A Predicted Atlas of Metal-Binding Sites Across the Protein Universe

Feng Zhang[1,2,6], Linghao Kong[1,3,6], Yize Liu[4], Tong Jiang[5], Zihua Liu[1,3], Yuan Liu[1,2,*], Chu Wang[1,2,3,*]

[1]Beijing National Laboratory for Molecular Sciences, Synthetic and Functional Biomolecules Center, Key Laboratory of Bioorganic Chemistry and Molecular Engineering of Ministry of Education, College of Chemistry and Molecular Engineering, Peking University, Beijing, China.

[2]AI for Science (AI4S)-Preferred Program, Peking University Shenzhen Graduate School, Shenzhen, China

[3]Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China.

[4]State Key Laboratory of Elemento-Organic Chemistry, College of Chemistry, Nankai University, Tianjin, China.

[5]School of Life Sciences, Peking University, Beijing, China.

[6]These authors contributed equally to this work.

Corresponding Author* E-mail: chuwang@pku.edu.cn; wendao@pku.edu.cn;

Abstract

Metal-binding proteins (MBPs) play crucial roles in biological systems, including signal transduction, transcription, and catalysis. Recently, the AlphaFold Protein Structure Database (AFDB) has released approximately 214 million predicted protein structures, which provides a vast resource for the discovery of MBPs with novel structures and functions. However, systematic exploration of this massive dataset requires highly efficient computational approaches. Here, we present a machine learning model based on single-sequence input to predict metal-binding sites systematically across the whole AFDB protein universe. Using this approach, we established a comprehensive database termed “MetalDB” that contains approximately 38 million predicted MBPs with high confidence, and identified nearly 3 million sequence clusters that lack existing metal-binding annotations. Meta-analysis of MetalDB revealed a large number of metal-binding sites from previously unannotated protein families, structural domains as well as protein-protein interfaces, and uncovered metal-dependent functional proteins. Collectively, MetalDB provides a valuable resource for exploring MBPs across the whole protein universe and will significantly accelerate the discovery of proteins with novel functions in the post-genomic era.

Introduction

Metal-binding proteins (MBPs) constitute one of the most fundamental classes of biomolecules throughout the entire evolution. Even in the earliest stages of life, organisms had already evolved mechanisms to absorb metal elements from the environment and utilize them as cofactors[1]. To date, many essential biological processes in higher organisms remain critically dependent on the involvement of metal ions, such as cellular respiration, photosynthesis, and gene regulation[2]. Theoretically, metal ions can coordinate with amino acid sidechains to stabilize the three-dimensional structure of proteins or the transition states of catalytic reactions[3]. Such stabilization reduces the activation energy required for these reactions to proceed. Moreover, the variable oxidation states of metal ions allow them to function as electron transfer centers, which play indispensable roles in electron transport chains[4]. These unique chemical properties

not only underscore the physiological roles of metal ions but also inspire the rational design of novel MBPs to overcome the limitations imposed by the Irving-Williams series[5] and catalyze reactions such as hydrolysis[6].

To this end, systematic investigation of metal-binding proteins (MBPs) in living organisms has become an endeavor of considerable importance. Although various experimental techniques have been developed for this purpose[7–9], their widespread applications are often hindered by restriction in time, cost, and scalability. The release of AlphaFold Protein Structure Database (AFDB)[10,11], which contains approximately 214 million predicted protein structures across more than one million species, has opened new avenues for exploring the metalloproteome at an unprecedented computational scale and resolution. Such a comprehensive repository provides unique opportunities for the large-scale structural and functional characterization of proteins, such as identifying previously "hidden" structural clusters[12] and expanding knowledge of protein domain diversity[13]. However, the current version of AFDB lacks information regarding whether and how proteins interact with metal ions or small molecules, which are critical clues for inferring protein function. To fill this gap, methods of homology-based ligand transplantation have been proposed to reintroduce missing ligands into AlphaFold2 (AF2) models[14,15], but these approaches are inherently constrained by the availability of annotated homologous sequences. Meanwhile, recent advances such as AlphaFold3 (AF3)[16] and RoseTTAFold All-Atom (RFAA)[17] offer the potential to model biomolecular interactions directly. Nevertheless, they are mainly used as a "modeling" rather than "predicting" tool for inferring metal-binding sites given a specific target protein.

We previously developed a series of AI-based approaches, including MetalNet[18] and MetalNet2[19], to achieve accurate prediction of metal-binding sites in proteomes from a series of prokaryotic and eukaryotic species by leveraging coevolution signals from multiple sequence alignments (MSA). However, the dependence on MSA makes these prediction tools computationally intensive for large-scale mining across AFDB. For example, with a typical runtime of one minute per sequence for MSA generation and pair encoding, prediction of the entire AFDB database would require approximately

400 years of cumulative computational time. To tackle this challenge, we here presented "MetalNet3", a single-sequence-based method that achieves extremely high speed with even greater accuracy for large-scale prediction of MBPs from 214 million protein sequences recorded in AFDB. Combining with a structure-based filtering pipeline, we established a high-confidence database named "MetalDB" with 38 million MBPs and 46 million metal-binding sites (**Fig. 1a)**. We integrated multiple data sources to provide a comprehensive meta-analysis of metal-binding sites across the whole protein universe and identified unannotated MBPs from novel protein families, structural domains, protein binding interfaces as well as enzyme classes.

