SuperMetal: A Generative AI Framework for Rapid and Precise Metal Ion Location Prediction in Proteins

Anonymous Author(s) Affiliation Address email

Abstract

Metal ions serve as essential cofactors in numerous proteins, playing a critical 1 role in enzymatic activities and protein interactions. Given their importance, 2 accurately identifying metal-binding sites is fundamental to understanding their З biological functions, with significant implications for protein engineering and 4 drug discovery. To address this challenge, we present SuperMetal, a generative 5 AI framework that combines a score-based diffusion model, confidence model, 6 and clustering mechanism to predict metal-binding sites with high accuracy and 7 efficiency. Using zinc ions as an example, SuperMetal outperforms existing state-8 of-the-art tools, achieving a precision of 94% and coverage of 90%, with zinc 9 ions localization within 0.52 ± 0.55 Å of experimentally determined positions. 10 Furthermore, SuperMetal delivers rapid predictions and is minimally affected by 11 increases in protein size. Notably, SuperMetal predicts metal-binding locations 12 without needing prior knowledge of ions numbers, unlike AlphaFold3, which 13 requires this information for its predictions. While currently trained exclusively 14 on zinc ions, SuperMetal's framework can be easily adapted to predict the binding 15 sites of other metal ions by adjusting the training data. 16

17 **1 Introduction**

The Protein Data Bank (PDB) contains nearly 200,000 structures, and approximately one-third of
these proteins contain metal ions [1]. Many proteins require the binding of one or more metal ions
to perform their functions. Zinc, a vital biologically active metal, is particularly noteworthy as it
binds to approximately 10% of all human proteins [2]. These proteins rely on zinc for their biological
function, structural stability, or regulation of activities. [3, 4, 5, 6, 4].

Given the importance and unique functionality of zinc in proteins, accurately identifying zinc-binding 23 24 sites is crucial. Consequently, Many computational methods have been developed to predict zincbinding sites [7, 8, 9, 10, 11]. Current state of the art predictors for metal location is Metal3D 25 [11], a structure-based method that employs 3D Convolutional Neural Networks (CNNs) to predict 26 the positions of metal ions, such as zinc. Despite its success, Metal3D faces challenges similar to 27 other 3D CNN models [12, 13, 14, 15], such as the need for fine grid spacing (voxelization). The 28 computational cost for these voxel-based models increases cubically with the resolution of the input, 29 making scaling up difficult [16]. Moreover, these CNN-based models are sensitive to the orientation 30 of the input structure, requiring data augmentation to increase the number of training samples and 31 reduce the risk of overfitting [11, 17, 18, 19]. 32

In recent years, diffusion models have emerged as powerful generative AI tools [20, 21], leading to significant advancements across various areas of bioinformatics. [22, 23, 24, 25, 26, 27, 28]. Inspired by these advancements, we present SuperMetal, a novel generative AI approach that integrates a score-based diffusion model, equivariant graph neural networks, and a clustering mechanism to

accurately predict zinc ion positions within protein structures. Instead of directly approximating the 37 probability distribution of zinc ions, our model estimates the gradient of this distribution and generates 38 zinc positions from a normal distribution. These positions are then refined using a confidence model 39 and clustered to deliver precise predictions of both the number and locations of zinc ions in a protein. 40 SuperMetal surpasses existing methods in terms of coverage and precision, while providing rapid 41 predictions, offering significant potential for applications in structural biology, multi-body docking, 42 and metalloprotein engineering. 43

2 **Methods** 44

The SuperMetal framework operates in three general stages as shown in Fig. 1. The detailed method-45 ology for each of these steps is provided in the supplemental information. The key contributions of 46 SuperMetal include (1) a score-based diffusion model that processes geometric graphs of protein 47 structures, enabling the sampling of metal positions within proteins, (2) an equivariant graph neural 48 network that accurately evaluates each sampled point and filters out low-confidence positions, and 49 (3) postprocessing operations designed to optimize prediction accuracy by clustering the predicted 50 51 positions.



Figure 1: Workflow of SuperMetal.

