Accurate prediction of transition metal ion location via deep learning

Simon L. Dürr[1], **Andrea Levy**[1], **Ursula Rothlisberger**[1]

[1]Laboratory of Computational Chemistry and Biochemistry, Institute of Chemical Sciences and Engineering, Swiss Federal Institute of Technology (EPFL) CH-1015 Lausanne, Switzerland

***For correspondence:**
ursula.roethlisberger@epfl.ch (UR)

Github:
lcbc-epfl/metal-site-prediction
Webapp:
hf.space/simonduerr/metal3d
Interactive manuscript:
lcbc-epfl.github.io/metal-site-prediction

Abstract

Metal ions are essential cofactors for many proteins. In fact, currently, about half of the structurally characterized proteins contain a metal ion. Metal ions play a crucial role for many applications such as enzyme design or design of protein-protein interactions because they are biologically abundant, tether to the protein using strong interactions, and have favorable catalytic properties e.g. as Lewis acid. Computational design of metalloproteins is however hampered by the complex electronic structure of many biologically relevant metals such as zinc that can often not be accurately described using a classical force field. In this work, we develop two tools - Metal3D (based on 3D convolutional neural networks) and Metal1D (solely based on geometric criteria) to improve the identification and localization of zinc and other metal ions in experimental and computationally predicted protein structures. Comparison with other currently available tools shows that Metal3D is the most accurate metal ion location predictor to date outperforming geometric predictors including Metal1D by a wide margin using a single structure as input. Metal3D outputs a confidence metric for each predicted site and works on proteins with few homologes in the protein data bank. The predicted metal ion locations for Metal3D are within 0.70 ± 0.64 Å of the experimental locations with half of the sites below 0.5 Å. Metal3D predicts a global metal density that can be used for annotation of structures predicted using e.g. AlphaFold2 and a per residue metal density that can be used in protein design workflows for the location of suitable metal binding sites and rotamer sampling to create novel metalloproteins. Metal3D is available as easy to use webapp, notebook or commandline interface.

Introduction

Metalloproteins are ubiquitous in nature and are present in all major enzyme families.[1,2]The metals predominantly found in biological systems are the first and second row alkali and earth alkali metals and the first row transition metals such as zinc and copper. Zinc is the most common transition metal (present in ~10% of deposited structures) and can fulfill both a structural (e.g. in zinc finger proteins) or a catalytic role in up to trinuclear active sites. Zn^{2+} is an excellent Lewis acid and is most often found in tetrahedral, pentavalent, or octahedral coordination. About 10 % of all reactions catalyzed by enzymes use zinc as cofactor[3].

Metalloproteins are well studied because metal cofactors are essential for the function of many proteins and loss of this function is an important cause of diseases.[4] Industrial applications for metalloproteins capitalize on the favorable catalytic properties of the metal ion where the protein environment dictates (stereo)-selectivity.[5–7] To crystallize proteins, metal salts are also often added to the crystallization buffer as they can help in the formation of protein crystals overcoming the enthalpic cost of association of protein surfaces. Metal ion binding sites can be used to engineer protein-protein interactions (PPI)[10–12] and the hypothesis has been put forward that one origin of macromolecular complexity is the superficial binding of metal ions in early single domain proteins.[12]

While simple metal ion binding sites can be rapidly engineered because initial coordination on a protein surface can for example be achieved by creating an i, i+4 di-histidine site on an alphahelix[13] or by placing cysteines in spatial proximity,[14] the engineering of complex metal ion binding sites e.g. in the protein interior is considerably more difficult[2,11] as such sites are often supported by a network of hydrogen bonds. A complication for computational design of metalloproteins is the unavailability of good (non-bonded) force fields for zinc and other transition metals that accurately reproduce (e.g. tetrahedral) coordination with the correct coordination distances which renders design using e.g. Rosetta very difficult.[2,15] In fact, the latest parametrization of the Rosetta energy function (ref2015)[16] did not refit the parameters for the metal ions which originally are from CHARMM27 with empirically derived Lazaridis-Karplus solvation terms. To adequately treat metal sites in proteins quantum mechanical treatments such as in hybrid quantum mechanics/molecular mechanics (QM/MM) simulations[17,18] is needed whose computational cost is prohibitive for regular protein design tasks. QM/MM simulations can however be used to verify coordination chemistry for select candidate proteins.[19] On the other hand, neural network potentials have been developed for zinc however those require the experimental zinc location as input.[20]

Many tools exist to predict whether a protein contains metals (e.g. ZincFinder[21]), which residues in the protein bind a metal (e.g. IonCom,[22] MIB[23]) and where the metal is bound (AlphaFill,[24] FindsiteMetal,[25] BioMetAll,[26]MIB[23]). The input for these predictors is based on sequence and/or structure information. Sequence-based predictors use pattern recognition to identify the amino acids which might bind a metal.[27] Structure-based methods use homology to known structures (MIB, Findsite-metal, AlphaFill) or distance features (BioMetAll) to infer the location of metals. Some tools like Findsite-metal or ZincFinder employ machine learning based approaches such as support vector machines.

Structure based deep learning approaches have been used in the field of protein research for a variety of applications such as protein structure prediction,[28] prediction of identity of masked residues[30–32], functional site prediction,[33,34] for ranking of docking poses,[35,36] prediction of the location of ligands,[36–40] and prediction of effects of mutations for stability and disease.[4] Current state of the art predictors for metal location are MIB,[23,42] which combines structural and sequence information in the “Fragment Transformation Method” to search for homologous sites in its database, and BioMetAll,[26] a geometrical predictor based on backbone preorganization. Both methods have significant drawbacks: MIB excludes metal sites with less than 2 coordination partners from its analysis and is limited by the availability of templates in its database. BioMetAll does not use templates but provides many possible locations for putative binding sites on a regular grid. The individual probes in BioMetAll do not have a confidence metric therefore only allowing to rank sites by the number of probes found, which results in a large uncertainty in the position. Both tools suffer from many false positives. In this work, we present two metal ion location predictors that do not suffer from these drawbacks. The deep learning based Metal3D predictor operates on a voxelized representation of the protein environment and predicts a per residue metal density that can be averaged to get a smooth metal probability density over the whole protein. The distance based predictor Metal1D predicts the location of metals using distances mined from the protein data bank (PDB) directly predicting coordinates of the putative metal binding site. These tools pave the way to perform in silico design of metal ion binding sites without relying on predefined geometrical rules or expensive quantum mechanical calculations.

Results

A dataset of experimental high resolution crystal structures (2085 structures/252324 voxelized environments) containing zinc sites was used for training of the geometric predictor Metal1D and the deep learning predictor Metal3D (Figure 1). For training, we used the crystal environment including crystal contacts. For predictions, the biological assembly was used.

Metal3D

Metal3D takes a protein structure and a set of residues as input, voxelizes the environment around each of the residues and predicts the per residue metal density. The predicted per residue densities

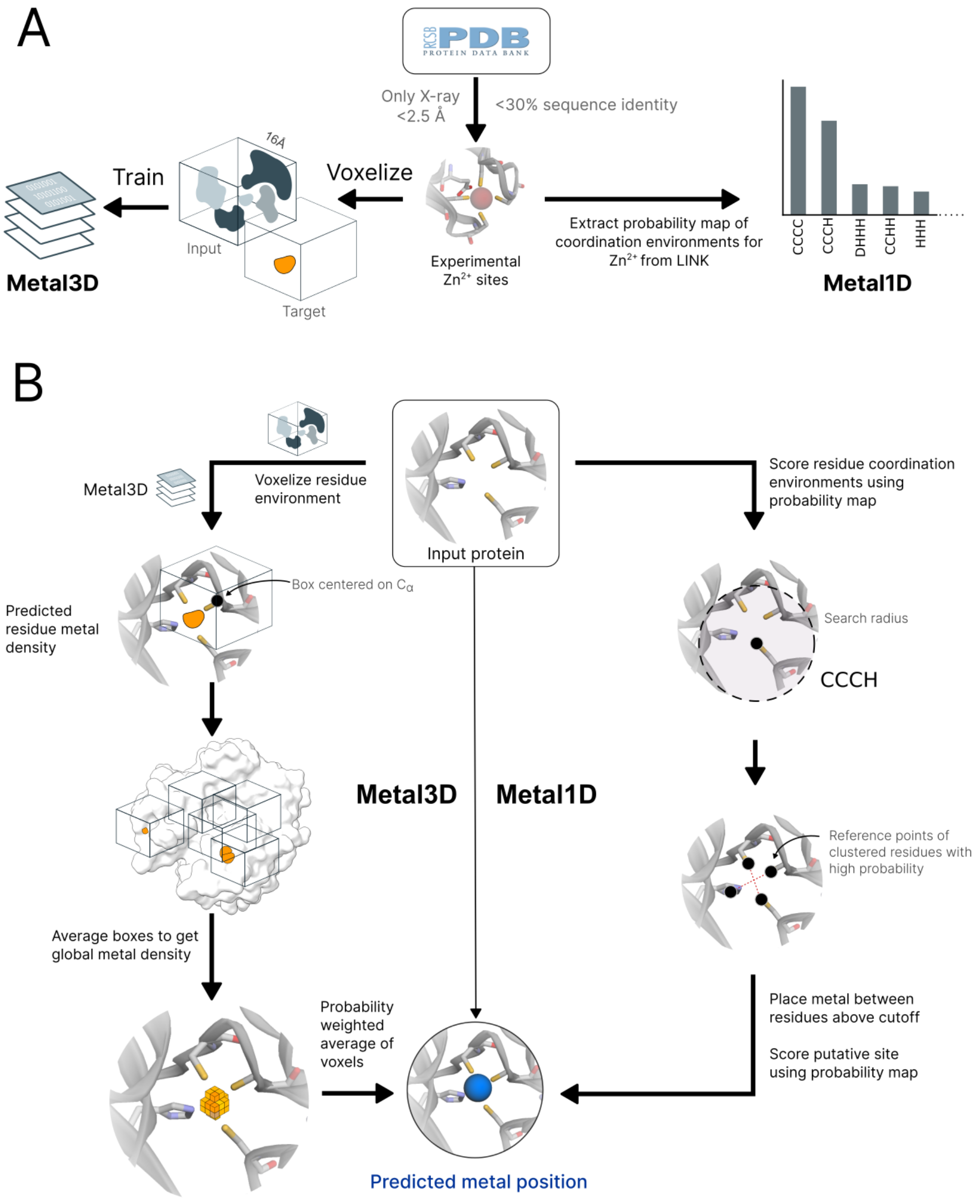

Figure 1. Workflow of Metal3D and Metal1D A Training of Metal3D and Metal1D is based on experimental Zn^{2+} sites. Metal1D extracts coordination environments from LINK records, Metal3D is a fully convolutional 3DCNN trained to predict the metal density from voxelized protein environments. **B** In inference mode Metal3D predicts the location of a metal ion by computing per residue metal densities and then averaging them to obtain a global metal density for the input proteins. The ions can then be placed using the weighted average of voxels above a cutoff. For Metal1D all residues in the protein are scanned for compatibility with the probability map. Metals are placed at the geometric center of residues with high scores according to the probability map. A final ranking of sites is obtained using the probability map.

(within a 16 x 16 x 16 Å ^{3} volume) can then be averaged to yield a zinc density for the whole protein. At high probability cutoffs the predicted metal densities are spherical (Figure 2 E), at low probability cutoffs the predicted densities are non-regular (Figure 2 A).

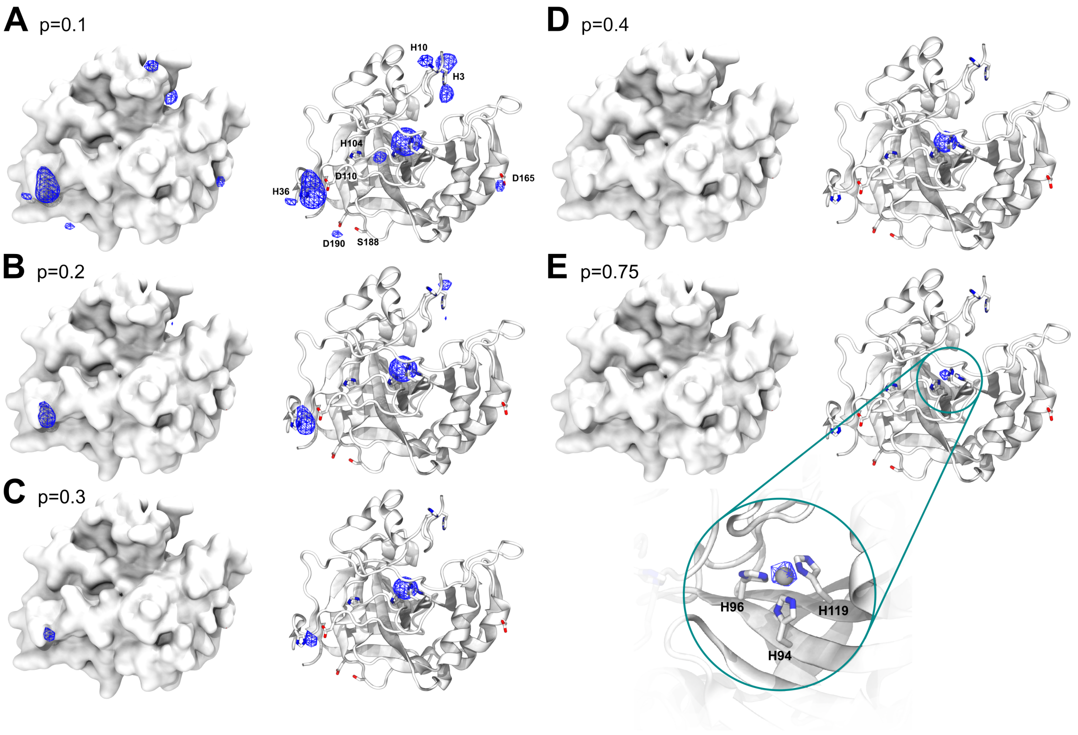

Figure 2. Metal3D probability density Probability evolution in HCA2 (PDB 2CBA) for different probability cutoffs A) p=0.1 B) p=0.2 C) p=0.3 D) p=0.4 E) p=0.75.

