ($2F_o - F_c$ or EM) are shown in gray mesh with optimal σ cutoff in the proximity of the metal sites. M method, Res resolution.

to routinely model missing MBSs or annotate candidate MBSs in cryo-EM structures.

For example, PMM predicts a zinc site on 50 s ribosomal protein L36 (6xz7, Chain e)⁴⁴, coordinated by residues C11, C14, C27, and H33, and is supported by an observed peak in the charge density map (Fig. 3c). In mammalian RNA polymerase II subunit RPB7 (6exv)⁴⁵, PMM predicts a zinc site coordinated by residues C20, C39, and C42. Though not experimentally modeled, it gives an educated estimation of the candidate zinc binding site that is not contradictory to the charge density map. Conversely, a nearby zinc binding site modeled by the experimenter is not reasonably coordinated and lacks experimental support (Fig. 3d). These discrepancies underscore the challenges in cryo-EM structural determination, while PMM's prediction suggests its potential in supplementing MBS modeling. Additionally, PMM predicts a zinc site on the SMG8 protein (7pw5)⁴⁶, coordinated by C566, C576, H581, and H601 (Fig. 3e). While the insufficient resolution may not support the direct atomic modeling of metal ions in this model, PMM provides an alternative approach to model coordination bonds pertaining to metal ions in medium-to-low-resolution cryo-EM structures.

Moreover, PMM accurately predicts metal MBSs not only in protein structures but also in complex macromolecular assemblies, such as the Cryo-EM structure of the CasPhi-2 (Cas12j) complex with crRNA and Phosphorothioate-DNA (7lyt)⁴⁷. PMM successfully predicted the zinc binding site coordinated by residues C670, C667, C685, and C688

(Fig. 3f), with a minimal distance deviation of 0.025 \AA from the experimentally determined site.

Comparison with other predictors

The comparison of PMM with other MBS predictors focuses on key features such as input data requirements, prediction methods, output data, and response times. Table 1 summarized these aspects, showcasing that PMM stands out for its ability to predict both metal ion coordinates and binding residues, offering fast response times (5–50 seconds) while incorporating a geometry and machine learning-based approach. Each step in the PMM process, from dataset construction to result validation, is rigorously verified to ensure accuracy. Unlike several other predictors like Metal3D and AlphaFill, PMM provides detailed ligand information and can predict sites with CHED ≥ 2 ligands. Additionally, PMM delivers structural models and metal ion locations with higher accuracy than predictors like ZincBindDB and znMachine, which do not provide metal ion locations or structural models.

PMM is compared with representative predictors from each of the three categories in more detail, including Category I predictors ZincBindDB, znMachine, CHED; Category II predictors Metal3D, AlphaFill; and Category III predictors GRE4Zn, TEMSP. For an apple-to-apple comparison, the same TP and FN definition and the corresponding datasets used in Metal3D and TEMSP are also used to evaluate PMM.

Table 1 | Comparison with other metal predictors

Predictor	Category	Input data	Method	Output data	Type and number of ligands	Provide metal ion location	Provide a structural model	Typical response time	Year of publication
PMM	III	Structure, Uni-prot ID	Geometry, ML	PDB file, Structure	CHED ≥ 2	Yes	Yes	5–50 s	2023
Metal3D	II	Structure	CNN	Zinc ion location	N/A	Yes	No	3–60 min	2023
AlphaFill	II	Structure, Uni-prot ID	Structure homology	PDB file, Structure	N/A	Yes	Yes	5–50 s	2023
ZincBindDB	I	Structure, Sequence	ML	Predicted sites	CHED ≥ 2	No	No	3–10 min	2021
znMachine	I	Sequence	ML	Predicted sites	CHED ≥ 3	No	No	Unavailable	2019
GRE4Zn	III	Structure	Geometric restriction	PDB file	CHED ≥ 3	Yes	No	5–30 s	2014
TEMSP	III	Structure	ML	PDB file	CHED ≥ 3	Yes	No	Unavailable	2011
CHED	I	Structure	ML	Predicted sites	CHED	No	Yes	Unavailable	2007

Categories (I) binding site predictors for metal binding residues; (II) binding position predictors for metal ion coordinates; (III) predictors that identify both residues and coordinates. Metal3D's runtime can be accelerated using GPU processing.

When comparing PMM, ZincBindDB, GRE4Zn, TEMSP, and CHED, the evaluation uses a dataset comprising 136 experimentally determined zinc binding sites derived from 100 protein structures²⁰. These data were excluded from the training set of the PMM algorithm to avoid bias. Nevertheless, PMM identified 134 out of the 136 actual zinc-binding sites using the same 0.5 IoUR cutoff, achieving a recall value of 98.5%, which notably exceeds the recall predicted by ZincBindDB (84.6%), GRE4Zn (74.3%), TEMSP (86.0%), and CHED (82.4%). PMM also scores a smaller deviation of 0.237 Å, compared with the average deviations of zinc positions predicted by GRE4Zn (0.267 Å) and TEMSP (0.38 Å). Additionally, PMM achieves a precision of 86.5%, outperforming ZincBindDB (29.6%), while closely matching or slightly underperforming relative to GRE4Zn (95.3%), TEMSP (95.9%) and CHED (91.1%). Though PMM's 21 false positive sites contributed to its lower precision, these false positives included 9 Zn, 3 Mn, 5 Cu, and 4 VIII (Fe, Co, Ni), while the other predictors, except CHED, only focus on zinc sites (Fig. 4a and Supplementary Table 4).

Preprocessing of the dataset of 189 zinc-binding sites reported by Metal3D results in the removal of some redundancies and errors, and 178 valid zinc-binding sites remain for further evaluation (Supplementary Fig. 5). In comparing PMM and Metal3D for predicting zinc-binding sites at different thresholds ($p = 0.5$ and $p = 0.75$, where p represents the probability threshold), subtle yet significant differences emerge. At $p = 0.5$, PMM shows a lower average error with greater variability (0.670 ± 1.104 Å) compared to Metal3D (0.729 ± 0.657 Å) (Supplementary Fig. 6a, c). PMM has a notably smaller median error (0.191 Å vs. 0.544 Å), indicating higher accuracy in most predictions but with more outliers beyond 1.5 Å. This suggests that while PMM is generally more accurate, it is more prone to larger deviations in cases with diverse ligand environments, where coordinate errors increase. Metal3D, on the other hand, performs better in recall (0.550 vs. 0.494) and F1-score (0.631 vs. 0.597), achieving a better balance between precision and recall. At $p = 0.75$, the trend persists—Metal3D achieves a higher F1-score and precision, while PMM maintains a lower median error and slightly lower average error (Supplementary Fig. 6b, d). At their respective recommended thresholds, PMM at $p = 0.5$ has lower deviation and median error, making it more reliable for predicting zinc ion position, with a slightly higher recall. In contrast, Metal3D at $p = 0.75$ demonstrates higher precision, particularly in reducing false positives. Overall, PMM at $p = 0.5$ offers a more balanced approach with lower error and strong identification, while Metal3D at $p = 0.75$ excels in precision and overall performance balance (Supplementary Table 5).

A dataset of 979 MBSSs from 503 structures is used to compare the selectivity for common transition metals in PMM versus Metal3D using a precision-recall scatterplot. We derive the 503 structures from Metal3D reported structures with 30% sequence similarity, which encompass 100, 100, 57, 30, 93, 68, 59 PDB structures that contain Mn, Fe, Fe2, Co, Ni, Cu, Zn ions, respectively. None of these metal-containing structures are included in the training sets of PMM or Metal3D. These metal sites bind to at least 3 unique protein ligands and have occupancy >0.5 . Using the same threshold of $p = 0.75$, the precision-recall scatterplot shows that PMM demonstrates stable performance, with high precision and recall across most metals, particularly for Zn, Co, and Fe. In contrast, Metal3D, trained primarily on zinc, exhibits more variability in performance for other metals, with lower recall and precision for some metals like Fe2 and Cu (Fig. 4b). PMM demonstrates more precise positional prediction for most metals, with a median error value of around 0.172–0.328 Å compared to that of around 0.39–0.818 Å in Metal3D (Supplementary Table 5).

Tools like AlphaFill use structural homology to transplant metals from similar PDB structures to the predicted structure and may limit their ability to predict potential MBSSs. For example, in the tryptophan RNA-binding attenuator protein-inhibitory protein 2zp9, five zinc-binding sites (Zn1–Zn5, Fig. 4c) are present. Both PMM and Metal3D

