Predict metal-binding proteins and structures through integration of evolutionary-scale and physics-based modeling

Xin Dai[1,*], Max Henderson[2], Shinjae Yoo[1], and Qun Liu[2,3,*]

[1]Computational Science Initiative, Brookhaven National Laboratory, Upton, NY, USA
[2]Department of Biochemistry and Cell Biology, Stony Brook University, Stony Brook, NY, USA
[3]Biology Department, Brookhaven National Laboratory, Upton, NY, USA
[*]Correspondence: xdai@bnl.gov or qunliu@bnl.gov

ABSTRACT

Metals are essential elements in all living organisms, binding to approximately 50% of proteins. They serve to stabilize proteins, catalyze reactions, regulate activities, and fulfill various physiological and pathological functions. While there have been many advancements in determining the structures of protein-metal complexes, numerous metal-binding proteins still need to be identified through computational methods and validated through experiments. To address this need, we have developed the ESMBind-based workflow, which combines evolutionary scale modeling (ESM) for metal-binding prediction and physics-based protein-metal modeling. Our approach utilizes the ESM-2 and ESM-IF models to predict metal-binding probability at the residue level. In addition, we have designed a metal-placement method and energy minimization technique to generate detailed 3D structures of protein-metal complexes. Our workflow outperforms other models in terms of residue and 3D-level predictions. To demonstrate its effectiveness, we applied the workflow to 142 uncharacterized fungal pathogen proteins and predicted metal-binding proteins involved in fungal infection and virulence.

Keywords: Evolutionary scale modeling, protein function prediction, energy minimization, metal binding

Introduction

Protein-metal ion interactions are essential in numerous biological processes, including enzyme catalysis, signal transduction, structural stabilization, and regulation of physiological and pathological activities. It is estimated that approximately 50% of proteins bind to metals [1] Predicting and studying these interactions at the atomic level is critical to understanding protein function and enabling applications such as drug discovery and protein engineering. However, accurately predicting metal binding sites and the precise positioning of metals within 3D protein structures remains challenging. Recent advances in deep learning have opened new possibilities for improved metal binding site prediction by leveraging large-scale protein data. Pretrained foundation models like AlphaFold2[2], ESM-2[3], ESM-IF[4], and ProtTrans[5] have shown great promise in capturing intricate patterns within protein sequences and structures. These models enable more accurate predictions of protein structures, even for proteins with limited homologies to known structures.

Several deep-learning approaches have been developed specifically for metal binding site prediction at residue levels. LMetalSite[6] and M-Ionic[7] utilize pretrained language models to predict binding sites for multiple metal types. While both methods demonstrate the utility of pretrained protein language

models, they are restricted to residue-level predictions and do not provide atomic-level details about the placement of metals and atomic-level protein-metal interactions. Dürr et al. introduced Metal3D[8], a deep learning framework that uses 3D convolutional neural networks to predict zinc locations with high spatial accuracy. Metal3D demonstrates impressive performance for zinc, predicting its locations within 0.70 ± 0.64 Å of experimental positions. However, its performance on other metals and ability to distinguish between different metal types remains to be fully explored. To address the limitations of structure-based prediction methods in handling metal ions, Hekkelman et al. developed AlphaFill[9], an algorithm that enriches AlphaFold models with ligands and cofactors, including metals. While this approach can provide valuable insights, it is inherently limited by the availability of related experimental structures and may not accurately predict novel metal-binding proteins and sites.

In this study, we present an ESMBind-based workflow, a novel approach that combines deep learning and physics-based metal placement and modeling to predict metal-binding proteins and structures at the atomic level for seven essential-to-life metals: Zn^{2+}, Ca^{2+}, Mg^{2+}, Mn^{2+}, Fe^{3+}, Co^{2+}, and Cu^{2+}. Our method addresses the limitations of existing approaches by extending residue-level predictions to precise 3D coordinates while maintaining physical realism. ESMBind integrates sequence information from the pretrained ESM-2 model and structural information from the pretrained ESM-IF model to predict metal binding probabilities at the residue level.

The key innovation of the workflow lies in its ability to transform residue-level predictions into accurate 3D structures of protein-metal complexes. We developed a sophisticated local grid search strategy that addresses challenges such as determining the number of ions present and optimizing their positions within the protein structure. This initial placement is followed by energy minimization that further optimizes the positions of the metal ions, revealing physical interactions between the proteins and metal ions. We trained and evaluated our approach using high-quality protein-metal binding data from the BioLip database[10], which curates information from the Protein Data Bank (PDB). By comparing our predictions with the ground truth data, we demonstrate the accuracy and effectiveness of our method in identifying protein-metal binding sites and determining the optimal positions of metal ions within protein structures.