We evaluated the quality of the metal densities generated by the model with the discretized Jaccard similarity (Figure S1) for all environments in the test set. We noticed that at the edges of the residue-centered output densities often spurious density is predicted wherefore we evaluated the similarity of the test set metal density and the predicted metal probability density taking into account a smaller box with zeroed outer edges. Figure S1 shows that the similarity of the boxes does not depend much on the probability cutoff chosen with higher cutoffs yielding slightly higher discretized Jaccard similarity values (0.02 - 0.04 difference between p=0.5 and p=0.9). Reducing the size of the analyzed boxes (i.e trimming of the edges) increases the Jaccard similarity from ≈ 0.64 to 0.88 showing that the metal density in the center of the box is more accurate than the density at the edges.

Metal3D is available as self-contained notebook on Google Colab and on Huggingface Spaces.

Metal1D

The statistical analysis for the geometric predictor uses the `LINK` records present in deposited PDB structures. A probability map for all zinc coordination motifs was extracted from all training structures (Figure 1 A). The mean coordination distance in the training set was found to be 2.2 ± 0.2 Å, and the default search radius for the predictions was therefore set to 5.5 Å (Table S1). In total 208 different environments with more than 5 different proteins (at 30 % sequence identity) were identified. Metal1D is available as self-contained notebook on Google Colab

Comparison of Metal1D, Metal3D, MIB and BioMetAll

Existing metal ion predictors can be subdivided into two categories: binding site predictors and binding location predictors. The former identify only the residues binding the ion, the latter predict the coordinates of the metal ion itself. Both Metal1D and Metal3D can predict the coordinates of putative binding sites. We therefore assessed their performance by comparing to recent binding location predictors with available code/webserver: BioMetAll[26] and MIB.[23] The main tuning parameter of MIB is the template similarity t, with higher values requiring higher similarity of the templates available for the search in structurally homologous metalloproteins. BioMetAll on the other hand was calibrated on available protein structures and places probes on a regular grid at all sites where the criteria for metal binding are fulfilled. The main adjustable parameter for BioMetAll is the cluster cutoff c, which indicates how many probes in reference to the largest cluster a specific cluster has. We used the recommended cutoff of 0.5 requiring all chosen clusters to have at least 50% of the probes of the most populous cluster and used the cluster center to compute distances.

We first investigated the potential of all tools to detect the location of a zinc ion binding site in a binary fashion (zinc site or no zinc site). We defined a correctly identified binding site (true positive, TP) as a prediction within 5 Å of an experimental zinc site. In case a tool predicted no metal within the 5 Å radius, we counted this site as false negative (FN). False positive (FP) predictions, i.e sites where a metal was placed spuriously, were clustered in a 5 Å radius and counted once per cluster. All tools were assessed against the held out test biological assemblies for Metal3D and Metal1D. When the performance of MIB (t=1.25) and BioMetAll is compared against Metal3D with probability cutoff p=0.75 we find that Metal3D identifies more sites (85) than MIB (78) or BioMetAll (75) with a much lower number of false positives (Figure 3). MIB predicts 180 false positive sites, BioMetAll 134 sites whereas Metal3D only predicts 9 false positive sites at the p=0.75 cutoff. Metal1D (t=0.5) offers similar detection capabilities (78 sites detected) with a lower number of false positives (47) compared to MIB and BioMetAll. We removed 56 sites from the list of zinc sites in the test set (189 total) that had less than 2 unique protein ligands within 2.8 Å of the experimental zinc location. The amount of correct predictions in this reduced set is almost unchanged for all tools (Figure 3) indicating that most tools correctly predict sites if they have 2 or more protein ligands. For Metal3D at p=0.75 and p=0.9 as well as MIB with t=1.9 all sites that are correctly predicted to contain a metal are sites with more than 2 protein ligands. The number of false negatives is reduced for all tools by about 50 sites indicating that most tools do not predict these crystallographic artifacts that might depend on additional coordinating residues from an adjacent molecule in the crystal. Of all tools, Metal3D has the least false positives (1 FP at p=0.9) and the highest number of detected sites (110 at p=0.25). The single false positive at p=0.9 does not contain a zinc ion but is a calcium binding site with three aspartates and one backbone carbonyl ligand (Figure S8).

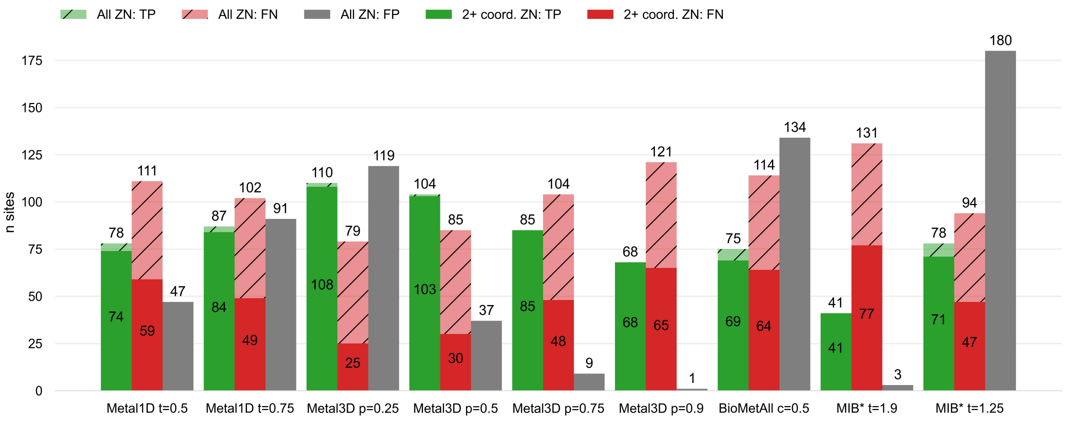

Figure 3. Identification of metal sites Comparison of Metal1D, Metal3D, BioMetAll and MIB on the test set held out from training of Metal1D and Metal3D. Predicted sites are counted as true positives (TP) if they are within 5 Å of a true metal location and as false negatives (FN) otherwise. False positive (FP) probes are clustered and counted once per cluster. *For MIB we used 2 structures less because the server did not accept these structures.

After assessment of how many sites the tools predict, another crucial metric is the spatial precision of the predictions. For the correctly identified sites (TP) we measured the mean absolute distance (MAD) between experimental and predicted position (Figure 4). The MAD for Metal3D at p=0.9 is 0.70 ± 0.64 Å and 0.74 ± 0.66 Å at p=0.25 indicating that low confidence predictions are still accurately placed inside the protein. The median MAD of predictions for Metal3D at p=0.9 is 0.52 Å indicating that for half of the predictions the model predicts at or better than the grid resolution of 0.5 Å.

BioMetAll is not very precise with a MAD for correctly identified sites of 2.80 ± 1.30 Å. BioMetAll predicts many possible locations per cluster with some of them much closer to the experimental metal binding site than the cluster center. However, it does not provide any ranking of the probes within a cluster and therefore the cluster center was used for the distance calculation. Metal1D (MAD 2.06 ± 1.33 Å) which identifies more sites than BioMetAll is also more precise than BioMetAll. MIB t=1.9 detects sites with high precision (MAD 0.77 ± 1.09 Å) but it relies on the existence of homologous sites to align the found sites.

Selectivity for other metals

Both Metal3D and Metal1D were exclusively trained on zinc and we assessed their performance on sodium (Na^+, PDB code NA), potassium (K^+, PDB code K), calcium (Ca^{2+}, PDB code CA), magnesium (Mg^{2+}, PDB code MG), and various transition metals (Fe^{2+}, Fe^{3+}, Co^{2+}, Cu^{2+}, Cu^+, Mn^{2+} with corresponding PDB codes FE2, FE, CO, CU, CU1, MN, respectively) from 25 randomly drawn structures from the clustered PDB at 30% identity. Only sites with at least 3 unique protein ligands were used for the analysis to exclude crystallographic artifacts and use only highly defined sites which should exhibit most selectivity towards a specific metal. Figure 4 B shows that recall for Metal3D is high for all transition metals, meaning that the model correctly finds most sites even though it was only trained on zinc. For the alkali and earth alkali metals recall is much lower as the model only finds some sites. The mean probability for found zinc structures (ZN p=0.95 ± 0.10) in the test set is higher than for the other transition metals (Figure S6) and significantly higher than the probability for alkali metals (NA p=0.61 ± 0.10, K p=0.79 ± 0.16) while the probability for the earth alkali metals is slightly higher with MG (p=0.77 ± 0.16) similar to CA (p=0.73 ± 0.16). The MAD for each found metal site is again lowest for zinc (0.56 ± 0.59 Å). The MAD for the found sodium (n=2) and potassium (n=5) sites are as low as for the other transition metals. The only metal with significantly higher MAD (1.45 ± 0.93 Å) is CU1 (Figure S5).

The only two structures where a sodium is detected by Metal3D (2OKQ[44], 6KFN[45]) have at least 2 side chain coordinating ligand atoms and only one backbone (2OKQ) or no backbone ligand atom (6KFN). Canonical sodium binding sites e.g. such as in PDB 4I0W[46] with two coordinating backbone carbonyl oxygen atoms and one asparagine side chain have probabilities around 5 % and are basically indistinguishable from background noise of the model. For Metal1D overall recall is lower with similar differences in the detection of main group metals versus transition metals (Figures S3 and S4).

Applications

After having evaluated the accuracy of Metal3D on held out test structures we also investigated possible uses in downstream applications such as protein function annotation and protein design.

Alpha Fold

AlphaFold2 often predicts side chains in metal ion binding sites in the holo conformation.[28] Tools like AlphaFill[24]use structural homology to transplant metals from similar PDB structures to the predicted structure. Metal3D does not require explicit homology based on sequence or structural alignment like AlphaFill so it is potentially suited to annotate the dark proteome that is now accessible from the AlphaFold database with metal binding sites. Metal3D identifies both the catalytic site (1) and the zinc finger (2) for the example (PDB 3RZV[47] , Figure 5 A) used in ref[24] with high probability (p=0.99)

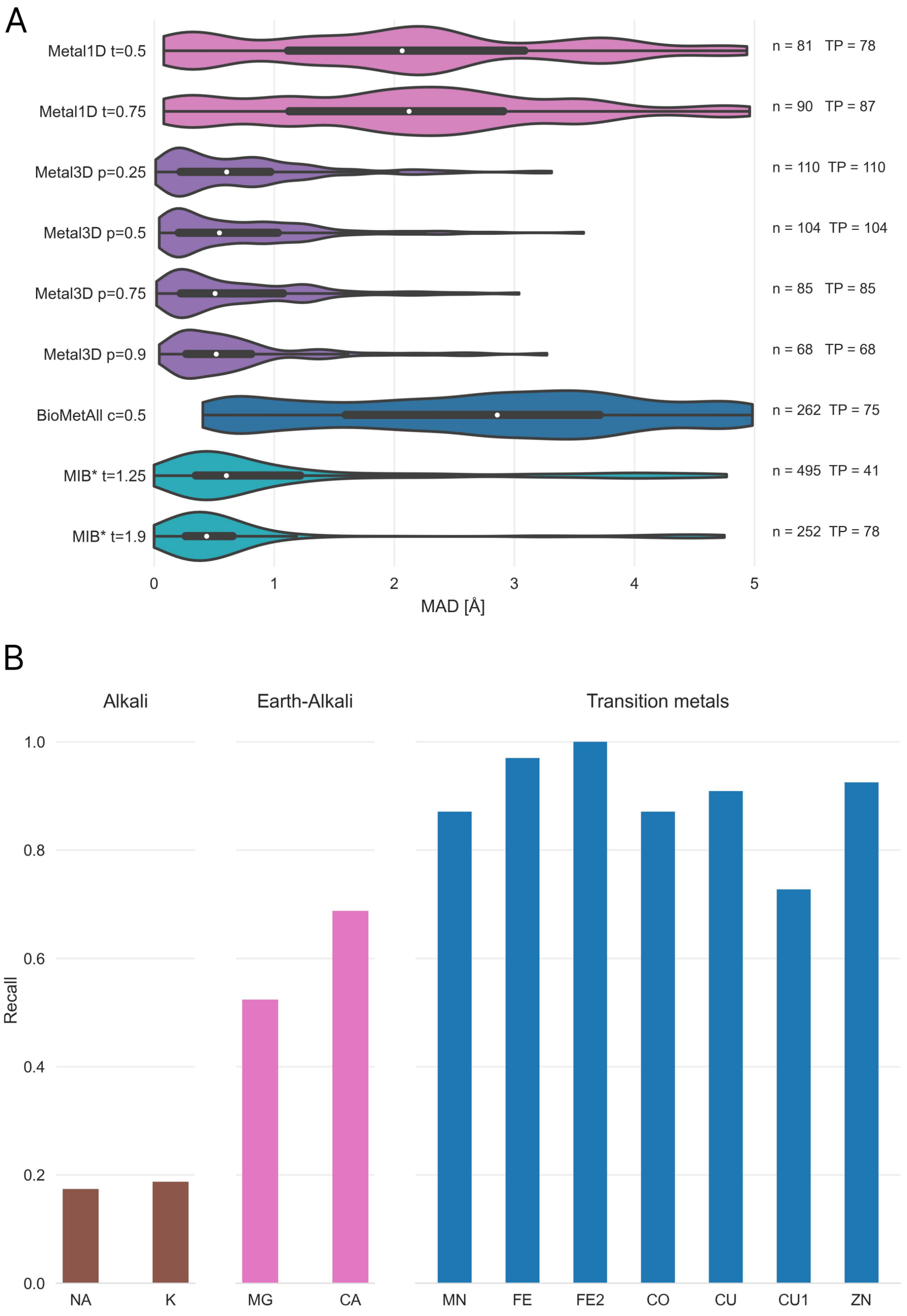

Figure 4. Precision of predicted sites and selectivity for Zn^{2+} **A** Mean absolute deviation (MAD) of predicted zinc ion locations using Metal1D, Metal3D, BioMetAll and MIB on the test set used to train Metal1D and Metal3D for all correctly identified (TP) sites. n is the number of sites predicted by the tool. *For MIB we used 2 structures less because the server did not accept these structures. For each tool the whisker plot indicates the median (white dot) and the first quartiles (black box). **B** Recall for the zinc test set and 25 randomly drawn structures for other transition, alkali and earth-alkali metal ions for Metal3D using p=0.5 as cutoff.

even though one of the sites in the AlphaFold model is slightly disordered with one of the binding residues in the solvent facing conformation (D309). The distances between predicted and modeled metal locations for Metal3D are 0.22 Å and 0.37 Å, for AlphaFill they are 0.21 Å and 0.41 Å.