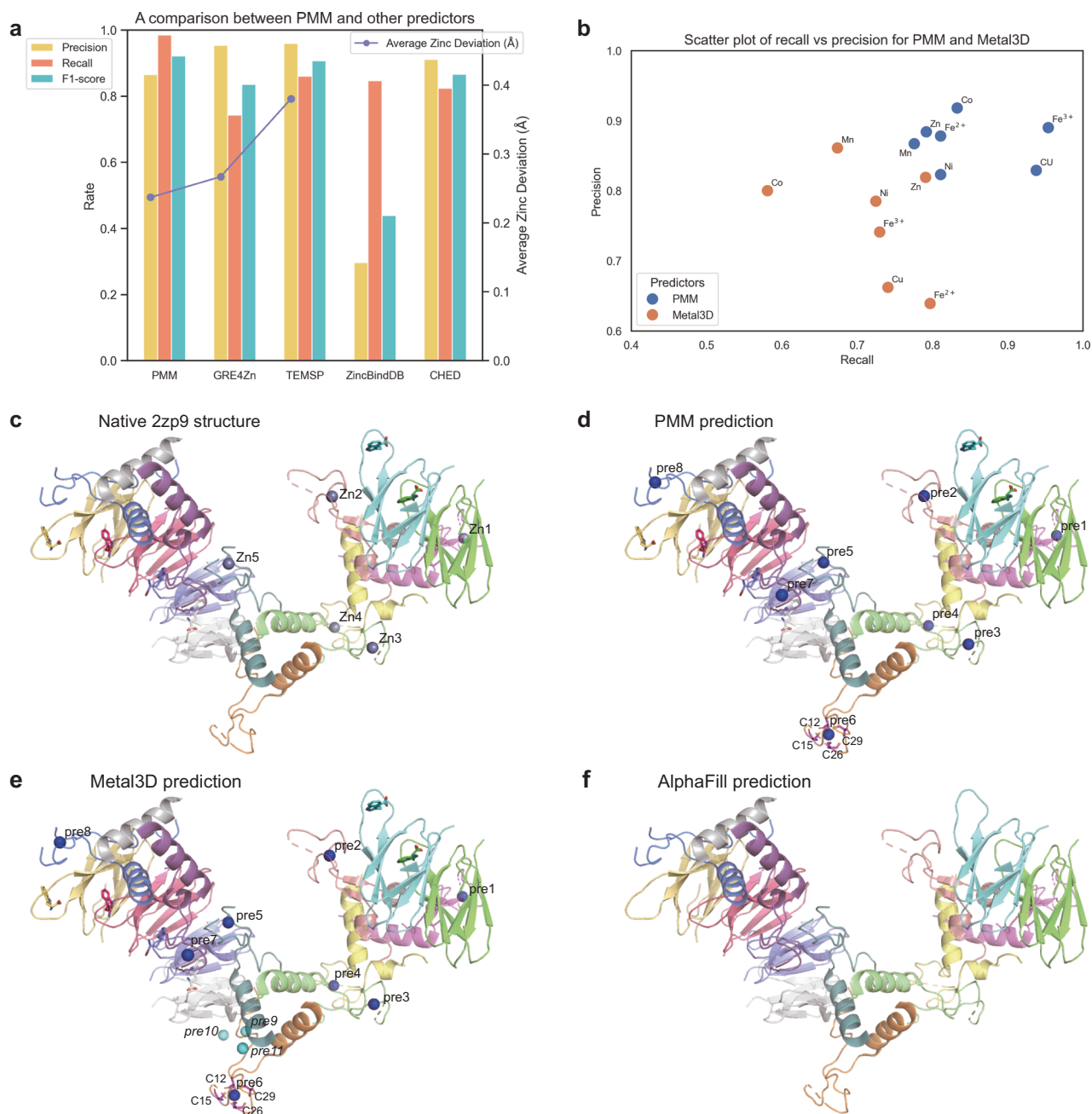


Fig. 4 | Prediction results of PMM and other predictors for transition metals.

a Performance comparison of five zinc-binding site predictors (PMM, GRE4Zn, TEMSP, ZincBindDB, CHED). Bar plots represent precision, recall, and F1-score, while the line plot shows the average zinc deviation (Å) for available methods.

b Comparison between Metal3D (orange) and PMM (blue) for transition metal binding sites in the test set. **c–f** Comparison of native and predicted metal binding sites in 2zp9. **c** Native structure of 2zp9, with five observed zinc sites (Zn1–Zn5, gray spheres). **d** Predicted metal binding sites by PMM. **e** Predicted metal binding sites by Metal3D. Both PMM and Metal3D identified the five known zinc sites (pre1–pre5)

and predicted three potential sites (pre6–pre8, blue spheres) at their thresholds (PMM: 0.5, Metal3D: 0.75). Metal3D also found three low-confidence sites (*pre9–pre11*, cyan spheres) with certainty scores above 0.5 but below 0.75. The certainty scores for PMM predictions (pre1–pre8) were 0.94, 0.93, 0.93, 0.92, 0.96, 0.94, 0.94, and 0.93, respectively. The certainty scores for Metal3D predictions (pre1–pre8) were all 1.0, with pre9–pre11 certainty scores of 0.53, 0.50, and 0.51. No statistical hypothesis testing was performed. **f** AlphaFill Predictor: no metal ions were predicted by AlphaFill. Predicted sites for PMM and Metal3D were obtained at a threshold of 0.5.

accurately identified these sites (pre1–pre5) at their recommended thresholds, achieving comparable accuracy with prediction errors ranging from 0.3 to 0.9 Å. They also predicted three potential MBSs (pre6–pre8, Fig. 4d, e), with one site (pre6) validated by electron density mapping, where four cysteine residues coordinate the metal (see Fig. 3b). In contrast, AlphaFill failed to predict any metal sites in 2zp9 because it could not identify homologous structures containing similar MBSs (Fig. 4f). Metal3D also identified three new low-

confidence sites (pre9–pre11, Fig. 4e), with confidence values below 0.75 and showing no observed electron density.

Biological implication of PMM in predicting MBS with diverse coordination geometries

Although metal ligands and coordination geometries are largely different among regulatory, catalytic, and structural sites, PMM achieves high accuracy with commendable biological implications in all

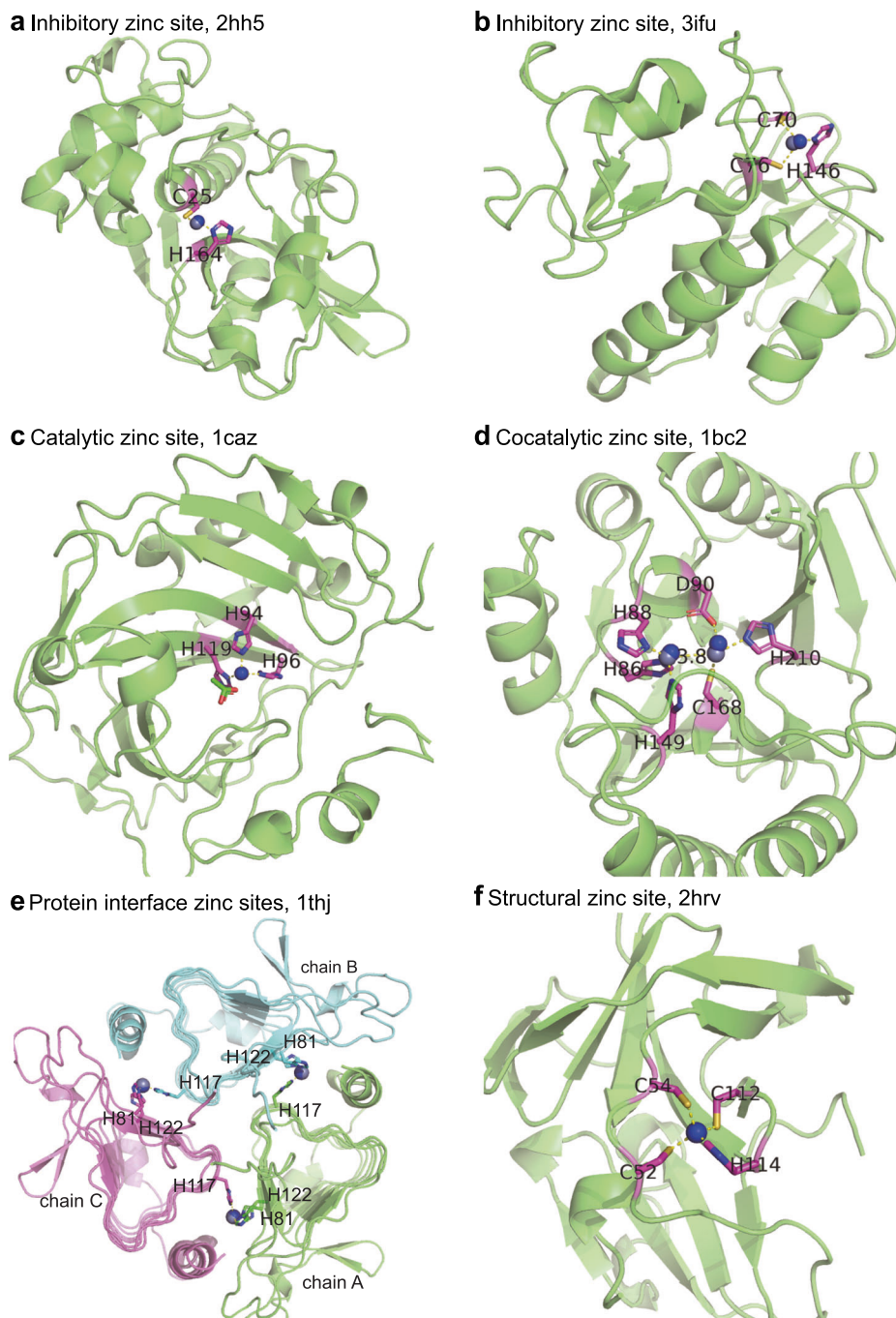


Fig. 5 | PMM predicts zinc binding sites for different ligands or functions.

a Inhibitory zinc site, 2hh5, C25-H164, 0.26 Å. **b** Inhibitory zinc site, 3ifu, C70-C76-H146, 0.67 Å. **c** Catalytic zinc site, 1caz, H94-H96-H119, 0.15 Å. **d** Cocatalytic zinc site, 1bc2, Zn1: H88-H86-H149, 0.40 Å. Zn2: D90-C168-H210, 0.78 Å. **e** Protein

interface zinc sites, 1thj, H-81-H117-H122, 0.13/0.10/0.09 Å. **f** Structural zinc site, 2hrv, C52-C54-C112-H114, 0.06 Å. The blue spheres represent the predicted zinc site, in agreement with the gray spheres depicting the zinc site modeled by the experimenter.

scenarios, as exemplified by its performance in predicting zinc-binding sites. (Fig. 5). Zinc ions at inhibitory and catalytic sites in zinc-containing enzymes require two or three coordinating ligands for full activity (Fig. 5a–c). PMM can accurately predict zinc ion location at cocatalytic sites containing two or three metals in proximity with two of the metals bridged by a side chain moiety of a single residue, such as Asp, Glu, or His and sometimes a water molecule (Fig. 5d). The application of PMM is not limited to a single polypeptide chain, but also includes protein interface zinc sites formed from ligands supplied from amino acid residues residing in the binding surface of two polypeptide chains (Fig. 5e). Similar to other zinc ions, zinc-binding sites on

the protein interface can be regulatory, catalytic, or structural. Additionally, PMM effectively predicts structural zinc sites coordinated by four cysteine ligands (Fig. 5f).