Prediction of MBPs in AFDB

During the development of MetalNet2, we observed that co-evolution signals were not strictly required for predicting metal-binding residue pairs composed of cysteine, histidine, glutamate, and aspartate when an encoding scheme based on protein language model (PLM) was employed[19] (**Table S1**). This observation motivated us to explore whether individual metal-binding residues could be predicted directly from protein sequences using a PLM-based approach, which in principle can cut down the most time-consuming step for computing MSA. We therefore disassembled the metal-binding residue pairs from the original MetalNet2 dataset into individual residues and split them into the training set (221,266 sample points) and test set (24,640 sample points) with each residue labeled as either metal-binding or not. Metal-binding residues ("positive") account for only approximately 8% of the samples in both the training and test sets, indicating a highly imbalanced class distribution. We systematically explored a range of classifier model architectures as well as PLM encoding strategies in the training set, and evaluated their performances on the test set (**Fig. 1b, Fig. S1**). To strike an effective balance between computational efficiency and predictive accuracy, we ultimately selected a support vector machine (SVM) architecture paired with the Ankh Base encoding scheme[20]. The resulting model, named as "MetalNet3", achieved an F1-score of 0.80 on the test set, which represents an improvement of 5% over MetalNet2 (**Fig. 1b, Table S2**). It also exhibited robust generalization on an independent dataset of

MBPs constructed from recently released PDB structures that were not included in the original MetalNet2 dataset with an F1-score of 0.81 **(Fig. S2)**. In terms of throughput, MetalNet3 can process 10,000 sequences in 5 minutes and run approximately 2,000 times faster than MetalNet2 (**Fig. 1b, Table S2**). The full-scale prediction across all the 214 million AFDB sequences was completed in approximately four days using 16 NVIDIA L40 GPUs in parallel, resulting in an initial dataset of 96,827,100 predicted MBPs (**Fig. 1a**). Given that the type of metal bound provides critical insights into protein function, we additionally trained a separate classifier to predict the metal-binding types following a similar training protocol. The classifier achieved an average F1-score of 0.605 in a dataset of MBPs containing 11 metal types and demonstrated strong performance in identification of calcium, zinc, and Fe-S clusters (**Table S3**). In contrast, a classifier trained to distinguish metal groups (e.g., alkaline earth metals, transition metals, and Fe-S clusters) achieved a much higher F1-score of 0.932 on average (**Table S4**), which suggests that the primary challenge lies in discriminating between different metal-binding types within each group.

To obtain a high-confidence subset of predicted MBPs by MetalNet3, we required that the predicted metal-binding residues should be spatially adjacent to form a feasible metal-binding site in the target protein (**Supplementary Methods**). Application of this filter cut the initial dataset down to 38,361,041 proteins with 46,143,213 predicted metal-binding sites, which we collectively refer as "MetalDB" (**Fig. 1a**). Although no explicit pLDDT threshold was applied during the structure-based filtering process, 93.7% of the predicted metal-binding sites exhibit high-confidence structural predictions (pLDDT > 70) as recorded in AFDB (**Fig. 1c**). Additionally, when a randomly selected subset of MetalDB (20 proteins in triplicates) from each pLDDT category were fed to Protenix[21] for structural modeling of the metal-binding complex, we observed excellent overlaps (94.8±4.9%) between the metal-binding sites predicted by MetalNet3 and those modeled by Protenix in the group of pLDDT > 90 (**Fig. 1d**).

To assess the coverage of existing metal-binding annotations in MetalNet3 predictions, we developed a computational pipeline to crosscheck MetalDB with four established databases (**Fig. S3**) including the Protein families database (Pfam)[22], the

Encyclopedia of Domains (TED)[13], the AFDB sequence clusters (AFDB50) and the AFDB structure clusters (AFDB clusters)[12]. For each database, we counted the number of MetalDB predictions with metal-binding annotations as inferred from sequence or structural similarity (**Fig. 1e**). For example, about 22.7% predictions (8,709,939 out of 38,361,041) belong to the Pfam families with known metal-binding activities and this number goes up to 42.5% if they were mapped to different AFDB50 clusters. In total, 72.8% MetalDB predictions (27,927,308) were found to have putative metal-binding annotations collectively from all these four databases (**Fig. 1e**). Taken together, these results substantiate the reliability of MetalDB with support from both computational structural modeling and existing literature annotations.

Analysis of sequence clusters in MetalDB

Sequence clusters in MetalDB are analyzed based on AFDB50[12] using the clustering criteria of 0.9 for sequence coverage and a minimum sequence identity of 0.5. These thresholds were imposed to ensure that proteins within the same cluster share identical or similar biological functions[23]. Accordingly, the 38 million predictions in MetalDB were mapped to 7,415,834 AFDB50 clusters. We hypothesized that if metal-binding sites are functionally critical, all members within a given cluster tend to be predicted with consistent metal-binding activity at both protein and residue levels in MetalDB (defined as “consistency”, see **Fig. 2a** and **Supplementary Methods**). Although the protein or residue consistency marginally decreased as the cluster size increased (**Fig. S4**), the average value remained robust above 0.8 (**Fig. 2b**). This result suggests that MetalNet3 captures evolutionarily conserved (thus likely functionally important) features of metal-binding sites within individual sequence clusters.

