

¹ Accurate prediction of transition metal ion location via deep learning

³ **Simon L. Dürr¹, Andrea Levy¹, Ursula Rothlisberger¹**

*For correspondence:

ursula.rothlisberger@epfl.ch (UR) ⁴ ¹Laboratory of Computational Chemistry and Biochemistry, Institute of Chemical Sciences and Engineering, Swiss Federal Institute of Technology (EPFL) CH-1015 Lausanne, Switzerland

Github:

lcbc-epfl/metal-site-prediction

Webapp:

hf.space/simonduerr/metal3d

Interactive manuscript:

lcbc-epfl.github.io/metal-site-prediction

⁷ **Abstract**

⁸ Metal ions are essential cofactors for many proteins. In fact, currently, about half of the structurally ⁹ characterized proteins contain a metal ion. Metal ions play a crucial role for many applications ¹⁰ such as enzyme design or design of protein-protein interactions because they are biologically ¹¹ abundant, tether to the protein using strong interactions, and have favorable catalytic properties ¹² e.g. as Lewis acid. Computational design of metalloproteins is however hampered by the complex ¹³ electronic structure of many biologically relevant metals such as zinc that can often not be accurately ¹⁴ described using a classical force field. In this work, we develop two tools - Metal3D (based on 3D ¹⁵ convolutional neural networks) and Metal1D (solely based on geometric criteria) to improve the ¹⁶ identification and localization of zinc and other metal ions in experimental and computationally ¹⁷ predicted protein structures. Comparison with other currently available tools shows that Metal3D is ¹⁸ the most accurate metal ion location predictor to date outperforming geometric predictors including ¹⁹ Metal1D by a wide margin using a single structure as input. Metal3D outputs a confidence metric ²⁰ for each predicted site and works on proteins with few homologues in the protein data bank. The ²¹ predicted metal ion locations for Metal3D are within $0.70 \pm 0.64 \text{ \AA}$ of the experimental locations ²² with half of the sites below 0.5 \AA . Metal3D predicts a global metal density that can be used for ²³ annotation of structures predicted using e.g. AlphaFold2 and a per residue metal density that can ²⁴ be used in protein design workflows for the location of suitable metal binding sites and rotamer ²⁵ sampling to create novel metalloproteins. Metal3D is available as easy to use webapp, notebook or ²⁶ commandline interface.

²⁷ **Introduction**

²⁸ Metalloproteins are ubiquitous in nature and are present in all major enzyme families.^{1,2} The metals ²⁹ predominantly found in biological systems are the first and second row alkali and earth alkali metals ³⁰ and the first row transition metals such as zinc and copper. Zinc is the most common transition ³¹ metal (present in ~10% of deposited structures) and can fulfill both a structural (e.g. in zinc finger ³² proteins) or a catalytic role in up to trinuclear active sites. Zn^{2+} is an excellent Lewis acid and is most ³³ often found in tetrahedral, pentavalent, or octahedral coordination. About 10 % of all reactions ³⁴ catalyzed by enzymes use zinc as cofactor³.

³⁵ Metalloproteins are well studied because metal cofactors are essential for the function of many ³⁶ proteins and loss of this function is an important cause of diseases.⁴ Industrial applications for ³⁷ metalloproteins capitalize on the favorable catalytic properties of the metal ion where the protein ³⁸ environment dictates (stereo)-selectivity.⁵⁻⁷ To crystallize proteins, metal salts are also often added ³⁹ to the crystallization buffer as they can help in the formation of protein crystals overcoming the ⁴⁰ enthalpic cost of association of protein surfaces. Metal ion binding sites can be used to engineer ⁴¹ protein-protein interactions (PPI)¹⁰⁻¹² and the hypothesis has been put forward that one origin of ⁴² macromolecular complexity is the superficial binding of metal ions in early single domain proteins.¹²

⁴³ While simple metal ion binding sites can be rapidly engineered because initial coordination on ⁴⁴ a protein surface can for example be achieved by creating an $i, i+4$ di-histidine site on an alpha-

45 helix¹³ or by placing cysteines in spatial proximity,¹⁴ the engineering of complex metal ion binding
46 sites e.g. in the protein interior is considerably more difficult^{2,11} as such sites are often supported
47 by a network of hydrogen bonds. A complication for computational design of metalloproteins
48 is the unavailability of good (non-bonded) force fields for zinc and other transition metals that
49 accurately reproduce (e.g. tetrahedral) coordination with the correct coordination distances which
50 renders design using e.g. Rosetta very difficult.^{2,15} In fact, the latest parametrization of the Rosetta
51 energy function (ref2015)¹⁶ did not refit the parameters for the metal ions which originally are from
52 CHARMM27 with empirically derived Lazaridis-Karplus solvation terms. To adequately treat metal
53 sites in proteins quantum mechanical treatments such as in hybrid quantum mechanics/molecular
54 mechanics (QM/MM) simulations^{17,18} is needed whose computational cost is prohibitive for regular
55 protein design tasks. QM/MM simulations can however be used to verify coordination chemistry
56 for select candidate proteins.¹⁹ On the other hand, neural network potentials have been developed
57 for zinc however those require the experimental zinc location as input.²⁰

58 Many tools exist to predict whether a protein contains metals (e.g. ZincFinder²¹), which residues
59 in the protein bind a metal (e.g. IonCom,²² MIB²³) and where the metal is bound (AlphaFill,²⁴
60 Findsitemetal,²⁵ BioMetAll,²⁶ MIB²³). The input for these predictors is based on sequence and/or
61 structure information. Sequence-based predictors use pattern recognition to identify the amino
62 acids which might bind a metal.²⁷ Structure-based methods use homology to known structures
63 (MIB, Findsitemetal, AlphaFill) or distance features (BioMetAll) to infer the location of metals. Some
64 tools like Findsitemetal or ZincFinder employ machine learning based approaches such as support
65 vector machines.

66 Structure based deep learning approaches have been used in the field of protein research for
67 a variety of applications such as protein structure prediction,²⁸ prediction of identity of masked
68 residues³⁰⁻³², functional site prediction,^{33,34} for ranking of docking poses,^{35,36} prediction of the
69 location of ligands,³⁶⁻⁴⁰ and prediction of effects of mutations for stability and disease.⁴ Current
70 state of the art predictors for metal location are MIB,^{23,42} which combines structural and sequence
71 information in the "Fragment Transformation Method" to search for homologous sites in its database,
72 and BioMetAll,²⁶ a geometrical predictor based on backbone preorganization. Both methods have
73 significant drawbacks: MIB excludes metal sites with less than 2 coordination partners from its
74 analysis and is limited by the availability of templates in its database. BioMetAll does not use
75 templates but provides many possible locations for putative binding sites on a regular grid. The
76 individual probes in BioMetAll do not have a confidence metric therefore only allowing to rank sites
77 by the number of probes found, which results in a large uncertainty in the position. Both tools
78 suffer from many false positives. In this work, we present two metal ion location predictors that
79 do not suffer from these drawbacks. The deep learning based Metal3D predictor operates on a
80 voxelized representation of the protein environment and predicts a per residue metal density that
81 can be averaged to get a smooth metal probability density over the whole protein. The distance
82 based predictor Metal1D predicts the location of metals using distances mined from the protein
83 data bank (PDB) directly predicting coordinates of the putative metal binding site. These tools
84 pave the way to perform in silico design of metal ion binding sites without relying on predefined
85 geometrical rules or expensive quantum mechanical calculations.

86 **Results**

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

91 **Metal3D**

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

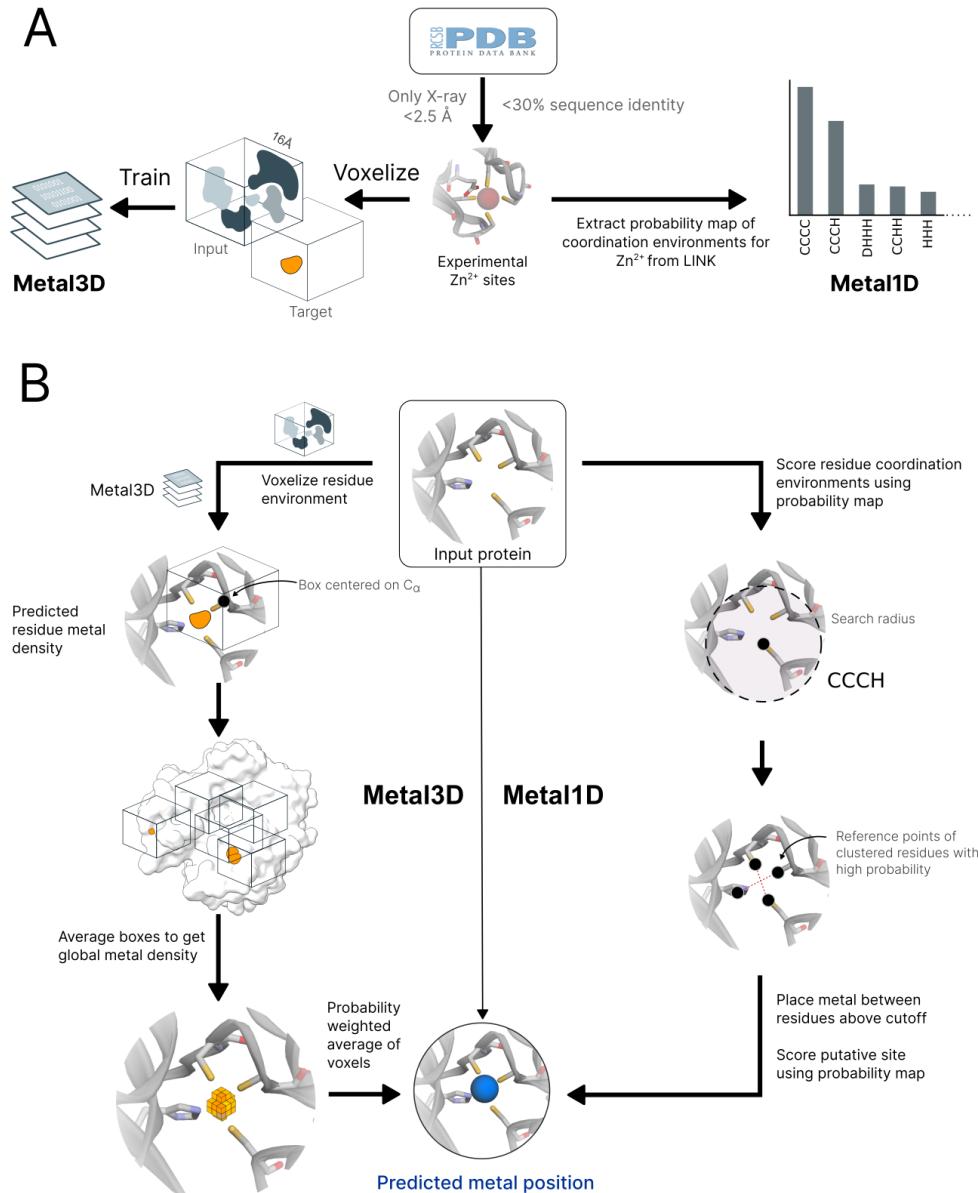


Figure 1. Workflow of Metal3D and Metal1D **A** Training of Metal3D and Metal1D is based on experimental Zn^{2+} sites. Metal1D extracts coordination environments from LINK records, Metal3D is a fully convolutional 3DCNN trained to predict the metal density from voxelized protein environments. **B** 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.