Results

Overview of the workflow

Our workflow is illustrated in Figure 1a. The workflow uses protein sequence and structural information to enhance prediction accuracy through the integration of ESMBind-based prediction followed by modeling of metal placement and energy minimization. The workflow is designed to predict protein-metal binding sites for seven common metal ions in metalloproteins: Zn^{2+}, Ca^{2+}, Mg^{2+}, Mn^{2+}, Fe^{3+}, Co^{2+}, and Cu^{2+}. The workflow begins with the input of protein sequence and structure data. ESMBind processes the input information to predict binding probabilities for each residue to each of the seven metals, effectively performing a binary classification at the residue level to determine whether a residue binds to a metal ion. Following this residue-level prediction, the workflow proceeds to the metal placement stage. Here, the binding probabilities are used to guide the initial placement of metals within the protein structure. This step transforms the residue-level predictions into a 3D structure of the protein-metal complex. The final step involves energy minimization, which optimizes the protein structure and metal positions according to physical principles using the AMBER force field. This step optimizes the initial ion placements, ensuring they are physically and stereochemically plausible within the context of the protein structure. The output of this stage is a 3D structure of the protein modeled with a predicted metal-binding site.

The architecture of the ESMBind model (Figure 1b) consists of several components. It processes input features from ESM-2 (650M variant) and ESM-IF, which provide sequence and structure embeddings,

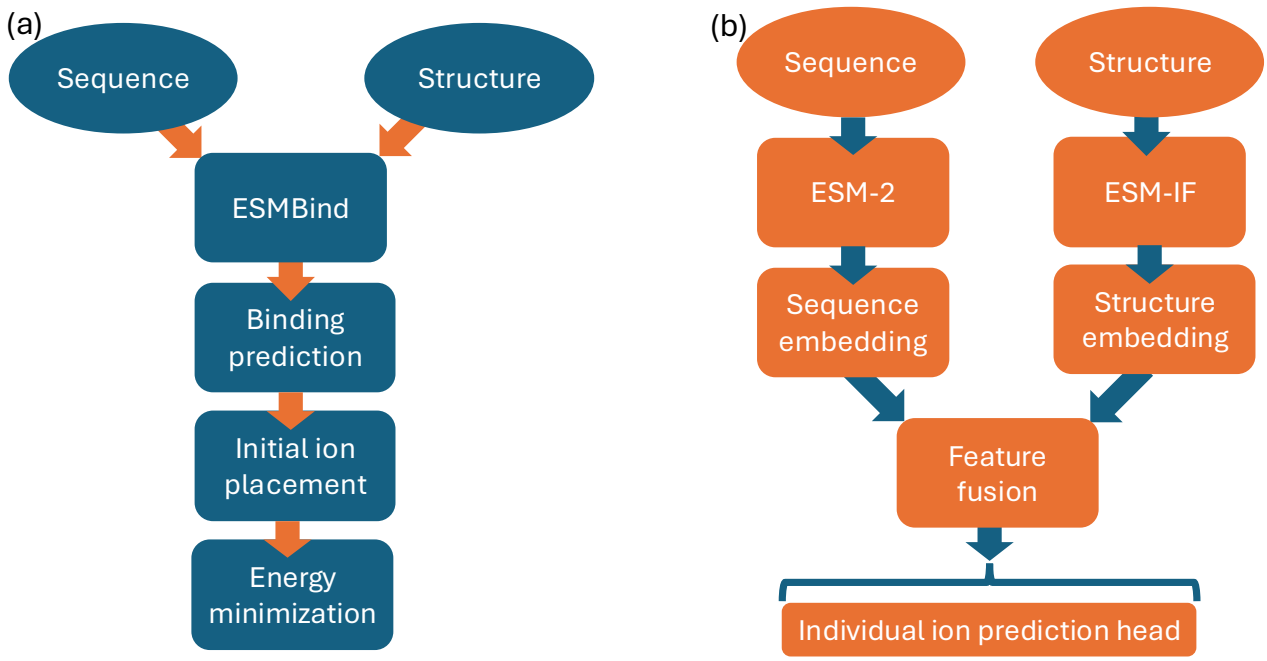

Figure 1. ESMBind workflow and architecture. (a) Workflow of ESMBind for metal-binding site prediction. The process begins with protein sequence and structure as inputs to ESMBind, which predicts binding probabilities at the residue level. These probabilities are used for ion placement followed by energy minimization to produce the modeled 3D structure with placed ions. (b) Architecture of the ESMBind model. The model combines sequence embeddings from ESM-2 and structure embeddings from ESM-IF through feature fusion. The fused features are then processed by individual ion prediction heads to generate binding probabilities for each metal type.

respectively. These embeddings are processed through separate input blocks, which normalize and transform them, reducing their dimensionality. The outputs from these input blocks are then normalized and concatenated, resulting in a feature vector that combines sequence and structure information. This combined feature vector is further processed and fed into metal-specific classifier heads, which output binding probabilities for each metal at the residue level. This approach, combining deep learning predictions for residue-level binding with subsequent 3D metal placement and physical modeling, enables the ESMBind-based workflow to provide accurate and physically consistent predictions of metal ion locations in protein structures.