AlphaFill uses a 25% sequence identity cutoff which can be problematic for certain proteins with no structurally characterized homologues. For human palmitoyltransferase ZDHHC23 (Uniprot Q8IYP9) a high confidence AlphaFold2 prediction exists but AlphaFill cannot place the zinc ions because the sequence identity is 24% to the closest PDB structure (PDB 6BMS[48]), i.e below the 25% cutoff. For the identical site in another human palmitoyltransferase ZDHHC15 (Uniprot Q96MV8) AlphaFill is able to place the metal because of higher sequence identity to 6BMS (64%) (Figure 5 B). For ZDHHC23 Metal3D is able to place the metal with high confidence (MAD 0.75 Å for site 1 and 0.48 Å for site 2, p>0.99) based on the single input structure alone.

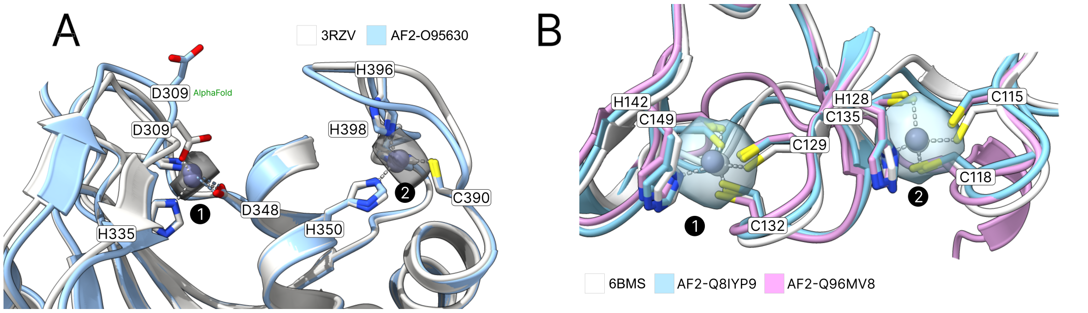

Figure 5. Annotation of AlphaFold2 structures A Predicted metal binding sites (a and b) from Metal3D, respectively AlphaFill compared to the experimentally found zinc positions for Uniprot O95630. Metal3D places the metal with high accuracy even if sidechains are not perfectly predicted by AlphaFold for site 1 **B** Palmitoyltransferase ZDHHC23 (Uniprot Q8IYP9) and ZDHHC15 (Uniprot Q96MV8). AlphaFill can only place the metal for ZDHHC15 because sequence identity for ZDHHC23 is only 24 %. Probability isosurfaces from Metal3D for both structures at p=0.6, colored in gray.

Metalloprotein engineering

Human carbonic anhydrase II (HCA2) is a well studied metalloenzyme with a rich amount of mutational data available. For the crystal structure of the wildtype enzyme (PDB 2CBA[49]), Metal3D recapitulates the location of the active site metal with a RMSD to the true metal location of 0.21 Å with a probability of p=0.99. At lower probability cutoffs (p<0.4) the probability map indicates further putative metal ion binding sites with interactions mediated by surface residues (e.g. H36, D110, p=0.22) (Figure 2).

To investigate the capabilities for protein engineering we used mutational data for first and second shell mutants of the active site residues in HCA2 with corresponding K_d values from a colorimetric assay.[50] For most mutants no crystal structures are available so we used the structure builder in the EVOLVE package to choose the most favorable rotamer for each single point mutation based on the EVOLVE-ddg energy function with explicit zinc present (modeled using a dummy atom approach[51]). The analysis was run for every single mutant and the resulting probability maps from Metal3D were analyzed. For the analysis we used the maximum predicted probability as a surrogate to estimate relative changes in K_d. For mutants that decrease zinc binding drastically we observe a drop in the maximum probability predicted by Metal3D (Figure 6).The lowest probability mutants are H119N and H119Q with p=0.23 and 0.38. The mutant with the largest loss in zinc affinity H94A has a zinc binding probability of p=0.6. Conservative changes to the primary coordination motif (e.g. H → C) reduce the predicted probability by 10 - 30 %. For second shell mutants the influence of the mutations is less drastic with only minor changes in the predicted probabilities.

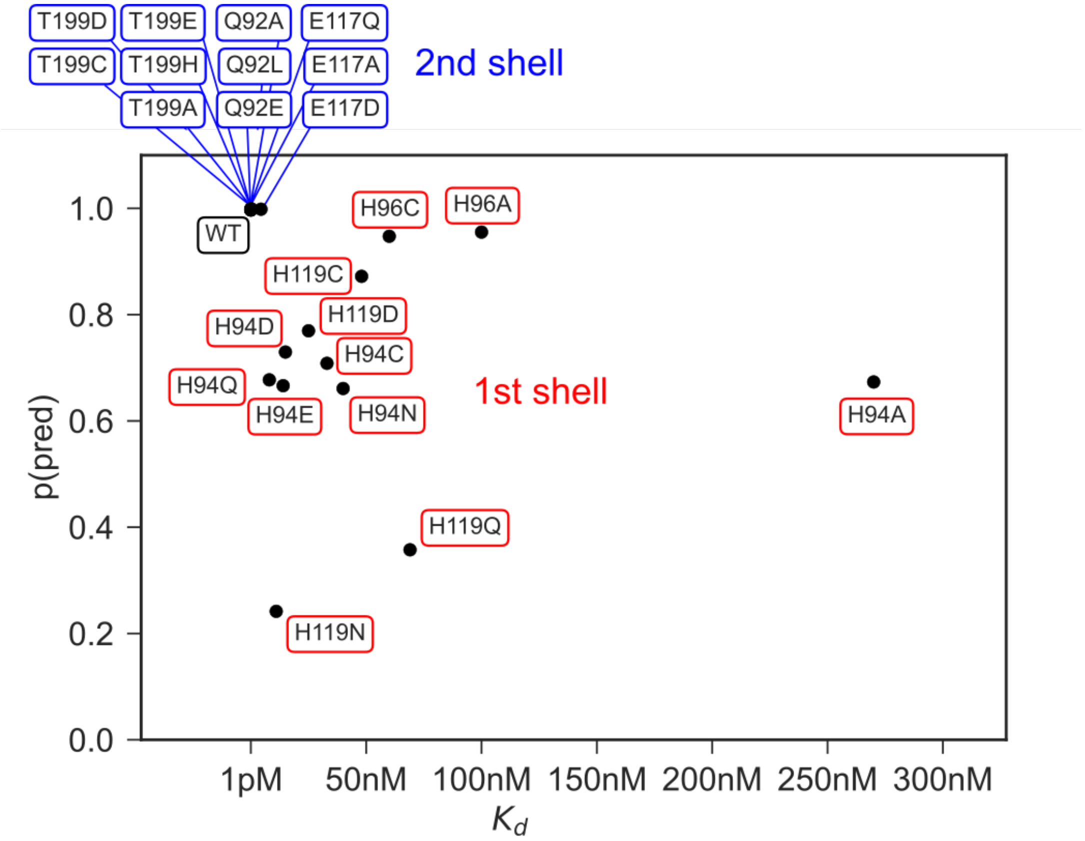

Figure 6. Protein design application Experimentally measured K_d values[52–56] for 1st and 2nd shell active site mutants of HCA2 and predicted max probability for zinc using Metal3D.

Discussion

Metal3D predicts the probability distribution of zinc ions in protein crystal structures based on a neuronal network model trained on natural protein environments. The model performs a segmentation task to determine if a specific point in the input space contains a zinc ion or not. Metal3D predicts zinc ion sites with high accuracy making use of high resolution crystal structures (<2.5 Å). The use of high resolution structures is necessary because at resolutions greater than the average zinc ligand coordination distance (2.2 Å) the uncertainty of the zinc location noticeably increases[43] which would likely hamper the accuracy of the site prediction.

In contrast to currently available tools, for Metal3D, it is not necessary to filter the training examples for certain coordination requirements (i.e only sites with at least 2 protein ligands). The model thus sees the whole diversity of zinc ion sites present in the PDB. Such a model is advantageous since metalloprotein design workflows require models to score the full continuum of zinc sites starting from a suboptimal binding site only populated at high metal concentration to a highly organized zinc site in an enzyme with nanomolar metal affinity. The predicted probability can be used as a confidence metric or as an optimization target where mutations are made to increase probability of zinc binding.

Site quality

The fraction of artifactual zinc binding sites in the PDB is estimated to be about 1/3[43] similar to our test set used with 70% (133) well coordinated zinc sites with at least 2 distinct protein ligands. To reduce the amount of artifactual sites in the training set we presented the model with as many complete sites as possible by using crystal symmetry to add adjacent coordinating protein chains. The frequency of artifacts in the training set is therefore much lower than 30 %. The sites which still remain incomplete or that are wrongly modeled and not excluded through the resolution cutoffs and filtering procedures likely present only a small fraction of the training set and their signal is drowned out by the numerical superiority of the correctly modeled sites. If the model is used on artifactual sites or partially disordered ones it can still predict the metal location with high spatial accuracy but often indicates a lower confidence for the prediction (Figures 2 and 5).

Metal ion locators that rely on homology such as MIB perform worse on partial binding sites because reducing the quality of the available templates by including 1- or 2-coordinate sites would yield many false positives (similar to including less homologous structures for the template search). The deep learning based Metal3D can likely circumvent this because it does not require any engineered features to predict the location of the metal and learns directly from a full representation of the environment surrounding the binding site. This allows looking at low confidence sites in the context of a given environment.

Influence of non-protein ligands

Exogenous ligands play an important role for metals in biology as all empty coordination sites of metals are filled with water molecules in case there is no other exogenous ligand with higher affinity present (e.g. a thiol). Like other predictors, both Metal1D and Metal3D do not consider water molecules or other ligands in the input as the quality of ligand molecules in the PDB varies.[40,57] In addition, other potential sources of input such as AlphaFold do not provide explicit waters wherefore models should not rely on water as an input source. It is also not possible to use in silico water predictions because common water placement algorithms to place deep waters[40,58,59] either rely on metal ions being present in the input or ignore them completely. Moreover, in protein design algorithms, water is usually only implicitly modeled (e.g. in Rosetta).

For Metal3D, the input channel that encodes the total heavy atom density also encodes an implicit water density where all empty space can be interpreted as the solvent. For Metal1D, the contribution of water molecules is considered in an implicit way when the score is assigned to a site by considering coordinations including water compatible with the one observed (e.g. a $\mathrm{HIS_3Wat}$ site is equivalent to a $\mathrm{His_3}$ site for the scoring).

Choice of architecture

This work is the first to report a modern deep learning based model destined for identification of metal ligands in proteins. Similar approaches have been used in the more general field of protein-ligand docking where a variety of architectures and representations have been used. 3D CNN based approaches such as LigVoxel[38] and DeepSite[37] commonly use a resolution of 1 Å and similar input features as our model to predict the ligand density. However, predicting the density of a multi-atomic ligand is more complex than predicting the density of mononuclear metal ions. We therefore did not deem it necessary to include a conditioning on how many metal ions are present in the box and rather chose to reflect this in the training data where the model needs to learn that only about half of the environments it sees contain one or more metals. This choice is validated by the fact that the output probability densities at sufficiently high probability cutoffs are spherical with their radius approximately matching the van der Waals radius of zinc.

Mesh convolutional neural networks trained on a protein surface representation[36] also have been used to predict the location and identity of protein ligands but this approach can only label the regions of the surface that bind the metal ion and is conceptually not able to return the exact location of the metal. Some metal ion binding sites are also heavily buried inside proteins as they mediate structural stability rendering them inaccessible to a surface based approach. The most recent approaches such as EquiBind[39] use equivariant neural networks such as En-Transformer[60] to predict binding keypoints (defined as 1/2 distance between the Cα of the binding residue and a ligand atom). Explicit side chains are still too expensive for such models and these models assume a fixed known stoichiometry of the protein and ligand. Metal3D can also deal with proteins that do not bind a metal and does not assume that the amount of ions is known. The lack of explicit side chain information renders equivariant models unsuitable for the design of complex metal ion binding sites supported by an intricate network of hydrogen bonds that need to be positioned with sub-angstrom accuracy. Our model in contrast is less data- and compute-efficient than approaches representing the protein as graph due to the need to voxelize the input and provide different rotations of the input environment in training but the overall processing time for our model is still low taking typically 25 seconds for a 250 residue protein on a multicore GPU workstation (20 CPUs, GTX2070). Sequence based models[61,62] can only use coevolution signals to infer residues in spatial proximity that can bind a metal. This might be difficult when it comes to ranking similar amino acids such as aspartate and glutamate or even ranking different rotamers where sub-angstrom level precision is needed to identify the mutant with the highest affinity for zinc.