2.1 **Data Set and Preprocessing** 52

We utilized the ZincBind database [29], a high-quality, non-redundant collection of 19,154 zinc-53 binding sites from 19,103 PDB files. ZincBind clusters sites based on structural and sequence 54 similarity, accounting for protein symmetry to avoid mislabeling surface zincs [30]. From this, we 55 56 extracted 10,253 PDB files, excluding structures with over 3000 residues and removing exogenous 57 ligands. For structures with multiple models, only the first was used. We randomly selected 1,000 structures for validation and 350 for testing, ensuring no binding site overlap between testing and 58 training/validation datasets. 59

2.2 Evaluation and Comparison 60

We evaluate predicted metal ion positions using precision, coverage, and mean absolute deviation 61 (MAD). Precision is the ratio of true positives (TP) to total predicted sites (TP + FP): Precision = $(1 + 1)^{1/2}$ 62 $\frac{TP}{TP+FP}$. Coverage measures the percentage of correctly predicted sites relative to all true sites (TP + 63 FN): Coverage = $\frac{TP}{TP+FN}$. A site is correctly predicted if it is within 5 Å of the experimental position. 64 MAD quantifies positional accuracy as the average absolute difference between predicted $\hat{\mathbf{x}}_i$ and true 65 positions \mathbf{x}_i : MAD = $\frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|$, where *n* is the number of sites. These metrics provide a 66 comprehensive evaluation of SuperMetal compared to other methods. 67

Forward and Reverse Diffusion 2.3 68

As illustrated in Fig. 2, the forward step of the diffusion process is governed by a forward stochastic 69 70 differential equation (SDE), described as:

$$d\mathbf{x} = f(\mathbf{x}, t) \, dt + g(t) \, d\mathbf{w},\tag{1}$$

where \mathbf{x} represents the positions of all metal ions, t denotes time, \mathbf{w} refers to Gaussian noise 72

or Brownian motion, g(t) is the diffusion coefficient, and $f(\mathbf{x},t)$ is the drift coefficient. In our 73

case, $f(\mathbf{x}, t) = 0$, and g(t) is given by $\sqrt{\frac{d\sigma^2(t)}{dt}}$. The variance $\sigma^2(t)$ evolves based on a model hyperparameter, where σ_t is expressed as $\sigma_{\min}^{(1-t)} \cdot \sigma_{\max}^t$, leading to the forward SDE $d\mathbf{x} = \sqrt{\frac{d\sigma^2(t)}{dt}} d\mathbf{w}$. 74

75

For a given protein-metal complex, as Gaussian noise $\mathbf{w}(t)$ is added, and starting from the initial 76 metal ion distribution at t = 0, denoted $\mathbf{x}(0)$, the metal position at time t, $\mathbf{x}(t)$, can be numerically 77 determined using this equation. The translation perturbation vectors, $\Delta \mathbf{r}$, also follow a Gaussian 78 distribution with mean $\mu(t)$ and variance $\sigma^2(t)$, which allows us to compute the gradient of the log 79 probability of metal translations over the protein structure y using the equation $\nabla \log p_t(\Delta \mathbf{r} \mid \mathbf{y}) =$ 80 $-\frac{\Delta \mathbf{r} - \mu(t)}{\sigma^2(t)}$. Meanwhile, the score function $S_{\theta}(\mathbf{x})$ is predicted by a neural network, which takes the 81 metal ion locations, the protein structure, and the time t as inputs. The parameters of the neural 82 network, θ , are optimized by minimizing the loss function L_{θ} : 83

$$L_{\theta} = \mathbb{E}_{p(\mathbf{x})} \left[\|\nabla \log p_t(\Delta \mathbf{r} \mid \mathbf{y}) - S_{\theta}(\mathbf{x}, \mathbf{y}, t)\|_2^2 \right]$$
(2)

85 The expectation value $\mathbb{E}_{p(\mathbf{x})}$ is calculated by averaging the L_2 -norm between the true vectors

⁸⁶ $\nabla \log p_t(\Delta \mathbf{r} \mid \mathbf{y})$ and the predicted vectors $S_{\theta}(\mathbf{x}, \mathbf{y}, t)$ across all metal ions for each protein in

87 the training data.



Figure 2: Fundamental theory of the score-based generative diffusion model for metal ions in proteins.