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

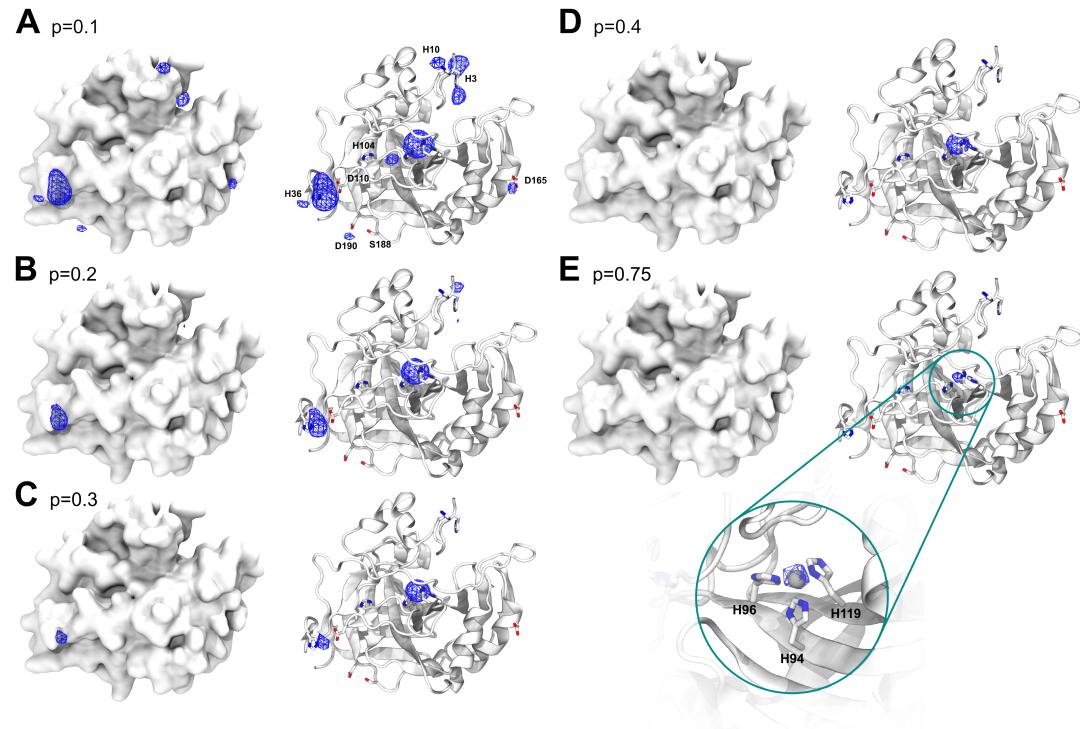


Figure 2. Metal3D probability density Probability evolution in HCA2 (PDB 2CBA) for different probability cutoffs A) $p=0.1$ B) $p=0.2$ C) $p=0.3$ D) $p=0.4$ E) $p=0.75$.

97 We evaluated the quality of the metal densities generated by the model with the discretized
 98 Jaccard similarity (Figure S1) for all environments in the test set. We noticed that at the edges of
 99 the residue-centered output densities often spurious density is predicted wherefore we evaluated
 100 the similarity of the test set metal density and the predicted metal probability density taking into
 101 account a smaller box with zeroed outer edges. Figure S1 shows that the similarity of the boxes
 102 does not depend much on the probability cutoff chosen with higher cutoffs yielding slightly higher
 103 discretized Jaccard similarity values (0.02 - 0.04 difference between $p=0.5$ and $p=0.9$). Reducing the
 104 size of the analyzed boxes (i.e trimming of the edges) increases the Jaccard similarity from ≈ 0.64 to
 105 0.88 showing that the metal density in the center of the box is more accurate than the density at
 106 the edges.

107 Metal3D is available as self-contained notebook on [Google Colab](#) and on [Huggingface Spaces](#).

108 Metal1D

109 The statistical analysis for the geometric predictor uses the LINK records present in deposited
 110 PDB structures. A probability map for all zinc coordination motifs was extracted from all training
 111 structures (Figure 1 A). The mean coordination distance in the training set was found to be 2.2 ± 0.2
 112 \AA , and the default search radius for the predictions was therefore set to 5.5 \AA (Table S1). In total
 113 208 different environments with more than 5 different proteins (at 30 % sequence identity) were
 114 identified. Metal1D is available as self-contained notebook on [Google Colab](#)

115 Comparison of Metal1D, Metal3D, MIB and BioMetAll

116 Existing metal ion predictors can be subdivided into two categories: binding site predictors and
 117 binding location predictors. The former identify only the residues binding the ion, the latter predict

118 the coordinates of the metal ion itself. Both Metal1D and Metal3D can predict the coordinates of
 119 putative binding sites. We therefore assessed their performance by comparing to recent binding
 120 location predictors with available code/webserver: BioMetAll²⁶ and MIB.²³ The main tuning pa-
 121 rameter of MIB is the template similarity t , with higher values requiring higher similarity of the
 122 templates available for the search in structurally homologous metalloproteins. BioMetAll on the
 123 other hand was calibrated on available protein structures and places probes on a regular grid at all
 124 sites where the criteria for metal binding are fulfilled. The main adjustable parameter for BioMetAll
 125 is the cluster cutoff c , which indicates how many probes in reference to the largest cluster a specific
 126 cluster has. We used the recommended cutoff of 0.5 requiring all chosen clusters to have at least
 127 50% of the probes of the most populous cluster and used the cluster center to compute distances.

128 We first investigated the potential of all tools to detect the location of a zinc ion binding site in a
 129 binary fashion (zinc site or no zinc site). We defined a correctly identified binding site (true positive,
 130 TP) as a prediction within 5 Å of an experimental zinc site. In case a tool predicted no metal within
 131 the 5 Å radius, we counted this site as false negative (FN). False positive (FP) predictions, i.e sites
 132 where a metal was placed spuriously, were clustered in a 5 Å radius and counted once per cluster.
 133 All tools were assessed against the held out test biological assemblies for Metal3D and Metal1D.
 134 When the performance of MIB ($t=1.25$) and BioMetAll is compared against Metal3D with probability
 135 cutoff $p=0.75$ we find that Metal3D identifies more sites (85) than MIB (78) or BioMetAll (75) with a
 136 much lower number of false positives (Figure 3). MIB predicts 180 false positive sites, BioMetAll
 137 134 sites whereas Metal3D only predicts 9 false positive sites at the $p=0.75$ cutoff. Metal1D ($t=0.5$)
 138 offers similar detection capabilities (78 sites detected) with a lower number of false positives (47)
 139 compared to MIB and BioMetAll. We removed 56 sites from the list of zinc sites in the test set (189
 140 total) that had less than 2 unique protein ligands within 2.8 Å of the experimental zinc location.
 141 The amount of correct predictions in this reduced set is almost unchanged for all tools (Figure 3)
 142 indicating that most tools correctly predict sites if they have 2 or more protein ligands. For Metal3D
 143 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
 144 are sites with more than 2 protein ligands. The number of false negatives is reduced for all tools by
 145 about 50 sites indicating that most tools do not predict these crystallographic artifacts that might
 146 depend on additional coordinating residues from an adjacent molecule in the crystal. Of all tools,
 147 Metal3D has the least false positives (1 FP at $p=0.9$) and the highest number of detected sites (110
 148 at $p=0.25$). The single false positive at $p=0.9$ does not contain a zinc ion but is a calcium binding site
 149 with three aspartates and one backbone carbonyl ligand (Figure S8).

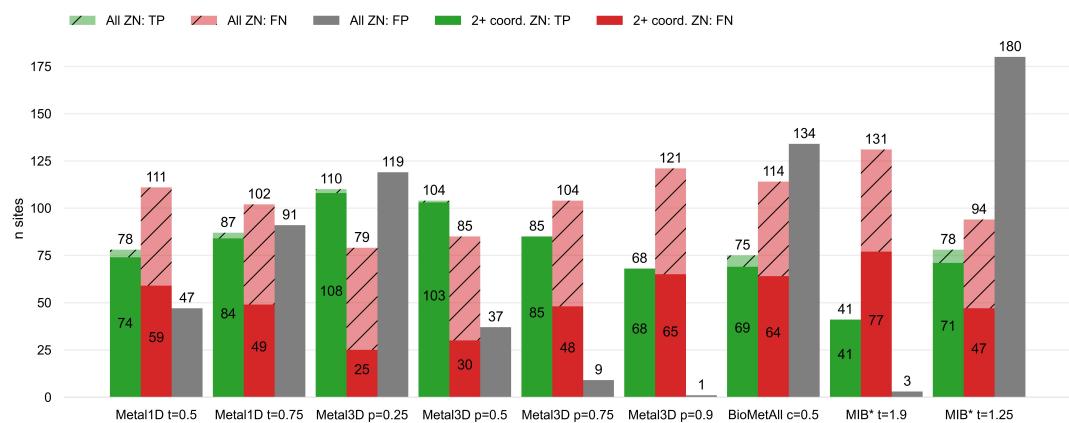


Figure 3. Identification of metal sites Comparison of Metal1D, Metal3D, BioMetAll and MIB on the test set held out from training of Metal1D and Metal3D. Predicted sites are counted as true positives (TP) if they are within 5 Å of a true metal location and as false negatives (FN) otherwise. False positive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

150 After assessment of how many sites the tools predict, another crucial metric is the spatial
151 precision of the predictions. For the correctly identified sites (TP) we measured the mean absolute
152 distance (MAD) between experimental and predicted position (Figure 4). The MAD for Metal3D at
153 $p=0.9$ is $0.70 \pm 0.64 \text{ \AA}$ and $0.74 \pm 0.66 \text{ \AA}$ at $p=0.25$ indicating that low confidence predictions are still
154 accurately placed inside the protein. The median MAD of predictions for Metal3D at $p=0.9$ is 0.52 \AA
155 indicating that for half of the predictions the model predicts at or better than the grid resolution of
156 0.5 \AA .

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

164 Selectivity for other metals

165 Both Metal3D and Metal1D were exclusively trained on zinc and we assessed their performance on
166 sodium (Na^+ , PDB code NA), potassium (K^+ , PDB code K), calcium (Ca^{2+} , PDB code CA), magnesium
167 (Mg^{2+} , PDB code MG), and various transition metals (Fe^{2+} , Fe^{3+} , Co^{2+} , Cu^{2+} , Cu^+ , Mn^{2+} with corre-
168 sponding PDB codes FE2, FE, CO, CU, CU1, MN, respectively) from 25 randomly drawn structures
169 from the clustered PDB at 30% identity. Only sites with at least 3 unique protein ligands were used
170 for the analysis to exclude crystallographic artifacts and use only highly defined sites which should
171 exhibit most selectivity towards a specific metal. Figure 4 B shows that recall for Metal3D is high
172 for all transition metals, meaning that the model correctly finds most sites even though it was only
173 trained on zinc. For the alkali and earth alkali metals recall is much lower as the model only finds
174 some sites. The mean probability for found zinc structures ($\text{ZN } p=0.95 \pm 0.10$) in the test set is higher
175 than for the other transition metals (Figure S6) and significantly higher than the probability for alkali
176 metals ($\text{NA } p=0.61 \pm 0.10$, $\text{K } p=0.79 \pm 0.16$) while the probability for the earth alkali metals is slightly
177 higher with $\text{MG } (p=0.77 \pm 0.16)$ similar to $\text{CA } (p=0.73 \pm 0.16)$. The MAD for each found metal site is
178 again lowest for zinc ($0.56 \pm 0.59 \text{ \AA}$). The MAD for the found sodium ($n=2$) and potassium ($n=5$) sites
179 are as low as for the other transition metals. The only metal with significantly higher MAD ($1.45 \pm$
180 0.93 \AA) is CU1 (Figure S5).

181 The only two structures where a sodium is detected by Metal3D (2OKQ⁴⁴, 6KFN⁴⁵) have at least
182 2 side chain coordinating ligand atoms and only one backbone (2OKQ) or no backbone ligand
183 atom (6KFN). Canonical sodium binding sites e.g. such as in PDB 4IOW⁴⁶ with two coordinating
184 backbone carbonyl oxygen atoms and one asparagine side chain have probabilities around 5 %
185 and are basically indistinguishable from background noise of the model. For Metal1D overall recall
186 is lower with similar differences in the detection of main group metals versus transition metals
187 (Figures S3 and S4).