Among the 6.8 million predicted MBP sequence clusters that show 100% protein-level consistency, 9.6% did not show full consistency at the residue level (**Fig. S5**). Analysis of this discrepancy further revealed suggestive evidence of metal-binding site migration (**Fig. 2c**), a process where the primary metal-coordination sphere in one member relocates to a different region within the same structural scaffold in other members. For instance, the sequence cluster of YdhG-like domain-containing proteins

(UniProt ID of the cluster representative protein: A0A2D0NET4) can be further partitioned into two sub-clusters based on their metal-binding site topologies: one contains two members with predicted N-terminal metal-binding sites, and the other contains seven members with C-terminal sites (**Fig. 2d**). Structural modeling indicates that both sub-clusters utilize a flexible terminal loop to coordinate a [2Fe-2S] cluster (**Fig. 2e**). Through AFDB structural clustering, this specific binding mode can be traced back to the CDGSH-type iron-binding zinc finger (UniProt ID: A0A7L4ZFW4), highlighting a conserved evolutionary origin despite the topological shift (**Fig. S6**).

To identify any novel MBP sequence cluster, we filtered the MetalDB dataset to search for functionally unannotated clusters with at least 10 members and a protein-level prediction consistency of at least 0.7, which yielded a total of 88,207 high-confidence candidates. Gene Ontology (GO) enrichment analysis of these novel clusters demonstrated a pronounced enrichment toward various catalytic activities, suggesting these previously uncharacterized MBPs largely tend to play enzymatic roles (**Fig. S7**). We selected the "dark cluster"[12] A0A090IEJ0 that lacks any existing functional annotations for experimental validation. MetalNet3 predicted that a member of this cluster, the uncharacterized protein (UniProt ID: A0A6C9M6E0) in *E. coli*, contains two zinc-binding sites with eight coordinating residues (**Fig. 2f**). Subsequent ICP-MS analysis of the wild-type protein and corresponding mutants confirmed both the sites and stoichiometry of zinc binding (**Fig. 2g**).

We further catalogued the taxonomic composition of all the sequence clusters mapped in MetalDB (**Fig. 2h**), which essentially spans all major domains of cellular life, with the majority of sequences originating from Bacteria (27.3 million), followed by Eukaryota (9.1 million) and Archaea (1.36 million). Notably, although Archaea constitute a relatively small fraction of MetalDB, they exhibited the highest proportion (22%~26%) of high-confidence predictions relative to the proteome size compared with representative species from Bacteria and Eukaryota (**Table S5**). This observation suggests that archaeal proteomes are relatively enriched in metal-binding proteins, potentially reflecting a greater utilization of metal-associated functions in archaeal cellular processes. In addition, we sampled 200,000 proteins from distinct sequence

clusters across three domains of life to estimate the proportions of metal groups (i.e., alkaline metals, transition metals and Fe-S clusters) within each domain. While transition metal-binding proteins constitute the dominant fraction across all life domains, the focus of metal utilization appears to have shifted from Fe-S clusters to more diverse transition metal types along the evolutionary trajectory from archaea to bacteria to eukaryota (**Fig. S8**). Lastly, we identified 638 sequence clusters of metal-binding proteins that are shared across all three domains of life. KEGG functional annotation indicated that these MBPs were broadly distributed across diverse metabolic pathways, primarily associated with conserved core metabolic processes such as central carbon metabolism and biosynthetic pathways (**Fig. S9**). These results collectively suggest the centrality of metal-associated biochemistry to early life.

Discovery of Novel MBP Families

Pfam is a widely used, comprehensive database of protein families and domains, which classifies proteins based on conserved functional regions[22]. By crosschecking MetalDB with the Pfam database, we obtained 9570 protein families with predicted MBPs, among which 846 families lack existing metal-binding annotations (**Fig. 3a**). To obtain a list of novel MBP families with high confidence, we further filtered the dataset for any family with at least 10 members and a predicted consistency $\geqslant 0.5$ (e.g. at least 5 members predicted as MBPs in a family with 10 members). Among the resulting 181 families, 148 were functionally uncharacterized (e.g., domain of unknown functions or DUFs) and the rest of 33 families with functional annotations could be classified into seven major categories (**Fig. 3b**).

The functional category related to “cell structure and motility” is particularly noteworthy because it comprises the largest number of proteins overall (2,108 in total) from only three families (PF06904, PF14245, and PF21811), among which the Extensin-like_C family (PF06904) alone accounts for 1,892 proteins (**Fig. 3b, c**). The protein-level prediction consistency for this category reaches 0.98 and most of the family members contain three histidine residues and one aspartate residue as the

putative metal-binding residues (**Fig. 3c**). Structural alignment via FoldSeek[24] identified a local match to the zinc-binding site in the Zinc D-Ala-D-Ala carboxypeptidase from Streptomyces albus G (UniProt ID: P00733; PDB ID: 1LBU; **Fig. 3d**), suggesting possible enzymatic activity and a conserved metal-binding site within this family.

