Accurate prediction of metalsites using deep learning

Authors

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

† — To whom correspondence should be addressed: ursula.roethlisberger@epfl.ch

Affiliations

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

Abstract

Introduction

Methods

Results

Metal1D

Metal3D

The model was trained using a train/test/val split based on sequence identity. We used the MMSeqs2 clustered PDB at 30% sequence identity and used the highest resolution structure from each cluster

that contained a zinc, did not contain DNA and had resolution <2.5A. In case no structure was found the cluster was discarded.

The training examples were sampled from the chosen structure by choosing a balanced number of boxes from each protein that contained or did not contain a zinc. Each box was randomly rotated such that the model is insensitive to rotation.

Metal3D predicts a per residue score that can then be averaged over all residues or used individually (e.g for protein design).

Hyperparameter tuning using Ray tune

Evaluation of quality of predictions per box using discretized Jaccard Score (similarity of two sets). We noticed that at the edges often spurious density is predicted .

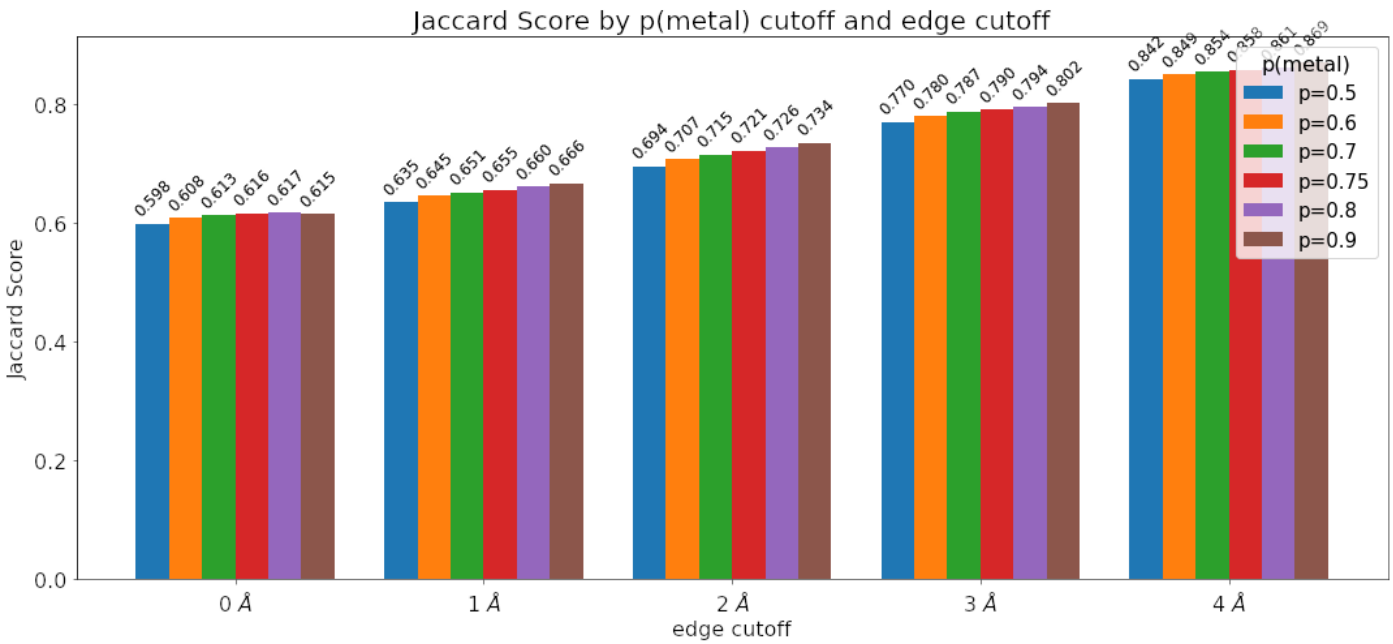


Figure 1: Discretized Jaccard index using different edge cutoffs and different probability cutoff

Selectivity for other metals

- Trained exclusively on zinc, predictions are similar for metals with different binding modes e.g Copper