Open-Source PMM predictor: local and web access

The PMM predictor code is open source, allowing peers to download, run, and compile locally. Additionally, an online version is provided for convenient web-based predictions, enhancing flexibility and ease of use in practical applications. The PMM web server is publicly available and freely accessible at <https://PMM.biocloud.top>. Even though PMM is a structural-based method, it implements an automated structure-

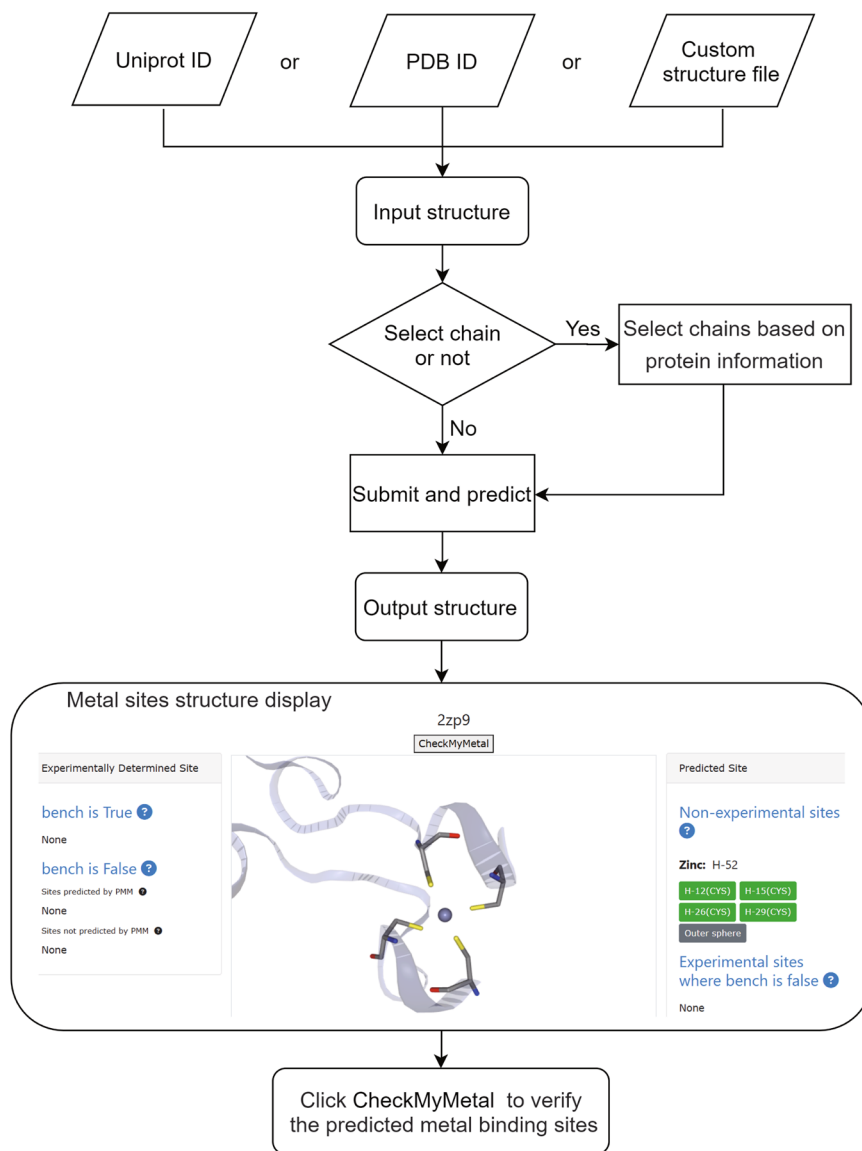


Fig. 6 | PMM web prediction flow chart.

retrieval interface that allows users to search by protein name or sequence as identified by Uniprot ID. The server provides three input methods to acquire protein structures for MBS prediction: (1) PDB id from the PDB website; (2) Uniprot ID of the target protein, which will be used to retrieve protein structures from the Uniprot database for further analysis. If multiple experimentally determined structures are found for the same Uniprot entry, structure with the highest sequence completeness and highest resolution is chosen. If no experimentally determined structure is found, a computational model from AlphaFold2 is selected; and (3) PDB or CIF format coordinate file is uploaded by the user (Fig. 6). Preprocessing of the protein structure prompts a chain selection page containing the chain ID, name, source organism, and length for each chain, allowing the user to choose one or more chains of interest to conduct MBS prediction.

After submission, users can typically expect to receive a response in about 40 s or less. The submitted protein structure, along with all experimental and predicted MBSs, will be displayed on an interactive NGL 3D view page (Fig. 6). The output of PMM is divided into two panels: the right panel features predicted metal ion location and coordinating amino acid type and residue sequence number (resseq), while the left panel features experimentally determined metal ion location and

coordinating amino acid annotated with whether or not it passes the validation criteria. Experimental MBSs that have not passed the validation criteria are compared with predicted MBSs using IoUR ≥ 0.5 as the criteria to determine if they are the same site as defined in section 2.2. A CheckMyMetal button is provided on the PMM output interface to allow the seamless validation of the predicted MBS on the sister CMM website, with an '@' indicating predicted sites. The experimenter may download the coordinate in PDB format, with the predicted sites annotated in the ATOM and LINK records. A certainty score between the range of 0 to 1, indicating the confidence value of the MBS, is provided in the occupancy field. The NGL interface also allows the visualization of other non-CH/EDH amino acids or small molecule ligands within 4 Å of the metal center. Careful examination of the interactions of the metal-coordinating ligands beyond the first coordination sphere could reveal other global characteristics of the protein structure.

Discussion

Metal binding site classification and extended ligand detection in PMM

The PMM classification scheme identifies MBSs based on geometric preferences using the CH and EDH algorithms. For four-coordinated

metals (e.g., zinc, copper), the CH algorithm predicts binding sites by identifying 2 to 4 cysteine (C) and histidine (H) ligand combinations, while for six-coordinated metals (e.g., iron, manganese, nickel), the EDH algorithm focuses on 2 to 6 glutamate (E), aspartate (D), and histidine (H) ligands. This approach reduces the number of classes when constructing templates, ensuring sufficient training data for each class of coordination motifs.

This metal ion classification approach fundamentally differs from the principles used in existing metal coordination motif classifiers such as ZincBindDB²². ZincBindDB considers all CHED combinations and can only predict sites with a sufficient number of cases, such as the top 10 most populated classes (C2H1, C2H2, C3, C3H1, C4, DIH1, DIH2, E1H1, E1H2, H3). For CHED combinations with fewer experimentally determined structures, ZincBindDB cannot build a prediction model or may suffer from significantly compromised accuracy. Compared to databases like ZincBindDB, PMM's streamlined classification strategy addresses the issue of rare ligand combinations and delivers more accurate predictions. Additionally, it accounts for non-amino acid ligands, particularly for six-coordinated metals, where small molecules like water often serve as ligands. PMM extends its prediction range by examining a 4 Å radius around the metal ion to identify additional ligands. For example, if two cysteine and two histidine ligands are initially predicted, any glutamate or water molecule detected within the 4 Å radius is also considered part of the binding site.

PMM classifies sites based on CH and EDH, using a straightforward approach with sufficient training samples for higher prediction accuracy. It also includes ligands within a 4 Å radius when presenting metal site structures. PMM does not overlook the auxiliary measure (Glutamate/Aspartate residues in case of tetrahedral, and Cysteine residues in the case of octahedral) but rather postpones its consideration after the location of the metal ion is determined. For example, the metalloproteinase structure (2qvp) contains a CH site B498 coordinated by 2 histidine residues, while a third and fourth coordinating ligands Glu and water is also identified after the location of the metal ion is predicted (Supplementary Fig. 7).

The biological implications of this classification scheme are validated through the analysis of zinc-containing enzyme structures from the PDB. For example, zinc is a ubiquitous cofactor for all six major classes of enzymes and zinc-containing enzyme structures from the PDB are analyzed. Sites from CH₄ group lack catalytic capability and are considered as structural sites, featuring cysteine as the most prominent coordinating ligand, followed by histidine, with the most common combinations being C4 and C3H1. Zinc may contribute to the catalytic activity in sites from CH₃ or CH₂ group, featuring histidine as the most prominent coordinating ligand, followed by cysteine, with many common CH combinations in different scenarios (Supplementary Table 6). Catalytic zinc generally forms complexes with any three nitrogen, oxygen, and sulfur donors from CHED residues, with histidine (usually the Nε2 nitrogen) being the predominant amino acid because of its capacity to disperse charge through H-bonding of the other non-liganding nitrogen (usually the Nδ1 nitrogen)¹⁶.

Identifying metal types: the predictive power of PMM

In recent years, significant progress has been made in predicting MBSs, particularly with the model like MetalSiteHunter⁴⁸, which utilizes 3D convolutional neural networks. This model performs well in distinguishing common transition metals, particularly for Zn and Fe binding sites, but has limitations with metals like Co, Ni, and Cu. Moreover, most MBS predictors only determine if a site binds a metal, without specifying the metal type. MetalSiteHunter, as a region-specific predictor, focuses on locations that may bind to metals rather than evaluating all areas of a protein. This

specificity makes a systematic comparison with the PMM predictor challenging.

The widespread binding of transition metal ions, as described by the Irving-Williams series⁴⁹, presents challenges in distinguishing these ions due to their similar properties. Most Zn binding sites could also bind Cu in competitive binding conditions, and that selectivity in such cases is not determined solely by the binding site, while the contributions of environmental factors, such as chaperones or compartmentalization, should not be underestimated or overlooked. For instance, in the carbonic anhydrase II variants 1fr4⁵⁰ and 2fos⁵¹, the MBSs at positions 94, 96, and 119 on chain A are different: 1fr4 binds copper, while 2fos binds zinc (Supplementary Fig. 8a). PMM predictor also shows inaccuracies in identifying metal ions, such as Fe being incorrectly assigned as Zn (Supplementary Fig. 8b, 1jyb)⁵², while Cu is misidentified as Zn, although another Cu site is correctly identified (Supplementary Fig. 8c, 3mnd)⁵³. This difficulty in uniquely determining the specific identity of transition metals at certain MBSs arises from the natural versatility and lack of physiological selectivity of the MBSs. Protein predicted by PMM to bind zinc could also bind multiple metals *in vivo* due to various environmental factors, highlighting the ongoing challenges faced by MBS predictors in distinguishing between transition metals.

In conclusion, the PMM predictor performs well in identifying Mn and Zn binding sites, with accuracies of 90.3% and 73.8%, respectively, based on fivefold cross-validation. By grouping Fe2, Fe, Co, and Ni into the VIII category, PMM improves classification accuracy and reduces ambiguity, achieving 73.3% accuracy for VIII. However, challenges remain in differentiating between VIII, Mn, and Cu, mostly due to the common physiochemical properties shared by transition metal binding sites that allow for interchangeability of different metal types in these sites (Fig. 2).