Another MBP family with high prediction consistency is the Activator of Apoptosis Harakiri ("HRK") family with the Pfam ID of PF15196, whose members are known to promote apoptosis by binding Bcl-2 and inhibiting its anti-apoptotic activity[25]. This family contains 56 proteins according to Pfam database, among which 53 members were predicted with metal-binding sites by MetalNet3 (**Table S6**). The human HRK protein (UniProt ID: O00198) is predicted with four residues (2C, 4C, 16C, and 18C) for metal coordination, which is fully supported by the structure modeling result (**Fig. 3e**). Satisfyingly, ICP-MS experiments confirmed the zinc-binding properties of HRK, and mutation of the four predicted chelating residues substantially reduced the protein's metal-binding capacity (**Fig. 3f**).

Structural domains of MBPs

While the coordination chemistry of metals imposes strict constraints on the geometry and residue types within the first coordination sphere, the protein backbones that can accommodate metal-binding sites exhibit remarkable diversity[26]. To date, the characterization of metal-binding domains has been limited to the structural repertoire experimentally recorded in PDB. By crosschecking MetalDB with the structural domain annotations from "The Encyclopedia of Domains" (TED) database[13] that is derived from AFDB, we are now able to fully navigate the landscape of metal-binding domains across a significantly expanded structural space. Mapping the predictions from MetalNet3 to the TED database yielded 11,746,515 domain clusters that contain putative metal-binding sites. Among them, 8,116,142 clusters have annotations from the CATH database[27] and can be assigned into 1,178 distinct CAT label groups (**Fig. 4a**). Within these CAT label groups, 688 have been documented with metal-binding annotations, which cover 98.6% of the known metal-binding domains (**Fig. 4b**). The

remaining 490 groups represent potentially novel metal-binding scaffolds (**Fig. 4b**) whose CAT labels encompass all four primary structural classes (**Fig. S10**), namely mainly-alpha, mainly-beta, alpha/beta, and few secondary structure domains.

A structural comparison of the MetalNet3-predicted MBPs with those PDB entries sharing the same CAT labels reveals that localized discrepancies in key coordinating residues often preclude the formation of viable metal-binding environments in the PDB structures although the global scaffolds align closely. For example, the pyrimidine dimer DNA glycosylase from Deltaproteobacteria bacterium (UniProt ID: A0A7V4K303) with the CAT label of 1.10.440 was predicted by MetalNet3 to contain a high-confidence metal-binding site containing C12, H15, C105, C108 (**Fig. 4c**), which could be confirmed experimentally to bind zinc by ICP-MS (**Fig. 4d**). Its homologue from Tequatrovirus T4 shares the same CAT label and was structurally characterized with a dominantly helical scaffold[28] (PDB ID: 2fcc). However, the protein lacks the metal-coordinating chemical environment for this site according to the crystal structure with three substitutions including C11A, C105Q, and C108T (**Fig. 4c**). Similar examples can also be found in other CAT groups of mainly-beta, alpha/beta, and few secondary structure domains (**Fig. S11**). Collectively, these results underscore that site-level predictive precision by MetalNet3 is critical for distinguishing metal-binding structural features among closely related scaffolds, which identify molecular variations with functional implications that remain undetected in conventional structural databases.

We further analyzed the metal-binding sites predicted by MetalNet3 in terms of their locations in structural domains (**Fig. 4e**). While the majority (70.0%) are located within single domains, there are still 3.6% of the metal-binding sites (1,649,455 cases) that are composed of residues from multiple domains, which reflects potential prototypes of protein-protein interactions mediated by metal binding. The remaining 26.4% of the sites are mapped to structurally ambiguous regions where the domain-parsing algorithm of TED fails to reach high or medium agreement. This observation suggests that a substantial fraction of metal-binding sites may involve flexible inter-domain hinges or intrinsically disordered regions, where metal coordination could

potentially induce the formation of localized structural conformations.

In order to investigate metal-binding sites located outside well-defined structural domains, particularly those in regions with low structural prediction confidence, we examined the high-confidence MetalNet3 predictions (probability≥0.8) with their metal-binding sites located in regions with pLDDT≤50. Among the resulting 100,764 sites from 97,855 proteins, 89.9% are cysteines (**Fig. S12**), which is consistent with previous observations that cysteine-rich and metal-coordinating motifs are frequently associated with structurally flexible or conditionally folded regions[29]. We further cross-referenced these predictions with the DisProt database[30] that contains 3,201 experimentally annotated disordered proteins, and identified 16 MBPs with the predicted metal-binding sites located within structurally disordered regions. Notably, when we modeled one of these MBPs (UniProt ID: J7JU64) by Protenix with or without the metal, we observed that the presence of zinc could induce the folding of an otherwise low-confidence region into an alpha helix structure (**Fig. 4f**). These results highlight the ability of MetalNet3 to reveal potential functional sites embedded within poorly resolved regions and to assist in refining structural models in AFDB.