- For sodium (binds via backbone O) not predicted
- Some sites for Mg are predicted with high but not super high probability

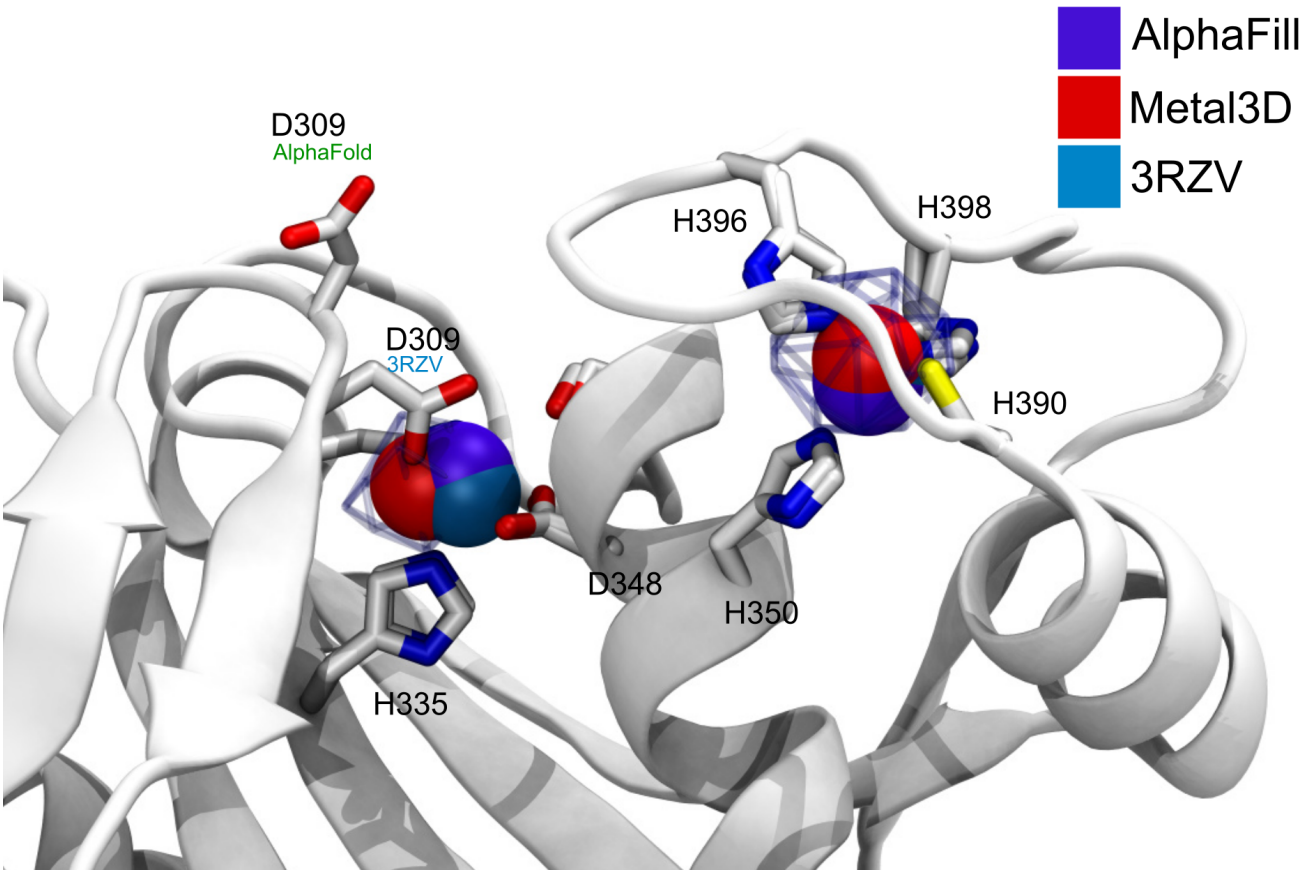
Figure 4 panels Copper protein, False positive Calcium Zinc and Zinc->Calcium site, Dimetallic zinc site

Discussion: Lack in selectivity could be related to smoothing the gaussian quite a bit when training (anything >0.05) is a hit.