88 Now, for reverse diffusion, we use the trained score function $S_{\theta}(\mathbf{x})$ to solve the reverse stochastic

89 differential equation (SDE) and compute the favorable positions of metal ions from a random

90 distribution:

$$d\mathbf{x} = \left[f(\mathbf{x}, t) - g(t)^2 S_{\theta}(\mathbf{x}, \mathbf{y}, t)\right] dt + g(t) d\mathbf{w},\tag{3}$$

where $S_{\theta}(\mathbf{x}, t)$ is the learned score function from the training phase, and $f(\mathbf{x}, t)$ and g(t) represent the drift and diffusion terms from the forward SDE. For each protein structure, we run the diffusion model's inference procedure to generate 100 candidate metal ion positions. These ions are guided by

the score model $S_{\theta}(\mathbf{x})$, ultimately reaching their most favorable positions within the protein structure.

96 **3 Results and Discussion**

In this section, we present the evaluation and comparative results of SuperMetal against the state-ofthe-art Metal3D in predicting zinc ion positions within protein structures. Both methods utilize the same testing dataset to ensure a fair comparison. As shown in Fig. S3, during SuperMetal's inference step, 100 metal ions are sampled at random positions across the system and denoised via reverse diffusion over their translational degrees of freedom. The sampled positions are then filtered using the trained confidence model to retain only the most probable metal positions. Finally, these positions are clustered to obtain the final predicted metal positions.

104 3.1 Comparison between SuperMetal and Metal3D

Figure 3 shows the precision versus coverage curves for SuperMetal and Metal3D based on varying probability thresholds. SuperMetal demonstrates higher precision across a wider range of coverage compared to Metal3D. For example, when Metal3D achieves 100% precision, its coverage is around 30%, whereas SuperMetal reaches approximately 70% coverage at the same level of precision—more than double the coverage of Metal3D. Similarly, at 77% coverage, SuperMetal maintains near 100% accuracy, while Metal3D's accuracy drops to around 93%. Moreover, at 88% coverage, Metal3D's precision is $\sim 84\%$, whereas SuperMetal achieves $\sim 95\%$, marking a significant improvement.

112 These results clearly indicate that SuperMetal outperforms Metal3D, providing higher precision

even at greater coverage. This demonstrates SuperMetal's ability to maintain high precision while significantly expanding the scope of its predictions.



Figure 3: Precision vs. coverage for SuperMetal and Metal3D at different probability cutoffs. Labels beside the curves represent probability cutoffs, p for SuperMetal and t for Metal3D.

115 114 116 In addition to evaluating the precision and coverage of metal site predictions, assessing the spatial accuracy of the predicted positions is crucial. For each true positive (TP), we measured the mean absolute deviation (MAD) between the experimentally determined and predicted metal ion positions 117 (Fig. 4). At a probability threshold of p = 0.1, SuperMetal achieves a MAD of 0.61 ± 0.66 Å, which 118 improves to 0.44 \pm 0.58 Å as the threshold increases to p = 0.9. This trend demonstrates that higher 119 probability cutoffs lead to greater spatial precision, with the median MAD decreasing from 0.37 Å at 120 p = 0.1 to 0.23 Å at p = 0.999. The relatively small difference between these two median values 121 suggests that even low-confidence predictions are spatially accurate within the protein structure. In 122 contrast, Metal3D exhibits an increasing median MAD, rising from 0.36 Å at t = 0.7 to 0.87 Å at 123 t = 0.99, indicating a greater deviation from ground-truth positions as the probability cutoff increases. 124 Additionally, the spread of MAD values in *SuperMetal* decreases with higher probability cutoffs, 125 opposite to the trend observed in Metal3D, where the spread increases. These results indicate that 126 SuperMetal consistently provides spatially precise predictions across different probability thresholds, 127 and its improved prediction accuracy is accompanied by enhanced spatial precision. 128