Prediction of metal-binding sites at homomeric protein interfaces

It has been estimated that approximately 30%-50% of proteins naturally form homomeric complexes in proteomes[31] that are composed of two or more identical polypeptide chains (“subunits”). While many of these proteins bind metal ions within each subunit to stabilize their structures or to function as enzymes (e.g., the homo-tetramer of manganese superoxide dismutase), we are motivated to repurpose MetalNet3 for predicting metal-binding sites across homomeric protein interfaces because metals might play important roles to regulate protein assembly. We reason that our PLM-based prediction model using single sequences should be able to capture certain evolutionarily conserved metal-binding features at the homomeric interface because other PLM-based methods have been successfully applied to predict protein quaternary structure[32], homo-oligomer symmetries[33] as well as protein-protein interactions[34].

As AFDB provides structural models exclusively for monomeric proteins, we refer to a set of 8,195 homo-oligomeric assemblies compiled by Schweke *et al*[35] from four representative proteomes. We refined our structural filtering criteria (**Supplementary Methods**) in MetalNet3 to specifically target interface-associated regions and detected 386 targets with putative metal-binding sites located at their homomeric interfaces (**Fig. 5a**). The majority (333 cases, 86.3%) of these complexes involve two chains, while 51 (13.2%) and 2 (0.5%) cases involve three and four chains, respectively (**Fig. S13**).

A total of 320 predicted metal-binding homomers can be supported by existing annotations. For example, the human OTCase (UniProt ID: P00480) was predicted to bind transition metal ion at its trimeric interface (**Fig. 5b**), and the result can be supported by the crystal structure of a remote homologue, the putrescine carbamoyltransferase from Enterococcus faecalis (UniProt ID: Q837U7; PDB ID: 4A8H), which shows a nickel ion binding at the homomeric interfaces with a C3 axis (**Fig. S14a**). Interestingly, our experimental characterization demonstrated that the recombinant human OTCase purified from *E. coli* cells binds zinc (**Fig. 5c**), and the result highlights differences in metal-binding preferences among homologous proteins across species.

Similar to the example of OTCase, we also predicted a sword-shaped trimeric protein, namely the phage baseplate assembly protein, from *E. coli* (UniProt ID: A0A0H3JK28) as an iron-binding protein, with H187 and H189 from each subunit as the metal-coordinating residues (**Fig. 5d**). Although a region of this protein (residues 123–175) belongs to the Phage spike trimer family (PF18715) that is annotated with iron-binding activities[36](**Fig. S14b**), the ICP-MS analysis confirmed that mutation of the two histidine residues in the monomer abolished the zinc binding of the complex (**Fig. 5e**). Another noteworthy case is the human glutathione reductase (UniProt ID: P00390), which was predicted by MetalNet3 with a metal-binding site across the homo-dimeric interface that contains C107 and H511. The prediction was timely echoed by a chemoproteomic study to map zinc-binding cysteines[7] and is supported by the Protenix modeling (**Fig. 5f**). Although the protein is classified at the family level as a member of the pyridine nucleotide-disulfide oxidoreductase dimerization domain (Pfam ID:

PF02852), no other family members are reported with zinc-binding activity at this exact interface. The result highlights a limitation of family-level annotations that may fail to capture functionally important sites, and underscores the need for more fine-grained and site-resolved annotation to fully explore the metal-binding landscape in protein universe.

Discovery of novel metal-binding enzymes

To explore the impact of metal ions on protein function, particularly enzymatic activities beyond structural stabilization, we implemented a computational pipeline to screen for functionally conserved human proteins in the MetalDB with no prior metal-binding annotation (**Fig. 6a**). Among the resulting 26 candidate proteins (**Table S7**), we selected the human Ribose-5-phosphate isomerase (RPIA, UniProt ID: P49247) to investigate the influence of metal ions on its enzyme activity. RPIA is an essential enzyme in the non-oxidative branch of the pentose phosphate pathway (PPP), which catalyzes the reversible conversion between ribose-5-phosphate and ribulose-5-phosphate[37]. Although the crystal structure of its homolog in *E. coli* has been solved[37] (PDB ID: 1KS2), its metal-binding activity has never been reported in literature to our best knowledge. MetalNet3 predicted that 160D, 163D, and 182E of RPIA are involved in metal binding and are likely to bind transition metals. These three residues align with the key residues (81D, 84D, and 103E) in the active site of the *E. coli* homolog (**Fig. 6b**) and are highly conserved across representative eukaryotic species (**Fig. S15**). ICP-MS analysis demonstrated that the protein binds magnesium and zinc with 50% occupancy for each metal, respectively, and mutation of the three predicted metal-chelating residues or treatment with EDTA completely abolished the metal-binding activity of RPIA (**Fig. 6c**).