Innovation and validation of the PMM algorithm

Existing metal prediction tools that screen the hydrophobicity contrast function at dense grid points to determine candidate metal ion location in the protein structure require much computational resources. PMM introduces an innovative algorithm that significantly reduces the computational resources required for screening the hydrophobicity contrast function and determining candidate metal ion locations. By deducing the most probable location before applying the contrast function, PMM maintains accuracy while enhancing efficiency, making it a powerful tool for predicting optimal metal ion locations within protein structures. However, it is important to consider the inherent uncertainties in experimental structures when interpreting these results⁵⁴. Typical positional uncertainties in protein structures range from 0.1–0.2 Å, and metal location accuracy is measured as the deviation from these experimental positions⁵⁵. Given this, performance differences in the hundredths of an angstrom may not be statistically significant, especially when accounting for the experimental error margins.

Considering metal ions in macromolecular structures requires a multidisciplinary approach, coherently considering chemical, crystallographic, biological, and experimental aspects²⁷. PMM's validation procedure, specifically the CMM validation, effectively identifies incorrect metal assignments and suboptimal modeling of MBSs. Addressing potential complications, such as geometric distortions of the first coordination sphere, the quality of the diffraction data (e.g., the resolution), and sample preparation concerns, ensure the robustness of PMM in predicting MBSs⁵⁶.

PMM predictor's role in zinc regulation of enzyme activity

As a signal transduction messenger, zinc regulates protein activities, including the inhibition of enzymatic activities, yet this occurs only when the concentration of zinc ions elevates to a certain level. Nevertheless, the regulatory sites at the active or allosteric sites of

enzymes share similar coordination environments with catalytic zinc in zinc metalloenzymes. The only notable distinction is a tendency for lower coordination numbers in regulatory zinc sites. While the K_d for zinc ion can range from milli-molar concentration to micro- or nano-molar concentration, how zinc regulates enzyme activity is not clearly defined from the structural perspective. low-coordination CH algorithm provides a one-stop solution to propose a hypothetical mechanism for such inhibition by predicting candidate regulatory zinc sites and other zinc-binding sites coordinated by two CH residues. Many enzyme active sites feature two metal-binding amino acid side chains, such as Cys-Cys, His-His, Cys-His, Glu(Asp)-His, and Cys-Glu(Asp), forming a catalytic dyad. Yet not all of them contain two catalytic cysteine or histidine residues, as seen in enzymes like cysteine proteases, protein tyrosine phosphatases (PTPs), aldehyde dehydrogenases, and glyceraldehyde 3-phosphate dehydrogenase¹⁴. Therefore, even if a zinc-binding site is not predicted due to the lack of two CH residues (such as Cys-Cys, His-His, or Cys-His), zinc may still exert its inhibitory or regulatory role through alternative mechanisms, such as coordinating with other residues like Cys-Glu(Asp) or a single C residue.

In conclusion, PMM can predict metal ion locations and coordinating ligands based on local geometrical and chemical micro-environments. The application of PMM in MBSs exhibits superior accuracy and efficiency performance compared to other predictors, providing a quick way for the scientific community to predict MBSs with easy accessibility, high confidence, and minimal latency. The high efficiency also prompts PMM to excel in the large-scale prediction of MBS for the superfamily of metal-binding proteins or genomic-scale prediction of MBSs. PMM also specializes in predicting regulatory (transient) MBSs (2-residue predominate) not specifically handled in any other metal predictors and exhibits much superior prediction accuracy than other metal predictors. Experimentally determined protein structures generally represent a single snapshot of the protein, while the metal binding state may not be observed under a specific experimental condition. Therefore, the absence of MBSs in a given crystal structure does not warrant its absence in the associated biological processes. In this sense, PMM opens up a window of opportunity to examine candidate metal-binding proteins from a perspective not accessible using any known experimental or computational methodologies. We have also demonstrated the effective routine use of PMM to annotate MBSs in cryo-EM structures with limited resolution. PMM offers a complementary and accurate solution to model metal ions in cryo-EM structures which would otherwise be challenging due to the limitations of electron penetration depth and scattering effects.

Methods

Concept and strategy

Coordination prediction based on CH and EDH. In macromolecular structures, transition metal ligands are mostly oxygen-, nitrogen-, and sulfur-containing groups, with coordination numbers typically ranging from four to six, forming tetrahedral or octahedral geometries. The CHED residues are the most common ligands for transition metal ions⁵⁷. In contrast, donor atoms from other amino acids, such as mainchain oxygen, serine, threonine, or lysine, are much less frequent and accounts for only a small fraction of transition metal-ligand interactions^{27,58}. Different metals exhibit distinct coordination preferences. High-quality non-redundant analysis of transition metal coordination ligands shows that H residues are the primary coordinating ligands for most transition metals (e.g., Fe, Co, Ni, Cu, Zn), indicating a strong coordination preference with nitrogen atoms. Additionally, Zn and Cu also coordinate with the sulfur atoms of C. In contrast, acidic residues, such as carboxylate oxygen's from E and D, are more common in Mn, Fe, Co, and Ni coordination, but less frequent in Zn and Cu (Supplementary Fig. 9).

C, due to its weaker ligand field, tends to form tetrahedral structures with metal ions, whereas E and D, which generate stronger ligand fields, are more suited for octahedral structures. In metals with coordination numbers of two or more, the distribution of ligand-metal angles for E, D, and H is concentrated around 90° and 180°, indicating that octahedral structures are more common in complexes involving these ligands. In contrast, for C and H, ligand-metal-ligand angles cluster around the typical 109.5° angle in tetrahedral structures. This pattern is especially pronounced for zinc and copper, suggesting that their coordination with C and H ligands tends to favor tetrahedral geometry. Even though Zn and Cu may be coordinated by either ED residues or C residues, the presence of ED residues and C residues are mutually exclusive. Therefore, the presence of ED residues to coordinate Zn and Cu are typically correlated with octahedral geometry and the lack of C residues in the first coordination sphere (Supplementary Fig. 10).

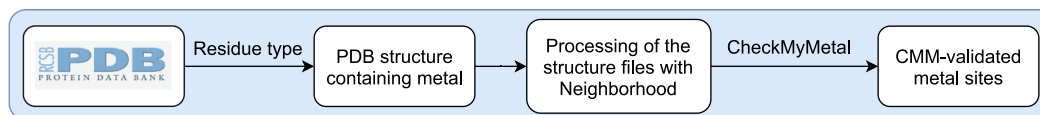
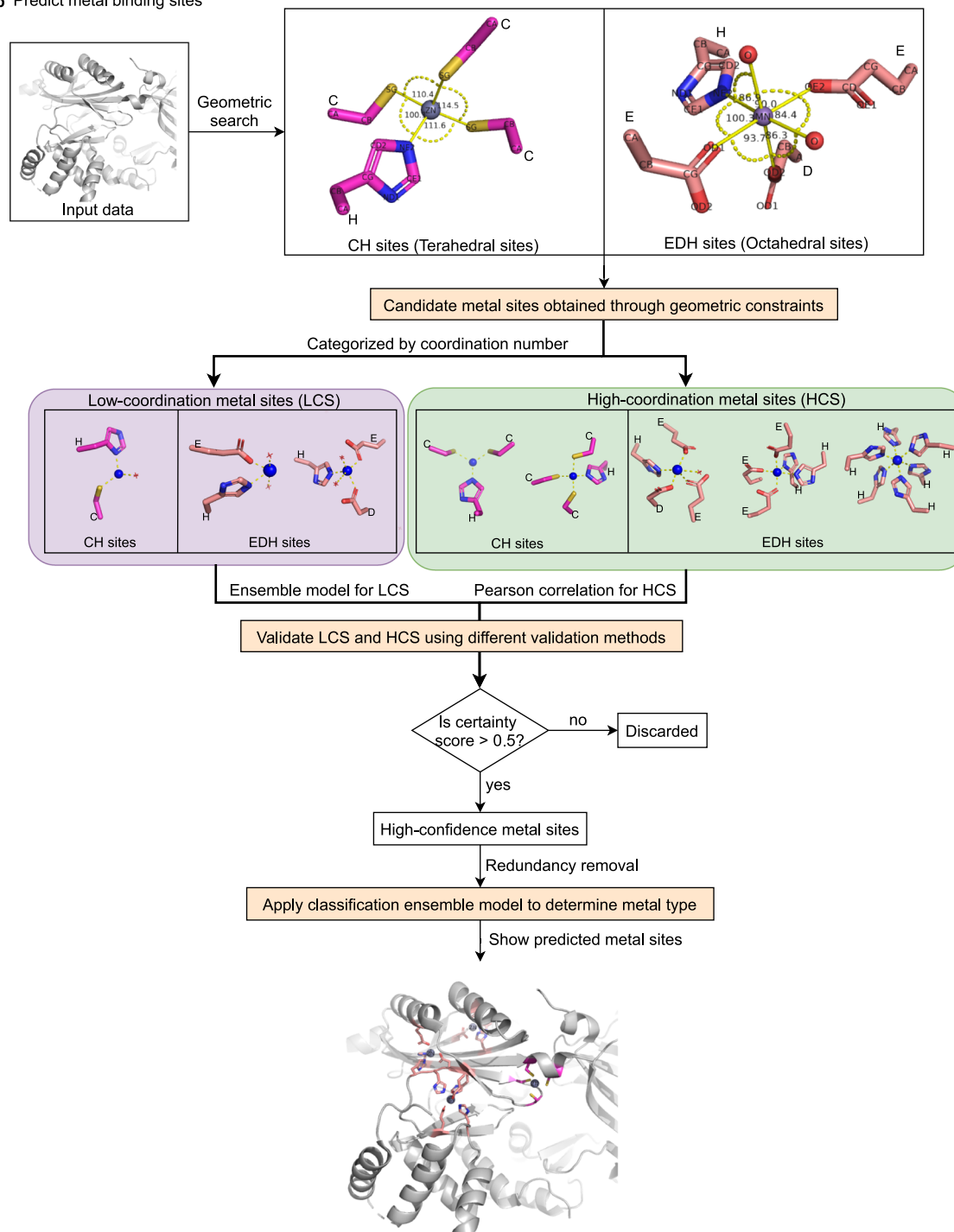
Metals with different coordination geometries exhibit varying ligand preferences. PMM uses an EDH-based algorithm to consider combinations of 2–6 coordinating E, D, and H residues for all transition metals. For four-coordinated tetrahedral sites typically observed in zinc or copper, an additional CH-based algorithm is introduced to focus on geometric searches involving combinations of 2, 3, or 4 coordinating C and H residues. Our tailored approach accommodates metals with different coordination geometries, and reduces the number of classifications when constructing templates for predicting potential MBSs, ensuring sufficient training data for each class.

hybrid learning system. PMM identifies metal sites for transition metals through three key steps tailored to different coordination geometries. In Step 1, geometric constraints are used to predict candidate metal binding sites. For tetrahedral metal sites, ligands composed of C and/or H residues are identified, while for octahedral metal sites, ligands composed of E, D, and/or H residues are identified. Based on ligand preferences, tetrahedral metal sites are also called CH sites, and octahedral metal sites are also called EDH sites. Metal sites with coordination numbers \leq half of the full structure (e.g., 2 residues in tetrahedral, 2–3 in octahedral) are low-coordination metal sites (LCS), while those with greater coordination or fully coordinated sites (e.g., 3–4 residues in tetrahedral, 4–6 in octahedral) are high-coordination metal sites (HCS).