Selectivity

In terms of selectivity both of our methods have a clear preference for transition metals over main group metals after having been trained exclusively on zinc binding sites. The only sites that Metal3D identifies for sodium in the test set are the ones that have side chain ligands. Many sodium and potassium sites are using backbone carbonyl coordination exclusively, which is not common for zinc and those sites are therefore not detected. Both of our methods could be rapidly adapted to predict not only location but also the identity of the metal. In the framework of Metal3D even a semantic metal prediction would be possible where the same model predicts different output channels for each metal it was trained on. To achieve perfect selectivity using such a model will be difficult because sometimes non-native metals are used for crystallization experiments. In this work we chose to work exclusively with zinc because it is the most redox stable transition metal and because many training examples are available.

Application for protein design

Protein design using 3DCNNs trained on residue identity has been successfully demonstrated and we anticipate that our model could be seamlessly integrated into such a workflow[32] to enable fully deep learning based design of metalloproteins. We are currently also investigating the combination of Metal3D combined with a classic energy-based genetic algorithm-based optimization to make design of metalloproteins[19] easier without having to explicitly model the metal to compute the stability of the protein. As the model computes a probability density per residue it can be readily integrated into established software like Rosetta relying on rotamer sampling.

The HCA2 application demonstrates the utility of Metal3D for protein engineering (Figure 2). The thermodynamics of metal ion binding to proteins are complicated[63] and there are currently no high-throughput based experimental approaches that could generate a dataset large enough to train a model directly on predicting K_d. The data we use were obtained from a colorimetric assay with very high affinity of zinc in the picomolar range.[52–56] More recent studies using ITC[63] instead of the colorimetric assay indicate lower K_d values in the nanomolar range for wild type HCA2. We can therefore only use the colorimetric data to estimate how well the model can recapitulate relative changes in the K_d for different mutations in the first and second shell of a prototypical metalloprotein.

Metal3D allows moving away from using rational approaches such as the i, i + 4 di His motifs used for the assembly and stabilization of metalloproteins to a fully automated approach where potential metal binding configurations can be scored computationally.[64–66]

Metal1D vs. Metal3D

Metal1D is inferior to Metal3D for the prediction of metal ion binding sites because it produces more false positives while at the same time detecting fewer metal sites. Also the positioning of sites is somewhat imprecise. This demonstrates the inherent limitation of using solely distance based features for prediction of metal location. BioMetAll which is the tool most similar to Metal1D also suffers from many false positive predictions. In contrast, Metal1D is more data-efficient than Metal3D and provides predictions faster.

Conclusion

We present two metal ion location predictors: Metal3D based on 3D convolutional neural networks and Metal1D based on distances and amino acid propensity maps. Metal3D is the first tool with sub-angstrom level precision to predict the location of metal ions in proteins that does not rely on searching for structurally homologous proteins in a database. We therefore anticipate different applications such as protein-function annotation of structures predicted using AlphaFold2,[67] integration in protein design software and detection of cryptic metal binding sites that can be used to engineer PPIs. Such cryptic metal ion binding sites in common drug targets could also be used to engineer novel metallodrugs. Many of these applications will allow us to explore the still vastly untapped potential of proteins as large multi-dentate metal ligands with programmable surfaces.

Materials and Methods

Dataset

The input PDB files for training were obtained from the RCSB[68] protein databank (downloaded 5th March 2021). We use a clustering of the structures at 30% sequence identity using mmseqs2[69] to largely remove sequence and structural redundancy in the input dataset. For each cluster, we check whether a zinc is contained in one of the structures, whether the resolution of these structures is better than 2.5 Å, if the experimental method is x-ray crystallography and whether the structure does not contain nucleic acids. If there are multiple structures fulfilling these criteria, the highest resolution structure is used. All structures larger than 3000 residues are discarded. We always use the first biological assembly to sample the training environments. The structures were stripped of all exogenous ligands except for zinc . If there are multiple models with e.g. alternative residue conformations for a given structure, the first one is used. For each biological assembly we used the symmetry of the asymmetric unit to generate a protein structure that contains all neighboring copies of the protein in the crystal such that metal sites at crystal contacts are fully coordinated.

The train/val/test split was performed based on sequence identity using `easy-search` in mmseqs2. All proteins that had no (partial) sequence overlap with any other protein in the dataset were put into the test/val set (85 proteins) which we further split into a test set of 59 structures and a validation set of 26 structures. The training set contained 2085 structures. (Supplemental Data 1).

For the analysis, we always used the biological assembly and not the symmetry augmented structure. For the specificity analysis with respect to other transition metals, clusters from the PDB were randomly sampled to extract 25 biological assemblies per metal.

By default all zinc sites in the test and validation set were used for the analysis. Since some of the sites might be affected by the crystallization conditions, we also created a subset of all sites that contained at least 2 amino acid ligands to largely exclude crystallization artifacts. To analyze metal ion selectivity, we selected sites with at least 3 unique protein ligands to only use biologically significant sites with a high degree of metal preorganization as such sites should exhibit more selectivity for specific metals compared to sites with only 2 unique protein-ligands.

Metal 1D

Metal1D uses a probability map derived from `LINK` records in protein structures (Figure 1). The `LINK` section of a PDB file specifies the connectivity between zinc (or any other ligand) and the amino acids of the protein, and each `LINK` record specifies one linkage. This is an extension of the approach by Barber-Zucker *et al.*,[70] in which `LINK` records were used to investigate the propensity of transition metals to bind different amino acids.

Using the training set we generated a probability map for the propensity of different coordination environments to bind a zinc (e.g CCCC, CCHH etc.). For each zinc ion the coordination is extracted from the`LINK` records excluding records involving only single amino acids (weak binding sites). Also, `LINK` records containing water molecules are excluded because of the difficulties in placing water molecules a posteriori in 3D structures when metal ions are present and because data quality of modelled water molecules varies. The probability map contains the counts of coordination environments found.

Making a prediction using Metal1D consists of two steps (Figure 1): Identification of possible metal coordinating residues in the structure via the scoring of each amino acid, and scoring of the likelihood of coordination for putative sites predicted by placing a metal between the identified coordinating residues.

The protein structure is analyzed using the BioPandas python library.[71] To identify coordinating residues, a per residue score is assigned by performing a geometrical search from a reference point, defined as the coordinate of the most probable metal binding atom, within a search radius considered as roughly twice the typical distance between the metal ion and the binding atom of amino acids in proteins (2.2 ± 0.2 Å as determined from `LINK` records). The search radius used was 5.5 Å in order to be able to take into account also deviations from the ideal coordination. In the case of amino acids which present more than one putative coordinating atom, such as e.g. histidine, the mid-point between the donor atoms is used as reference point and the search radius is enlarged accordingly. The atoms used as reference points for each amino acid and the increase in the search radius are reported in Supplemental Table S1 . The score is assigned to each amino acid considering all the other reference points of other amino acids within the search radius, and summing the probabilities in the probability map for coordinations compatible with the one observed. In the ideal case, a score of 1 corresponds to an amino acid surrounded by all possible coordinating amino acids observed in the probability map. In practice, scores result between 0 and < 1. Once all amino acids in the chain are scored, the metal location predictions are made grouping the highest-scored amino acids in clusters (defined as the ones within the chosen threshold with respect to the highest-scored one) based on distance. This is done using `scipy.spatial.distance_matrix` and grouping together highest-scored amino acids closer than twice the search radius. For each cluster, a site prediction is made as a weighted average between the coordinates of the reference point of each amino acid, using as weighting factor the amino acid score. For isolated amino acids with a high score (e.g. a single histidine) the same score is assigned to the closest reference point from another amino acid, to be able to compute the position of the metal as before. Possible artifacts resulting from this fictitious score are resolved in the final step of the prediction.

After the metal has been placed the likelihood of the putative sites can be assessed by performing a geometrical search centered on the predicted metal coordinates (within 60% of the search radius, i.e 3.3 Å) and a final score is now assigned to the site. The final score is assigned in the same way as the amino acid scores based on the probability map, and has the advantage of being able to sort the predicted metal sites based on their frequency in the training set. A cutoff parameter is used to exclude sites with a probability lower than a certain threshold with respect to the highest-scored one. This final scoring also mitigates the errors which can be introduced by calculating the coordinates of the site simply as a weighted average excluding or assigning a low probability to the site ending in unfavorable positions in space.

Metal 3D

Voxelization

We used the moleculekit python library[38,72] to voxelize the input structures into 3D grids. 8 different input channels are used: aromatic, hydrophobic, positive ionizable, negative ionizable, hbond donor, hbond acceptor, occupancy, and metal ion binding site chain (Supplemental Table S2). The channels are assigned using AutoDockVina atom names and a boolean mask. For each atom matching one of the categories a pair correlation function centered on the atom is used to assign the voxel value.[38] For the target tensor only the zinc ions were used for the voxelization. The target tensor was discretized setting any voxel above 0.05 to 1 (true location of zinc), all other to 0 (no zinc). We used a box size of 16 Å centered on the Cα atom of a residue, rotating each environment randomly for training before voxelization. The voxel grid used a 0.5 Å resolution for the input and target tensors. Any alternative side chain conformations modeled were discarded keeping only the highest occupancy. For the voxelization only heavy atoms were used. For all structures selected for the respective sets we partitioned the residues of the protein into residues within 12 Å of a zinc ion and those further away (based on the distance to the Cα atom). A single zinc site will therefore be present many times in the dataset but each time translated and rotated in the box. A balanced set of examples was used sampling equal numbers of residues that are close to a zinc and residues randomly drawn from the non-zinc binding residues. The sampling of residues is based on the biological assembly of the protein, the voxelization is based on the full 3D structure including neighboring asymmetric units in the crystal structure. The environments are precomputed and stored using lxf compression in HDF5 files for concurrent access during training. In total, 252324 environments were voxelized for the training set, 6550 for the test set, 3067 for the validation set. The voxelization was implemented using ray.[73]

Model training

We used PyTorch 1.10[74] to train the model. All layers of the network are convolutional layers with filter size 1.5 Å except for the fifth layer where a 8 Å filter is used to capture long range interactions. We use zero padding to keep the size of the boxes constant. Models were trained on a workstation with NVIDIA GTX3090 GPU and 32 CPU cores. Binary Cross Entropy[75] loss is used to train the model. The rectified linear unit (ReLU) non-linearity is used except for the last layer which uses a sigmoid function that yields the probability for zinc per voxel. A dropout layer (p = 0.1) was used between the 5th and 6th layers. The network was trained using AdaDelta employing a stepped learning rate (lr=0.5, γ=0.9), a batch size of 150, and 12 epochs to train.

Hyperparameter tuning

We used the ray[tune] library[73] to perform a hyperparameter search choosing 20 different combinations between the following parameters with the best combination of parameters in bold.

- filtersize: **3**,4 (in units of 0.5 Å)
- dropout : **0.1**, 0.2, 0.4, 0.5
- learning rate : **0.5**, 1.0, 2.0

- gamma: 0.5, 0.7, 0.8, **0.9**
- largest dimension 80, 100, **120**

Grid Averaging

The model takes as input a `(8,32,32,32)` tensor and outputs a `(1,32,32,32)` tensor containing the probability density for zinc centered on the Cα atom of the input residue. Predictions for a complete protein were obtained by voxelizing select residues of the protein (default all cysteines, histidines, aspartates, glutamates) and averaging the boxes using a global grid (Figure 1 B). 98 % of the metal sites in the training data have at least one of those residues closeby wherefore this significant decrease in computational cost seems appropriate for most uses. The global grid is obtained by computing the bounding box of all points and using a regular spaced (0.5 Å) grid. For each grid point in the global grid the predicted probability maps within 0.25 Å of the grid point are averaged. The search is sped up using the KD-Tree implementation in scipy.[76]

Metal ion placement The global probability density is used to perform clustering of voxels above a certain probability threshold (default p=0.15, cutoff 7 Å) using AgglomerativeClustering implemented in scikit-learn.[77] For each cluster the weighted average of the voxels in the cluster is computed using the probabilities for each point as the weight. This results in one metal placed per cluster.

Visualization We make available a command line program and interactive notebook allowing the user to visualize the results. The averaged probability map is stored as a `cube` file. The most likely metal coordinates for use in subsequent processing are stored in a `pdb` file. The command line program uses VMD[78] to visualize the input protein and the predicted density, for the jupyter notebook 3Dmol.js/py3Dmol[79] is used.

Evaluation

Comparison In order to standardize the evaluation between different tools, we always used the same test set used for the training of Metal1D and Metal3D. In order to compute standard metrics such as precision and recall, we chose to assess the performance of all assessed tools (Metal1D, Metal3D, BioMetAll, MIB) in a binary fashion. Any prediction within 5 Å of an experimental metal site is counted as true positive (TP). Multiple predictions by the same tool for the same site are counted as 1 TP. Any experimental site that has no predicted metal within 5 Å is counted as false negative (FN). A false positive (FP) prediction is a prediction that is not within 5 Å of a zinc site and also not within 5 Å of any other false positive prediction. If two or more false positive predictions are within 5 Å, they are counted as a single false positive prediction for the same site. In practice we first evaluate the true positive and false negative predictions and remove those from the set of predicted positions. The remaining predictions are all false positives and are clustered using AgglomerativeClustering with a radius of 5 Å. The number of false positives is determined from the number of clusters. Using the binary metric we assessed how good the models are at discovering sites and how much these predictions can be trusted.

In order to assess the quality of the predictions, we additionally compute for all the true positive predictions the mean of the Euclidean distance between the true and predicted site (mean absolute deviation MAD). For Metal1D, MIB, and BioMetAll, MAD was computed for all predictions above the threshold within 5 Å of a true zinc site where $\sum$ predicted sites $\geq \sum$ TP. This was done as some tools predict the same site for different residue combinations and we wanted to assess the general performance for all predicted sites above a certain cutoff and not just for the best predicted site above the cutoff. For Metal3D the weighted average of all voxels above the cutoff was used.