We next assessed the impact of metals on the catalytic activity of RPIA by measuring the characteristic absorbance of the product ribulose-5-phosphate at 280 nm (**Supplementary Methods**). The enzyme kinetics profile revealed that the zinc-treated enzyme showed significantly enhanced reaction rates compared to the apo enzyme (EDTA-treated), with the addition of 1 or 5 μM zinc reaching near-maximal substrate

turnover within 10 minutes (**Fig. 6d, Fig. 6e**). In contrast, the addition of Mg^{2+} did not result in a significant increase in catalytic activity under the same conditions (**Fig. S16**), highlighting the distinct zinc specificity in regulating RPIA's function. Our discovery of RPIA as a zinc-dependent enzyme may provide a novel mechanistic link between cellular metal homeostasis and PPP. While zinc is known to competitively inhibit glucose-6-phosphate dehydrogenase (G6PD, the entry point of the oxidative PPP) at higher concentrations[38], its potent activation of RPIA may suggest a sophisticated regulatory mechanism where zinc fluctuations may function as a metabolic 'rheostat', funneling glucose-derived carbons toward ribose-5-phosphate synthesis. This coordination ensures that nucleotide precursor production is tightly coupled with the availability of zinc, an essential cofactor for DNA replication and repair machinery.

Discussion

The exponential growth of protein sequence and structural data demands powerful analytical tools to decipher biological functions, among which ligand interaction serves as a critical key. In this study, we addressed the challenge of predicting metal-binding proteins and sites in proteomes by developing a robust protein language model named "MetalNet3" with single sequences as the input. By applying MetalNet3 to the entire AlphaFold Protein Structure Database, we established MetalDB, the first comprehensive database of metal-binding sites that covers the whole protein universe, and systematically characterized the site-specific landscape of metal-binding proteins across diverse families and structural domains.

Our results demonstrate that protein language models can efficiently and accurately identify metal-binding sites in proteomes. Notably, this sequence-based approach exhibits remarkable speed and robustness, and is able to capture functional sites in challenging scenarios, such as complex homomeric interfaces and intrinsically disordered regions where structural modeling often fails to yield confident hits. While integrating structural information could potentially improve the prediction performance, as evidenced by structure-aware PLMs[39], it presents a trade-off between the potential predictive gains and the reliability of the input structures. Our findings suggest that for

mining vast datasets, a structure-independent approach can effectively uncover metal-binding sites that might otherwise be overlooked due to structural uncertainties. However, how to accurately predict specific metal ion types given a target site remains a significant challenge. This difficulty arises not only from the severe imbalance in current training datasets but also from the intrinsic complexity of metal coordination, including the promiscuity of metal binding and the strict geometric constraints required by specific ions.

Analysis of millions of predictions in MetalDB offers fresh insights into the evolutionary history of metalloproteins through cross-species analysis and residue-level alignments. We observed site-specific differentiation that likely corresponds to functional divergence and environmental adaptation that warrant further investigation (e.g., the migration of metal-binding sites between the N- and C-termini in YdhG-like domain-containing proteins). Furthermore, the high-confidence predictions in MetalDB help clarify ambiguous annotations in existing databases and provide residue-level evidences for Pfam families that previously lacked specific binding site information. Such data will serve as valuable resources for knowledge distillation to train more powerful models to predict binding sites of other metal cofactors or metallodrug. Lastly, the newly identified metal-binding domains significantly expand the known structural diversity of MBPs, and offer potential templates for rational design of artificial metalloproteins beyond the symmetrical helical motifs[26]. Collectively, we anticipate that MetalNet3 together with MetalDB will accelerate research into understanding the physiological functions of metal-binding proteins and fuel the efforts of rational design of novel biological functions.

Author contribution

C.W. and Y.L. conceived the project; F.Z. performed computational analysis with the help of Y.L.; L.H.K. performed experimental validation with the help from Y.Z.L., T.J. and Z.H.L.; F.Z., Y.L. and C.W. analyzed the data and wrote the manuscript with input from all authors.

Acknowledgements

We thank the Computing Platform of the Center for Life Science for providing the computing power and the Analytical Instrumentation Center of College of Chemistry and Molecular Engineering, Peking University, for ICP-MS analysis. We thank the National Center for Protein Sciences (Beijing) – Protein and Gene Platform at Peking University for their assistance with cDNA library preparation, experimental support, and quality control. Figure 2h (Sankey plot) was created by tools provided by https://github.com/nowthis/sankeymatic. Logo of species in Figure 5a and Figure 6a was created by tools provided at https://gdp.renlab.cn/.

Funding

C.W acknowledges the support from Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM507), National Natural Science Foundation of China (22525071, 22361142837 and 92578206), Ministry of Science and Technology of China (2022YFA1304700), and Beijing National Laboratory for Molecular Sciences (Grant No. BNLMS-CXTD-202401). This work is also supported by National Natural Science Foundation of China (22477008 and 22277007) and the National Key Research and Development Projects (2022YFA1300008, 2023YFA1506500) to Y.L.

Conflict of Interest

The authors declare no competing financial interest.

Code availability

The MetalNet3 code is available as open source at https://github.com/wangchulab/MetalNet3. The MetalNet3 server with prediction data for each species can be accessed at https://www.chem.pku.edu.cn/wangchulab/metalnet under the “Database” tab.