Residue-level prediction

The residue-level prediction is the foundation of ESMBind's capabilities. For each metal type, the model predicts the probability of each residue in the protein sequence binding to that metal. This task is formulated as a binary classification problem, where each residue is classified as either binding or non-binding for a specific metal type. To evaluate the performance of ESMBind in this task, we focus on the Area Under the Precision-Recall Curve (AUPRC) as the key metric. AUPRC is particularly suitable for this task due to the highly imbalanced nature of the dataset, where non-binding residues far outnumber binding residues. Figure 2 presents the AUPRC scores for ESMBind compared to two other deep-learning-based methods, M-Ionic and LMetalsite, across the seven metal types studied. The results demonstrate that ESMBind consistently outperforms the other methods for most metals. For Zn^{2+}, which has the largest

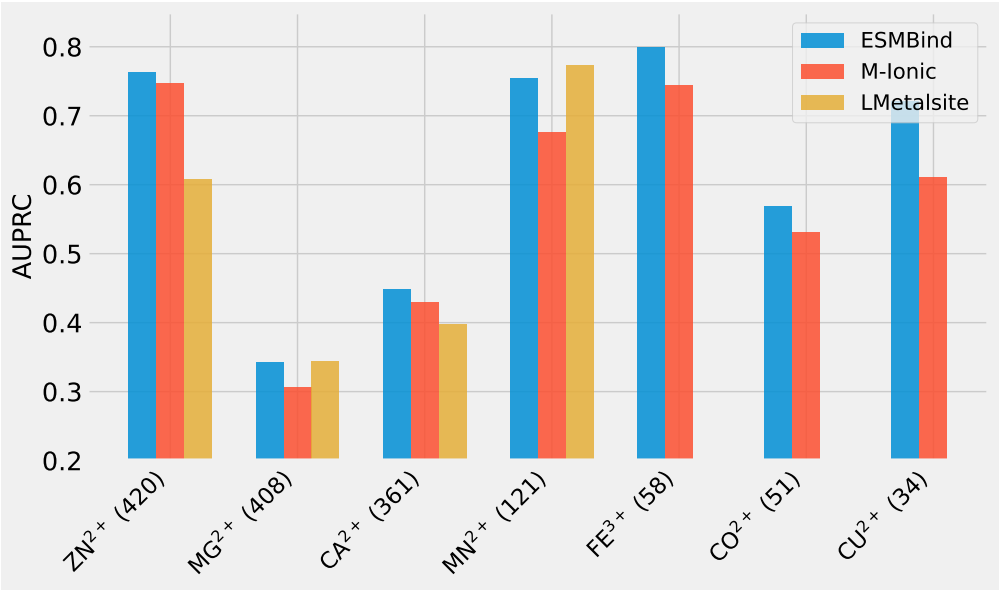

Figure 2. Comparison of Area Under the Precision-Recall Curve (AUPRC) scores for residue-level metal ion binding prediction across different methods. The plot shows AUPRC scores for ESMBind, M-Ionic, and LMetalsite for seven metal ions: Zn^{2+}, Mg^{2+}, Ca^{2+}, Mn^{2+}, Fe^{3+}, Co^{2+}, and Cu^{2+}. The number of examples for each ion is indicated in parentheses. LMetalsite results are not available for Fe^{3+}, Co^{2+}, and Cu^{2+} as it was not trained on these ions.

number of test examples (420), ESMBind achieves an AUPRC of 0.763, outperforming M-Ionic (0.747) and LMetalsite (0.608). ESMBind maintains its robust performance for metals with fewer training data in the dataset. For instance, it achieves the highest AUPRC for Fe^{3+} (0.800) and Cu^{2+} (0.722), significantly outperforming M-Ionic. It is noted that LMetalsite is not trained on Fe and Cu and thus was not used for comparison. The consistent and superior performance across all metal types suggests that ESMBind effectively leverages the combined sequence and structural information to make accurate predictions, even in cases where data is limited. This strong residue-level prediction provides a solid foundation for the subsequent steps in the workflow, i.e., the initial metal placement and energy minimization.

Atomic Level Modeling

Transitioning from residue-level predictions to accurate 3D metal placement presents a significant challenge. The deep-learning model predicts high probabilities for most transition metals, complicating the determination of the specific metal type. To simplify this issue, we consider only one metal type per protein. Additionally, although ESMBind predicts binding residues accurately, it doesn't predict the number of metals present in the structure. We thus devised an effective metal placement strategy to tackle these challenges, optimizing metal positions based on geometric and energetic considerations (see Materials and Methods for details). To evaluate the performance of our method in predicting metal binding sites at the atomic level, we first define what constitutes a correct prediction. In our evaluation, a true positive is defined as a predicted metal within 5 Å of an experimental metal position in the protein structure. It's important to note that this definition differs from the residue-level predictions discussed earlier; here, we are concerned about the accuracy of the 3D positions of the metals. We use several metrics to assess the performance of our method: precision, recall, F-1 score, and median deviation. Precision measures the proportion of correctly predicted metal sites among all predictions, while recall indicates the proportion of actual metal sites that were correctly identified. The F-1 score is the harmonic mean of precision and