Resolution of grid might be an issue Might be improved by improving the grid resolution to 0.5 A

Alpha Fold

AlphaFold often predicts sidechains in metal ion binding sites in the holo conformation. Services like AlphaFill use homology to transplant metals from similar PDB structures to the AlphaFold structure. Metal3D does not use homology can even deal with metal sites that are only partially in the holo conformation.

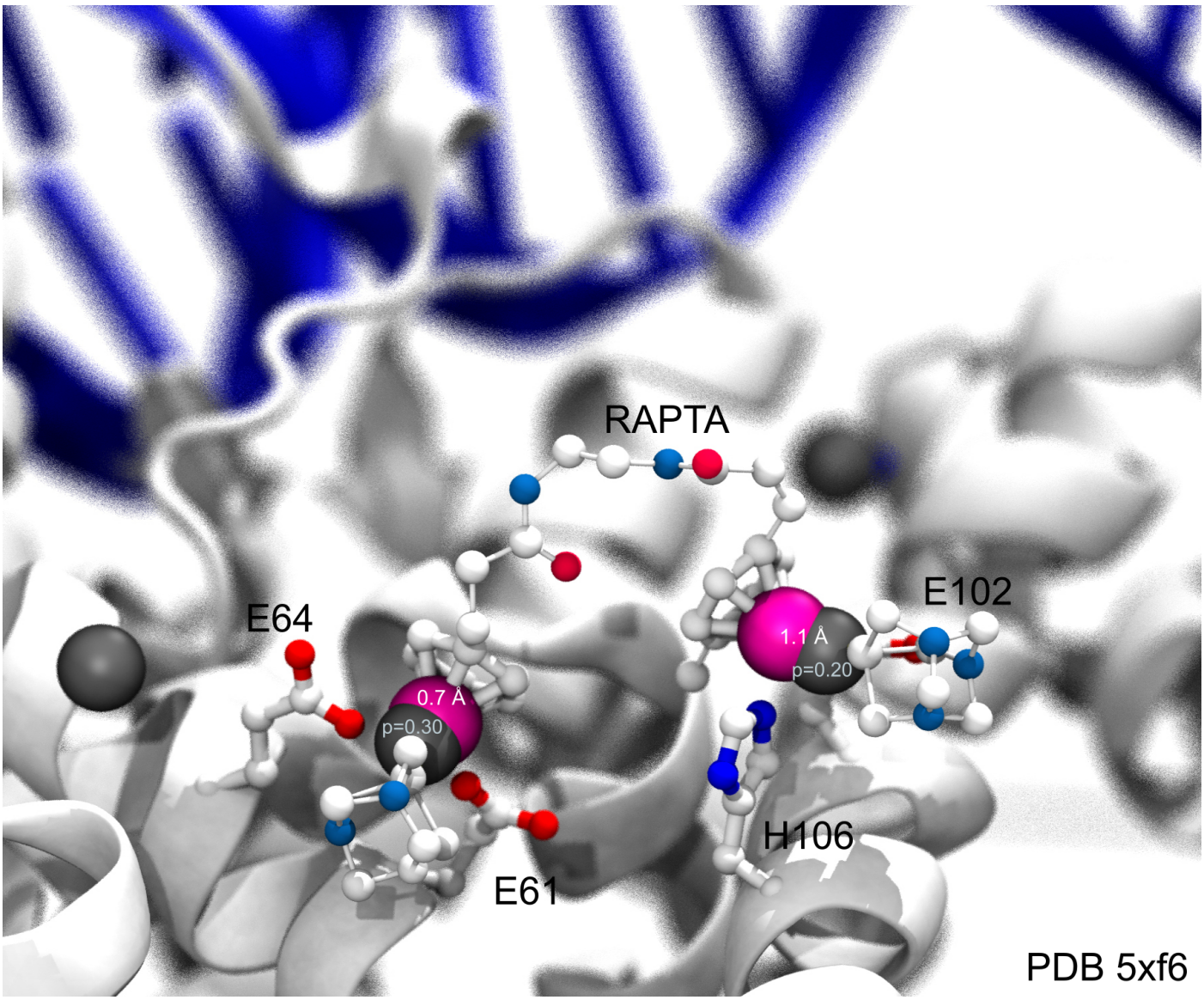


Metal3D, AlphaFill and 3RZV zinc positions. Metal3D places the metal with high accuracy even if coordination is not perfectly predicted. Probability map colored at p=0.6