188 Applications

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

191 Alpha Fold

192 AlphaFold2 often predicts side chains in metal ion binding sites in the holo conformation.²⁸ Tools like
193 AlphaFill²⁴ use structural homology to transplant metals from similar PDB structures to the predicted
194 structure. Metal3D does not require explicit homology based on sequence or structural alignment
195 like AlphaFill so it is potentially suited to annotate the dark proteome that is now accessible from the
196 AlphaFold database with metal binding sites. Metal3D identifies both the catalytic site (1) and the
197 zinc finger (2) for the example (PDB 3RZV⁴⁷, Figure 5 A) used in ref²⁴ with high probability ($p=0.99$)

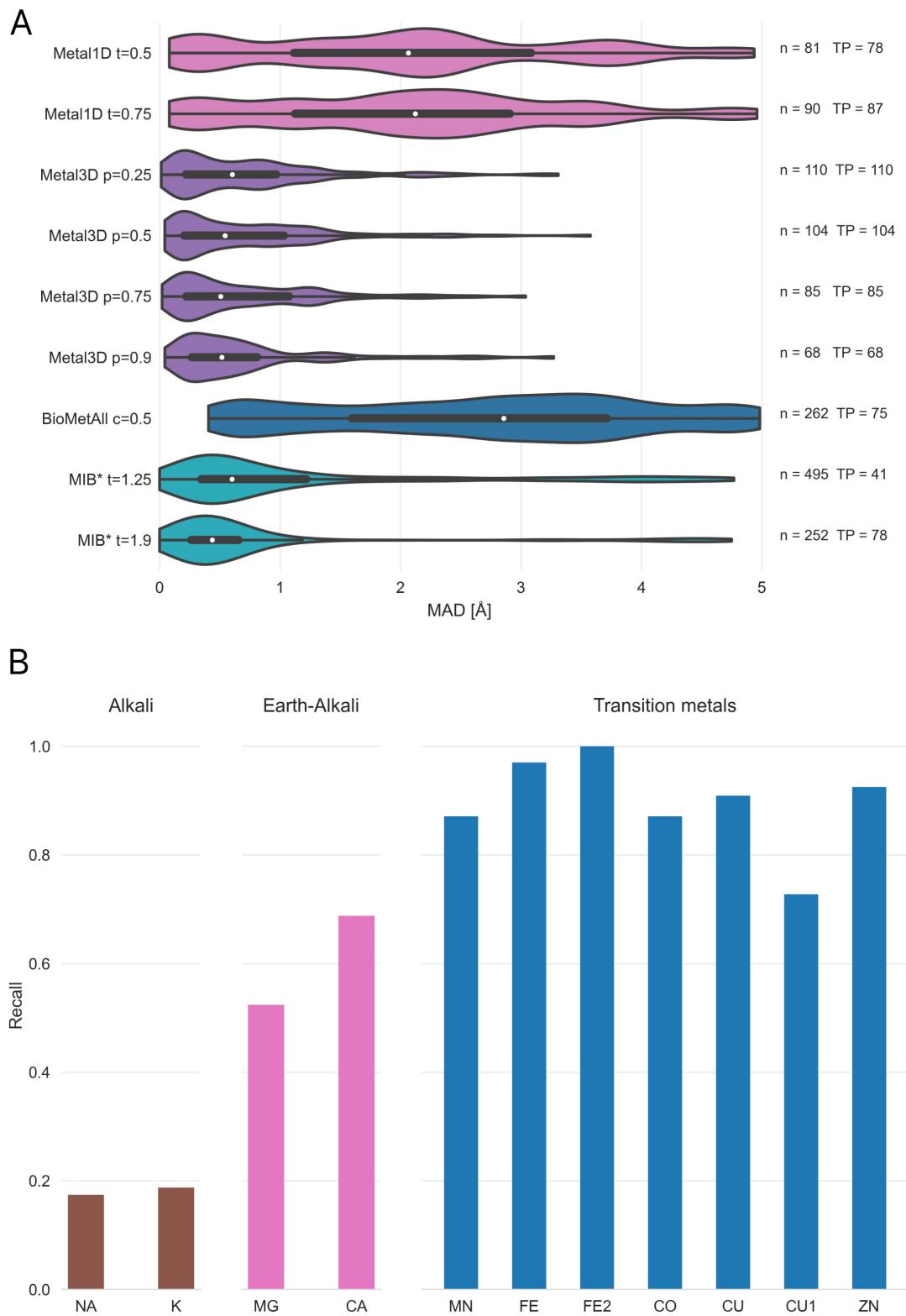


Figure 4. Precision of predicted sites and selectivity for Zn^{2+} **A** Mean absolute deviation (MAD) of predicted zinc ion locations using Metal1D, Metal3D, BioMetAll and MIB on the test set used to train Metal1D and Metal3D for all correctly identified (TP) sites. n is the number of sites predicted 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.

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

201 AlphaFill uses a 25% sequence identity cutoff which can be problematic for certain proteins with
 202 no structurally characterized homologues. For human palmitoyltransferase ZDHC23 (Uniprot
 203 Q8IYP9) a high confidence AlphaFold2 prediction exists but AlphaFill cannot place the zinc ions
 204 because the sequence identity is 24% to the closest PDB structure (PDB 6BMS⁴⁸), i.e. below the 25%
 205 cutoff. For the identical site in another human palmitoyltransferase ZDHC15 (Uniprot Q96MV8)
 206 AlphaFill is able to place the metal because of higher sequence identity to 6BMS (64%) (Figure 5 B).
 207 For ZDHC23 Metal3D is able to place the metal with high confidence (MAD 0.75 Å for site 1 and
 208 0.48 Å for site 2, p>0.99) based on the single input structure alone.

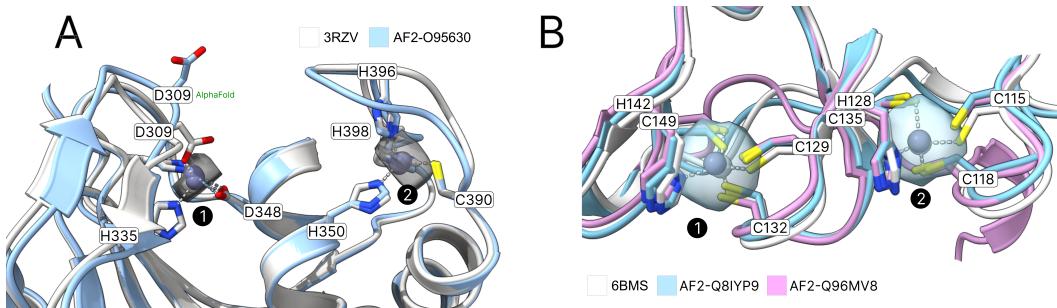


Figure 5. Annotation of AlphaFold2 structures **A** Predicted metal binding sites (a and b) from Metal3D, respectively AlphaFill compared to the experimentally found zinc positions for Uniprot O95630. Metal3D places the metal with high accuracy even if sidechains are not perfectly predicted by AlphaFold for site 1 **B** Palmitoyltransferase ZDHC23 (Uniprot Q8IYP9) and ZDHC15 (Uniprot Q96MV8). AlphaFill can only place the metal for ZDHC15 because sequence identity for ZDHC23 is only 24 %. Probability isosurfaces from Metal3D for both structures at p=0.6, colored in gray.

209 **Metalloprotein engineering**

210 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 2CBA⁴⁹), Metal3D
 211 recapitulates the location of the active site metal with a RMSD to the true metal location of 0.21
 212 Å with a probability of p=0.99. At lower probability cutoffs (p<0.4) the probability map indicates
 213 further putative metal ion binding sites with interactions mediated by surface residues (e.g. H36,
 214 D110, p=0.22) (Figure 2).

215 To investigate the capabilities for protein engineering we used mutational data for first and
 216 second shell mutants of the active site residues in HCA2 with corresponding K_d values from a
 217 colorimetric assay.⁵⁰ For most mutants no crystal structures are available so we used the structure
 218 builder in the EVOLVE package to choose the most favorable rotamer for each single point mutation
 219 based on the EVOLVE-ddg energy function with explicit zinc present (modeled using a dummy atom
 220 approach⁵¹). The analysis was run for every single mutant and the resulting probability maps from
 221 Metal3D were analyzed. For the analysis we used the maximum predicted probability as a surrogate
 222 to estimate relative changes in K_d . For mutants that decrease zinc binding drastically we observe a
 223 drop in the maximum probability predicted by Metal3D (Figure 6). The lowest probability mutants
 224 are H119N and H119Q with p=0.23 and 0.38. The mutant with the largest loss in zinc affinity H94A
 225 has a zinc binding probability of p=0.6. Conservative changes to the primary coordination motif
 226 (e.g. H → C) reduce the predicted probability by 10 - 30 %. For second shell mutants the influence of
 227 the mutations is less drastic with only minor changes in the predicted probabilities.

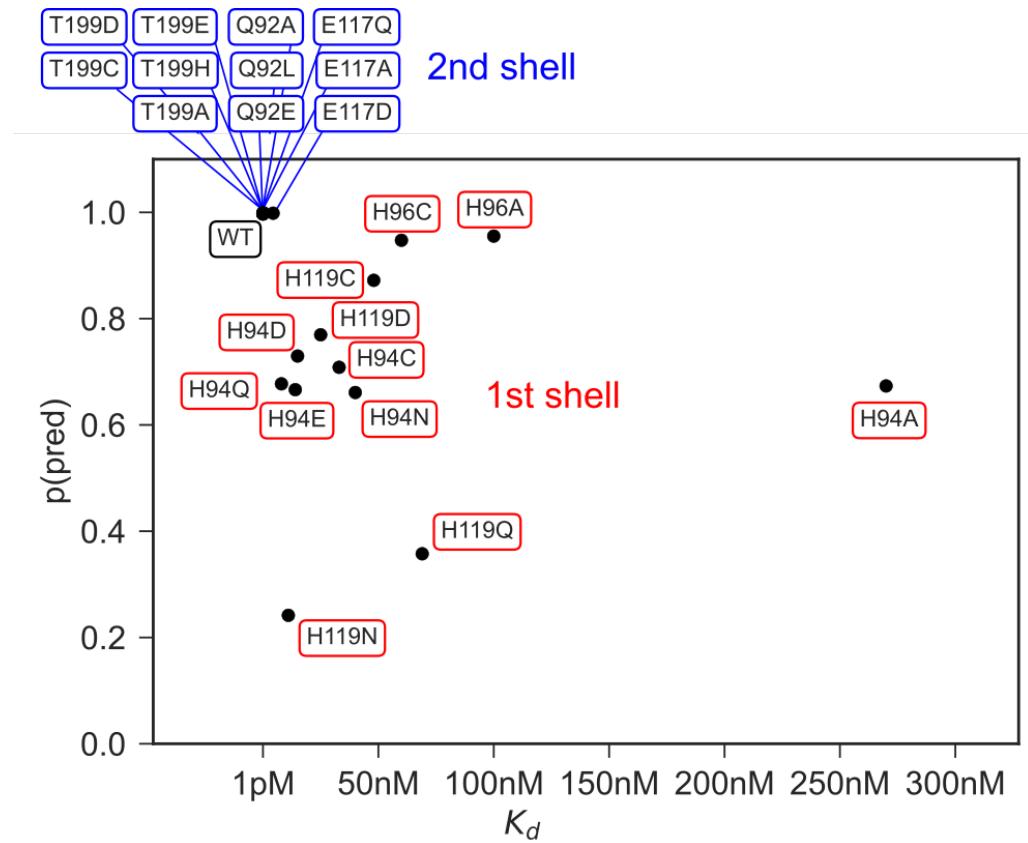


Figure 6. Protein design application Experimentally measured K_d values⁵²⁻⁵⁶ for 1st and 2nd shell active site mutants of HCA2 and predicted max probability for zinc using Metal3D.

229 **Discussion**

230 Metal3D predicts the probability distribution of zinc ions in protein crystal structures based on a
231 neuronal network model trained on natural protein environments. The model performs a segmen-
232 tation task to determine if a specific point in the input space contains a zinc ion or not. Metal3D
233 predicts zinc ion sites with high accuracy making use of high resolution crystal structures (<2.5 Å).
234 The use of high resolution structures is necessary because at resolutions greater than the average
235 zinc ligand coordination distance (2.2 Å) the uncertainty of the zinc location noticeably increases⁴³
236 which would likely hamper the accuracy of the site prediction.

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

245 **Site quality**

246 The fraction of artifactual zinc binding sites in the PDB is estimated to be about 1/3⁴³ similar to our
247 test set used with 70% (133) well coordinated zinc sites with at least 2 distinct protein ligands. To
248 reduce the amount of artifactual sites in the training set we presented the model with as many
249 complete sites as possible by using crystal symmetry to add adjacent coordinating protein chains.
250 The frequency of artifacts in the training set is therefore much lower than 30 %. The sites which still
251 remain incomplete or that are wrongly modeled and not excluded through the resolution cutoffs
252 and filtering procedures likely present only a small fraction of the training set and their signal is
253 drowned out by the numerical superiority of the correctly modeled sites. If the model is used on
254 artifactual sites or partially disordered ones it can still predict the metal location with high spatial
255 accuracy but often indicates a lower confidence for the prediction (Figures 2 and 5).

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

263 **Influence of non-protein ligands**

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

273 For Metal3D, the input channel that encodes the total heavy atom density also encodes an
274 implicit water density where all empty space can be interpreted as the solvent. For Metal1D, the
275 contribution of water molecules is considered in an implicit way when the score is assigned to a
276 site by considering coordinations including water compatible with the one observed (e.g. a HIS₃Wat
277 site is equivalent to a His₃ site for the scoring).