Machine learning methods are selectively applied in Step 2 and Step 3 of the workflow, forming a hybrid strategy for metal ion validation and classification. In Step 2, predicted sites are validated to determine whether they are indeed metal binding sites. For LCS, two ensemble learning models are employed: one for the CH sites, and another for the EDH sites. These models integrate geometric and hydrophilicity features to compute certainty scores, while for HCS, certainty scores are calculated using the Pearson correlation coefficient for each site, which assesses the similarity between predicted sites and experimental data based on hydrophilicity patterns. In Step 3, for sites with certainty scores greater than a given threshold (0.5), a single ensemble learning model is used to determine the most probable metal type for all sites. This model incorporates a hydrophobicity profile, seven chemical features, and 27 features from the NEIGHBORHOOD database. The use of different prediction strategies for various types and coordination numbers of MBSs at different stages is referred to as a hybrid learning system (Fig. 7).

PinMyMetal workflow

Data acquisition, validation, and redundancy elimination. We process the April 2024 version of the PDB⁵⁹ containing metal binding asymmetric unit structures using the Neighborhood database as described earlier²⁷. The program CONTACT from the Collaborative

a Obtain the metal site dataset used for training and validating PMM**b** Predict metal binding sites**Fig. 7 | Workflow of PMM. a** Obtain CMM-validated experimental metal sites.**b** Predict metal binding sites. The three orange background boxes represent the three key steps of the algorithm: Step 1: Candidate metal sites obtained through

geometric constraints; Step 2: Validate LCS and HCS using different validation methods; Step 3: Apply ensemble model to predict metal type for high-certainty sites.

Computational Project No. 4 (CCP4) suite is used to apply crystallographic symmetry operations on the asymmetric units and identify symmetry-related ligands to complete metal coordination sphere. The intermolecular interaction between metal ions and proteins is stored in the form of coordination bonds represents the MBS (Fig. 7a).

$$Q_v = \min\left(\left|\frac{\sum V_i}{V_{ox}}\right|, \left|\frac{V_{ox}}{\sum V_i}\right|\right) \quad (1)$$

$$Q_c = 1 - \frac{|\mathbf{v}_1 + \mathbf{v}_2 + \dots + \mathbf{v}_n|}{\sum V_i} \quad (2)$$

$$O_e = \frac{\sum \mathbf{v}_i O_i}{\sum \mathbf{v}_i} \quad (3)$$

$$B_e = \frac{\sum \mathbf{v}_i B_i}{\sum \mathbf{v}_i} \quad (4)$$

$$Q_e = \min(2 \times \min(O_m, O_e), 1) \times \min\left(\frac{B_m/O_m}{B_e/O_e}, \frac{B_e/O_e}{B_m/O_m}\right) \quad (5)$$

The quality of MBS is evaluated using CheckMyMetal (CMM)³⁶, with modification based on the previously described algorithm used to validate magnesium binding sites in nucleic acid structures⁶⁰. Since the previous algorithm was tested for magnesium ions, the validation parameters are adapted to be applicable to other MBSs. Three parameters were used to quantitatively evaluate the agreement with expected valence (oxidation state) (Q_v)⁽²⁾, completeness of the first coordination sphere (Q_c)⁽³⁾, and experimental agreement (B factor and occupancy) with the environment (Q_e)⁽⁶⁾. In all formulas, \mathbf{v}_i represents the bond valence vector of coordination bond i . In formulas (1)–(3), V_i represents the magnitude of bond valence vector \mathbf{v}_i ; V_{ox} represents the expected oxidation state. In formulas (4) and (5), B_m and B_e represent the B factor of metal (m) or environment (e); while O_m and O_e represent occupancy of metal (m) or environment (e). Each of the three validation parameters Q_c , Q_v , and Q_e has a valid range of 0 and 1, with 1 indicating the best quality and 0 indicating the worst quality.

The validation procedure is fine-tuned based on the number of coordinating ligands, assuming that four ligands comprise a stable zinc coordination sphere that adopts a tetrahedral coordination geometry⁶¹. For tetrahedral MBSs with 3 or 4 coordinating ligands, a threshold of half of the optimal quality was set as the validation criteria: $Q_v > 0.5$ and $Q_c > 0.5$ and $Q_e > 0.5$. For tetrahedral MBSs with two coordinating ligands, while the expected oxidation state V_{ox} stays at 2, the optimal theoretical bond valence summation ($\sum V_i$) is 1, and the optimal theoretical vector sum is $|\mathbf{v}_1 + \mathbf{v}_2| = 0.58$. Therefore, the optimal Q_v would be 0.5 according to formula (1), and the optimal Q_c would be 0.71 according to formula (2). Using a threshold of half of the optimal quality would result in different validation criteria: $Q_v > 0.25$, $Q_c > 0.355$ and $Q_e > 0.5$. Even though the thresholds are deduced using tetrahedral MBSs as an illustration, the same set of thresholds ($Q_v > 0.25$, $Q_c > 0.355$ and $Q_e > 0.5$) would also apply for octahedral MBSs.

A total of 73,418 CMM-validated metal sites are identified. The polypeptide chain contributing the highest number of ligands is designated as the metal binding chain. Polypeptide chains are clustered using MMseqs2⁶² with a sequence identity threshold of 30% and a coverage threshold of 80% to quantify redundancy. For MBSs on homologous protein chains, the criterion of ligand overlap ≥ 0.5 is used to identify and eliminate redundant sites, retaining non-redundant metal sites. Finally, representative structures are selected based on the highest number of MBSs for each combination of cluster ID and metal

type, while preference is given to higher-resolution structures in case of multiple candidate structures with equal number of MBSs. This process results in a benchmark dataset of 4984 non-redundant MBSs from 2778 structures (Supplementary Table 1).

Prediction of candidate metal binding sites. According to the geometric characteristics of metal sites, PMM searching throughout the protein structure to identify candidate MBSs based on the type of coordinating residues and atoms, and interatomic distances (Fig. 7b). The specific geometric restrictions are as follows:

For four-coordinated metal sites, the coordinating atoms are limited to SG from Cysteine, and ND1, NE2, CE1, or CD2 from Histidine. For six-coordinated sites, they are restricted to OD1 or OD2 from Aspartate, OE1 or OE2 from Glutamate, and the same options from Histidine. The delta and epsilon carbon atoms from the histidine side chain are also included due to the possible presence of alternative conformation or mislabeling⁶³. The presence of proximal SG atoms from cysteine side chains may implicate the presence of either zinc-binding sites or disulfide bonds, depending on the distances between SG atoms. A survey of the distance between SG atoms in protein structures reveals the presence of two peaks, with the smaller peak below 2.2 Å indicating a disulfide bond and the larger peak above 2.8 Å indicating MBSs (Supplementary Fig. 11). The disulfide bond peak ($\mu = 2.058$ Å, $\sigma = 0.133$) is excluded using a p -value cutoff of 0.01, corresponding to a Z value of 2.575. The upper limit of the confidence interval is determined as $\mu_0 = 2.058$ Å + 0.133 Å * $2.575 = 2.400$ Å using two tail t -test. Therefore, pairs of cysteine residues with a distance of SG atoms below 2.4 Å are excluded from further analysis. Pairs of cysteine residues with distance longer than 4.5 Å are considered sub-optimal to coordinate the same metal ion and are also excluded from further analysis. Therefore, the interatomic distance is restricted to the range of 2.4–4.5 Å.

Sites are identified as candidate MBSs if they contain two or more ligands for the calculation of interatomic distances, and that the interatomic distance falls between the abovementioned range criteria. For four-coordinated sites, 2, 3, or 4 residues incur 1, 3, 6 interatomic distances that must meet these conditions, while for six-coordinated sites, 2–6 residues incur 1–15 interatomic distances that must meet these conditions. The accuracy of predictions is measured using the intersection over union ratio (IoUR), which quantifies results by balancing the numbers of correctly and incorrectly predicted ligand residues for a specific binding site⁽⁶⁾. An IoUR of 1 indicates that the predicted ligands match the actual ligands precisely, while a threshold of IoUR ≥ 0.5 is used to evaluate prediction accuracy.

$$\text{IoUR} = \frac{N(\text{predicted ligand residues} \cap \text{actual ligand residues})}{N(\text{predicted ligand residues} \cup \text{actual ligand residues})} \quad (6)$$

Determination of metal ion location. PMM uses an intuitive algorithm to deduce the most probable location of metal ions prior to the application of the hydrophobicity contrast function, greatly reducing the number of evaluations needed without compromising the accuracy. First, the binding atoms are determined from metal-coordinating amino acids (CHED) by selecting the atom pair with the smallest distance between them. Next, based on the reference point coordinates (a1, b1, c1, d1), the Biopython Superimposer function is applied to perform geometric transformations (rotation and translation) that align the midpoint e1 with the origin of a new coordinate system e2, transforming the reference coordinates into (a2, b2, c2, d2) (Fig. 8). After computing the virtual metal position f2 based on molecular geometry, the reverse transformation is applied to return f2 to the original coordinate system, yielding the final metal binding location f1.