Precision was calculated as

$$\text{Precision} = \frac{\text{\# correct metal sites}}{\text{\# correct metal sites} + \text{\# false positive clustered}} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall was calculated as

$$\text{Recall} = \frac{\text{\# correct metal sites}}{\text{\# correct metal sites} + \text{\# not found metal sites}} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Model assessment Metal3D To evaluate the trained models we monitored loss and how accurately the model predicts the metal density of the test set. We used a discretized version of the Jaccard index setting each voxel either as 0 (no metal) or 1 (zinc present). We tested multiple different decision boundaries (0.5, 0.6, 0.75, 0.9) and also compared a slightly smaller centered box to remove any spurious density at the box edges, where the model has only incomplete information to make predictions.

The Jaccard index is computed as

$$J = \frac{\# \left| V_p \cap V_{exp} \right|}{\# \left| V_p \cup V_{exp} \right|},$$

where $V p$ is the array of voxels with predicted probability above the decision boundary and V_{exp} is the array of voxels with the true metal locations also discretized at the same probability threshold.

HCA2 mutants The data for human carbonic anhydrase 2 (HCA2) mutants was extracted from refs[52–56] and the crystal structure 2CBA[49,80] was used. The zinc was modeled using the zinc cationic dummy model forcefield[51] and we verified that energy minimization produced the correct coordination environment. The Richardson rotamer library[81] was used with the EVOLVE-ddG energy function to compute the most stable rotamer for a given mutation with the zinc present. The lowest-energy mutant was used for the prediction of the location of metals using Metal3D.

Additional information

Acknowledgement

Supported by Swiss National Science Foundation Grant Number 200020-185092 with computational resources from the Swiss National Computing Centre CSCS. We thank the developers of 3Dmol.js and manubot that made the interactive version of this manuscript possible.

Supplemental Data

Code and training data are available on Github and on Zenodo.

Conflict of interest

None declared

Author contributions

S.L.D and A.L designed research, S.L.D, A.L, U.R conceptualized research, S.L.D and A.L developed methodology and software, S.L.D and A.L wrote first draft, S.L.D, A.L, U.R revised and edited draft, U.R supervised research and acquired funding.

Supplement

Table S1. Atoms used as reference points for each amino acid in Metal1D. In the case of amino acids with more than one possible ligand atom, the search radius is enlarged, the increase is computed from the midpoint between all ligating atoms. Typical values computed for structure data files downloaded from the PDB are reported.

Amino acid	Residue name	Label(s)	Search radius increase (Å)
Alanine	ALA	O	0
Arginine	ARG	NH1, NH2	1.2

Amino acid	Residue name	Label(s)	Search radius increase (Å)
Asparagine	ASN	OD1	0
Aspartic acid	ASP	OD1, OD2	1.105
Cysteine	CYS	SG	0
Glutamic acid	GLU	OE1, OE2	1.105
Glutamine	GLN	OE1	0
Glycine	GLY	O	0
Histidine	HIS	ND1, ND2	1.08
Isoleucine	ILE	O	0
Leucine	LEU	O	0
Lysine	LYS	NZ	0
Methionine	MET	SD	0
Phenylalanine	PHE	O	0
Proline	PRO	O	0
Serine	SER	OG	0
Threonine	THR	OG1	0
Tryptophan	TRP	O	0
Tyrosine	TYR	OH	0
Valine	VAL	OH	0

Table S2. Atom selections used for voxelization of proteins using moleculekit

Channel name	Selected atoms
aromatic	HIS TRP TYR PHE sidechain without CB
hydrophobic	element C
occupancy	all protein heavy atoms
hbond donor	(ASN GLN TRP MSE SER THR MET CYS and name ND2 NE2 NE1 SG SE OG OG1) and name N
hbond acceptor	(resname ASP GLU HIS SER THR MSE CYS MET and name ND2 NE2 OE1 OE2 OD1 OD2 OG OG1 SE SG) or name O
metalbinding	(name ND1 NE2 SG OE1 OE2 OD2) or (protein and name O N)
positive charge	resname LYS ARG HIS and name NZ NH1 NH2 ND1 NE2 NE
negative charge	resname ASP GLU and name OD1 OD2 OE1 OE2

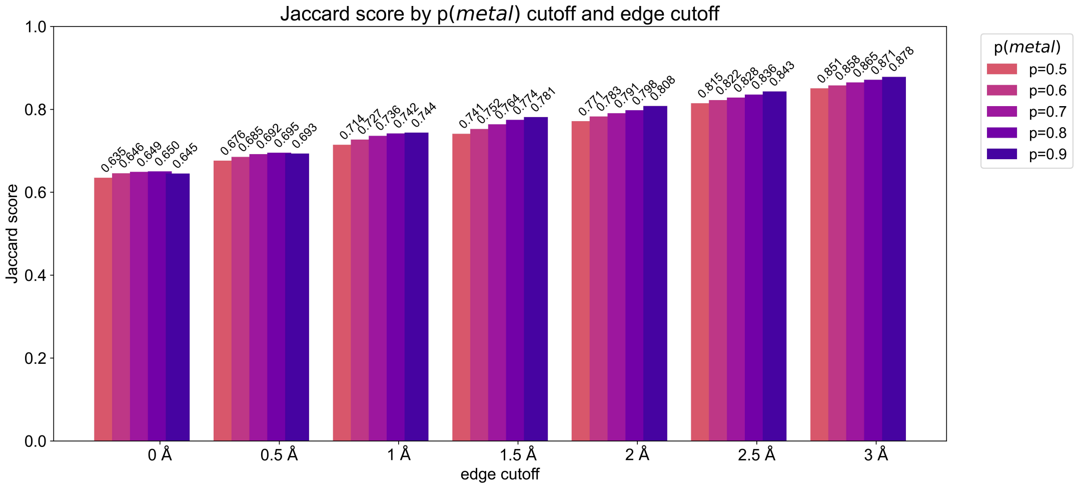

Figure S1. Discretized Jaccard indices using different cutoffs for edge trimming and different probability cutoffs (p(metal)) showing that Metal3D predictions well reproduce the target environments in the test set.

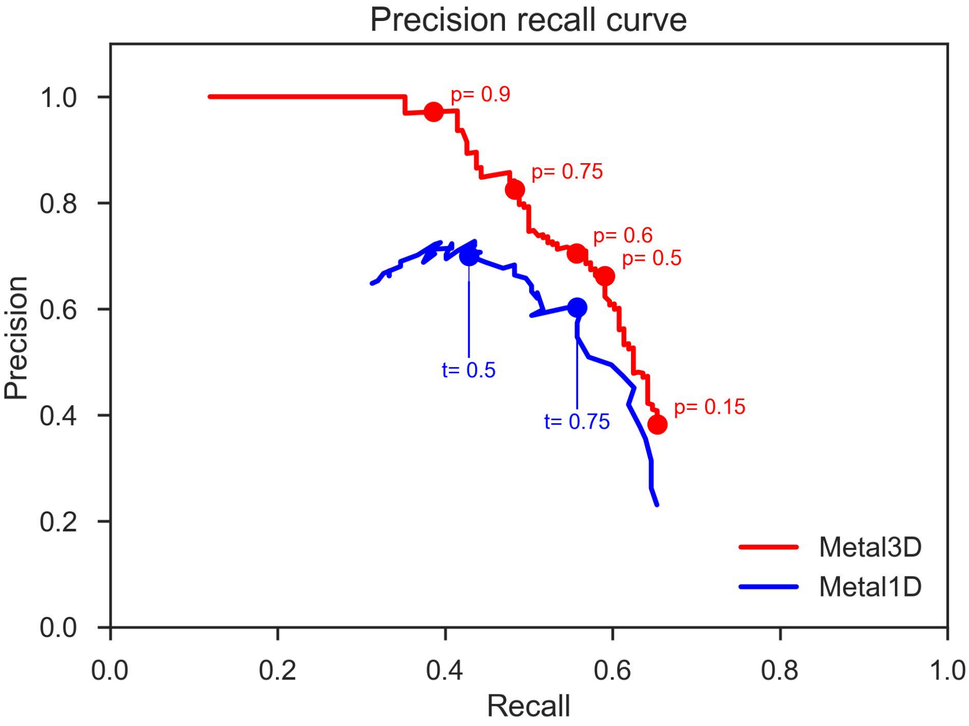

Figure S2. Precision recall curve for Metal1D and Metal3D with the probability cutoffs (Metal3D) or thresholds (Metal1D) used in the analysis.

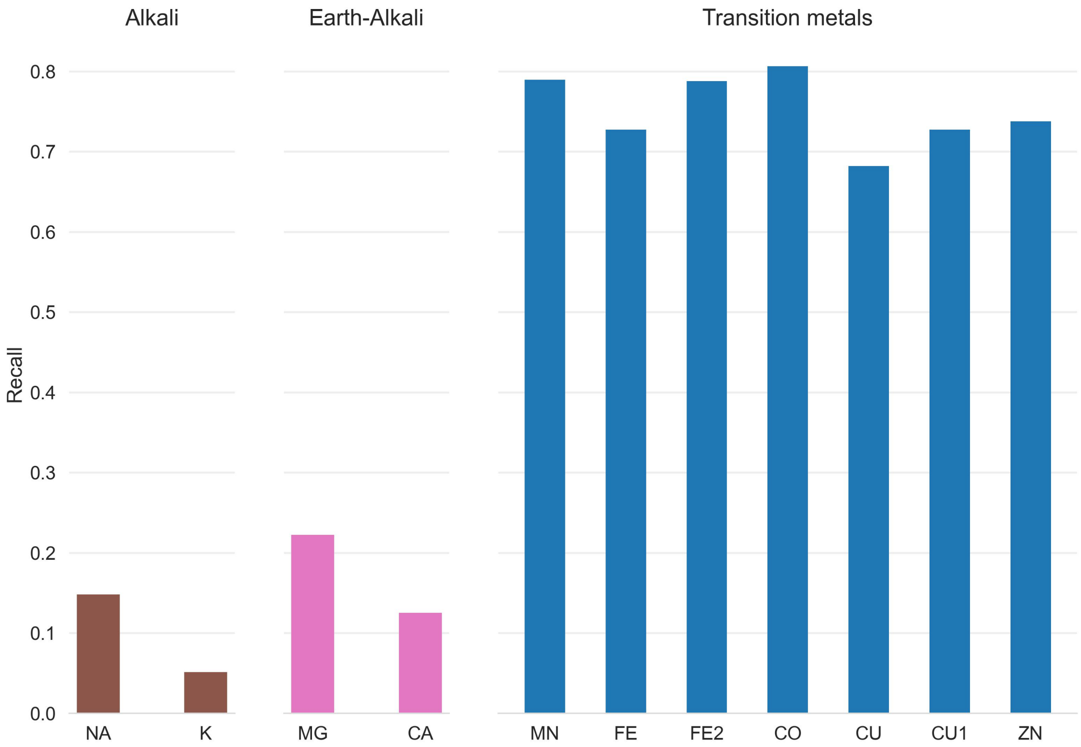

Figure S3. Recall for zinc testset and a 25 randomly drawn structures for other transition, alkali and earth-alkali metals for Metal1D

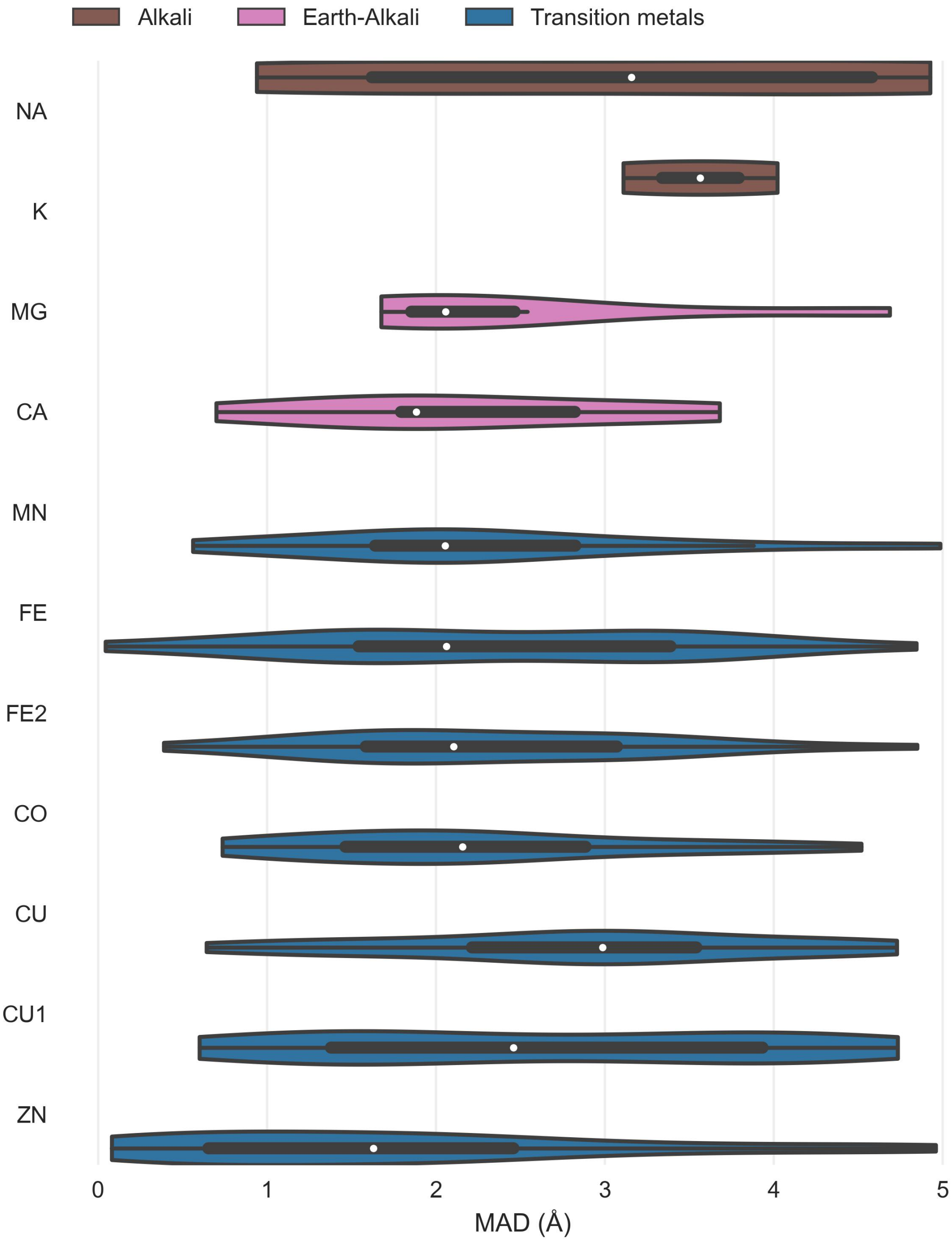

Figure S4. Distance distribution Metal1D. For each ion the whisker plot indicates the median (white dot) and the first quartiles (black box).