278 **Choice of architecture**

279 This work is the first to report a modern deep learning based model destined for identification
280 of metal ligands in proteins. Similar approaches have been used in the more general field of
281 protein-ligand docking where a variety of architectures and representations have been used. 3D
282 CNN based approaches such as LigVoxel³⁸ and DeepSite³⁷ commonly use a resolution of 1 Å and
283 similar input features as our model to predict the ligand density. However, predicting the density of
284 a multi-atomic ligand is more complex than predicting the density of mononuclear metal ions. We
285 therefore did not deem it necessary to include a conditioning on how many metal ions are present
286 in the box and rather chose to reflect this in the training data where the model needs to learn that
287 only about half of the environments it sees contain one or more metals. This choice is validated by
288 the fact that the output probability densities at sufficiently high probability cutoffs are spherical
289 with their radius approximately matching the van der Waals radius of zinc.

290 Mesh convolutional neural networks trained on a protein surface representation³⁶ also have
291 been used to predict the location and identity of protein ligands but this approach can only label
292 the regions of the surface that bind the metal ion and is conceptually not able to return the exact
293 location of the metal. Some metal ion binding sites are also heavily buried inside proteins as they
294 mediate structural stability rendering them inaccessible to a surface based approach. The most
295 recent approaches such as EquiBind³⁹ use equivariant neural networks such as En-Transformer⁶⁰
296 to predict binding keypoints (defined as 1/2 distance between the C α of the binding residue and a
297 ligand atom). Explicit side chains are still too expensive for such models and these models assume
298 a fixed known stoichiometry of the protein and ligand. Metal3D can also deal with proteins that
299 do not bind a metal and does not assume that the amount of ions is known. The lack of explicit
300 side chain information renders equivariant models unsuitable for the design of complex metal ion
301 binding sites supported by an intricate network of hydrogen bonds that need to be positioned with
302 sub-angstrom accuracy. Our model in contrast is less data- and compute-efficient than approaches
303 representing the protein as graph due to the need to voxelize the input and provide different
304 rotations of the input environment in training but the overall processing time for our model is still
305 low taking typically 25 seconds for a 250 residue protein on a multicore GPU workstation (20 CPUs,
306 GTX2070). Sequence based models^{61,62} can only use coevolution signals to infer residues in spatial
307 proximity that can bind a metal. This might be difficult when it comes to ranking similar amino
308 acids such as aspartate and glutamate or even ranking different rotamers where sub-angstrom
309 level precision is needed to identify the mutant with the highest affinity for zinc.

310 **Selectivity**

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

322 **Application for protein design**

323 Protein design using 3DCNNs trained on residue identity has been successfully demonstrated and
324 we anticipate that our model could be seamlessly integrated into such a workflow³² to enable fully
325 deep learning based design of metalloproteins. We are currently also investigating the combination
326 of Metal3D combined with a classic energy-based genetic algorithm-based optimization to make

327 design of metalloproteins¹⁹ easier without having to explicitly model the metal to compute the
328 stability of the protein. As the model computes a probability density per residue it can be readily
329 integrated into established software like Rosetta relying on rotamer sampling.

330 The HCA2 application demonstrates the utility of Metal3D for protein engineering (Figure 2).
331 The thermodynamics of metal ion binding to proteins are complicated⁶³ and there are currently
332 no high-throughput based experimental approaches that could generate a dataset large enough
333 to train a model directly on predicting K_d . The data we use were obtained from a colorimetric
334 assay with very high affinity of zinc in the picomolar range.⁵²⁻⁵⁶ More recent studies using ITC⁶³
335 instead of the colorimetric assay indicate lower K_d values in the nanomolar range for wild type HCA2.
336 We can therefore only use the colorimetric data to estimate how well the model can recapitulate
337 relative changes in the K_d for different mutations in the first and second shell of a prototypical
338 metalloprotein.

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

342 Metal1D vs. Metal3D

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

349 Conclusion

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

359 Materials and Methods

360 Dataset

361 The input PDB files for training were obtained from the RCSB⁶⁸ protein databank (downloaded 5th
362 March 2021). We use a clustering of the structures at 30% sequence identity using mmseqs2⁶⁹ to
363 largely remove sequence and structural redundancy in the input dataset. For each cluster, we check
364 whether a zinc is contained in one of the structures, whether the resolution of these structures is
365 better than 2.5 Å, if the experimental method is x-ray crystallography and whether the structure
366 does not contain nucleic acids. If there are multiple structures fulfilling these criteria, the highest
367 resolution structure is used. All structures larger than 3000 residues are discarded. We always use
368 the first biological assembly to sample the training environments. The structures were stripped
369 of all exogenous ligands except for zinc. If there are multiple models with e.g. alternative residue
370 conformations for a given structure, the first one is used. For each biological assembly we used
371 the symmetry of the asymmetric unit to generate a protein structure that contains all neighboring
372 copies of the protein in the crystal such that metal sites at crystal contacts are fully coordinated.

373 The train/val/test split was performed based on sequence identity using easy-search in mm-
374 seqs2. All proteins that had no (partial) sequence overlap with any other protein in the dataset

375 were put into the test/val set (85 proteins) which we further split into a test set of 59 structures and
376 a validation set of 26 structures. The training set contained 2085 structures. (Supplemental Data 1).

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

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

386 Metal 1D

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

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

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

403 The protein structure is analyzed using the BioPandas python library.⁷¹ To identify coordinating
404 residues, a per residue score is assigned by performing a geometrical search from a reference
405 point, defined as the coordinate of the most probable metal binding atom, within a search radius
406 considered as roughly twice the typical distance between the metal ion and the binding atom of
407 amino acids in proteins (2.2 ± 0.2 Å as determined from LINK records). The search radius used was
408 5.5 Å in order to be able to take into account also deviations from the ideal coordination. In the case
409 of amino acids which present more than one putative coordinating atom, such as e.g. histidine, the
410 mid-point between the donor atoms is used as reference point and the search radius is enlarged
411 accordingly. The atoms used as reference points for each amino acid and the increase in the search
412 radius are reported in Supplemental Table S1. The score is assigned to each amino acid considering
413 all the other reference points of other amino acids within the search radius, and summing the
414 probabilities in the probability map for coordinations compatible with the one observed. In the ideal
415 case, a score of 1 corresponds to an amino acid surrounded by all possible coordinating amino acids
416 observed in the probability map. In practice, scores result between 0 and < 1. Once all amino acids
417 in the chain are scored, the metal location predictions are made grouping the highest-scored amino
418 acids in clusters (defined as the ones within the chosen threshold with respect to the highest-scored
419 one) based on distance. This is done using `scipy.spatial.distance_matrix` and grouping
420 together highest-scored amino acids closer than twice the search radius. For each cluster, a site
421 prediction is made as a weighted average between the coordinates of the reference point of each
422 amino acid, using as weighting factor the amino acid score. For isolated amino acids with a high
423 score (e.g. a single histidine) the same score is assigned to the closest reference point from another
424 amino acid, to be able to compute the position of the metal as before. Possible artifacts resulting

425 from this fictitious score are resolved in the final step of the prediction.

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

435 Metal 3D

436 Voxelization

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

458 Model training

459 We used PyTorch 1.10⁷⁴ to train the model. All layers of the network are convolutional layers with
460 filter size 1.5 Å except for the fifth layer where a 8 Å filter is used to capture long range interactions.
461 We use zero padding to keep the size of the boxes constant. Models were trained on a workstation
462 with NVIDIA GTX3090 GPU and 32 CPU cores. Binary Cross Entropy⁷⁵ loss is used to train the model.
463 The rectified linear unit (ReLU) non-linearity is used except for the last layer which uses a sigmoid
464 function that yields the probability for zinc per voxel. A dropout layer (p = 0.1) was used between
465 the 5th and 6th layers. The network was trained using AdaDelta employing a stepped learning rate
466 (lr=0.5, γ =0.9), a batch size of 150, and 12 epochs to train.

467 Hyperparameter tuning

468 We used the ray[tune] library⁷³ to perform a hyperparameter search choosing 20 different combi-
469 nations between the following parameters with the best combination of parameters in bold.

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

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

475 Grid Averaging

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

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

490 **Visualization** We make available a command line program and interactive notebook allowing
491 the user to visualize the results. The averaged probability map is stored as a cube file. The most
492 likely metal coordinates for use in subsequent processing are stored in a pdb file. The command
493 line program uses VMD⁷⁸ to visualize the input protein and the predicted density, for the jupyter
494 notebook 3Dmol.js/py3Dmol⁷⁹ is used.

495 Evaluation

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

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

517 Precision was calculated as

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

518 Recall was calculated as

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

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

525 The Jaccard index is computed as

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

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

528 **HCA2 mutants** The data for human carbonic anhydrase 2 (HCA2) mutants was extracted
529 from refs⁵²⁻⁵⁶ and the crystal structure 2CBA^{49,80} was used. The zinc was modeled using the zinc
530 cationic dummy model forcefield⁵¹ and we verified that energy minimization produced the correct
531 coordination environment. The Richardson rotamer library⁸¹ was used with the EVOLVE-ddG
532 energy function to compute the most stable rotamer for a given mutation with the zinc present.
533 The lowest-energy mutant was used for the prediction of the location of metals using Metal3D.

534 Additional information

535 Acknowledgement

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

539 Supplemental Data

540 Code and training data are available on [Github](#) and on Zenodo.

541 Conflict of interest

542 None declared

543 Author contributions

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

547 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

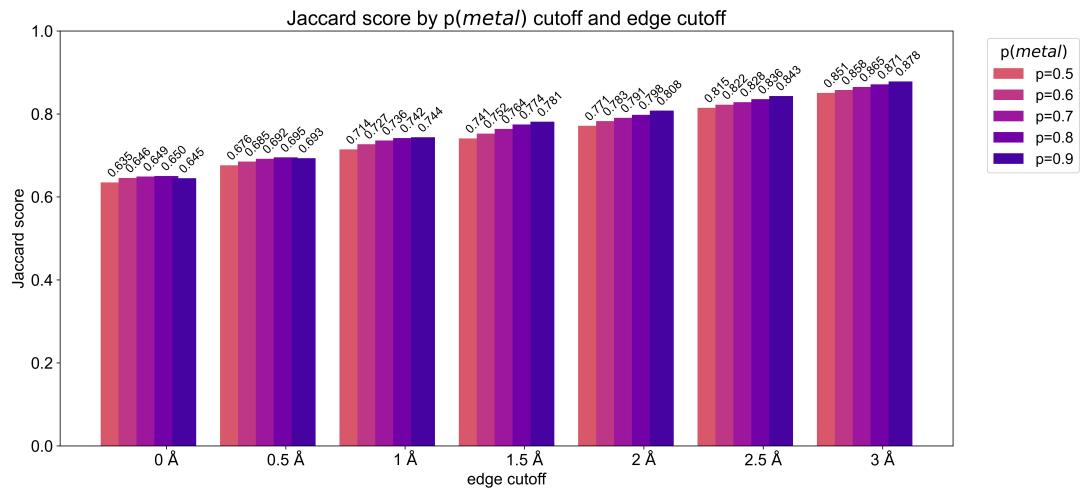


Figure S1. Discretized Jaccard indices using different cutoffs for edge trimming and different probability cutoffs ($p(\text{metal})$) showing that Metal3D predictions well reproduce the target environments in the test set.