Fully coordinated sites have the maximum number of ligands, like four in tetrahedral or six in octahedral geometry, while incomplete sites have fewer ligands. The metal position is inferred using different

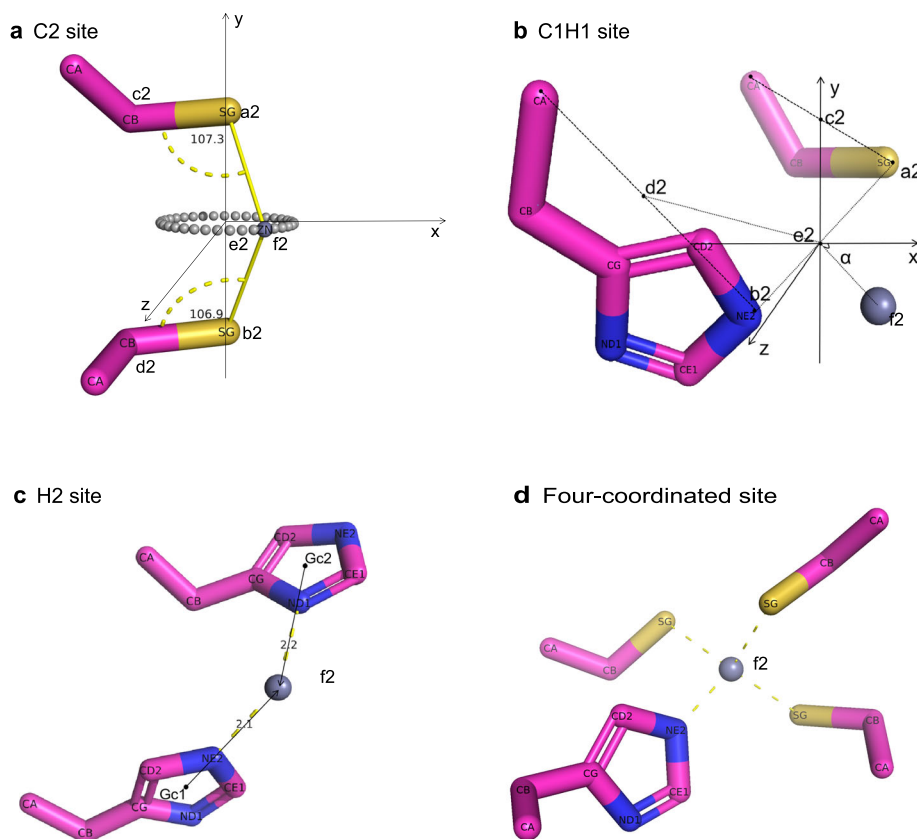


Fig. 8 | Schematic diagram of zinc ion coordinate prediction algorithm. **a** C2 site. **b** C1H1 site. **c** H2 site. **d** Four-coordinated site. a1, b1, c1, d1 refer to the actual sites, while a2, b2, c2, d2 represent the transformed coordinates corresponding to these actual sites.

methods depending on the number and type of ligands. Using zinc as an example, a total of seven strategies used to cover all scenarios are described in more detail below.

$$f_1^* = \arg \min_{f_1} (|\theta_{f1a1c1} - 109^\circ| + |\theta_{f1b1d1} - 109^\circ|) \quad (7)$$

- (a) C2 sites: A segment a_2b_2 is drawn between the two Sy atoms with the coordinate of the midpoint marked as e_2 , which is also on a plane p_2 perpendicular to the segment a_2b_2 (Fig. 8a). With a coordinating bond distance between the metal and atom of about 2.1 Å and an a-metal-b bond angle of $\sim 109^\circ$ in a tetrahedral site, the theoretical distance between zinc ion and e_2 is 1.2 Å, based on isosceles triangle geometry (Supplementary Fig. 10). The optimal location of zinc ion is restricted on the plane p_2 and has a theoretical distance of 1.2 Å from e_2 , resulting in a collection of points forming a circle. The two Sy atoms coordinating the zinc ion should feature a Zn-Sy-Cβ angle of 109° and be on the distal side of Cβ to dodge possible clash (Fig. 8a and Supplementary Fig. 12). In the actual coordinate system, the final Zn coordinate (f_1^*) is determined by summing the deviations of the angles between two sets of Zn-Sy-Cβ ($f1a1c1$, $f1b1d1$) from the expected angle of 109° , selecting the position with the minimum score from the points on the circle⁽⁷⁾.
- (b) C1H1 sites: After the original coordinates undergo transformation, the midpoint e_2 between the two metal-coordinating atoms, C (a_2) and H (b_2), is set as the origin of the coordinate system. The midpoints of the lines connecting each atom to the Cα atom are designated as c_2 and d_2 , with c_2 located on the y-axis. The distance from the metal f_2 to e_2 is 1.48 Å (for an

octahedral metal site, the a-metal-b bond angle is $\sim 90^\circ$) (Supplementary Fig. 10). Using the coordinates of the midpoints in the coordinate system and their relationship to the axes, the angle α is derived geometrically, which then allows for the calculation of the f_2 coordinates. (Fig. 8b).

- (c) H2 sites: The gravity centers G_{c1} and G_{c2} are calculated using the five atoms forming the corresponding five-member ring. All four atoms Cδ2, Nε2, Cε1, Nδ1 on the five-member ring of the histidine sidechain are considered as candidate coordinating atoms. Four rays G_{c1} -Cδ2, G_{c1} -Nε2, G_{c1} -Cε1, G_{c1} -Nδ1 are drawn for the first five-member ring, with 2.1 Å segments $G_{c1}Z_1$, $G_{c1}Z_2$, $G_{c1}Z_3$, $G_{c1}Z_4$ aligned with each ray, and Z_1 , Z_2 , Z_3 , Z_4 being the candidate zinc location, respectively. The candidate zinc location for the second five-member ring is deduced using the same procedure and denoted as y_1 , y_2 , y_3 , y_4 . The distance between each candidate zinc location from Z_1 , Z_2 , Z_3 , Z_4 and each candidate zinc location from y_1 , y_2 , y_3 , y_4 are calculated to determine the closest pair of candidate zinc ions (Fig. 8c). The average coordinate of this pair is chosen as the optimal zinc location.
- (d) E2, D2, E1D1, E1H1 or D1H1 sites: These sites represent ligand combinations used for predicting six-coordinated metal sites, where the theoretical ligand-metal angles are 90° or 180° . If the distance between two ligand atoms is less than 4 Å, the angle is 90° ; if greater than 4 Å, the angle is 180° . For 90° , the prediction follows method (b), but for glutamate (E) or aspartate (D), c_2 is the midpoint between the metal binding atom (OE1/OE2 for E, OD1/OD2 for D) and the CG atom. For 180° , the midpoint of the two metal-binding atoms is used as the predicted metal coordinate.
- (e) CH3 sites: Three candidate zinc locations are deduced using the strategies a-c for CC, CH, and HH subgroups. A voting mechanism is implemented in this scenario since cysteine is

more liable to adopt a conformation not suitable to coordinate metal when compared to histidine. Three distances are calculated from each pair of candidate zinc locations, with the shortest distance considered a major vote (2 out of 3). The average coordinate of these two candidate zinc locations is chosen as the optimal zinc location.

- (f) Other incomplete coordination sites: For a site with three or more ligands (e.g., EDH3, EDH4, EDH5 sites) but not fully occupied, generate all pairwise ligand combinations (e.g., ED2, ED1H1, H2), resulting in $n(n-1)/2$ pairs. Apply strategies c-d to deduce candidate zinc positions for each pair, generating coordinates. Average all candidate coordinates to obtain the final zinc position.
- (g) Fully coordinated sites: For four-coordinated sites, the center of the four metal-coordinating atoms is chosen as the optimal metal position (Fig. 8d). For six-coordinated sites, the center of the six coordinating atoms is used.

Handling redundant candidate metal locations. Two predicted metal ions are considered redundant if they are too close to each other to form a dinuclear site. Investigation of metal ion distance distribution reveals that most of the distance is between 3 Å and 4 Å, representing the presence of dinuclear metal sites (Supplementary Fig. 13a). While Metal3D uses a 5 Å threshold to assess occupancy redundancy²⁶, this distance may hinder the identification of dinuclear metal sites in some cases. Although the predicted density does include both binding sites, the placement algorithm may struggle to accurately identify and position two separate zinc ions due to overlapping peaks in the density map, which may be interpreted as a single composite peak. In contrast, PMM uses a 2.5 Å threshold to determine redundancy, effectively eliminating occupancy redundancy while accurately annotating dinuclear metal sites. For instance, in the 6jkw structure⁶⁴, PMM identifies two dinuclear metal sites at a distance of 3.2 Å (Supplementary Fig. 13b). Within this framework, true positives (TP) are defined as predicted sites within 2.5 Å of the experimental metal sites.

Calculation of hydrophobic profiles. Characterizing MBSs involves assessing features such as the ‘hydrophobicity contrast function,’ which quantifies the hydrophobicity difference between outer and inner atoms in a stabilizing shell. MBSs exhibit higher hydrophobicity contrast values, with the metal center coordinated by a hydrophilic atomic group shell (containing oxygen, nitrogen, or sulfur atoms) embedded within a larger hydrophobic atomic group shell (containing carbon atoms)⁶⁵. This qualitative observation can be described analytically by the hydrophobicity contrast function C , which is evaluated from the structure and characteristics of different types of metal ions.

The metal ion location is used as the center of the sphere to calculate the hydrophobicity contrast functions values (C) and mean atomic solvation parameters values ($\Delta\sigma$)⁶⁵. For each identified metal ion location, a series of 21 radii ranging from 2 Å to 7 Å, with a step size of 0.25 Å (2, 2.25, 2.5, ..., 7), are chosen to generate hydrophobicity contrast curves (Supplementary Fig. 2a, b) and mean atomic solvation parameter curves (Supplementary Fig. 2c, d). The hydrophobic profiles are used not only in calculating certainty score for each predicted metal ion, but also as features in the ensemble model.