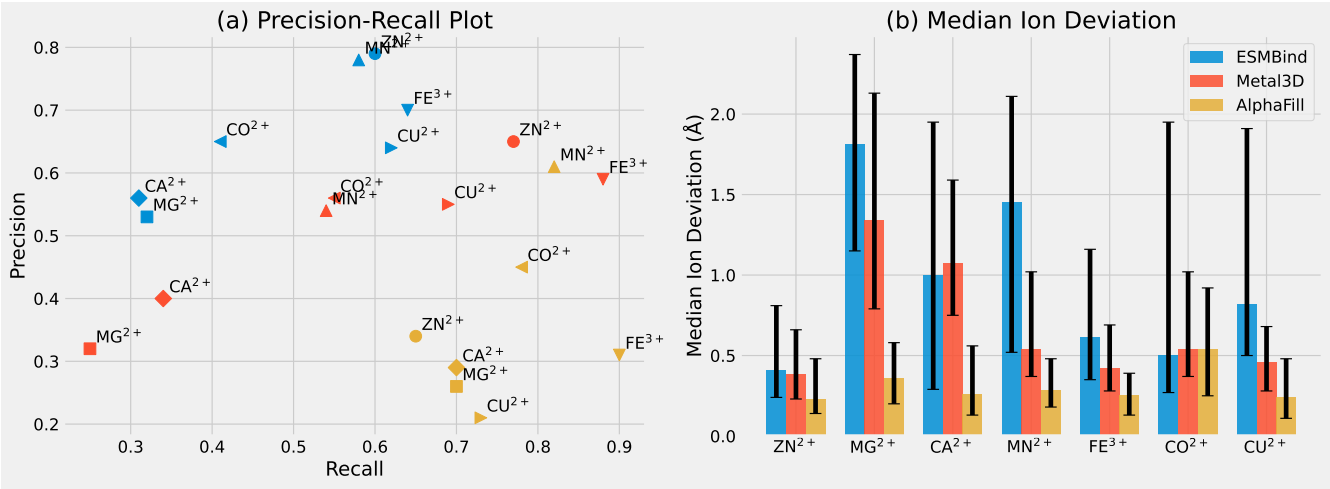

Figure 3. Performance comparison of ESMBind, Metal3D, and AlphaFill across various metal ions. (a) Precision-Recall plot. Different shapes represent different metal types. Different colors distinguish different methods. (b) Median deviation of predicted ion positions from their true locations. Error bars represent the 25th and 75th percentiles.

recall, providing a balanced measure of the two. The median deviation quantifies the spatial accuracy of the correctly predicted metal positions.

Figure 3(a) illustrates the precision-recall trade-off for each metal type. ESMBind demonstrates superior performance across various metal types, particularly for transition metals. For zinc (Zn^{2+}), we achieve a high precision of 0.79 and a recall of 0.60, resulting in an F-1 score of 0.68. This indicates that our method is particularly effective at identifying true zinc-binding proteins and sites while minimizing false positives. Similarly, for manganese (Mn^{2+}) and iron (Fe^{3+}), we obtain high precision (0.78 and 0.69 respectively) with good recall (0.58 and 0.64 respectively), resulting in F-1 scores of 0.67 and 0.66. The performance for lighter metals (element number < 20), calcium (Ca^{2+}) and magnesium (Mg^{2+}), is somewhat lower, with a precision of 0.56 and 0.53, and recall of 0.31 and 0.32 respectively, yielding F-1 scores of 0.40 for both. This lower performance might be attributed to the greater variability in binding sites for calcium and magnesium or their weaker and transient interactions compared to transition metals. The comparison of F-1 scores with Metal3D and AlphaFill is shown in Figure S1. To assess the overall performance across all metal types, we calculated the micro-averaged F-1 score, which gives equal weight to each prediction. ESMBind achieves a micro F-1 score of 0.54, outperforming Metal3D (0.51) and AlphaFill (0.46). This analysis demonstrates the robust overall performance of our method across different metal types.

Figure 3(b) presents the median deviation of correctly predicted ion positions from their true locations. ESMBind generally shows higher median deviations compared to Metal3D and AlphaFill, which is expected given that our approach predicts the initial metal positions at the residue level instead of 3D structures. This intermediate step introduces a certain loss in spatial precision. However, it's crucial to emphasize that these median deviations are calculated only for true positive predictions. When comparing ESMBind to Metal3D and AlphaFill, we observe distinct performance characteristics. ESMBind generally achieves higher precision across metal types, while AlphaFill tends to have higher recall but much lower precision. Metal3D often falls between the two in terms of precision and recall. The balance between precision and recall, as reflected in the F-1 score, is a key metric for evaluation. In this regard, ESMBind demonstrates competitive performance, achieving the highest micro-averaged F-1 score. This balanced

performance highlights ESMBind's robustness in handling diverse metal-binding scenarios, even though it may have slightly lower spatial accuracy for true positive predictions compared to structure-based methods.