SuperMetal not only outperforms Metal3D in terms of location prediction accuracy but also demon-129 strates a significant advantage in running speed. The inference runtime comparison as a function of 130 protein size (number of residues) for SuperMetal and Metal3D is shown in Fig. 5. For consistent 131 comparison, both models were executed using a single thread on one CPU core, with the same 132 GPU. We observed that Metal3D's runtime tends to increase exponentially as the protein size grows, 133 whereas SuperMetal maintains consistently low runtimes (under 10 seconds), even for larger proteins. 134 For example, when the protein size approaches 2000 residues, Metal3D requires approximately 500 135 seconds, which is around 60 times longer than SuperMetal. This large difference can be attributed to 136 the multi-scale approach used in SuperMetal. The graph and message passing between metal ions 137 and protein residues are constructed only when the residues fall within a certain radius of the metal 138 ions. Similarly, the message passing between metal ions and protein atoms is established only when 139 protein atoms are within an even smaller radius of the metal ions. This radius-based mechanism 140 ensures that only relevant metal-protein interactions are computed, thus optimizing SuperMetal's 141 efficiency. In contrast, Metal3D involves voxelization of the entire protein and grid averaging, which 142 results in significantly longer runtimes, especially as the number of protein residues increases. 143



Figure 4: MAD distribution for SuperMetal and Metal3D across various probability cutoffs. Kernel density estimation was used to illustrate distribution, highlighting medians (white circles), quartiles (black boxes), and data spread (whiskers up to 1.5x the interquartile range).



Figure 5: Computational runtime vs. protein sizes for superMetal and Metal3D. Polynomial regression curves (purple and green dashed lines) are only used to clarify the trends.

144 4 Conclusions

In this work, we introduced SuperMetal, a generative AI framework designed to predict metal ion positions within protein structures. Leveraging a score-based diffusion model, an equivariant graph neural network, and a clustering mechanism, SuperMetal achieves both high accuracy and efficiency in metal-binding site prediction. When compared to the state-of-the-art Metal3D, SuperMetal consistently outperformed across key metrics such as precision, recall, and MAD. It nearly doubled the coverage at 100% precision, maintained lower MAD values, and efficiently scaled to handle larger protein sizes.

Notably, SuperMetal does not require prior knowledge of the number of metal ions, providing greater
 flexibility than methods like AlphaFold3. Its ability to predict metal-binding locations with high
 accuracy, speed, and scalability paves the way for future advancements in metalloprotein research.

155 **References**

- [1] Nanjiang Shu, Tuping Zhou, and Sven Hovmöller. Prediction of zinc-binding sites in proteins
 from sequence. *Bioinformatics*, 24(6):775–782, 2008.
- [2] Claudia Andreini, Lucia Banci, Ivano Bertini, and Antonio Rosato. Counting the zinc-proteins
 encoded in the human genome. *Journal of proteome research*, 5(1):196–201, 2006.
- [3] Keith A McCall, Chih-chin Huang, and Carol A Fierke. Function and mechanism of zinc
 metalloenzymes. *The Journal of nutrition*, 130(5):1437S–1446S, 2000.
- [4] Christos T Chasapis, Ariadni C Loutsidou, Chara A Spiliopoulou, and Maria E Stefanidou. Zinc
 and human health: an update. *Archives of toxicology*, 86:521–534, 2012.
- [5] Jeremy M Berg and Yigong Shi. The galvanization of biology: a growing appreciation for the
 roles of zinc. *Science*, 271(5252):1081–1085, 1996.
- [6] Maria Inês Costa, Ana Bela Sarmento-Ribeiro, and Ana Cristina Gonçalves. Zinc: from
 biological functions to therapeutic potential. *International Journal of Molecular Sciences*,
 24(5):4822, 2023.
- [7] Maarten L Hekkelman, Ida de Vries, Robbie P Joosten, and Anastassis Perrakis. Alphafill:
 enriching alphafold models with ligands and cofactors. *Nature Methods*, 20(2):205–213, 2023.

[8] Yu-Feng Lin, Chih-Wen Cheng, Chung-Shiuan Shih, Jenn-Kang Hwang, Chin-Sheng Yu, and
 Chih-Hao Lu. Mib: metal ion-binding site prediction and docking server. *Journal of chemical information and modeling*, 56(12):2287–2291, 2016.