References

1. Sendra, K. M. *et al.* An ancient metalloenzyme evolves through metal preference modulation. *Nature Ecology & Evolution* **7**, 732–744 (2023).
2. Waldron, K. J., Rutherford, J. C., Ford, D. & Robinson, N. J. Metalloproteins and metal sensing. *Nature* **460**, 823–830 (2009).
3. Holm, R. H., Kennepohl, P. & Solomon, E. I. Structural and Functional Aspects of Metal Sites in Biology. *Chemical Reviews* **96**, 2239–2314 (1996).
4. Winkler, J. R. & Gray, H. B. Electron Flow through Metalloproteins. *Chemical Reviews* **114**, 3369–3380 (2014).
5. Choi, T. S. & Tezcan, F. A. Overcoming universal restrictions on metal selectivity by protein design. *Nature* **603**, 522–527 (2022).
6. Kim, D. *et al.* Computational design of metallohydrolases. *Nature* **649**, 246–253 (2026).
7. Burger, N. *et al.* The human zinc-binding cysteine proteome. *Cell* **188**, (2025).
8. Chen, L. *et al.* APEX2-based Proximity Labeling of Atox1 Identifies CRIP2 as a Nuclear Copper-binding Protein that Regulates Autophagy Activation. *Angewandte Chemie International Edition* **60**, 25346–25355 (2021).
9. Bak, D. W. & Weerapana, E. Monitoring Fe–S cluster occupancy across the E. coli proteome using chemoproteomics. *Nature Chemical Biology* **19**, 356–366 (2023).
10. Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research* **50**, D439–D444 (2021).

11. Varadi, M. *et al.* AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Research* **52**, D368–D375 (2023).

12. Barrio-Hernandez, I. *et al.* Clustering predicted structures at the scale of the known protein universe. *Nature* **622**, 637–645 (2023).

13. Lau, A. M. *et al.* Exploring structural diversity across the protein universe with The Encyclopedia of Domains. *Science* **386**, eadq4946 (2024).

14. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

15. Hekkelman, M. L., de Vries, I., Joosten, R. P. & Perrakis, A. AlphaFill: enriching AlphaFold models with ligands and cofactors. *Nature Methods* **20**, 205–213 (2023).

16. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).

17. Krishna, R. *et al.* Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science* **384**, eadl2528 (2024).

18. Cheng, Y. *et al.* Co-evolution-based prediction of metal-binding sites in proteomes by machine learning. *Nature Chemical Biology* **19**, 548–555 (2023).

19. Zhang, F. *et al.* MetalNet2: an enhanced server for predicting metal-binding sites in proteomes. *National Science Review* **11**, nwae391 (2024).

20. Elnaggar, A. *et al.* Ankh: Optimized Protein Language Model Unlocks General-Purpose Modelling. *arXiv preprint arXiv:2301.06568* (2023).

21. Team, B. A. A. *et al.* Protenix - Advancing Structure Prediction Through a Comprehensive AlphaFold3 Reproduction. *bioRxiv* https://doi.org/10.1101/2025.01.08.631967 (2025)

doi:10.1101/2025.01.08.631967.

22. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Research* **49**, D412–D419 (2020).

23. Suzek, B. E. *et al.* UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2014).

24. van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nature Biotechnology* **42**, 243–246 (2024).

25. Youle, R. J. & Strasser, A. The BCL-2 protein family: opposing activities that mediate cell death. *Nature Reviews Molecular Cell Biology* **9**, 47–59 (2008).

26. Chalkley, M. J., Mann, S. I. & DeGrado, W. F. De novo metalloprotein design. *Nature Reviews Chemistry* **6**, 31–50 (2022).

27. Sillitoe, I. *et al.* CATH: increased structural coverage of functional space. *Nucleic Acids Research* **49**, D266–D273 (2020).

28. Golan, G. *et al.* Structure of T4 Pyrimidine Dimer Glycosylase in a Reduced Imine Covalent Complex with Abasic Site-containing DNA. *Journal of Molecular Biology* **362**, 241–258 (2006).

29. Bhopatkar, A. A., Uversky, V. N. & Rangachari, V. Chapter Nine - Disorder and cysteines in proteins: A design for orchestration of conformational see-saw and modulatory functions. in *Dancing Protein Clouds: Intrinsically Disordered Proteins in Health and Disease, Part B* (ed. Uversky, V. N.) vol. 174 331–373 (Academic Press, 2020).

30. Nugnes, M. V. *et al.* DisProt in 2026: enhancing intrinsically disordered proteins accessibility, deposition, and annotation. *Nucleic Acids Research* gkaf1175 (2025) doi:10.1093/nar/gkaf1175.

31. Levy, E. D. & Teichmann, S. A. Chapter Two - Structural, Evolutionary, and Assembly Principles

of Protein Oligomerization. in *Oligomerization in Health and Disease* (eds Giraldo, J. & Ciruela, F.) vol. 117 25–51 (Academic Press, 2013).

32. Avraham, O., Tsaban, T., Ben-Aharon, Z., Tsaban, L. & Schueler-Furman, O. Protein language models can capture protein quaternary state. *BMC Bioinformatics* **24**, 433 (2023).

33. Kshirsagar, M. *et al.* Rapid and accurate prediction of protein homo-oligomer symmetry using Seq2Symm. *Nature Communications* **16**, 2017 (2025).