Determine metal binding site probability. We use different validation strategies based on geometric coordination. HCS typically involves more ligands that lead to greater geometric stability and more distinct hydrophobicity patterns. The core of the MBS is typically hydrophilic, surrounded by a pronounced hydrophobic contrast, which is especially prominent in high-coordination sites. As a result, hydrophobicity profile can effectively capture the features of HCS, achieving high predictive accuracy when compared with experimental data. The certainty score of the predicted site is determined by calculating the

Pearson correlation coefficient between the C values and $\Delta\sigma$ values curves of the predicted site and the corresponding curves obtained from the experimental site. HCS are rarer in our dataset when compared with LCS, and are characterized by a much higher number of parameters, increasing the complexity of the feature space and data-to-parameter ratio. This increase in geometric features raises the risk of model overfitting and makes model training more challenging. Therefore, use of the hydrophobicity alone becomes the optimal strategy to validate the presence of HCS.

LCS is characterized by fewer ligands that result in more flexible and irregular geometries, with less pronounced hydrophobicity patterns when compared with HCS. In such cases, hydrophobicity profile alone may not effectively distinguish MBSs from other similar regions. Introducing geometric features, such as ligand interatomic distances and ligand-metal-ligand (LML) angles, provide additional structural information, which can be integrated and analyzed through machine learning models to improve the accuracy of predictions for LCS. Therefore, an ensemble model consisting of five Base Learners is employed to distinguish between metal and non-MBSs for LCS, outputting a probability value for each site as a certainty score. The tetrahedral and octahedral metal sites are constructed using the same approach with different training and test datasets.

The five base learners include Logistic Regression (LR), Decision Tree (DT), MLPClassifier (MLP), Support Vector Machine (SVM), and a Feedforward Neural Network architecture (FFNN) deep learning model implemented using the Keras library. Randomized search cross-validation is used to randomly select several hyperparameter combinations from a predefined parameter range and assess each model's performance using 5-fold cross-validation on the training set. Subsequently, the best hyperparameter combination is chosen based on the weighted F1 score, and these best models are retrained on the entire training set. Finally, using the soft voting method, the average probability of each model predicting the positive class is taken as the certainty score for the candidate MBSs. The features and data used for training the models are as follows:

The features used in the ensemble model are categorized as ligand type, geometrical parameters, and hydrophobic profiles (Supplementary Table 7). Ligand type refers to the coordinating amino acid, encoded as numerical features using One-Hot Encoding. Geometrical parameters are divided into two categories: one represents the relative positions and orientations between side chain atoms ($C\alpha$ and $C\beta$) of the coordinating amino acids, while the others represent the relative positions and orientations between the side chains ($C\alpha$ and $C\beta$) or backbone oxygen atoms and the metal ion. Hydrophobic profiles include 21 hydrophobicity contrast function values (C) and 21 mean atomic solvation parameters values ($\Delta\sigma$). Detailed parameter information is provided in Supplementary Table 7. A compilation of the three categories of data result in a total of 61 predictors used for further model training.

The geometric features described in section 4.2.2, including interacting ligands, coordination atom types, coordination number, and distances between coordination atoms, are used to construct positive and negative datasets for model training. The positive dataset consists of experimentally resolved MBSs verified by CMM and derived from non-redundant protein structures. The negative dataset consists of sites in non-redundant protein structures that meet the geometric criteria outlined in section 4.2.2 but are located more than 5 Å away from any other metal ion, and that satisfy the conditions $Q_c < 0.355$ or $Q_v < 0.25$. The data are stratified and randomly split based on ligand and metal types to ensure that the proportion of each sample category remained consistent across the dataset. To ensure fairness when comparing with other MBS predictors (such as Metal3D and TEMSP), the training set excludes the datasets used by these predictors. Ultimately, 70% of the data are

allocated to the training set and 30% to the test set (Supplementary Table 8). Typical tetrahedral metal sites in transition metals are Zn sites and some Cu sites, so only Zn and Cu sites are used to train the CH sites. For octahedral metal sites, all common transition metal sites (Mn, Fe, Co, Ni, Cu, Zn) are used for training.

A certainty score higher than 0.5 is used as the criterion to further verify the identity of the MBS. The calculated certainty score is annotated in the occupancy field of each atom record for metal ion in the output coordinate file.

Metal type identification for validated metal binding sites. After identifying the MBSs, we predict their potential metal types by incorporating a hydrophobicity profile and seven chemical features, along with the 27 features implemented in the *NEIGHBORHOOD* database (Supplementary Table 9). The seven chemical features include aromaticity, hydrophobicity, positive ionizable, negative ionizable, hydrogen bond donor, hydrogen bond acceptor, and MBS chain²⁶. These features are related to atom types and have been utilized in the Metal3D and 3DCNN MBS prediction tools.

To represent the chemical features as machine learning-compatible feature data, we analyze atoms within a 2–5 Å radius of the metal coordinates, assessing their corresponding feature categories and distances from the metal site. The distances are categorized into three fixed ranges: [2–3 Å), [3–4 Å), and [4–5 Å). In each range, we count the number of atoms with different chemical properties to form structured feature data. This results in 24 numerical features, encompassing the seven biophysical properties and an additional category for atoms not belonging to these seven types.

When classifying MBSs, the number of newly added features remains unchanged regardless of the number of ligands, making the model applicable to all sites. Therefore, we merge the sites predicted by CH and EDH algorithms and remove redundancy by keeping only the sites with the highest certainty score if multiple sites are within 2.5 Å of each other. The training and test sets for the model also do not differentiate between sites based on their coordination numbers. After removing redundancy and splitting the MBS data into 70% training and 30% test sets, the training set has too few non-zinc transition metal sites. To address this, we simulate oversampling by adding metal sites from proteins with less than 30% structural similarity to any protein in the test set (Supplementary Table 10). Cobalt (Co) and nickel (Ni) are rare in biological systems and closely resemble iron (Fe) in binding sites. Thus, Fe2, Fe, Co, and Ni are grouped as a single category, VIII, for training, testing, and prediction, to minimize confusion and enhance accuracy.

Due to the significant differences in MBS counts, particularly the low number of Cu sites, the ensemble model for metal classification combines Logistic Regression (LR), MLPClassifier (MLP), and Support Vector Machine (SVM), along with Balanced Random Forest (RF) and Easy Ensemble Classifier (Easy) to better handle class imbalance and improve prediction accuracy. After optimizing each model, a soft voting approach is used to form the final ensemble.

Web service implementation

PMM web server is deployed using an Ubuntu Linux virtual machine running Nginx 1.14.0 and Unicorn 20.0.4. The interface components of the website are designed and implemented using the Django template engine 3.1.4. Molecular graphics on the view page use HTML5 as implemented in the NGL Javascript library. PMM has been tested in several popular web browsers, including Google Chrome 89.0.4389.82, Mozilla Firefox 87.0, Apple Safari 13.0.2 and Microsoft Edge 89.0.774.75. The styles of the web interface are optimized using the Bootstrap 4.5.0 library to accommodate both large computer screens and small screens on handheld devices. The PMM webserver is accessible via <https://PMM.biocloud.top>.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data used to train and test the model and source data are available with this paper under a deposition in Figshare under accession code (<https://doi.org/10.6084/m9.figshare.25011212>).

Code availability

Code is available under <https://github.com/hhz-lab/PinMyMetal>⁶⁶ and also on Zenodo under <https://doi.org/10.5281/zenodo.14830978>⁶⁷.

References

- Maret, W. New perspectives of zinc coordination environments in proteins. *J. Inorg. Biochem.* **111**, 110–116 (2012).
- Waldron, K. J., Rutherford, J. C., Ford, D. & Robinson, N. J. Metalloproteins and metal sensing. *Nature* **460**, 823–830 (2009).
- Andreini, C., Bertini, I., Cavallaro, G., Holliday, G. L. & Thornton, J. M. Metal-MACiE: a database of metals involved in biological catalysis. *Bioinformatics* **25**, 2088–2089 (2009).
- Andreini, C., Cavallaro, G., Lorenzini, S. & Rosato, A. MetalPDB: a database of metal sites in biological macromolecular structures. *Nucleic Acids Res.* **41**, D312–D319 (2013).
- Holm, R. H., Kennepohl, P. & Solomon, E. I. Structural and functional aspects of metal sites in biology. *Chem. Rev.* **96**, 2239–2314 (1996).
- Matthews, J. M., Loughlin, F. E. & Mackay, J. P. Designed metal-binding sites in biomolecular and bioinorganic interactions. *Curr. Opin. Struct. Biol.* **18**, 484–490 (2008).
- Koohi-Moghadam, M. et al. Predicting disease-associated mutation of metal-binding sites in proteins using a deep learning approach. *Nat. Mach. Intell.* **1**, 561–567 (2019).
- Chalkley, M. J., Mann, S. I. & DeGrado, W. F. De novo metalloprotein design. *Nat. Rev. Chem.* **6**, 31–50 (2022).
- Kakkis, A., Gagnon, D., Esselborn, J., Britt, R. D. & Tezcan, F. A. Metal-templated design of chemically switchable protein assemblies with high-affinity coordination sites. *Angew. Chem. Int. Ed. Engl.* **59**, 21940–21944 (2020).
- Maret, W. Zinc biochemistry: from a single zinc enzyme to a key element of life. *Adv. Nutr.* **4**, 82–91 (2013).
- Witkowska, D. & Rowińska-Żyrek, M. Biophysical approaches for the study of metal-protein interactions. *J. Inorg. Biochem.* **199**, 110783 (2019).
- Domagalski, M. J. et al. The quality and validation of structures from structural genomics. *Methods Mol Biol.* **1091**, 297–314 (2014).
- Turk, M. & Baumeister, W. The promise and the challenges of cryo-electron tomography. *FEBS Lett.* **594**, 3243–3261 (2020).
- Maret, W. Inhibitory zinc sites in enzymes. *BioMetals* **26**, 197–204 (2013).
- Liu, Z. et al. Computationally characterizing and comprehensive analysis of zinc-binding sites in proteins. *Biochim Biophys Acta.* **1844**, 171–180 (2014).
- Auld, D. S. Zinc coordination sphere in biochemical zinc sites. *Bio-Metals* **14**, 271–313 (2001).
- Patel, K., Kumar, A. & Durani, S. Analysis of the structural consensus of the zinc coordination centers of metalloprotein structures. *Biochim. Biophys. Acta.* **1774**, 1247–1253 (2007).
- Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
- Sánchez-Aparicio, J. E. et al. BioMetAll: Identifying Metal-Binding Sites in Proteins from Backbone Preorganization. *J. Chem. Inf. Model.* **61**, 311–323 (2021).