Ablation study

We evaluated the contribution of different components of ESMBind through an ablation study (Figure S2). The full model was compared to versions using only ESM-2 embeddings, only ESM-IF embeddings, and a non-ensemble approach. The full ESMBind model consistently outperformed the ablated versions across all seven metals, demonstrating the effectiveness of combining both sequence and structural information with an ensemble approach.

The ESM-2-only model performed well, suggesting that sequence information alone captures a significant portion of the signal for the metal binding prediction. This underscores the importance of evolutionary information encoded in protein sequences for identifying metal binding sites. In contrast, the ESM-IF-only model showed a notable decrease in performance, indicating that structural information alone is insufficient for accurate prediction. This result highlights the complementary nature of sequence and structure in predicting metal binding properties. The non-ensemble model showed only a slight decrease in performance compared to the full model. This suggests that while the ensemble approach improves performance, the majority of the predictive power comes from the model architecture and the combination of sequence and structural information. The ensemble approach likely helps reduce variance and improve generalization across different types of metal binding sites.

We also examined the impact of the energy minimization step on the 3D metal placement. This step slightly improved the overall performance, increasing the micro-averaged F1 score from 0.535 to 0.538. For Mg^{2+}, Ca^{2+}, and Mn^{2+}, which have relatively large initial deviations, energy minimization substantially reduced errors. This improvement is particularly notable for these metals, which often have more variable coordination geometries. These results indicate that energy minimization is mostly beneficial for refining less accurate initial predictions and may be particularly valuable when applied to predicted protein structures, such as those generated by AlphaFold. In such cases, the minimization step can help ensure that predicted metal positions are not only statistically likely based on the deep learning model's output but also chemically and energetically favorable within the context of the 3D atomic-level structure.

Predict novel metal-binding proteins and structures

To further evaluate the utility of the workflow in predicting metal binding proteins and structures, we applied it to 142 putative effector proteins secreted by *Colletotrichum sublineola* (*C. sublineola*), a fungal pathogen leading to Sorghum anthracnose[11,12]. These putative effector proteins could modulate the plant's immune system for pathogen's virulence and infection. However, none of their functions have been predicted and characterized. Using the sequences and Alphafold models of the 142 effectors as input to our workflow, we predict that, out of 142, 22 bind to Zn^{2+}, 18 bind to Mn^{2+}, and 21 bind to Fe^{3+}. We used the pLDDT scores from the AlphaFold-predicted models to rank the predicted metal-binding proteins and presented the top four in Figure 4. The highest pLDDT score is a zinc-finger protein (Uniprot# A0A066X8W9). A Zn^{2+} is coordinated by four cysteine residues (Figure 4a). A functional implication could be that this effector protein might modulate its host's RNA/DNA stability or activity in favor of infection and virulence. The prediction of zinc-finger protein is supported by a yeast homolog structure (PDB id 2J6A) of a Trm112 protein, which is a methyltransferase activator[13]. The superimposition of the two structures has an RMSD of only 0.87 Å with an almost identical Zn^{2+} location (Figure S3). Similarly, we predicted A0A066XCS5 is an iron-binding, Hce2 domain-containing effector

protein[14], A0A066XI99 is a manganese-binding, mannosylphosphate transferase [15], and A0A066XXV0 is a lysine-specific metalloendopeptidase domain-containing protein [16] (Figure 4b,c,d).