Hidden/transient metalsites

Metallodrugs are in important class of drugs that rely on binding inhibitors to a protein (or DNA). Metal3D can be used to screen the hidden metalloproteome by finding transient metal ion binding sites.

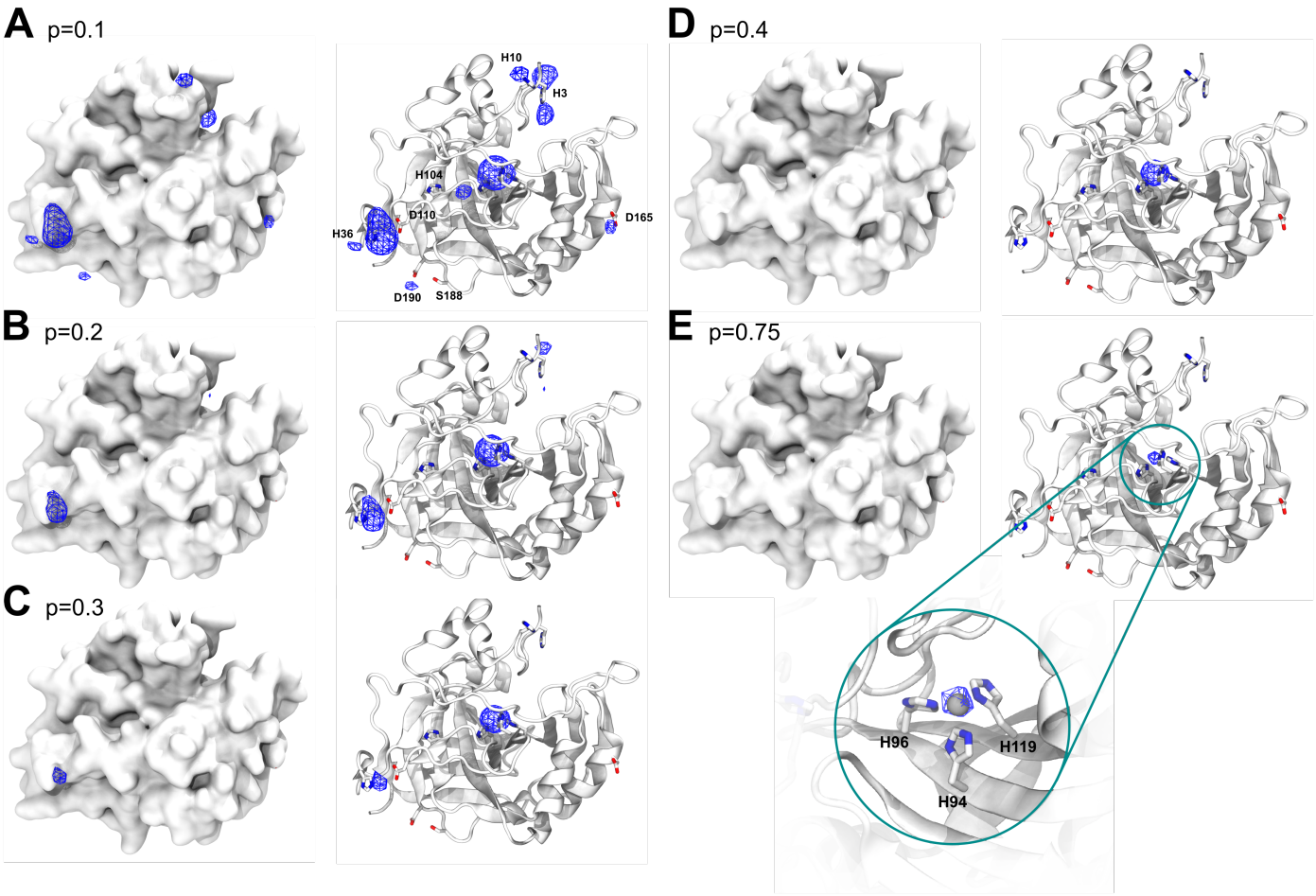
The site where Rapta binds is detected with p=0.3 but in a high resolution structure without (1KX4) there is a salt bridge with a lysine that might occlude metal detection. One could weight by the rotamer/do MD simulation e.g similar as for cryptic pockets.