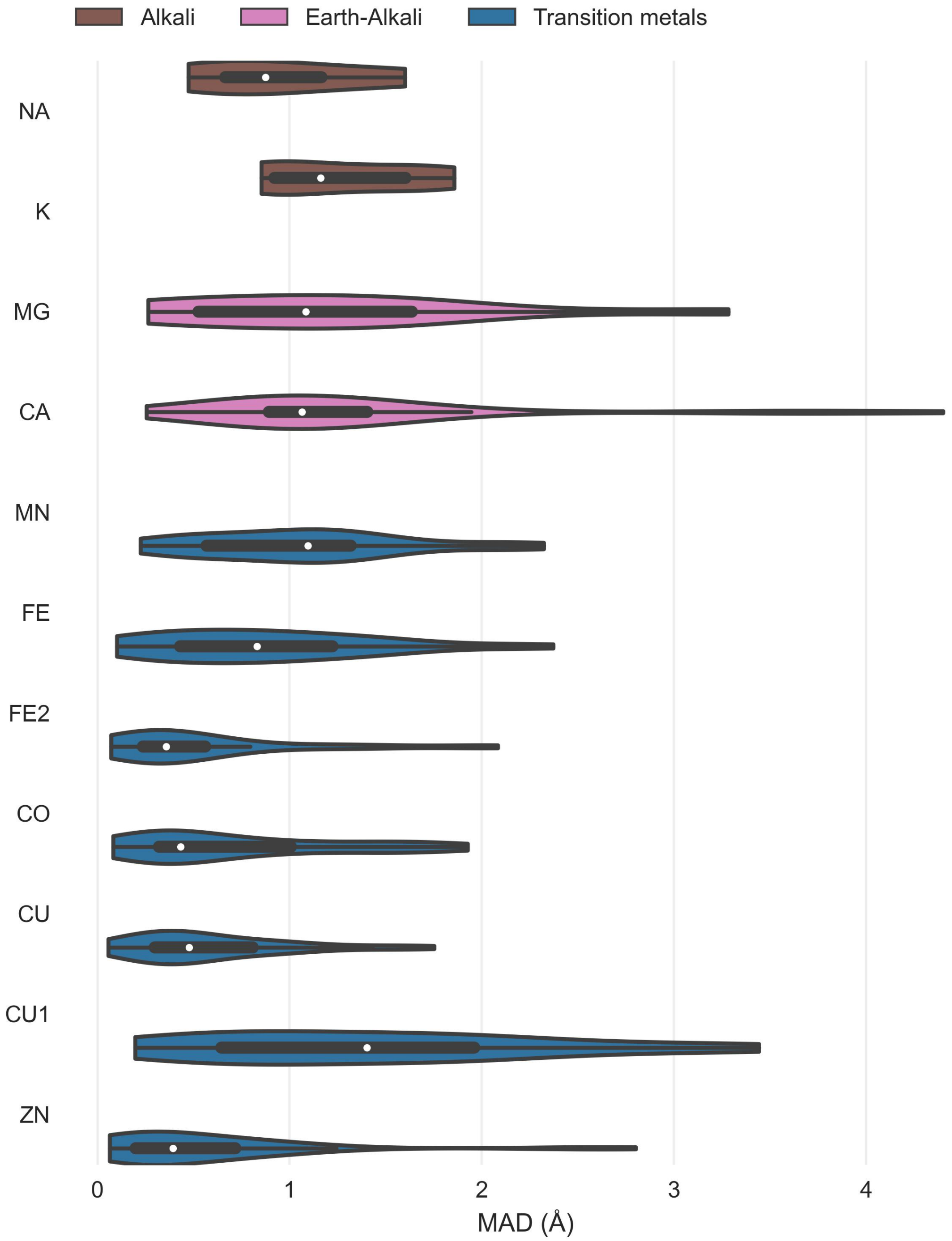

Figure S5. MAD for Metal3D for all sites with 3+ unique protein ligands in the test set and for the selected structures for the other metals. For each ion the whisker plot indicates the median (white dot) and the first quartiles (black box).

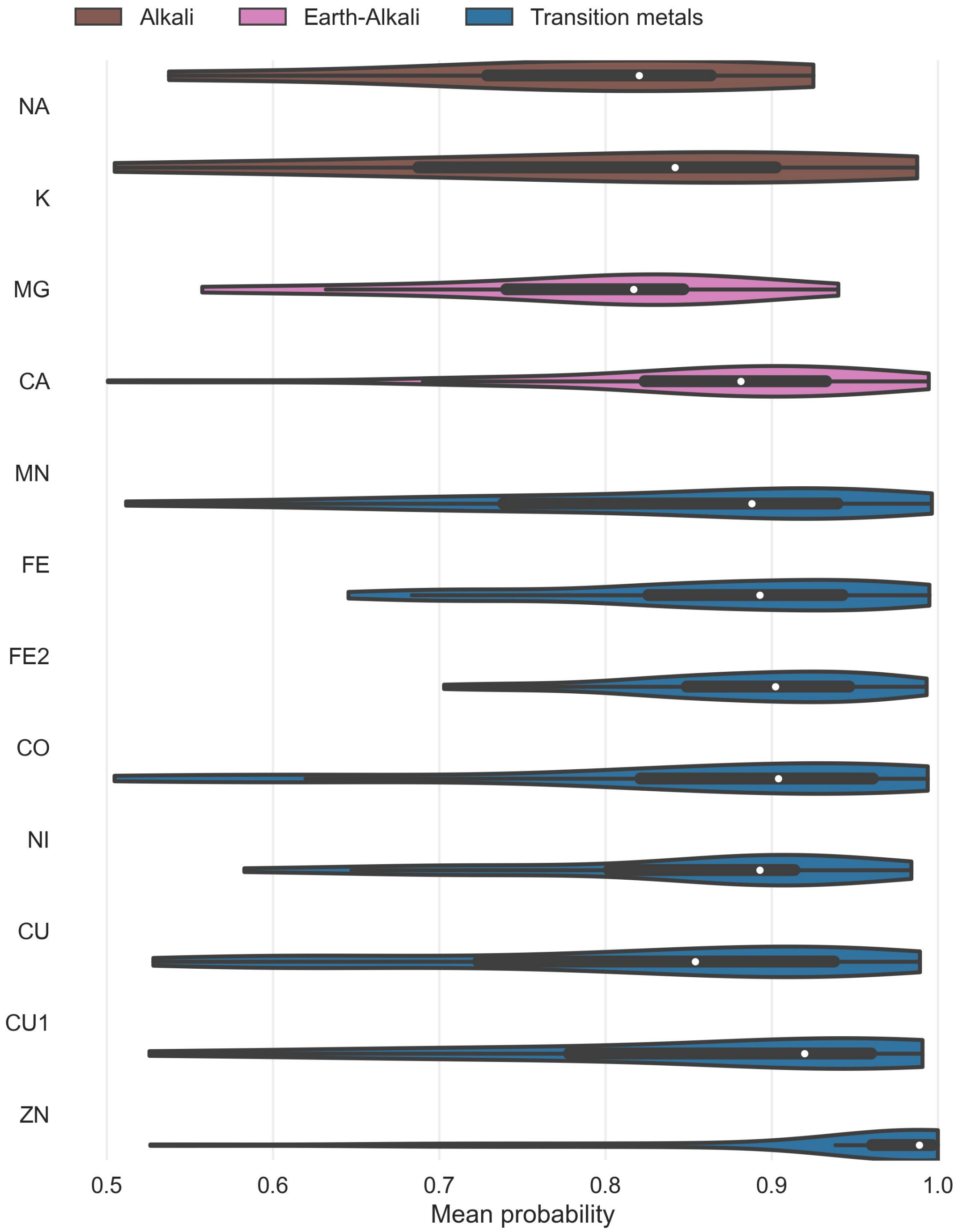

Figure S6. Probability distribution for Metal3D on sites with 3+ unique protein ligands in the test set and for the selected structures for the other metals. For each ion the whisker plot indicates the median probability (white dot) and the first quartiles (black box).

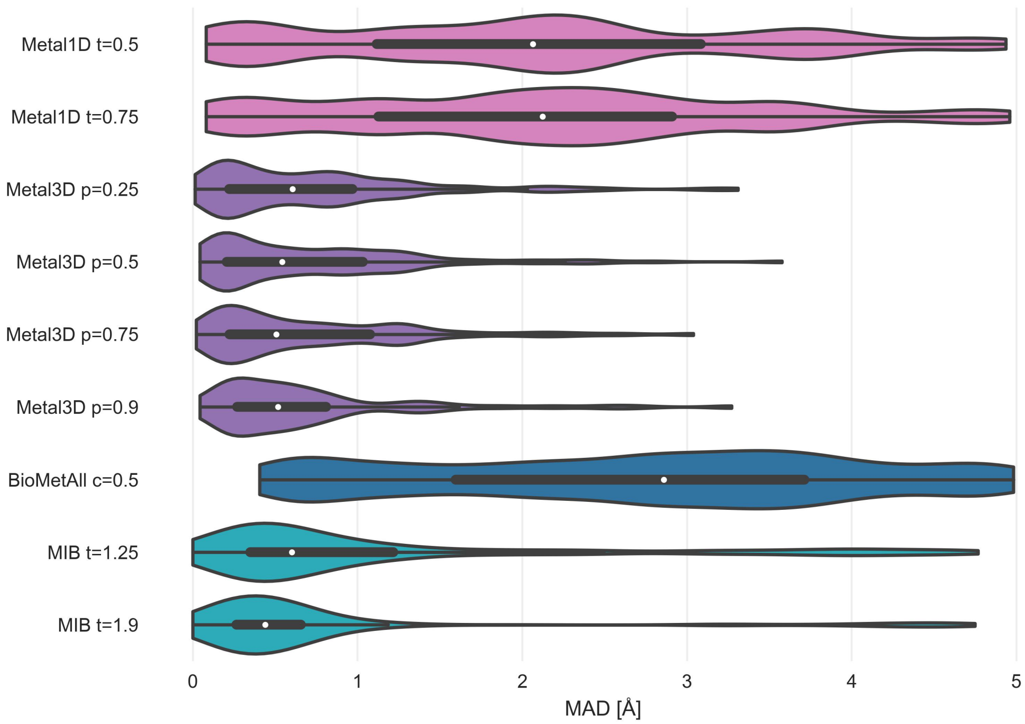

Figure S7. MAD only 2 residue coordinated zincs. For each tool the whisker plot indicates the median (white dot) and the first quartiles (black box).

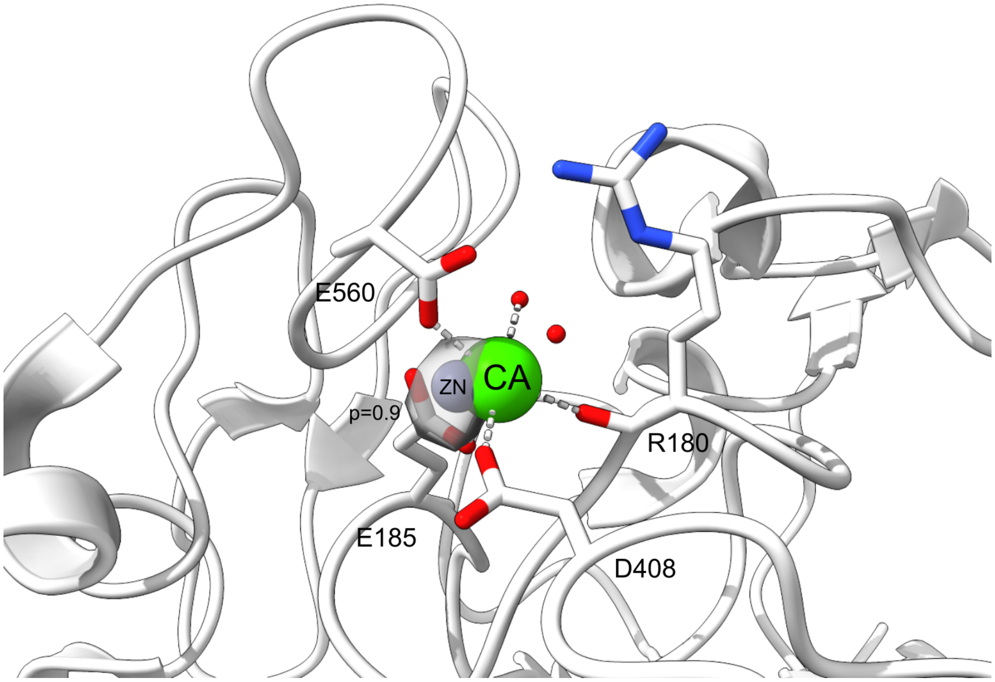

Figure S8. False positive for Metal3D at p=0.9 in PDB 4JJJ. A calcium site is misclassified as zinc site.