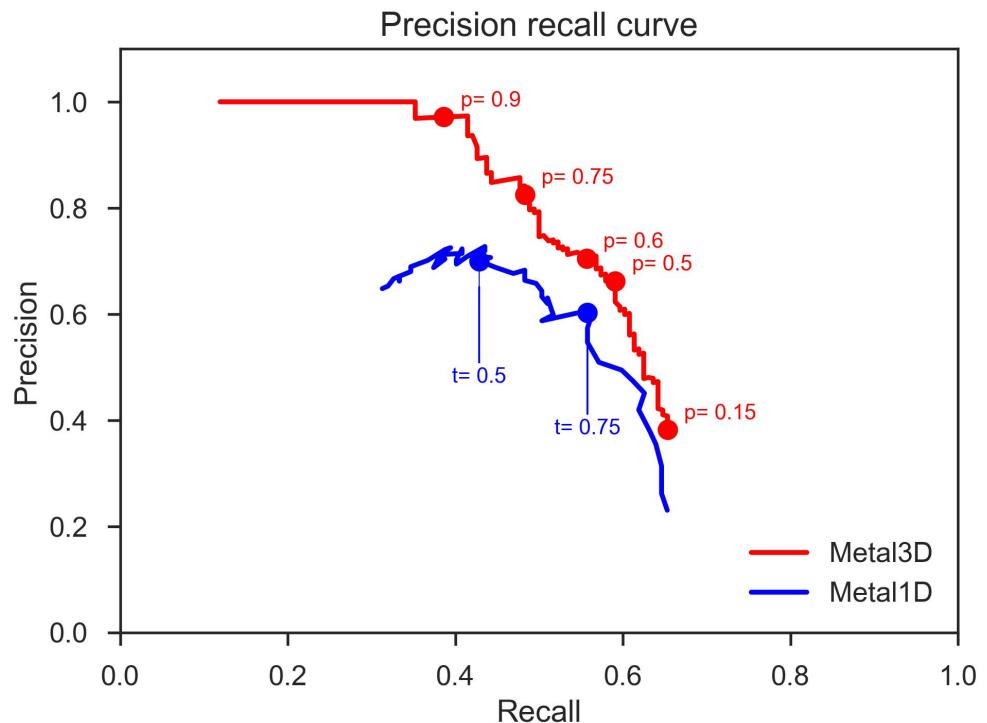


Figure S2. Precision recall curve for Metal1D and Metal3D with the probability cutoffs (Metal3D) or thresholds (Metal1D) used in the analysis.

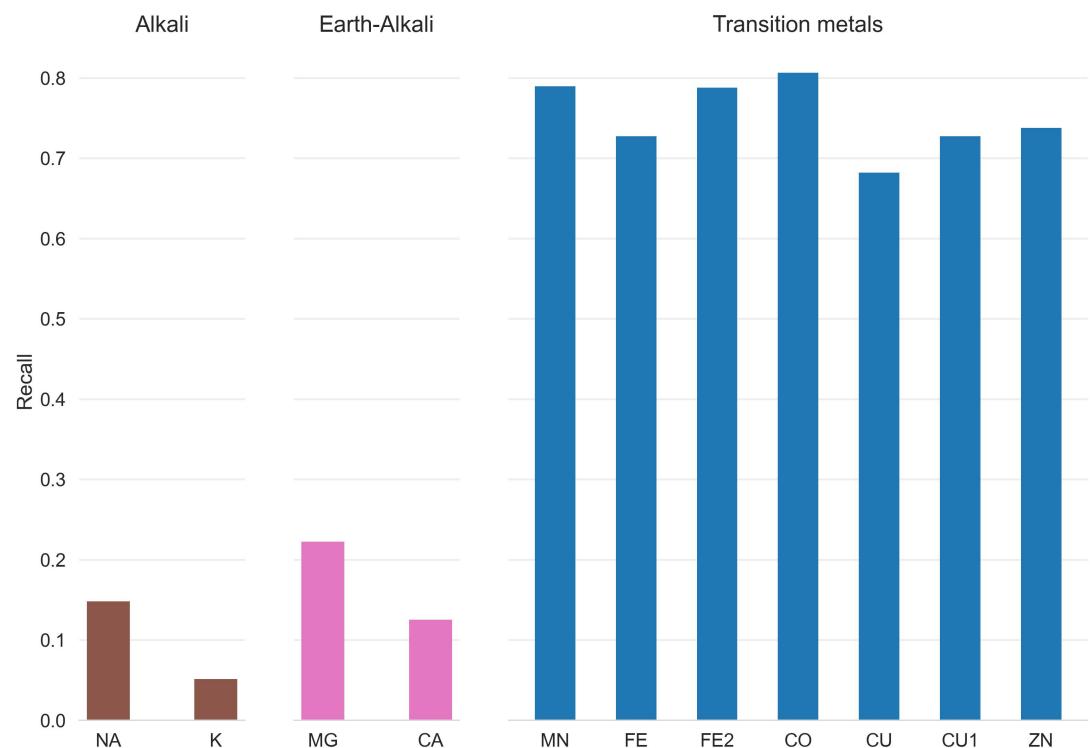


Figure S3. Recall for zinc testset and a 25 randomly drawn structures for other transition, alkali and earth-alkali metals for Metal1D

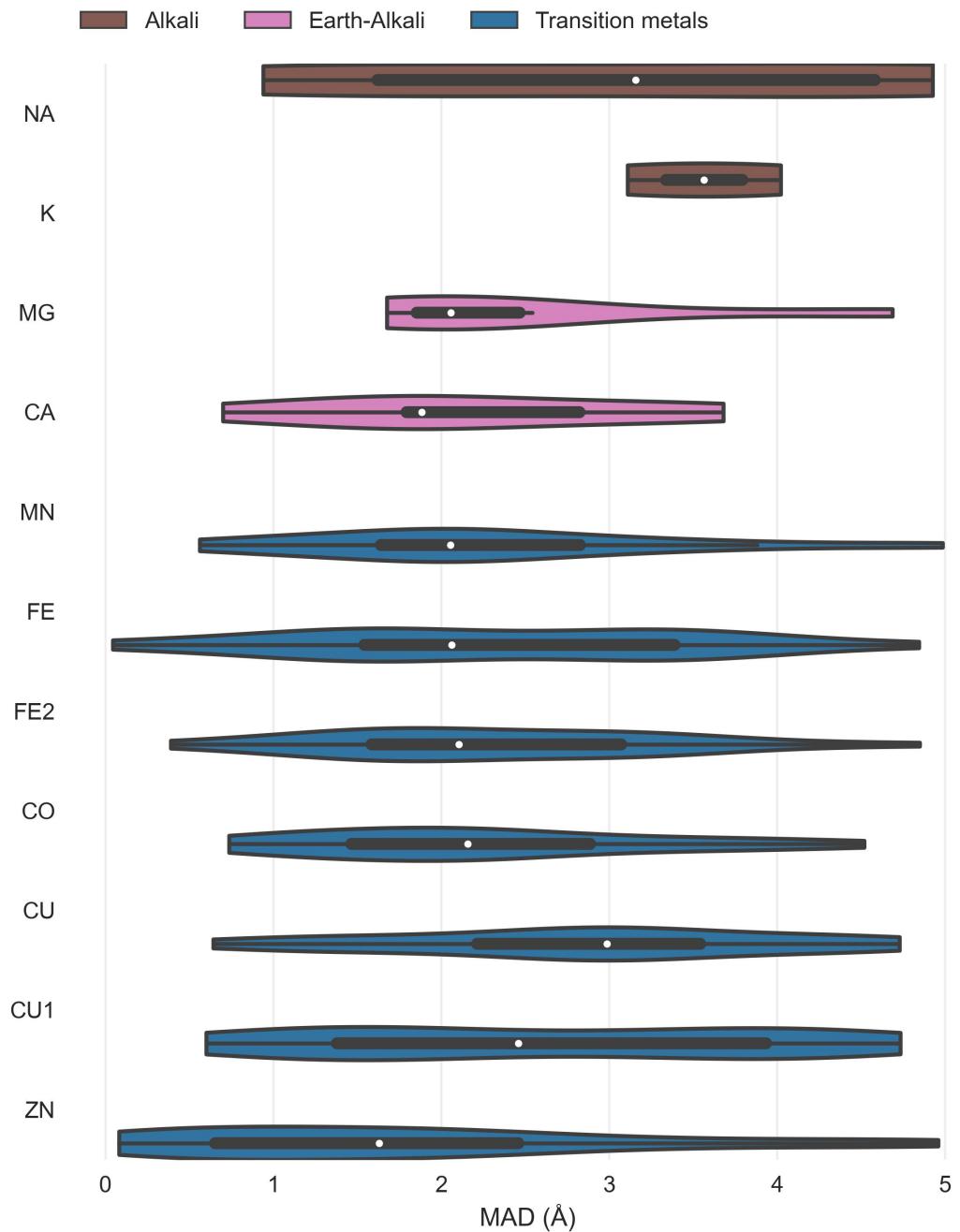


Figure S4. Distance distribution Metal1D. For each ion the whisker plot indicates the median (white dot) and the first quartiles (black box).

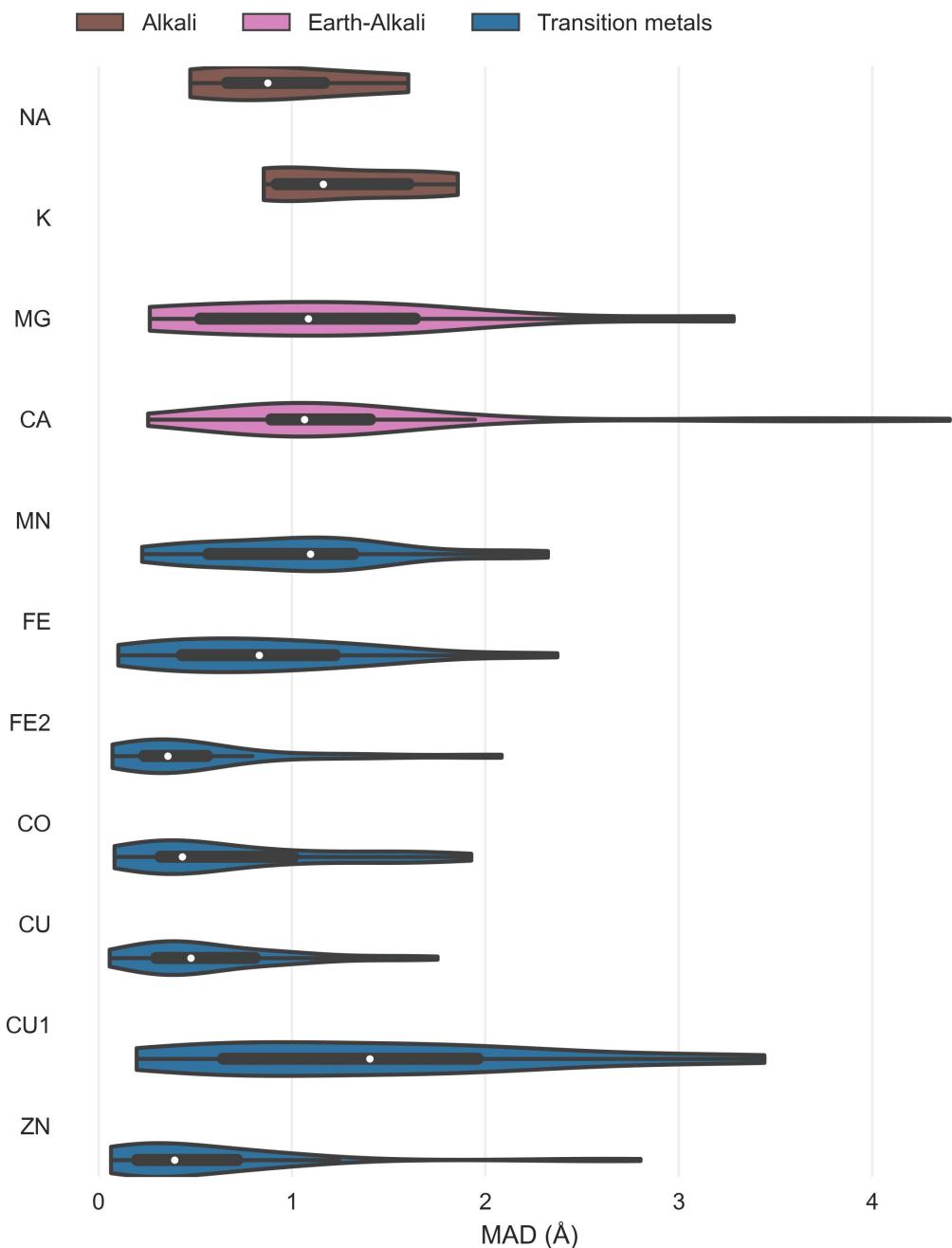


Figure S5. MAD for Metal3D for all sites with 3+ unique protein ligands in the test set and for the selected structures for the other metals. For each ion the whisker plot indicates the median (white dot) and the first quartiles (black box).