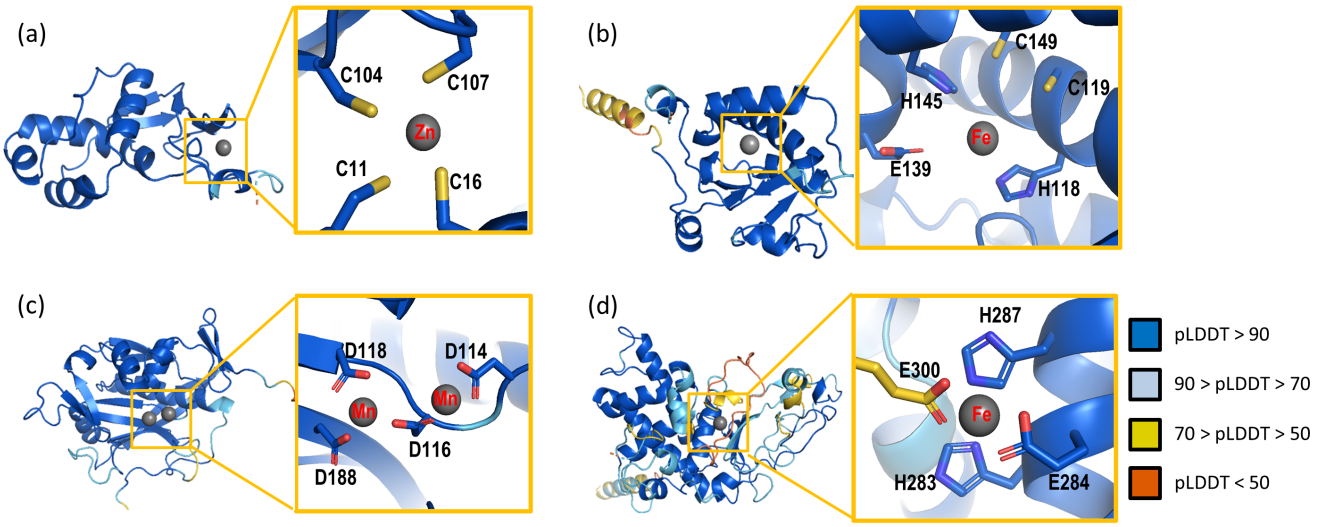

Figure 4. Predicted metal binding proteins and structures. The top four results from the application of ESMBind workflow to purtative C. sublineola effectors colored by pLDDT scores with metal ions shown in grey. (a) A0A066X8W9 (pLDDT 96.2) is predicted to be a Trm112p-like protein. (b) A0A066XCS5 (pLDDT 86.83) is predicted to be a Hce2 domain-containing effector protein (c) A0A066XI99 (pLDDT 85.5) is predicted to be a mannosylphosphate transferase. (d) A0A066XXV0 (pLDDT 83.58) is predicted to be a lysine-specific metallo-endopeptidase domain-containing protein.

Discussion

ESMBind leverages pretrained foundation models for metal binding protein prediction, distinguishing itself by incorporating sequence and structure information through ESM-2 and ESM-IF, respectively. This dual-input strategy, coupled with a multi-layer perceptron (MLP) architecture, outperforms the Transformer-based architectures of LMetalsite and M-Ionic, which use single protein language models (ProtTrans and ESM-2, respectively). The success of our MLP architecture suggests that the integration of both sequence and structure information provides richer features that can be effectively leveraged without complex downstream processing. A potential limitation of the ESMBind model is the computational requirement of two pre-trained models. Particularly, ESM-IF needs a protein structure as input. When experimental structures are not available, AlphaFold-predicted models may be used since ESM-IF has been trained on an extensive set of AlphaFold models.

The metal placement strategy serves as an effective bridge between residue-level predictions and atomic-level 3D localizations and structures. This approach successfully translates the probabilistic outputs of the deep learning model into precise 3D coordinates of protein-metal complexes. Importantly, this strategy is not limited to ESMBind and could, in principle, be adapted to enhance other residue-level deep learning models for metal binding protein prediction and structure modeling.

Compared to Metal3D and AlphaFill, ESMBind demonstrates superior precision in identifying binding metals, as evidenced by higher precision and recall scores. However, our spatial accuracy is somewhat lower, a trade-off resulting from our two-stage approach versus Metal3D's end-to-end 3D prediction or

AlphaFill's direct transplant method. We argue that accurate metal binding and site identification outweigh small differences in spatial precision for many applications. This is particularly true when predicting metal-binding proteins with low sequence homology to experimental structures. The ESMBind's success in balancing identification accuracy with reasonable spatial precision showcases the potential of using hybrid approaches integrating deep learning with physics-based modeling. Future work could focus on improving spatial accuracy through more sophisticated energy models or exploring end-to-end learning approaches that maintain the benefits of pre-trained foundation models.

Materials and methods

Data collection and process

We retrieved protein-metal binding data from the BioLip database[10,17]. BioLip parses protein structures from the Protein Data Bank (PDB) to identify biologically relevant ligands and their binding residues. For any biologically relevant ligand, it defines a residue as binding if the closest atomic distance between the residue and the ligand is within 0.5 Å plus the sum of Van der Waal's radii of the two atoms.

We downloaded data from BioLip on 05-30-2024. We filtered the data to only include sequences with a maximum length of 1500 residues and structures with an experimental resolution of 3.0 Å or better. To obtain a non-redundant set of sequences for each metal type, we used CD-HIT[18] with a 30% sequence identity cutoff. Finally, we randomly split the resulting sequences into training (80%) and test (20%) sets. The test set was reserved for evaluating the trained models.

Model training

The training process for ESMBind involves several key components and techniques to optimize performance and handle the challenges associated with the highly imbalanced dataset. We trained the model on multiple metal types simultaneously, with separate train and validation datasets prepared for each metal type, using the binary cross-entropy loss as the loss function. To address the issue of imbalanced data, where the number of negative samples (non-binding residues) greatly exceeds the number of positive samples (binding residues), we employ an imbalanced dataset sampler that adjusts the sampling probability of each class based on the desired positive ratio. Specifically, we boost the probability of the positive class to 3-fold of its original ratio, effectively oversampling the minority class. During training, we evaluate the model's performance using the Area Under the Precision-Recall Curve (AUPRC) metric, which is particularly suitable for evaluating imbalanced datasets. We also determined the optimal classification threshold for each metal type by finding the threshold that maximizes the F1 score on the validation set. We stored the best threshold for each metal type to use with binary predictions.