Table S3. MAD and median on zinc test set for all zinc sites (n=189).

tool	MAD [Å]	median [Å]
BioMetAll c=0.5	2.72 ± 1.33	2.86
MIB t=1.25	1.13 ± 1.24	0.60
MIB t=1.9	0.77 ± 1.09	**0.44**
Metal1D t=0.5	2.07 ± 1.33	2.06
Metal1D t=0.75	2.19 ± 1.26	2.12
Metal3D p=0.25	0.74 ± 0.66	0.61
Metal3D p=0.5	0.73 ± 0.66	0.54
Metal3D p=0.75	0.71 ± 0.64	0.51
Metal3D p=0.9	**0.70 ± 0.64**	0.52

Table S4. MAD and median on zinc test set for zinc sites with at least 2 protein ligands (n=133).

tool	MAD [Å]	median [Å]
BioMetAll c=0.5	2.68 ± 1.33	2.84
MIB t=1.25	1.09 ± 1.21	0.60
MIB t=1.9	0.77 ± 1.09	**0.44**
Metal1D t=0.5	1.97 ± 1.29	1.99
Metal1D t=0.75	2.06 ± 1.24	2.09
Metal3D p=0.25	0.69 ± 0.58	0.56
Metal3D p=0.5	0.69 ± 0.59	0.54
Metal3D p=0.75	0.71 ± 0.64	0.51
Metal3D p=0.9	**0.70 ± 0.64**	0.52

References

(1) Yu, F.; Cangelosi, V. M.; Zastrow, M. L.; Tegoni, M.; Plegaria, J. S.; Tebo, A. G.; Mocny, C. S.; Ruckthong, L.; Qayyum, H.; Pecoraro, V. L. Protein Design: Toward Functional Metalloenzymes. *Chem. Rev.* **2014**, *114* (7), 3495–3578. https://doi.org/10.1021/cr400458x.

(2) Guffy, S. L.; Der, B. S.; Kuhlman, B. Probing the Minimal Determinants of Zinc Binding with Computational Protein Design. *Protein Engineering, Design and Selection* **2016**, *29* (8), 327–338. https://doi.org/10.1093/protein/gzw026.

(3) Andreini, C.; Bertini, I.; Cavallaro, G.; Holliday, G. L.; Thornton, J. M. Metal Ions in Biological Catalysis: From Enzyme Databases to General Principles. *J Biol Inorg Chem* **2008**, *13* (8), 1205–1218. https://doi.org/10.1007/s00775-008-0404-5.

(4) Koohi-Moghadam, M.; Wang, H.; Wang, Y.; Yang, X.; Li, H.; Wang, J.; Sun, H. Predicting Disease-Associated Mutation of Metal-Binding Sites in Proteins Using a Deep Learning Approach. *Nat Mach Intell* **2019**, *1* (12), 561–567. https://doi.org/10.1038/s42256-019-0119-z.

(5) Studer, S.; Hansen, D. A.; Pianowski, Z. L.; Mittl, P. R. E.; Debon, A.; Guffy, S. L.; Der, B. S.; Kuhlman, B.; Hilvert, D. Evolution of a Highly Active and Enantiospecific Metalloenzyme from Short Peptides. *Science* **2018**, *362* (6420), 1285–1288. https://doi.org/10.1126/science.aau3744.

(6) Key, H. M.; Dydio, P.; Clark, D. S.; Hartwig, J. F. Abiological Catalysis by Artificial Haem Proteins Containing Noble Metals in Place of Iron. *Nature* **2016**, *534* (7608), 534–537. https://doi.org/10.1038/nature17968.

(7) Chalkley, M. J.; Mann, S. I.; DeGrado, W. F. De Novo Metalloprotein Design. *Nat Rev Chem* **2021**, *6* (1), 31–50. https://doi.org/10.1038/s41570-021-00339-5.

(8) Der, B. S.; Edwards, D. R.; Kuhlman, B. Catalysis by a de Novo Zinc-Mediated Protein Interface: Implications for Natural Enzyme Evolution and Rational Enzyme Engineering. *Biochemistry* **2012**, *51* (18), 3933–3940. https://doi.org/10.1021/bi201881p.

(9) Fujieda, N.; Schätti, J.; Stuttfeld, E.; Ohkubo, K.; Maier, T.; Fukuzumi, S.; Ward, T. R. Enzyme Repurposing of a Hydrolase as an Emergent Peroxidase Upon Metal Binding. *Chem. Sci.* **2015**, *6* (7), 4060–4065. https://doi.org/10.1039/c5sc01065a.

(10) Brodin, J. D.; Ambroggio, X. I.; Tang, C.; Parent, K. N.; Baker, T. S.; Tezcan, F. A. Metal-Directed, Chemically Tunable Assembly of One-, Two- and Three-Dimensional Crystalline Protein Arrays. *Nature Chem* **2012**, *4* (5), 375–382. https://doi.org/10.1038/nchem.1290.

(11) Der, B. S.; Machius, M.; Miley, M. J.; Mills, J. L.; Szyperski, T.; Kuhlman, B. Metal-Mediated Affinity and Orientation Specificity in a Computationally Designed Protein Homodimer. *J. Am. Chem. Soc.* **2011**, *134* (1), 375–385. https://doi.org/10.1021/ja208015j.

(12) Salgado, E. N.; Radford, R. J.; Tezcan, F. A. Metal-Directed Protein Self-Assembly. *Acc. Chem. Res.* **2010**, *43* (5), 661–672. https://doi.org/10.1021/ar900273t.

(13) 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.* **2020**, *59* (49), 21940–21944. https://doi.org/10.1002/anie.202009226.

(14) Zastrow, M. L.; Peacock, A. F. A.; Stuckey, J. A.; Pecoraro, V. L. Hydrolytic Catalysis and Structural Stabilization in a Designed Metalloprotein. *Nature Chem* **2011**, *4* (2), 118–123. https://doi.org/10.1038/nchem.1201.

(15) Song, L. F.; Sengupta, A.; Merz, K. M., Jr. Thermodynamics of Transition Metal Ion Binding to Proteins. *J. Am. Chem. Soc.* **2020**, *142* (13), 6365–6374. https://doi.org/10.1021/jacs.0c01329.

(16) Alford, R. F.; Leaver-Fay, A.; Jeliazkov, J. R.; O'Meara, M. J.; DiMaio, F. P.; Park, H.; Shapovalov, M. V.; Renfrew, P. D.; Mulligan, V. K.; Kappel, K.; Labonte, J. W.; Pacella, M. S.; Bonneau, R.; Bradley, P.; Dunbrack, R. L., Jr.; Das, R.; Baker, D.; Kuhlman, B.; Kortemme, T.; Gray, J. J. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **2017**, *13* (6), 3031–3048. https://doi.org/10.1021/acs.jctc.7b00125.

(17) Brunk, E.; Rothlisberger, U. Mixed Quantum Mechanical/Molecular Mechanical Molecular Dynamics Simulations of Biological Systems in Ground and Electronically Excited States. *Chem. Rev.*

2015, *115* (12), 6217–6263. https://doi.org/10.1021/cr500628b.

(18) Yang, Z.; Twidale, R. M.; Gervasoni, S.; Suardíaz, R.; Colenso, C. K.; Lang, E. J. M.; Spencer, J.; Mulholland, A. J. Multiscale Workflow for Modeling Ligand Complexes of Zinc Metalloproteins. *J. Chem. Inf. Model.* **2021**, *61* (11), 5658–5672. https://doi.org/10.1021/acs.jcim.1c01109.

(19) Bozkurt, E.; Perez, M. A. S.; Hovius, R.; Browning, N. J.; Rothlisberger, U. Genetic Algorithm Based Design and Experimental Characterization of a Highly Thermostable Metalloprotein. *J. Am. Chem. Soc.* **2018**, *140* (13), 4517–4521. https://doi.org/10.1021/jacs.7b10660.

(20). https://doi.org/10.3389/fchem.2021.692200.

(21) Passerini, A.; Andreini, C.; Menchetti, S.; Rosato, A.; Frasconi, P. Predicting Zinc Binding at the Proteome Level. *BMC Bioinformatics* **2007**, *8* (1). https://doi.org/10.1186/1471-2105-8-39.

(22) Hu, X.; Dong, Q.; Yang, J.; Zhang, Y. Recognizing Metal and Acid Radical Ion-Binding Sites by Integratingab Initiomodeling with Template-Based Transferals. *Bioinformatics* **2016**, *32* (21), 3260–3269. https://doi.org/10.1093/bioinformatics/btw396.

(23) Lin, Y.-F.; Cheng, C.-W.; Shih, C.-S.; Hwang, J.-K.; Yu, C.-S.; Lu, C.-H. MIB: Metal Ion-Binding Site Prediction and Docking Server. *J. Chem. Inf. Model.* **2016**, *56* (12), 2287–2291. https://doi.org/10.1021/acs.jcim.6b00407.

(24) Hekkelman, M. L.; de Vries, I.; Joosten, R. P.; Perrakis, A. AlphaFill: Enriching the AlphaFold Models with Ligands and Co-Factors, 2021. https://doi.org/10.1101/2021.11.26.470110.

(25) Brylinski, M.; Skolnick, J. FINDSITE-Metal: Integrating Evolutionary Information and Machine Learning for Structure-Based Metal-Binding Site Prediction at the Proteome Level. *Proteins* **2010**, *79* (3), 735–751. https://doi.org/10.1002/prot.22913.

(26) Sánchez-Aparicio, J.-E.; Tiessler-Sala, L.; Velasco-Carneros, L.; Roldán-Martín, L.; Sciortino, G.; Maréchal, J.-D. BioMetAll: Identifying Metal-Binding Sites in Proteins from Backbone Preorganization. *J. Chem. Inf. Model.* **2020**, *61* (1), 311–323. https://doi.org/10.1021/acs.jcim.0c00827.

(27) Haberal, İ.; Oğul, H. Prediction of Protein Metal Binding Sites Using Deep Neural Networks. *Mol. Inf.* **2019**, *38* (7), 1800169. https://doi.org/10.1002/minf.201800169.

(28) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.

(29) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; Millán, C.; Park, H.; Adams, C.; Glassman, C. R.; DeGiovanni, A.; Pereira, J. H.; Rodrigues, A. V.; van Dijk, A. A.; Ebrecht, A. C.; Opperman, D. J.; Sagmeister, T.; Buhlheller, C.; Pavkov-Keller, T.; Rathinaswamy, M. K.; Dalwadi, U.; Yip, C. K.; Burke, J. E.; Garcia, K. C.; Grishin, N. V.; Adams, P. D.; Read, R. J.; Baker, D. Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network. *Science* **2021**, *373* (6557), 871–876. https://doi.org/10.1126/science.abj8754.

(30) Torng, W.; Altman, R. B. 3D Deep Convolutional Neural Networks for Amino Acid Environment Similarity Analysis. *BMC Bioinformatics* **2017**, *18* (1). https://doi.org/10.1186/s12859-017-1702-0.

(31) Shroff, R.; Cole, A. W.; Diaz, D. J.; Morrow, B. R.; Donnell, I.; Annapareddy, A.; Gollihar, J.; Ellington, A. D.; Thyer, R. Discovery of Novel Gain-of-Function Mutations Guided by Structure-Based Deep Learning. *ACS Synth. Biol.* **2020**, *9* (11), 2927–2935. https://doi.org/10.1021/acssynbio.0c00345.

(32) Anand, N.; Eguchi, R.; Mathews, I. I.; Perez, C. P.; Derry, A.; Altman, R. B.; Huang, P.-S. Protein Sequence Design with a Learned Potential. *Nat Commun* **2022**, *13* (1). https://doi.org/10.1038/s41467-022-28313-9.

(33) Torng, W.; Altman, R. B. High Precision Protein Functional Site Detection Using 3D Convolutional Neural Networks. *Bioinformatics* **2018**, *35* (9), 1503–1512. https://doi.org/10.1093/bioinformatics/bty813.

(34) Feehan, R.; Franklin, M. W.; Slusky, J. S. G. Machine Learning Differentiates Enzymatic and Non-Enzymatic Metals in Proteins. *Nat Commun* **2021**, *12* (1). https://doi.org/10.1038/s41467-021-24070-3.

(35) Renaud, N.; Geng, C.; Georgievska, S.; Ambrosetti, F.; Ridder, L.; Marzella, D. F.; Réau, M. F.; Bonvin, A. M. J. J.; Xue, L. C. DeepRank: A Deep Learning Framework for Data Mining 3D Protein-Protein Interfaces. *Nat Commun* **2021**, *12* (1). https://doi.org/10.1038/s41467-021-27396-0.

(36) Gainza, P.; Sverrisson, F.; Monti, F.; Rodolà, E.; Boscaini, D.; Bronstein, M. M.; Correia, B. E. Deciphering Interaction Fingerprints from Protein Molecular Surfaces Using Geometric Deep Learning. *Nat Methods* **2019**, *17* (2), 184–192. https://doi.org/10.1038/s41592-019-0666-6.

(37) Jiménez, J.; Doerr, S.; Martínez-Rosell, G.; Rose, A. S.; De Fabritiis, G. DeepSite: Protein-Binding Site Predictor Using 3D-Convolutional Neural Networks. *Bioinformatics* **2017**, *33* (19), 3036–3042. https://doi.org/10.1093/bioinformatics/btx350.

(38) Skalic, M.; Varela-Rial, A.; Jiménez, J.; Martínez-Rosell, G.; De Fabritiis, G. LigVoxel: Inpainting Binding Pockets Using 3D-Convolutional Neural Networks. *Bioinformatics* **2018**, *35* (2), 243–250. https://doi.org/10.1093/bioinformatics/bty583.