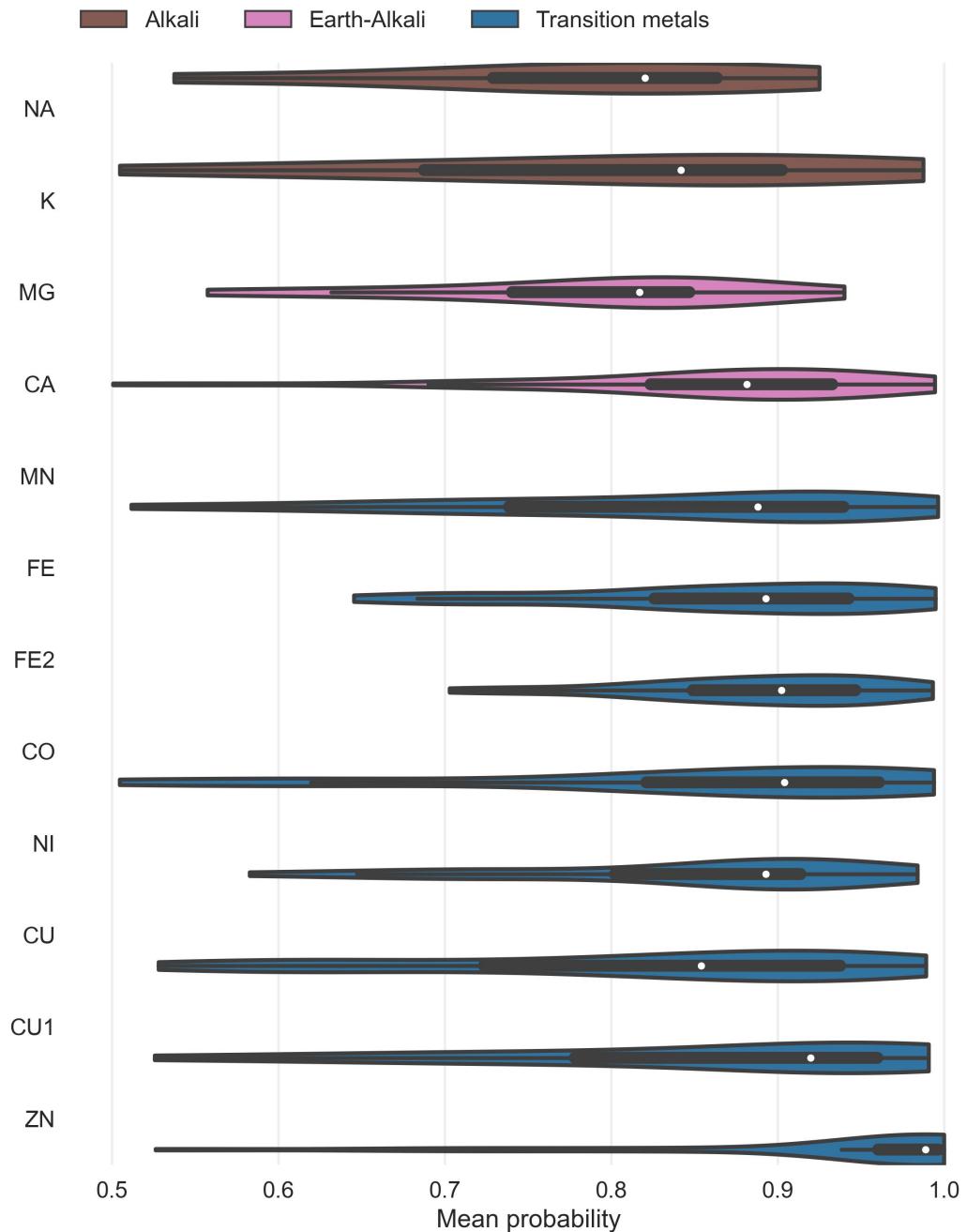


Figure S6. Probability distribution for Metal3D on sites with 3+ unique protein ligands in the test set and for the selected structures for the other metals. For each ion the whisker plot indicates the median probability (white dot) and the first quartiles (black box).

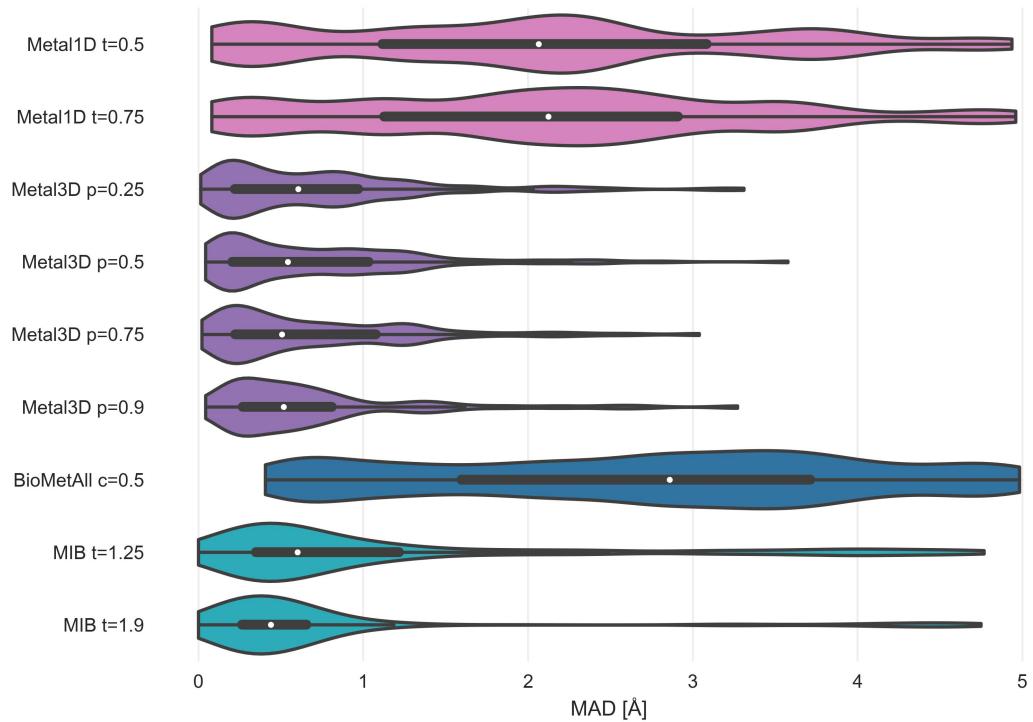


Figure S7. MAD only 2 residue coordinated zincs. For each tool the whisker plot indicates the median (white dot) and the first quartiles (black box).

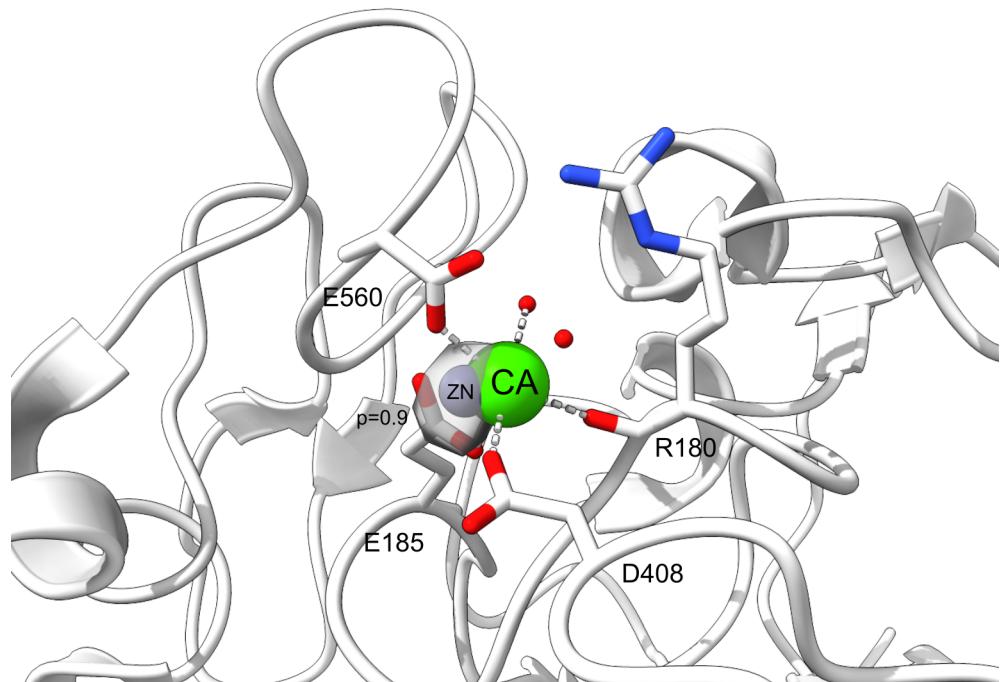


Figure S8. False positive for Metal3D at $p=0.9$ in PDB 4JJ. A calcium site is misclassified as zinc site.

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

548 **References**

- 549 (1) Yu, F.; Cangeli, V. M.; Zastrow, M. L.; Tegoni, M.; Plegaria, J. S.; Tebo, A. G.; Mocny, C. S.;
550 Ruckthong, L.; Qayyum, H.; Pecoraro, V. L. Protein Design: Toward Functional Metalloenzymes.
551 *Chem. Rev.* **2014**, 114 (7), 3495–3578. <https://doi.org/10.1021/cr400458x>.
- 552 (2) Guffy, S. L.; Der, B. S.; Kuhlman, B. Probing the Minimal Determinants of Zinc Binding with
553 Computational Protein Design. *Protein Engineering, Design and Selection* **2016**, 29 (8), 327–338.
554 <https://doi.org/10.1093/protein/gzw026>.
- 555 (3) Andreini, C.; Bertini, I.; Cavallaro, G.; Holliday, G. L.; Thornton, J. M. Metal Ions in Biological
556 Catalysis: From Enzyme Databases to General Principles. *J Biol Inorg Chem* **2008**, 13 (8), 1205–1218.
557 <https://doi.org/10.1007/s00775-008-0404-5>.
- 558 (4) Koohi-Moghadam, M.; Wang, H.; Wang, Y.; Yang, X.; Li, H.; Wang, J.; Sun, H. Predicting Disease-
559 Associated Mutation of Metal-Binding Sites in Proteins Using a Deep Learning Approach. *Nat Mach
560 Intell* **2019**, 1 (12), 561–567. <https://doi.org/10.1038/s42256-019-0119-z>.
- 561 (5) Studer, S.; Hansen, D. A.; Pianowski, Z. L.; Mittl, P. R. E.; Debon, A.; Guffy, S. L.; Der, B. S.;
562 Kuhlman, B.; Hilvert, D. Evolution of a Highly Active and Enantiospecific Metalloenzyme from Short
563 Peptides. *Science* **2018**, 362 (6420), 1285–1288. <https://doi.org/10.1126/science.aau3744>.
- 564 (6) Key, H. M.; Dydio, P.; Clark, D. S.; Hartwig, J. F. Abiological Catalysis by Artificial Haem Proteins
565 Containing Noble Metals in Place of Iron. *Nature* **2016**, 534 (7608), 534–537. [https://doi.org/10.103/nature17968](https://doi.org/10.103
566 /nature17968).
- 567 (7) Chalkley, M. J.; Mann, S. I.; DeGrado, W. F. De Novo Metalloprotein Design. *Nat Rev Chem*
568 **2021**, 6 (1), 31–50. <https://doi.org/10.1038/s41570-021-00339-5>.
- 569 (8) Der, B. S.; Edwards, D. R.; Kuhlman, B. Catalysis by a de Novo Zinc-Mediated Protein Interface:
570 Implications for Natural Enzyme Evolution and Rational Enzyme Engineering. *Biochemistry* **2012**, 51
571 (18), 3933–3940. <https://doi.org/10.1021/bi201881p>.
- 572 (9) Fujieda, N.; Schätti, J.; Stuttfeld, E.; Ohkubo, K.; Maier, T.; Fukuzumi, S.; Ward, T. R. Enzyme
573 Repurposing of a Hydrolase as an Emergent Peroxidase Upon Metal Binding. *Chem. Sci.* **2015**, 6 (7),
574 4060–4065. <https://doi.org/10.1039/c5sc01065a>.
- 575 (10) Brodin, J. D.; Ambroggio, X. I.; Tang, C.; Parent, K. N.; Baker, T. S.; Tezcan, F. A. Metal-Directed,
576 Chemically Tunable Assembly of One-, Two- and Three-Dimensional Crystalline Protein Arrays.
577 *Nature Chem* **2012**, 4 (5), 375–382. <https://doi.org/10.1038/nchem.1290>.
- 578 (11) Der, B. S.; Machius, M.; Miley, M. J.; Mills, J. L.; Szyperski, T.; Kuhlman, B. Metal-Mediated
579 Affinity and Orientation Specificity in a Computationally Designed Protein Homodimer. *J. Am. Chem.
580 Soc.* **2011**, 134 (1), 375–385. <https://doi.org/10.1021/ja208015j>.
- 581 (12) Salgado, E. N.; Radford, R. J.; Tezcan, F. A. Metal-Directed Protein Self-Assembly. *Acc. Chem.
582 Res.* **2010**, 43 (5), 661–672. <https://doi.org/10.1021/ar900273t>.
- 583 (13) Kakkis, A.; Gagnon, D.; Esselborn, J.; Britt, R. D.; Tezcan, F. A. Metal-templated Design of
584 Chemically Switchable Protein Assemblies with High-Affinity Coordination Sites. *Angew. Chem. Int.
585 Ed.* **2020**, 59 (49), 21940–21944. <https://doi.org/10.1002/anie.202009226>.
- 586 (14) Zastrow, M. L.; Peacock, A. F. A.; Stuckey, J. A.; Pecoraro, V. L. Hydrolytic Catalysis and
587 Structural Stabilization in a Designed Metalloprotein. *Nature Chem* **2011**, 4 (2), 118–123. [https://doi.org/10.1038/nchem.1201](https://
588 /doi.org/10.1038/nchem.1201).
- 589 (15) Song, L. F.; Sengupta, A.; Merz, K. M., Jr. Thermodynamics of Transition Metal Ion Binding to
590 Proteins. *J. Am. Chem. Soc.* **2020**, 142 (13), 6365–6374. <https://doi.org/10.1021/jacs.0c01329>.
- 591 (16) Alford, R. F.; Leaver-Fay, A.; Jeliazkov, J. R.; O'Meara, M. J.; DiMaio, F. P.; Park, H.; Shapovalov,
592 M. V.; Renfrew, P. D.; Mulligan, V. K.; Kappel, K.; Labonte, J. W.; Pacella, M. S.; Bonneau, R.; Bradley,
593 P.; Dunbrack, R. L., Jr.; Das, R.; Baker, D.; Kuhlman, B.; Kortemme, T.; Gray, J. J. The Rosetta All-Atom
594 Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **2017**, 13 (6),
595 3031–3048. <https://doi.org/10.1021/acs.jctc.7b00125>.
- 596 (17) Brunk, E.; Rothlisberger, U. Mixed Quantum Mechanical/Molecular Mechanical Molecular
597 Dynamics Simulations of Biological Systems in Ground and Electronically Excited States. *Chem. Rev.*