RAPTA (Ru-arene-phosphaadamantane) inhibitor bound to Nucleosome core particle. Metal3D identifies the two Ruthenium sites with low probabilities.

HCA2: case study

HCA2 is the first enzyme where a catalytic zinc was discovered and is therefore one of the best studied metalloenzymes to date with a rich throve of mutational data available. On the wildtype enzyme crystal structure (2CBA) Metal3D perfectly recapitulates the location of the active site metal when using a high probability cutoff (p>0.4). The sites predicted with lower probability all look like reasonable transient binding sites at the surface of the protein.



Probability evolution in HCA2 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 used in silico generated mutants matching mutants in the first and second shell of the active site zinc and probed the effect on the predicted metal probability. For mutants that decrease zinc binding also a drop in probability can be observed that correlates well with the experimentally measured K_d.

We used a consistent set of K_d values from the literature. Probability evolution in HCA2 for different probability cutoffs A) p=0.1 B) p=0.2 C) p=0.3 D) p=0.4 E) p= 0.75

Comparison of Metal1D, Metal3D, MIB and BioMetAll

Many metal ion predictors exists that can be subdivided in two categories: binding site predictors and binding location predictors. The former label only the residues binding the ion, the latter also predict a location of the ion.

In addition to Metal1D and Metal3D we also compared two recent predictors BioMetAll and MIB. MIB uses a fragment method to identify homologus binding sites to the motifs it finds in a given structure and will extract the location of the metal from the homologous structures in its database. The main performance regulator of MIB is the tscore cutoff which is a parameter for the template similiarity with higher values requiring higher similiarity. BioMetAll was calibrated on the PDB and places probes on a regular grid at all sites where they find the criteria to be fulfilled. For each collection of probes also a center of the probes is given which we used to assess performance as there is no individual ranking of the probes given by the program. The main parameter for BioMetAll is the cluster cutoff which indicates how many probes in reference to the largest cluster a specific cluster has. We used a cutoff of 0.5 requiring all chosen clusters to have at least 50% of the probes of the most popolous.

For both tools the recommended settings match the accuracy of Metal3D p=0.75 with a lot more false positives. Metal1D offers high detection capabilites but also with a high number of false positives. While MIB also offers high precision, BioMetAll (using the cluster center) is not very precise with a MAD for correctly identified sites of 2.8 +- XX. Metal1D which identifies more sites than BioMetAll is slightly more precise than BioMetAll. MIB detects less sites but does so with high precision because it can use homologues sites to correctly place the metal ligand. BioMetAll also often provides probes that correctly identify the metal but as there is no ranking of the probes any probe could be closest to the actual location.

Comparison of Metal1D, Metal3D, BioMetAll and MIB on the testset used to train Metal1D and Metal3D. Predicted sites are counted if within 5 A of true metal location. False positive probes are clustered and counted once per cluster.

Mean absolute deviation of predicted location of zincs using Metal1D, Metal3D, BioMetAll and MIB on the testset used to train Metal1D and Metal3D for all correctly identified sites.

Discussion

3D CNN model accuracy. Recent work EquiDock uses no sidechains at all, gets ligand RMSD where only 25% are under a 2 threshold, https://arxiv.org/pdf/2202.05146.pdf Mean RMSD 8.3 A, Centroid 42.4

Conclusion

Supplement

Metal site detection using Metal1D

Metal site detection using Metal3D

Figure MAD testset

Comparison

Figure MAD only good zincs comparison Figure only good zincs TP/FN/FP