(39) Stärk, H.; Ganea, O.-E.; Pattanaik, L.; Barzilay, R.; Jaakkola, T. EquiBind: Geometric Deep Learning for Drug Binding Structure Prediction. arXiv 2022. https://doi.org/10.48550/arxiv.2202.05146.

(40) Park, S.; Seok, C. GalaxyWater-CNN: Prediction of Water Positions on the Protein Structure by a 3D-Convolutional Neural Network. *J. Chem. Inf. Model.* **2022**, *62* (13), 3157–3168. https://doi.org/10.1021/acs.jcim.2c00306.

(41) Li, B.; Yang, Y. T.; Capra, J. A.; Gerstein, M. B. Predicting Changes in Protein Thermodynamic Stability Upon Point Mutation with Deep 3D Convolutional Neural Networks. *PLoS Comput Biol* **2020**, *16* (11), e1008291. https://doi.org/10.1371/journal.pcbi.1008291.

(42) Lu, C.-H.; Lin, Y.-F.; Lin, J.-J.; Yu, C.-S. Prediction of Metal Ion–Binding Sites in Proteins Using the Fragment Transformation Method. *PLoS ONE* **2012**, *7* (6), e39252. https://doi.org/10.1371/journal.pone.0039252.

(43) Laitaoja, M.; Valjakka, J.; Jänis, J. Zinc Coordination Spheres in Protein Structures. *Inorg. Chem.* **2013**, *52* (19), 10983–10991. https://doi.org/10.1021/ic401072d.

(44) Minasov, G.; Vorontsov, I. I.; Shuvalova, L.; Brunzelle, J. S.; Kiryukhina, O.; Collart, F. R.; Joachimiak, A.; Anderson, W. F.;. Crystal Structure of Unknown Conserved ybaA Protein from Shigella Flexneri, 2007. https://doi.org/10.2210/pdb2okq/pdb.

(45) Itoh, T.; Nakagawa, E.; Yoda, M.; Nakaichi, A.; Hibi, T.; Kimoto, H. Crystal Structure of Alginate Lyase from Paenibacillus Sp. Str. FPU-7, 2019. https://doi.org/10.2210/pdb6kfn/pdb.

(46) Adams, C. M.; Eckenroth, B. E.; Doublie, S. Structure of the Clostridium Perfringens CspB Protease, 2013. https://doi.org/10.2210/pdb4i0w/pdb.

(47) Davies, C. W.; Das, C. The Crystal Structure of a E280A Mutant of the Catalytic Domain of AMSH, 2011. https://doi.org/10.2210/pdb3rzv/pdb.

(48) Rana, M. S.; Kumar, P.; Lee, C.-J.; Verardi, R.; Rajashankar, K. R.; Banerjee, A. Fatty Acyl Recognition and Transfer by an Integral Membrane *S* -Acyltransferase. *Science* **2018**, *359* (6372). https://doi.org/10.1126/science.aao6326.

(49) Hakansson, K.; Carlsson, M.; Svensson, L. A.; Liljas, A. STRUCTURE OF NATIVE AND APO CARBONIC ANHYDRASE II AND SOME OF ITS ANION-LIGAND COMPLEXES, 1993. https://doi.org/10.2210/pdb2cba/pdb.

(50) Hunt, J. B.; Neece, S. H.; Ginsburg, A. The Use of 4-(2-Pyridylazo)resorcinol in Studies of Zinc Release from Escherichia Coli Aspartate Transcarbamoylase. *Anal Biochem* **1985**, *146* (1), 150–157. https://doi.org/10.1016/0003-2697(85)90409-9.

(51) Pang, Y. P.; Xu, K.; Yazal, J. E.; Prendergas, F. G. Successful Molecular Dynamics Simulation of the Zinc-Bound Farnesyltransferase Using the Cationic Dummy Atom Approach. *Protein Sci* **2000**, *9* (10), 1857–1865.

(52) Kiefer, L. L.; Fierke, C. A. Functional Characterization of Human Carbonic Anhydrase II Variants with Altered Zinc Binding Sites. *Biochemistry* **1994**, *33* (51), 15233–15240. https://doi.org/10.1021/bi00255a003.

(53) Kiefer, L. L.; Ippolito, J. A.; Fierke, C. A.; Christianson, D. W. Redesigning the Zinc Binding Site of Human Carbonic Anhydrase II: Structure of a His2Asp-Zn^{2}+ Metal Coordination Polyhedron. *J. Am. Chem. Soc.* **1993**, *115* (26), 12581–12582. https://doi.org/10.1021/ja00079a046.

(54) Ippolito, J. A.; Christianson, D. W. Structure of an Engineered His3 Cys Zinc Binding Site in Human Carbonic Anhydrase II. *Biochemistry* **1993**, *32* (38), 9901–9905. https://doi.org/10.1021/bi00089a005.

(55) Ippolito, J. A.; Baird, T. T., Jr; McGee, S. A.; Christianson, D. W.; Fierke, C. A. Structure-Assisted Redesign of a Protein-Zinc-Binding Site with Femtomolar Affinity. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92* (11), 5017–5021. https://doi.org/10.1073/pnas.92.11.5017.

(56) Huang, C.-c.; Lesburg, C. A.; Kiefer, L. L.; Fierke, C. A.; Christianson, D. W. Reversal of the Hydrogen Bond to Zinc Ligand Histidine-119 Dramatically Diminishes Catalysis and Enhances Metal Equilibration Kinetics in Carbonic Anhydrase II. *Biochemistry* **1996**, *35* (11), 3439–3446. https://doi.org/10.1021/bi9526692.

(57) Savage, H.; Wlodawer, A. Determination of Water Structure Around Biomolecules Using X-Ray and Neutron Diffraction Methods. *Methods Enzymol* **1986**, *127*, 162–183. https://doi.org/10.1016/0076-6879(86)27014-7.

(58) Morozenko, A.; Stuchebrukhov, A. A. Dowser++, a New Method of Hydrating Protein Structures. *Proteins* **2016**, *84* (10), 1347–1357. https://doi.org/10.1002/prot.25081.

(59) Sridhar, A.; Ross, G. A.; Biggin, P. C. Waterdock 2.0: Water Placement Prediction for Holo-Structures with a Pymol Plugin. *PLoS ONE* **2017**, *12* (2), e0172743. https://doi.org/10.1371/journal.pone.0172743.

(60) Satorras, V. G.; Hoogeboom, E.; Welling, M. E(n) Equivariant Graph Neural Networks. arXiv 2021. https://doi.org/10.48550/arxiv.2102.09844.

(61) Gligorijević, V.; Berenberg, D.; Ra, S.; Watkins, A.; Kelow, S.; Cho, K.; Bonneau, R. Function-Guided Protein Design by Deep Manifold Sampling, 2021. https://doi.org/10.1101/2021.12.22.473759.

(62) Greener, J. G.; Moffat, L.; Jones, D. T. Design of Metalloproteins and Novel Protein Folds Using Variational Autoencoders. *Sci Rep* **2018**, *8* (1). https://doi.org/10.1038/s41598-018-34533-1.

(63) Song, H.; Wilson, D. L.; Farquhar, E. R.; Lewis, E. A.; Emerson, J. P. Revisiting Zinc Coordination in Human Carbonic Anhydrase II. *Inorg. Chem.* **2012**, *51* (20), 11098–11105. https://doi.org/10.1021/ic301645j.

(64) Handel, T. M.; Williams, S. A.; DeGrado, W. F. Metal Ion-Dependent Modulation of the Dynamics of a Designed Protein. *Science* **1993**, *261* (5123), 879–885. https://doi.org/10.1126/science.8346440.

(65) Arnold, F. H.; Haymore, B. L. Engineered Metal-Binding Proteins: Purification to Protein Folding. *Science* **1991**, *252* (5014), 1796–1797. https://doi.org/10.1126/science.1648261.

(66) Krantz, B. A.; Sosnick, T. R. *Nat. Struct Biol.* **2001**, *8* (12), 1042–1047. https://doi.org/10.1038/nsb723.

(67) Tunyasuvunakool, K.; Adler, J.; Wu, Z.; Green, T.; Zielinski, M.; Žídek, A.; Bridgland, A.; Cowie, A.; Meyer, C.; Laydon, A.; Velankar, S.; Kleywegt, G. J.; Bateman, A.; Evans, R.; Pritzel, A.; Figurnov, M.; Ronneberger, O.; Bates, R.; Kohl, S. A. A.; Potapenko, A.; Ballard, A. J.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Clancy, E.; Reiman, D.; Petersen, S.; Senior, A. W.; Kavukcuoglu, K.; Birney, E.; Kohli, P.; Jumper, J.; Hassabis, D. Highly Accurate Protein Structure Prediction for the Human Proteome. *Nature* **2021**, *596* (7873), 590–596. https://doi.org/10.1038/s41586-021-03828-1.

(68) Berman, H. M. The Protein Data Bank. *Nucleic Acids Research* **2000**, *28* (1), 235–242. https://doi.org/10.1093/nar/28.1.235.

(69) Steinegger, M.; Söding, J. MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets. *Nat Biotechnol* **2017**, *35* (11), 1026–1028. https://doi.org/10.1038/nbt.

3988.

(70) Barber-Zucker, S.; Shaanan, B.; Zarivach, R. Transition Metal Binding Selectivity in Proteins and Its Correlation with the Phylogenomic Classification of the Cation Diffusion Facilitator Protein Family. *Sci Rep* **2017**, *7* (1). https://doi.org/10.1038/s41598-017-16777-5.

(71) Raschka, S. BioPandas: Working with Molecular Structures in Pandas DataFrames. *JOSS* **2017**, *2* (14), 279. https://doi.org/10.21105/joss.00279.

(72) Doerr, S.; Harvey, M. J.; Noé, F.; De Fabritiis, G. HTMD: High-Throughput Molecular Dynamics for Molecular Discovery. *J. Chem. Theory Comput.* **2016**, *12* (4), 1845–1852. https://doi.org/10.1021/acs.jctc.6b00049.

(73) Moritz, P.; Nishihara, R.; Wang, S.; Tumanov, A.; Liaw, R.; Liang, E.; Elibol, M.; Yang, Z.; Paul, W.; Jordan, M. I.; Stoica, I. Ray: A Distributed Framework for Emerging AI Applications. arXiv 2017. https://doi.org/10.48550/arxiv.1712.05889.

(74) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Köpf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; Chintala, S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. arXiv 2019. https://doi.org/10.48550/arxiv.1912.01703.

(75) de Boer, P.-T.; Kroese, D. P.; Mannor, S.; Rubinstein, R. Y. A Tutorial on the Cross-Entropy Method. *Ann Oper Res* **2005**, *134* (1), 19–67. https://doi.org/10.1007/s10479-005-5724-z.

(76) Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C. J.; Polat, İ.; Feng, Y.; Moore, E. W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; van Mulbregt, P.; Vijaykumar, A.; Bardelli, A. P.; Rothberg, A.; Hilboll, A.; Kloeckner, A.; Scopatz, A.; Lee, A.; Rokem, A.; Woods, C. N.; Fulton, C.; Masson, C.; Häggström, C.; Fitzgerald, C.; Nicholson, D. A.; Hagen, D. R.; Pasechnik, D. V.; Olivetti, E.; Martin, E.; Wieser, E.; Silva, F.; Lenders, F.; Wilhelm, F.; Young, G.; Price, G. A.; Ingold, G.-L.; Allen, G. E.; Lee, G. R.; Audren, H.; Probst, I.; Dietrich, J. P.; Silterra, J.; Webber, J. T.; Slavič, J.; Nothman, J.; Buchner, J.; Kulick, J.; Schönberger, J. L.; de Miranda Cardoso, J. V.; Reimer, J.; Harrington, J.; Rodríguez, J. L. C.; Nunez-Iglesias, J.; Kuczynski, J.; Tritz, K.; Thoma, M.; Newville, M.; Kümmerer, M.; Bolingbroke, M.; Tartre, M.; Pak, M.; Smith, N. J.; Nowaczyk, N.; Shebanov, N.; Pavlyk, O.; Brodtkorb, P. A.; Lee, P.; McGibbon, R. T.; Feldbauer, R.; Lewis, S.; Tygier, S.; Sievert, S.; Vigna, S.; Peterson, S.; More, S.; Pudlik, T.; Oshima, T.; Pingel, T. J.; Robitaille, T. P.; Spura, T.; Jones, T. R.; Cera, T.; Leslie, T.; Zito, T.; Krauss, T.; Upadhyay, U.; Halchenko, Y. O.; Vázquez-Baeza, Y.;. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat Methods* **2020**, *17* (3), 261–272. https://doi.org/10.1038/s41592-019-0686-2.

(77) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Müller, A.; Nothman, J.; Louppe, G.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *arXiv* **2012**. https://doi.org/10.48550/arxiv.1201.0490.

(78) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J Mol Graph* **1996**, *14* (1), 33–38, 27–28. https://doi.org/10.1016/0263-7855(96)00018-5.

(79) Rego, N.; Koes, D. 3Dmol.js: Molecular Visualization with WebGL. *Bioinformatics* **2014**, *31* (8), 1322–1324. https://doi.org/10.1093/bioinformatics/btu829.

(80) Håkansson, K.; Carlsson, M.; Svensson, L.; Liljas, A. Structure of Native and Apo Carbonic Anhydrase II and Structure of Some of Its Anion-Ligand Complexes. *Journal of Molecular Biology* **1992**, *227* (4), 1192–1204. https://doi.org/10.1016/0022-2836(92)90531-n.

(81) Lovell, S. C.; Word, J. M.; Richardson, J. S.; Richardson, D. C. The Penultimate Rotamer Library. *Proteins* **2000**, *40* (3), 389–408.