- 598 **2015**, *115* (12), 6217–6263. <https://doi.org/10.1021/cr500628b>.
- 599 (18) Yang, Z.; Twidale, R. M.; Gervasoni, S.; Suardíaz, R.; Colenso, C. K.; Lang, E. J. M.; Spencer, J.;
600 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>.
- 602 (19) Bozkurt, E.; Perez, M. A. S.; Hovius, R.; Browning, N. J.; Rothlisberger, U. Genetic Algorithm
603 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>.
- 605 (20). <https://doi.org/10.3389/fchem.2021.692200>.
- 606 (21) Passerini, A.; Andreini, C.; Menchetti, S.; Rosato, A.; Frasconi, P. Predicting Zinc Binding at
607 the Proteome Level. *BMC Bioinformatics* **2007**, *8* (1). <https://doi.org/10.1186/1471-2105-8-39>.
- 608 (22) Hu, X.; Dong, Q.; Yang, J.; Zhang, Y. Recognizing Metal and Acid Radical Ion-Binding Sites
609 by Integrating ab Initio Modeling with Template-Based Transfers. *Bioinformatics* **2016**, *32* (21),
610 3260–3269. <https://doi.org/10.1093/bioinformatics/btw396>.
- 611 (23) Lin, Y.-F.; Cheng, C.-W.; Shih, C.-S.; Hwang, J.-K.; Yu, C.-S.; Lu, C.-H. MIB: Metal Ion-Binding
612 Site Prediction and Docking Server. *J. Chem. Inf. Model.* **2016**, *56* (12), 2287–2291. <https://doi.org/10.1021/acs.jcim.6b00407>.
- 614 (24) Hekkelman, M. L.; de Vries, I.; Joosten, R. P.; Perrakis, A. AlphaFill: Enriching the AlphaFold
615 Models with Ligands and Co-Factors, 2021. <https://doi.org/10.1101/2021.11.26.470110>.
- 616 (25) Brylinski, M.; Skolnick, J. FINDSITE-Metal: Integrating Evolutionary Information and Machine
617 Learning for Structure-Based Metal-Binding Site Prediction at the Proteome Level. *Proteins* **2010**, *79*
618 (3), 735–751. <https://doi.org/10.1002/prot.22913>.
- 619 (26) Sánchez-Aparicio, J.-E.; Tiessler-Sala, L.; Velasco-Carneros, L.; Roldán-Martín, L.; Sciortino, G.;
620 Maréchal, J.-D. BioMetAll: Identifying Metal-Binding Sites in Proteins from Backbone Preorganization.
621 *J. Chem. Inf. Model.* **2020**, *61* (1), 311–323. <https://doi.org/10.1021/acs.jcim.0c00827>.
- 622 (27) Haberal, İ.; Oğul, H. Prediction of Protein Metal Binding Sites Using Deep Neural Networks.
623 *Mol. Inf.* **2019**, *38* (7), 1800169. <https://doi.org/10.1002/minf.201800169>.
- 624 (28) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool,
625 K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie,
626 A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.;
627 Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.;
628 Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction
629 with AlphaFold. *Nature* **2021**, *596* (7873), 583–589. <https://doi.org/10.1038/s41586-021-03819-2>.
- 630 (29) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.;
631 Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; Millán, C.; Park, H.; Adams, C.; Glassman, C. R.; DeGiovanni,
632 A.; Pereira, J. H.; Rodrigues, A. V.; van Dijk, A. A.; Ebrecht, A. C.; Opperman, D. J.; Sagmeister, T.;
633 Buhlheller, C.; Pavkov-Keller, T.; Rathinaswamy, M. K.; Dalwadi, U.; Yip, C. K.; Burke, J. E.; Garcia,
634 K. C.; Grishin, N. V.; Adams, P. D.; Read, R. J.; Baker, D. Accurate Prediction of Protein Structures
635 and Interactions Using a Three-Track Neural Network. *Science* **2021**, *373* (6557), 871–876. <https://doi.org/10.1126/science.abj8754>.
- 637 (30) Torng, W.; Altman, R. B. 3D Deep Convolutional Neural Networks for Amino Acid Environment
638 Similarity Analysis. *BMC Bioinformatics* **2017**, *18* (1). <https://doi.org/10.1186/s12859-017-1702-0>.
- 639 (31) Shroff, R.; Cole, A. W.; Diaz, D. J.; Morrow, B. R.; Donnell, I.; Annapareddy, A.; Gollighar, J.;
640 Ellington, A. D.; Thyer, R. Discovery of Novel Gain-of-Function Mutations Guided by Structure-Based
641 Deep Learning. *ACS Synth. Biol.* **2020**, *9* (11), 2927–2935. <https://doi.org/10.1021/acssynbio.0c00345>.
- 642 (32) Anand, N.; Eguchi, R.; Mathews, I. I.; Perez, C. P.; Derry, A.; Altman, R. B.; Huang, P.-S. Protein
643 Sequence Design with a Learned Potential. *Nat Commun* **2022**, *13* (1). <https://doi.org/10.1038/s41467-022-28313-9>.
- 645 (33) Torng, W.; Altman, R. B. High Precision Protein Functional Site Detection Using 3D Convolutional
646 Neural Networks. *Bioinformatics* **2018**, *35* (9), 1503–1512. <https://doi.org/10.1093/bioinformatics/bty813>.

- 648 (34) Feehan, R.; Franklin, M. W.; Slusky, J. S. G. Machine Learning Differentiates Enzymatic and
649 Non-Enzymatic Metals in Proteins. *Nat Commun* **2021**, 12 (1). <https://doi.org/10.1038/s41467-021-24070-3>.
- 650 (35) Renaud, N.; Geng, C.; Georgievska, S.; Ambrosetti, F.; Ridder, L.; Marzella, D. F.; Réau,
651 M. F.; Bonvin, A. M. J. J.; Xue, L. C. DeepRank: A Deep Learning Framework for Data Mining 3D
652 Protein-Protein Interfaces. *Nat Commun* **2021**, 12 (1). <https://doi.org/10.1038/s41467-021-27396-0>.
- 653 (36) Gainza, P.; Sverrisson, F.; Monti, F.; Rodolà, E.; Boscaini, D.; Bronstein, M. M.; Correia, B.
654 E. Deciphering Interaction Fingerprints from Protein Molecular Surfaces Using Geometric Deep
655 Learning. *Nat Methods* **2019**, 17 (2), 184–192. <https://doi.org/10.1038/s41592-019-0666-6>.
- 656 (37) Jiménez, J.; Doerr, S.; Martínez-Rosell, G.; Rose, A. S.; De Fabritiis, G. DeepSite: Protein-Binding
657 Site Predictor Using 3D-Convolutional Neural Networks. *Bioinformatics* **2017**, 33 (19), 3036–3042.
658 <https://doi.org/10.1093/bioinformatics/btx350>.
- 659 (38) Skalic, M.; Varela-Rial, A.; Jiménez, J.; Martínez-Rosell, G.; De Fabritiis, G. LigVoxel: Inpainting
660 Binding Pockets Using 3D-Convolutional Neural Networks. *Bioinformatics* **2018**, 35 (2), 243–250.
661 <https://doi.org/10.1093/bioinformatics/bty583>.
- 662 (39) Stärk, H.; Ganea, O.-E.; Pattanaik, L.; Barzilay, R.; Jaakkola, T. EquiBind: Geometric Deep
663 Learning for Drug Binding Structure Prediction. arXiv 2022. <https://doi.org/10.48550/arxiv.2202.05146>.
- 664 (40) Park, S.; Seok, C. GalaxyWater-CNN: Prediction of Water Positions on the Protein Structure
665 by a 3D-Convolutional Neural Network. *J. Chem. Inf. Model.* **2022**, 62 (13), 3157–3168. <https://doi.org/10.1021/acs.jcim.2c00306>.
- 666 (41) Li, B.; Yang, Y. T.; Capra, J. A.; Gerstein, M. B. Predicting Changes in Protein Thermodynamic
667 Stability Upon Point Mutation with Deep 3D Convolutional Neural Networks. *PLoS Comput Biol* **2020**,
668 16 (11), e1008291. <https://doi.org/10.1371/journal.pcbi.1008291>.
- 669 (42) Lu, C.-H.; Lin, Y.-F.; Lin, J.-J.; Yu, C.-S. Prediction of Metal Ion-Binding Sites in Proteins Using
670 the Fragment Transformation Method. *PLoS ONE* **2012**, 7 (6), e39252. <https://doi.org/10.1371/journal.pone.0039252>.
- 671 (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>.
- 672 (44) Minasov, G.; Vorontsov, I. I.; Shuvalova, L.; Brunzelle, J. S.; Kiryukhina, O.; Collart, F. R.;
673 Joachimiak, A.; Anderson, W. F.; Shigella Flexneri, 2007. <https://doi.org/10.2210/pdb2okq/pdb>.
- 674 (45) Itoh, T.; Nakagawa, E.; Yoda, M.; Nakaichi, A.; Hibi, T.; Kimoto, H. Crystal Structure of Alginate
675 Lyase from Paenibacillus Sp. Str. FPU-7, 2019. <https://doi.org/10.2210/pdb6kfn/pdb>.
- 676 (46) Adams, C. M.; Eckenroth, B. E.; Doublie, S. Structure of the Clostridium Perfringens CspB
677 Protease, 2013. <https://doi.org/10.2210/pdb4i0w/pdb>.
- 678 (47) Davies, C. W.; Das, C. The Crystal Structure of a E280A Mutant of the Catalytic Domain of
679 AMSH, 2011. <https://doi.org/10.2210/pdb3rzv/pdb>.
- 680 (48) Rana, M. S.; Kumar, P.; Lee, C.-J.; Verardi, R.; Rajashankar, K. R.; Banerjee, A. Fatty Acyl
681 Recognition and Transfer by an Integral Membrane S-Acyltransferase. *Science* **2018**, 359 (6372).
682 <https://doi.org/10.1126/science.aoa6326>.
- 683 (49) Hakansson, K.; Carlsson, M.; Svensson, L. A.; Liljas, A. STRUCTURE OF NATIVE AND APO
684 CARBONIC ANHYDRASE II AND SOME OF ITS ANION-LIGAND COMPLEXES, 1993. <https://doi.org/10.2210/pdb2cba/pdb>.
- 685 (50) Hunt, J. B.; Neece, S. H.; Ginsburg, A. The Use of 4-(2-Pyridylazo)resorcinol in Studies of Zinc
686 Release from Escherichia Coli Aspartate Transcarbamoylase. *Anal Biochem* **1985**, 146 (1), 150–157.
687 [https://doi.org/10.1016/0003-2697\(85\)90409-9](https://doi.org/10.1016/0003-2697(85)90409-9).
- 688 (51) Pang, Y. P.; Xu, K.; Yazal, J. E.; Prendergas, F. G. Successful Molecular Dynamics Simulation of
689 the Zinc-Bound Farnesyltransferase Using the Cationic Dummy Atom Approach. *Protein Sci* **2000**, 9
690 (10), 1857–1865.