20. Zhao, W. et al. Structure-based de novo prediction of zinc-binding sites in proteins of unknown function. *Bioinformatics* **27**, 1262–1268 (2011).
21. Babor, M., Gerzon, S., Raveh, B., Sobolev, V. & Edelman, M. Prediction of transition metal-binding sites from apo protein structures. *Proteins* **70**, 208–217 (2008).
22. Ireland S. M., Martin A. C. R. Zincbindpredict-prediction of zinc binding sites in proteins. *Molecules* **26**, (2021).
23. Lin, Y. F. et al. MIB: metal ion-binding site prediction and docking server. *J. Chem. Inf. Model.* **56**, 2287–2291 (2016).
24. Lu, C. H. et al. MIB2: metal ion-binding site prediction and modeling server. *Bioinformatics* **38**, 4428–4429 (2022).
25. Hekkelman, M. L., de Vries, I., Joosten, R. P. & Perrakis, A. AlphaFill: enriching AlphaFold models with ligands and cofactors. *Nat. Methods* **20**, 205–213 (2023).
26. Dürr, S. L., Levy, A. & Rothlisberger, U. Metal3D: a general deep learning framework for accurate metal ion location prediction in proteins. *Nat. Commun.* **14**, 2713 (2023).
27. Zheng, H., Chruszcz, M., Lasota, P., Lebioda, L. & Minor, W. Data mining of metal ion environments present in protein structures. *J. Inorg. Biochem.* **102**, 1765–1776 (2008).
28. Kochańczyk, T., Drozd, A. & Krężel, A. Relationship between the architecture of zinc coordination and zinc binding affinity in proteins—insights into zinc regulation. *Metallomics* **7**, 244–257 (2015).
29. Avvaru, B. S. et al. A short, strong hydrogen bond in the active site of human carbonic anhydrase II. *Biochemistry* **49**, 249–251 (2010).
30. Rulíšek, L. & Vondrášek, J. Coordination geometries of selected transition metal ions (Co²⁺, Ni²⁺, Cu²⁺, Zn²⁺, Cd²⁺, and Hg²⁺) in metalloproteins. *J. Inorg. Biochem.* **71**, 115–127 (1998).
31. Dokmanić, I., Sikić, M. & Tomić, S. 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 (2008).
32. Padjasek, M. et al. Structural zinc binding sites shaped for greater works: Structure-function relations in classical zinc finger, hook and clasp domains. *J. Inorg. Biochem.* **204**, 110955 (2020).
33. Andreini, C., Bertini, I. & Cavallaro, G. Minimal functional sites allow a classification of zinc sites in proteins. *PLoS ONE* **6**, e26325 (2011).
34. Vallee, B. L. & Auld, D. S. Active-site zinc ligands and activated H₂O of zinc enzymes. *Proc. Natl Acad. Sci. USA* **87**, 220–224 (1990).
35. Maret, W. Zinc in cellular regulation: the nature and significance of “Zinc Signals”. *Int. J. Mol. Sci.* **18**, 2285 (2017).
36. Zheng, H. et al. CheckMyMetal: a macromolecular metal-binding validation tool. *Acta Crystallogr. D Struct. Biol.* **73**, 223–233 (2017).
37. Gucwa, M. et al. CMM-An enhanced platform for interactive validation of metal binding sites. *Protein Sci.* **32**, e4525 (2023).
38. Ye, R., Tan, C., Chen, B., Li, R. & Mao, Z. J. FiC. Zinc-containing metalloenzymes: inhibition by metal-based anticancer agents. *Front. Chem.* **8**, 402 (2020).
39. Jomova, K. et al. Essential metals in health and disease. *Chem. Biol. Interact.* **367**, 110173 (2022).
40. Pohl, H. R., Wheeler, J. S. & Murray, H. E. Sodium and potassium in health and disease. *Met Ions Life Sci.* **13**, 29–47 (2013).
41. Virág, L., Erdodi, F. & Gergely, P. J. S., *Bioinorganic Chemistry for Medical Students*. 1–104 (University of Debrecen, Debrecen, 2016).
42. Bushnell, D. A. & Kornberg, R. D. Complete, 12-subunit RNA polymerase II at 4.1-Å resolution: implications for the initiation of transcription. *Proc. Natl Acad. Sci. USA* **100**, 6969–6973 (2003).
43. Watanabe, M. et al. The nature of the TRAP-Anti-TRAP complex. *Proc. Natl Acad. Sci. USA* **106**, 2176–2181 (2009).
44. Pichkur, E. B. et al. Insights into the improved macrolide inhibitory activity from the high-resolution cryo-EM structure of dirithromycin bound to the E. coli 70S ribosome. *RNA* **26**, 715–723 (2020).
45. Liu, X., Farnung, L., Wigge, C. & Cramer, P. Cryo-EM structure of a mammalian RNA polymerase II elongation complex inhibited by α-amanitin. *J. Biol. Chem.* **293**, 7189–7194 (2018).
46. Langer L. M., Bonneau F., Gat Y., Conti E. Cryo-EM reconstructions of inhibitor-bound SMG1 kinase reveal an autoinhibitory state dependent on SMG8. *ELife* **10**, e72353 (2021).
47. Pausch, P. et al. DNA interference states of the hypercompact CRISPR-CasΦ effector. *Nat. Struct. Mol. Biol.* **28**, 652–661 (2021).
48. Mohamadi A., et al. An ensemble 3D deep-learning model to predict protein metal-binding site. *Cell Rep. Phys. Sci.* **3**, 101046 (2022).
49. Waldron, K. J. & Robinson, N. J. How do bacterial cells ensure that metalloproteins get the correct metal? *Nat. Rev. Microbiol.* **7**, 25–35 (2009).
50. Cox, J. D., Hunt, J. A., Compher, K. M., Fierke, C. A. & Christianson, D. W. Structural influence of hydrophobic core residues on metal binding and specificity in carbonic anhydrase II. *Biochemistry* **39**, 13687–13694 (2000).
51. Jude, K. M. et al. Ultrahigh resolution crystal structures of human carbonic anhydrases I and II complexed with “two-prong” inhibitors reveal the molecular basis of high affinity. *J. Am. Chem. Soc.* **128**, 3011–3018 (2006).
52. Li, M. et al. Crystal structure studies on rubrerythrin: enzymatic activity in relation to the zinc movement. *J. Biol. Inorg. Chem.* **8**, 149–155 (2003).
53. Hernández-Santoyo, A. et al. Crystal structure of Cu/Zn superoxide dismutase from *Taenia solium* reveals metal-mediated self-assembly. *FEBS J.* **278**, 3308–3318 (2011).
54. Santhosh, R. et al. Diffraction precision index of macromolecular structures: a web based database. *Res. J. Life Sci. Bioinform. Pharm. Chem. Sci.* <https://doi.org/10.26479/2019.0502.10> (2019).
55. Luzzati, V. Traitement statistique des erreurs dans la détermination des structures cristallines. *Acta Crystallogr.* **5**, 802–810 (1952).
56. Handing, K. B. et al. Characterizing metal-binding sites in proteins with X-ray crystallography. *Nat. Protoc.* **13**, 1062–1090 (2018).
57. Li, J. et al. The metal-binding protein atlas (MbPA): an integrated database for curating metalloproteins in all aspects. *J. Mol. Biol.* **435**, 168117 (2023).
58. Sousa, S. F., Lopes, A. B., Fernandes, P. A. & Ramos, M. J. The Zinc proteome: a tale of stability and functionality. *Dalton Trans.* **14**, 7946–7956 (2009).
59. Burley, S. K. et al. Protein Data Bank (PDB): the single global macromolecular structure archive. *Methods Mol. Biol.* **1607**, 627–641 (2017).
60. Zheng, H., Shabalin, I. G., Handing, K. B., Bujnicki, J. M. & Minor, W. Magnesium-binding architectures in RNA crystal structures: validation, binding preferences, classification and motif detection. *Nucleic Acids Res.* **43**, 3789–3801 (2015).
61. Laitaoja, M., Valjakka, J. & Jänis, J. Zinc coordination spheres in protein structures. *Inorg. Chem.* **52**, 10983–10991 (2013).
62. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
63. McDonald, I. K. & Thornton, J. M. The application of hydrogen bonding analysis in X-ray crystallography to help orientate asparagine, glutamine and histidine side chains. *Protein Eng.* **8**, 217–224 (1995).
64. Park, K. S. et al. PNGM-1, a novel subclass B3 metallo-β-lactamase from a deep-sea sediment metagenome. *J. Glob. Antimicrob. Resist.* **14**, 302–305 (2018).

65. Yamashita, M. M., Wesson, L., Eisenman, G. & Eisenberg, D. Where metal ions bind in proteins. *Proc. Natl Acad. Sci. USA* **87**, 5648–5652 (1990).
66. Zhang, H. et al. *GitHub* <https://github.com/hhz-lab/PinMyMetal> (2024).
67. Zhang, H. et al. hhz-lab/PinMyMetal: V1.0 *Zenodo*, <https://doi.org/10.5281/zenodo.14830978> (2025).

Acknowledgements

This work is supported by the Science and Technology Planning Project of Guangdong Province of China A2020198 (N.W.); the Harrison Family Funds via University of Virginia and National Institute of General Medical Sciences grant R01-GM132595 (W.M.); and a fund from Suzhou Tributary Biologics Co., Ltd (H.Z.). We thank Yaowang Li and Yanxia Ru for their evaluation and suggestions regarding the application of the PMM system in cryo-EM structures.

Author contributions

Huihui Zhang (H.H.Z.), N.W., and Heping Zheng (H.Z.) conceived of the ideas implemented in this project; H.H.Z. and J.Z. investigated the data; H.H.Z., M.G., Y.Z., and H.M. developed methodology and web server; H.H.Z. and H.Z. drafted the manuscript; H.H.Z., L.D., L.M., W.M., N.W., and H.Z. revised and edited draft; M.G., N.W., W.M., and H.Z. tested PMM; H.Z. supervised the research.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-57637-5>.

Correspondence and requests for materials should be addressed to Wladek Minor, Nasui Wang or Heping Zheng.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025