We employ regularization techniques at both the model and training levels to prevent overfitting and improve the model's robustness. At the model level, we apply dropout and layer normalization and inject Gaussian noise into the embeddings during training. At the training level, we use weight decay as a regularization term in the optimizer and implement early stopping by monitoring the validation AUPRC, stopping training if there is no improvement after 7 epochs, and saving the model with the best validation AUPRC as the final model. We trained the model using the AdamW optimizer with a weight decay of 0.035 and employed a learning rate scheduler to adjust the learning rates during training by reducing the learning rate by a factor of 0.7 every 15 epochs. This helped the model converge to a better solution by taking smaller steps as the training progressed. Additionally, we applied Stochastic Weight Averaging[19] (SWA) to further improve generalization and stability, where we updated a separate SWA model with the average of the model parameters after a specified number of epochs, which acts as another regularization technique.

To further boost the performance, we employed an ensemble approach. We split the entire dataset into an 80:20 ratio, reserving the 20% split as the final test set. We divide the remaining 80% of the data into 5 folds for cross-validation and train each model in the ensemble on 4 folds while validating on the remaining fold. The final ensemble model comprises 5 models, each trained on a different combination of training and validation folds. During inference, we average the predictions from these 5 models to obtain the final output probabilities. This ensemble approach reduces overfitting and improves the robustness and generalization ability of the model.

Metal placement and energy minimization

Building upon the residue-level binding predictions from ESMBind, we developed a protocol to place metals in 3D protein structures and optimize their positions. This process involves several key steps: **Initial Metal Placement:** We use the binding probabilities output by ESMBind to identify potential binding residues. To determine binding residues, we employed a threshold set at half of the optimal threshold obtained during model validation. This optimal threshold is defined as the threshold that maximizes the F1 score on the validation set during the ESMBind training process. By using half of this optimal threshold, we address the model's tendency to produce more false negatives than false positives due to the highly imbalanced dataset distribution. For each predicted binding residue, we identified all potential binding atoms based on the metal type and then calculated the center of mass of these binding atoms to determine the initial metal position. This approach ensures that the initial metal placement considers all relevant binding atoms for each residue. The list of potential binding atoms for different residues is based on the study by Zheng et al.[20] and Bazayeva et al.[21]. **Binding Site Identification:** We employed agglomerative hierarchical clustering to group nearby binding metals after their initial placement. The process uses a dynamic distance threshold, starting with a predefined value (7Å) and gradually decreasing it until at least one cluster is identified. For each resulting cluster, we merge the binding atoms from all constituent placements and calculate a new metal position as the center of mass of these merged atoms. This approach allows for the consolidation of multiple potential binding sites into distinct metal ion locations. **Refined Metal Positioning:** We implemented a local grid search strategy to optimize the metal positions. This approach aims to place the predicted metals as close to their binding protein atoms as possible while avoiding steric clashes. The search considers ideal binding distances for different atom types and evaluates potential positions based on a scoring function that balances favorable interactions with binding atoms and unfavorable close contacts with non-binding atoms. This method allows for fine-tuning of metal positions within the local environment of the binding site. **System Construction:** We constructed the molecular system using the AMBER force field[22], incorporating a single protein chain, one metal type, and water as the solvent. The 12-6-4 Lennard-Jones non-bonded model[23] characterizes metal interactions, providing an accurate representation of the metal-protein and metal-solvent interactions. **Energy Minimization:** We used OpenMM[24] to perform a series of constrained energy minimizations: a) Apply restraints to all non-solvent atoms, allowing water molecules to adapt to the protein-metal complex. b) Restrain protein backbone atoms and metals, permitting side chain adjustment. c) Constrain only metals, enabling protein relaxation around optimized metal locations. The force constant is halved with each step, facilitating gradual structural adaptation. This multi-step minimization process allows for careful optimization of the metal positions while maintaining the overall protein structure.

This protocol leverages the strengths of deep learning for initial binding site prediction, employs a local search algorithm to refine metal locations, and uses energy minimization to further optimize the protein-metal complex. By addressing the challenges of transitioning from residue-level predictions to precise 3D ion placement, our method provides a robust and physically realistic approach to identifying and characterizing metal-binding proteins and structures.

Prediction of metal-binding pathogen effectors

The proteome of sorghum anthracnose fungal pathogen *Colletotrichum sublineola* (*C. sublineola*) was retried from Uniprot (www.uniprot.org). The proteome contains 12687 unique protein sequences. We used the program SignalP 6[25] to predict signal peptides, WolF-PSORT[26] to identify their subcellular localizations, and we obtained 1370 targets as secreted proteins. Using the effector prediction program EffectorP[27], we identified and selected 142 putative effectors that are secreted by *C. sublineola* and function in the host's cytoplasm. We used their AlphaFold structure models and sequences as input for the prediction of metal-binding proteins and modeling of the protein-metal complexes.