- 698 (52) Kiefer, L. L.; Fierke, C. A. Functional Characterization of Human Carbonic Anhydrase II
699 Variants with Altered Zinc Binding Sites. *Biochemistry* **1994**, 33 (51), 15233–15240. <https://doi.org/10.1021/bi00255a003>.
- 700 (53) Kiefer, L. L.; Ippolito, J. A.; Fierke, C. A.; Christianson, D. W. Redesigning the Zinc Binding Site
701 of Human Carbonic Anhydrase II: Structure of a His2Asp-Zn²⁺ Metal Coordination Polyhedron. *J.
702 Am. Chem. Soc.* **1993**, 115 (26), 12581–12582. <https://doi.org/10.1021/ja00079a046>.
- 703 (54) Ippolito, J. A.; Christianson, D. W. Structure of an Engineered His3 Cys Zinc Binding Site in
704 Human Carbonic Anhydrase II. *Biochemistry* **1993**, 32 (38), 9901–9905. <https://doi.org/10.1021/bi00089a005>.
- 705 (55) Ippolito, J. A.; Baird, T. T., Jr; McGee, S. A.; Christianson, D. W.; Fierke, C. A. Structure-Assisted
706 Redesign of a Protein-Zinc-Binding Site with Femtomolar Affinity. *Proc. Natl. Acad. Sci. U.S.A.* **1995**,
707 92 (11), 5017–5021. <https://doi.org/10.1073/pnas.92.11.5017>.
- 708 (56) Huang, C.-c.; Lesburg, C. A.; Kiefer, L. L.; Fierke, C. A.; Christianson, D. W. Reversal of the
709 Hydrogen Bond to Zinc Ligand Histidine-119 Dramatically Diminishes Catalysis and Enhances Metal
710 Equilibration Kinetics in Carbonic Anhydrase II. *Biochemistry* **1996**, 35 (11), 3439–3446. <https://doi.org/10.1021/bi9526692>.
- 711 (57) Savage, H.; Wlodawer, A. Determination of Water Structure Around Biomolecules Using
712 X-Ray and Neutron Diffraction Methods. *Methods Enzymol* **1986**, 127, 162–183. [https://doi.org/10.1016/0076-6879\(86\)27014-7](https://doi.org/10.1016/0076-6879(86)27014-7).
- 713 (58) Morozenko, A.; Stuchebrukhov, A. A. Dowser++, a New Method of Hydrating Protein Struc-
714 tures. *Proteins* **2016**, 84 (10), 1347–1357. <https://doi.org/10.1002/prot.25081>.
- 715 (59) Sridhar, A.; Ross, G. A.; Biggin, P. C. Waterdock 2.0: Water Placement Prediction for Holo-
716 Structures with a Pymol Plugin. *PLoS ONE* **2017**, 12 (2), e0172743. <https://doi.org/10.1371/journal.pone.0172743>.
- 717 (60) Satorras, V. G.; Hoogeboom, E.; Welling, M. E(n) Equivariant Graph Neural Networks. arXiv
718 2021. <https://doi.org/10.48550/arxiv.2102.09844>.
- 719 (61) Gligorijević, V.; Berenberg, D.; Ra, S.; Watkins, A.; Kelow, S.; Cho, K.; Bonneau, R. Function-
720 Guided Protein Design by Deep Manifold Sampling, 2021. <https://doi.org/10.1101/2021.12.22.473759>.
- 721 (62) Greener, J. G.; Moffat, L.; Jones, D. T. Design of Metalloproteins and Novel Protein Folds
722 Using Variational Autoencoders. *Sci Rep* **2018**, 8 (1). <https://doi.org/10.1038/s41598-018-34533-1>.
- 723 (63) Song, H.; Wilson, D. L.; Farquhar, E. R.; Lewis, E. A.; Emerson, J. P. Revisiting Zinc Coordination
724 in Human Carbonic Anhydrase II. *Inorg. Chem.* **2012**, 51 (20), 11098–11105. <https://doi.org/10.1021/ic301645j>.
- 725 (64) Handel, T. M.; Williams, S. A.; DeGrado, W. F. Metal Ion-Dependent Modulation of the
726 Dynamics of a Designed Protein. *Science* **1993**, 261 (5123), 879–885. <https://doi.org/10.1126/science.8346440>.
- 727 (65) Arnold, F. H.; Haymore, B. L. Engineered Metal-Binding Proteins: Purification to Protein
728 Folding. *Science* **1991**, 252 (5014), 1796–1797. <https://doi.org/10.1126/science.1648261>.
- 729 (66) Krantz, B. A.; Sosnick, T. R. *Nat. Struct. Biol.* **2001**, 8 (12), 1042–1047. <https://doi.org/10.1038/nsb723>.
- 730 (67) Tunyasuvunakool, K.; Adler, J.; Wu, Z.; Green, T.; Zielinski, M.; Žídek, A.; Bridgland, A.; Cowie,
731 A.; Meyer, C.; Laydon, A.; Velankar, S.; Kleywelt, G. J.; Bateman, A.; Evans, R.; Pritzel, A.; Figurnov, M.;
732 Ronneberger, O.; Bates, R.; Kohl, S. A. A.; Potapenko, A.; Ballard, A. J.; Romera-Paredes, B.; Nikolov,
733 S.; Jain, R.; Clancy, E.; Reiman, D.; Petersen, S.; Senior, A. W.; Kavukcuoglu, K.; Birney, E.; Kohli,
734 P.; Jumper, J.; Hassabis, D. Highly Accurate Protein Structure Prediction for the Human Proteome.
735 *Nature* **2021**, 596 (7873), 590–596. <https://doi.org/10.1038/s41586-021-03828-1>.
- 736 (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>.
- 737 (69) Steinegger, M.; Söding, J. MMseqs2 Enables Sensitive Protein Sequence Searching for the
738 Analysis of Massive Data Sets. *Nat Biotechnol* **2017**, 35 (11), 1026–1028. <https://doi.org/10.1038/nbt.1440>.

- 749 3988.
- 750 (70) Barber-Zucker, S.; Shaanan, B.; Zarivach, R. Transition Metal Binding Selectivity in Proteins
751 and Its Correlation with the Phylogenomic Classification of the Cation Diffusion Facilitator Protein
752 Family. *Sci Rep* **2017**, 7 (1). <https://doi.org/10.1038/s41598-017-16777-5>.
- 753 (71) Raschka, S. BioPandas: Working with Molecular Structures in Pandas DataFrames. *JOSS*
754 **2017**, 2 (14), 279. <https://doi.org/10.21105/joss.00279>.
- 755 (72) Doerr, S.; Harvey, M. J.; Noé, F.; De Fabritiis, G. HTMD: High-Throughput Molecular Dynamics
756 for Molecular Discovery. *J. Chem. Theory Comput.* **2016**, 12 (4), 1845–1852. <https://doi.org/10.1021/acs.jctc.6b00049>.
- 758 (73) Moritz, P.; Nishihara, R.; Wang, S.; Tumanov, A.; Liaw, R.; Liang, E.; Elibol, M.; Yang, Z.; Paul,
759 W.; Jordan, M. I.; Stoica, I. Ray: A Distributed Framework for Emerging AI Applications. arXiv 2017.
760 <https://doi.org/10.48550/arxiv.1712.05889>.
- 761 (74) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.;
762 Gimelshein, N.; Antiga, L.; Desmaison, A.; Köpf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.;
763 Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; Chintala, S. PyTorch: An Imperative Style, High-
764 Performance Deep Learning Library. arXiv 2019. <https://doi.org/10.48550/arxiv.1912.01703>.
- 765 (75) de Boer, P.-T.; Kroese, D. P.; Mannor, S.; Rubinstein, R. Y. A Tutorial on the Cross-Entropy
766 Method. *Ann Oper Res* **2005**, 134 (1), 19–67. <https://doi.org/10.1007/s10479-005-5724-z>.
- 767 (76) Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.;
768 Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K.
769 J.; Mayorov, N.; Nelson, A. R. J.; Jones, E.; Kern, R.; Larson, E.; Carey, C. J.; Polat, İ.; Feng, Y.; Moore,
770 E. W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris,
771 C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; van Mulbregt, P.; Vijaykumar, A.; Bardelli, A.
772 P.; Rothberg, A.; Hilboll, A.; Kloeckner, A.; Scopatz, A.; Lee, A.; Rokem, A.; Woods, C. N.; Fulton, C.;
773 Masson, C.; Häggström, C.; Fitzgerald, C.; Nicholson, D. A.; Hagen, D. R.; Pasechnik, D. V.; Olivetti, E.;
774 Martin, E.; Wieser, E.; Silva, F.; Lenders, F.; Wilhelm, F.; Young, G.; Price, G. A.; Ingold, G.-L.; Allen,
775 G. E.; Lee, G. R.; Audren, H.; Probst, I.; Dietrich, J. P.; Silterra, J.; Webber, J. T.; Slavič, J.; Nothman,
776 J.; Buchner, J.; Kulick, J.; Schönberger, J. L.; de Miranda Cardoso, J. V.; Reimer, J.; Harrington, J.;
777 Rodríguez, J. L. C.; Nunez-Iglesias, J.; Kuczynski, J.; Tritz, K.; Thoma, M.; Newville, M.; Kümmerer,
778 M.; Bolingbroke, M.; Tartre, M.; Pak, M.; Smith, N. J.; Nowaczyk, N.; Shebanov, N.; Pavlyk, O.;
779 Brodtkorb, P. A.; Lee, P.; McGibbon, R. T.; Feldbauer, R.; Lewis, S.; Tygier, S.; Sievert, S.; Vigna, S.;
780 Peterson, S.; More, S.; Pudlik, T.; Oshima, T.; Pingel, T. J.; Robitaille, T. P.; Spura, T.; Jones, T. R.;
781 Cera, T.; Leslie, T.; Zito, T.; Krauss, T.; Upadhyay, U.; Halchenko, Y. O.; Vázquez-Baeza, Y.; SciPy 1.0:
782 Fundamental Algorithms for Scientific Computing in Python. *Nat Methods* **2020**, 17 (3), 261–272.
783 <https://doi.org/10.1038/s41592-019-0686-2>.
- 784 (77) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.;
785 Müller, A.; Nothman, J.; Louppe, G.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos,
786 A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python.
787 *arXiv* **2012**. <https://doi.org/10.48550/arxiv.1201.0490>.
- 788 (78) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J Mol Graph* **1996**, 14
789 (1), 33–38, 27–28. [https://doi.org/10.1016/0263-7855\(96\)00018-5](https://doi.org/10.1016/0263-7855(96)00018-5).
- 790 (79) Rego, N.; Koes, D. 3Dmol.js: Molecular Visualization with WebGL. *Bioinformatics* **2014**, 31 (8),
791 1322–1324. <https://doi.org/10.1093/bioinformatics/btu829>.
- 792 (80) Håkansson, K.; Carlsson, M.; Svensson, L.; Liljas, A. Structure of Native and Apo Carbonic
793 Anhydrase II and Structure of Some of Its Anion-Ligand Complexes. *Journal of Molecular Biology*
794 **1992**, 227 (4), 1192–1204. [https://doi.org/10.1016/0022-2836\(92\)90531-n](https://doi.org/10.1016/0022-2836(92)90531-n).
- 795 (81) Lovell, S. C.; Word, J. M.; Richardson, J. S.; Richardson, D. C. The Penultimate Rotamer Library.
796 *Proteins* **2000**, 40 (3), 389–408.