- [9] Aditi Shenoy, Yogesh Kalakoti, Durai Sundar, and Arne Elofsson. M-ionic: prediction of metal ion-binding sites from sequence using residue embeddings. *Bioinformatics*, 40(1):btad782, 2024.
- I0] José-Emilio Sánchez-Aparicio, Laura Tiessler-Sala, Lorea Velasco-Carneros, Lorena Roldán Martín, Giuseppe Sciortino, and Jean-Didier Maréchal. Biometall: identifying metal-binding
 sites in proteins from backbone preorganization. *Journal of Chemical Information and Modeling*,
 61(1):311–323, 2020.
- [11] Simon L Dürr, Andrea Levy, and Ursula Rothlisberger. Metal3d: a general deep learning
 framework for accurate metal ion location prediction in proteins. *Nature Communications*,
 14(1):2713, 2023.
- [12] Andreas Zamanos, George Ioannakis, and Ioannis Z Emiris. Hydraprot: A new deep learning
 tool for fast and accurate prediction of water molecule positions for protein structures. *Journal of Chemical Information and Modeling*, 64(7):2594–2611, 2024.
- [13] Sangwoo Park and Chaok Seok. Galaxywater-cnn: Prediction of water positions on the protein
 structure by a 3d-convolutional neural network. *Journal of Chemical Information and Modeling*,
 62(13):3157–3168, 2022.
- [14] Ahmadreza Ghanbarpour, Amr H Mahmoud, and Markus A Lill. Instantaneous generation of
 protein hydration properties from static structures. *Communications Chemistry*, 3(1):188, 2020.
- [15] Kochi Sato, Mao Oide, and Masayoshi Nakasako. Prediction of hydrophilic and hydrophobic
 hydration structure of protein by neural network optimized using experimental data. *Scientific reports*, 13(1):2183, 2023.
- [16] Zhijian Liu, Haotian Tang, Yujun Lin, and Song Han. Point-voxel cnn for efficient 3d deep
 learning. Advances in neural information processing systems, 32, 2019.
- [17] Cheng Wang, Ming Cheng, Ferdous Sohel, Mohammed Bennamoun, and Jonathan Li. Normal net: A voxel-based cnn for 3d object classification and retrieval. *Neurocomputing*, 323:139–147, 2019.

- [18] Yang Zhang, Wenbing Huang, Zhewei Wei, Ye Yuan, and Zhaohan Ding. Equipocket: an e
 (3)-equivariant geometric graph neural network for ligand binding site prediction. *arXiv preprint arXiv:2302.12177*, 2023.
- [19] Sangwoo Park. Water position prediction with se (3)-graph neural network. *bioRxiv*, pages
 204 2024–03, 2024.
- [20] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and
 Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.
- [21] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data
 distribution. *Advances in neural information processing systems*, 32, 2019.
- [22] Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E
 Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo
 design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.
- [23] Emiel Hoogeboom, Victor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant
 diffusion for molecule generation in 3d. In *International conference on machine learning*, pages
 8867–8887. PMLR, 2022.
- [24] Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. Diffdock:
 Diffusion steps, twists, and turns for molecular docking. *arXiv preprint arXiv:2210.01776*, 2022.
- [25] Mohamed Amine Ketata, Cedrik Laue, Ruslan Mammadov, Hannes Stärk, Menghua Wu,
 Gabriele Corso, Céline Marquet, Regina Barzilay, and Tommi S Jaakkola. Diffdock-pp: Rigid
 protein-protein docking with diffusion models. *arXiv preprint arXiv:2304.03889*, 2023.
- [26] Karsten Kreis, Tim Dockhorn, Zihao Li, and Ellen Zhong. Latent space diffusion models of cryo-em structures. *arXiv preprint arXiv:2211.14169*, 2022.
- [27] Dominik JE Waibel, Ernst Röell, Bastian Rieck, Raja Giryes, and Carsten Marr. A diffusion
 model predicts 3d shapes from 2d microscopy images. In 2023 IEEE 20th International
 Symposium on Biomedical Imaging (ISBI), pages 1–5. IEEE, 2023.
- [28] Zhiye Guo, Jian Liu, Yanli Wang, Mengrui Chen, Duolin Wang, Dong Xu, and Jianlin Cheng.
 Diffusion models in bioinformatics and computational biology. *Nature reviews bioengineering*, 2(2):136–154, 2024.
- [29] Sam M Ireland and Andrew CR Martin. Zincbind—the database of zinc binding sites. *Database*,
 2019:baz006, 2019.
- [30] Geng-Yu Lin, Yu-Cheng Su, Yen Lin Huang, and Kun-Yi Hsin. Mespeus: a database of metal coordination groups in proteins. *Nucleic Acids Research*, 52(D1):D483–D493, 2024.