Acknowledgments

We thank Kriti Chopra for the helpful discussion. This work was supported in part by Brookhaven National Laboratory LDRD 23-06 and the U.S. Department of Energy (DOE), Office of Biological and Environmental Research (KP1601011).

Data and code availability

This study utilizes publicly available data from three main sources: protein structures from the Protein Data Bank (https://www.rcsb.org), AlphaFold predicted structures (https://alphafold.ebi.ac.uk/), and binding data from BioLiP (https://zhanggroup.org/BioLiP/index.cgi). ESMBind is developed as an open-source Python package and is freely available on GitHub (https://github.com/Structurebiology-BNL/ESMBind). This repository contains all the necessary code for both training and testing the model.

References

1. Thomson, A. J. & Gray, H. B. Bio-inorganic chemistry. *Curr. opinion chemical biology* **2**, 155–158 (1998).
2. Jumper, J. *et al.* Highly accurate protein structure prediction with alphafold. *nature* **596**, 583–589 (2021).
3. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
4. Hsu, C. *et al.* Learning inverse folding from millions of predicted structures. In *International conference on machine learning*, 8946–8970 (PMLR, 2022).
5. Elnaggar, A. *et al.* Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis machine intelligence* **44**, 7112–7127 (2021).
6. Yuan, Q., Chen, S., Wang, Y., Zhao, H. & Yang, Y. Alignment-free metal ion-binding site prediction from protein sequence through pretrained language model and multi-task learning. *Briefings bioinformatics* **23**, bbac444 (2022).
7. Shenoy, A., Kalakoti, Y., Sundar, D. & Elofsson, A. M-ionic: prediction of metal-ion-binding sites from sequence using residue embeddings. *Bioinformatics* **40**, btad782 (2024).
8. 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).
9. 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).

10. Yang, J., Roy, A. & Zhang, Y. Biolip: a semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic acids research* **41**, D1096–D1103 (2012).

11. Stutts, L. R. & Vermerris, W. Elucidating anthracnose resistance mechanisms in sorghum—a review. *Phytopathology®* **110**, 1863–1876 (2020).

12. Abreha, K. B., Ortiz, R., Carlsson, A. S. & Geleta, M. Understanding the sorghum–colletotrichum sublineola interactions for enhanced host resistance. *Front. plant science* **12**, 641969 (2021).

13. Heurgué-Hamard, V. *et al.* The zinc finger protein ynr046w is plurifunctional and a component of the erf1 methyltransferase in yeast. *J. Biol. Chem.* **281**, 36140–36148 (2006).

14. Zhang, M. *et al.* Hce2 domain-containing effectors contribute to the full virulence of valsa mali in a redundant manner. *Mol. plant pathology* **20**, 843–856 (2019).

15. Wildeman, A. S., Patel, N. K., Cormack, B. P. & Culotta, V. C. The role of manganese in morphogenesis and pathogenesis of the opportunistic fungal pathogen candida albicans. *PLoS pathogens* **19**, e1011478 (2023).

16. Healy, V., O'Connell, J., McCarthy, T. V. & Doonan, S. The lysine-specific proteinase from armillaria mellea is a member of a novel class of metalloendopeptidases located in basidiomycetes. *Biochem. Biophys. Res. Commun.* **262**, 60–63 (1999).

17. Zhang, C., Zhang, X., Freddolino, P. L. & Zhang, Y. Biolip2: an updated structure database for biologically relevant ligand–protein interactions. *Nucleic Acids Res.* **52**, D404–D412 (2024).

18. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. Cd-hit: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).

19. Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D. & Wilson, A. G. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407* (2018).

20. Zheng, H., Chruszcz, M., Lasota, P., Lebioda, L. & Minor, W. Data mining of metal ion environments present in protein structures. *J. inorganic biochemistry* **102**, 1765–1776 (2008).

21. Bazayeva, M., Andreini, C. & Rosato, A. A database overview of metal-coordination distances in metalloproteins. *Acta Crystallogr. Sect. D: Struct. Biol.* **80**, 362–376 (2024).

22. Case, D. A. *et al.* Ambertools. *J. chemical information modeling* **63**, 6183–6191 (2023).

23. Li, P., Song, L. F. & Merz Jr, K. M. Parameterization of highly charged metal ions using the 12-6-4 lj-type nonbonded model in explicit water. *The J. Phys. Chem. B* **119**, 883–895 (2015).

24. Eastman, P. *et al.* Openmm 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS computational biology* **13**, e1005659 (2017).

25. Teufel, F. *et al.* Signalp 6.0 predicts all five types of signal peptides using protein language models. *Nat. biotechnology* **40**, 1023–1025 (2022).

26. Horton, P. *et al.* Wolf psort: protein localization predictor. *Nucleic acids research* **35**, W585–W587 (2007).

27. Sperschneider, J. & Dodds, P. N. Effectorp 3.0: prediction of apoplastic and cytoplasmic effectors in fungi and oomycetes. *Mol. plant-microbe interactions* **35**, 146–156 (2022).