34. Ullanat, V., Jing, B., Sledzieski, S. & Berger, B. Learning the language of protein-protein interactions. *Nature Communications* **17**, 1199 (2026).

35. Schweke, H. *et al.* An atlas of protein homo-oligomerization across domains of life. *CELL* **187**, (2024).

36. Yamashita, E. *et al.* The host-binding domain of the P2 phage tail spike reveals a trimeric iron-binding structure. *Acta crystallographica. Section F, Structural biology and crystallization communications* **67**, 837—841 (2011).

37. Zhang, R. *et al.* Structure of Escherichia coli Ribose-5-Phosphate Isomerase: A Ubiquitous Enzyme of the Pentose Phosphate Pathway and the Calvin Cycle. *Structure* **11**, 31–42 (2003).

38. Tandogan, B. & Ulusu, N. N. Effects of cadmium and zinc ions on purified lamb kidney cortex glucose-6-phosphate dehydrogenase activity. *Journal of Enzyme Inhibition and Medicinal Chemistry* **21**, 225–230 (2006).

39. Su, J. *et al.* SaProt: Protein Language Modeling with Structure-aware Vocabulary. in *International Conference on Learning Representations* (eds Kim, B. et al.) vol. 2024 6987–7009 (2024).

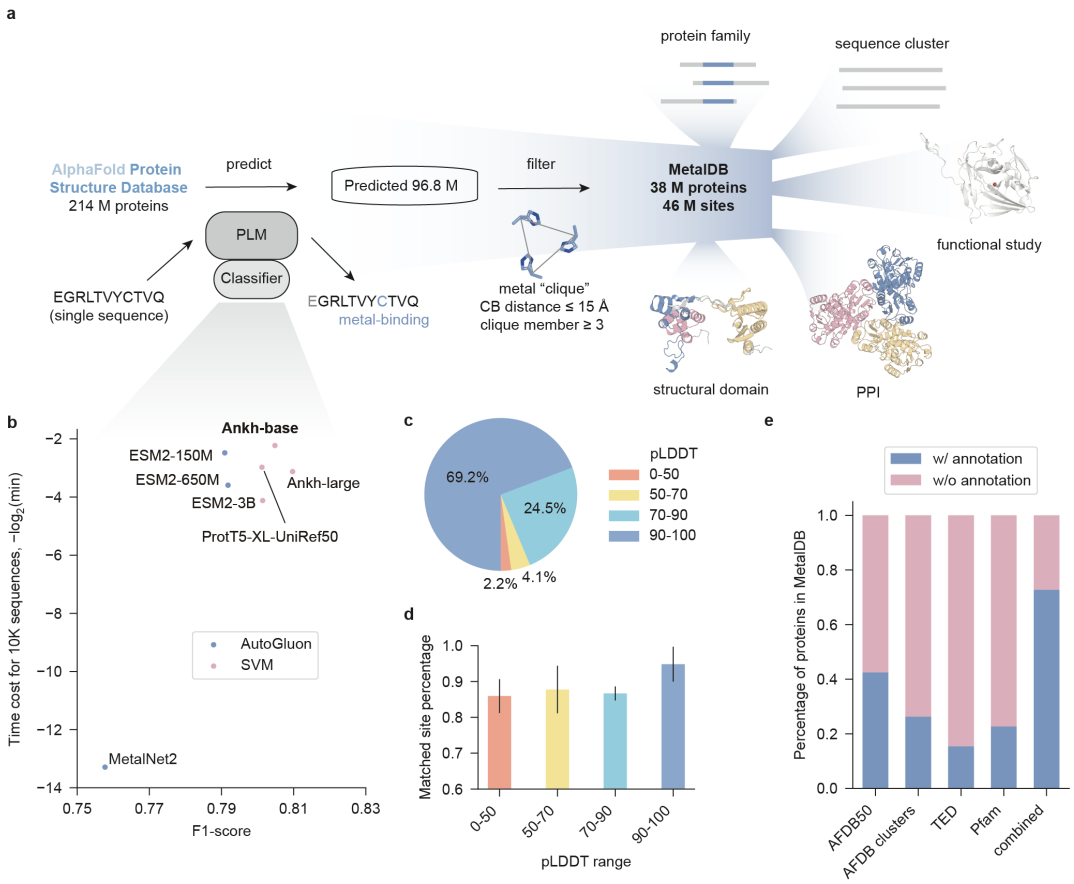

Figure 1: Prediction of metal-binding proteins and sites in AFDB by MetalNet3. **a**, Schematic workflow of predicting metal-binding proteins and sites in AFDB by MetalNet3 and constructing the MetalDB. **b**, Optimization of the performance of MetalNet3 with different combinations of PLM models and classifier architecture. The Ankh-base and SVM combination shows better F1 score (x-axis) and much faster speed (y-axis) than MetalNet2. **c**, Distribution of metal-binding sites in MetalDB based on their pLDDT scores. **d**, Percentage of matching between the modeled metal-binding sites by Protenix and the predicted metal-binding sites by MetalNet3 from each pLDDT group. **e**, Percentage of MetalDB predictions with annotated metal-binding activity according to each of four databases, including AFDB50, AFDB clusters, TED and Pfam, or with all of them "combined".

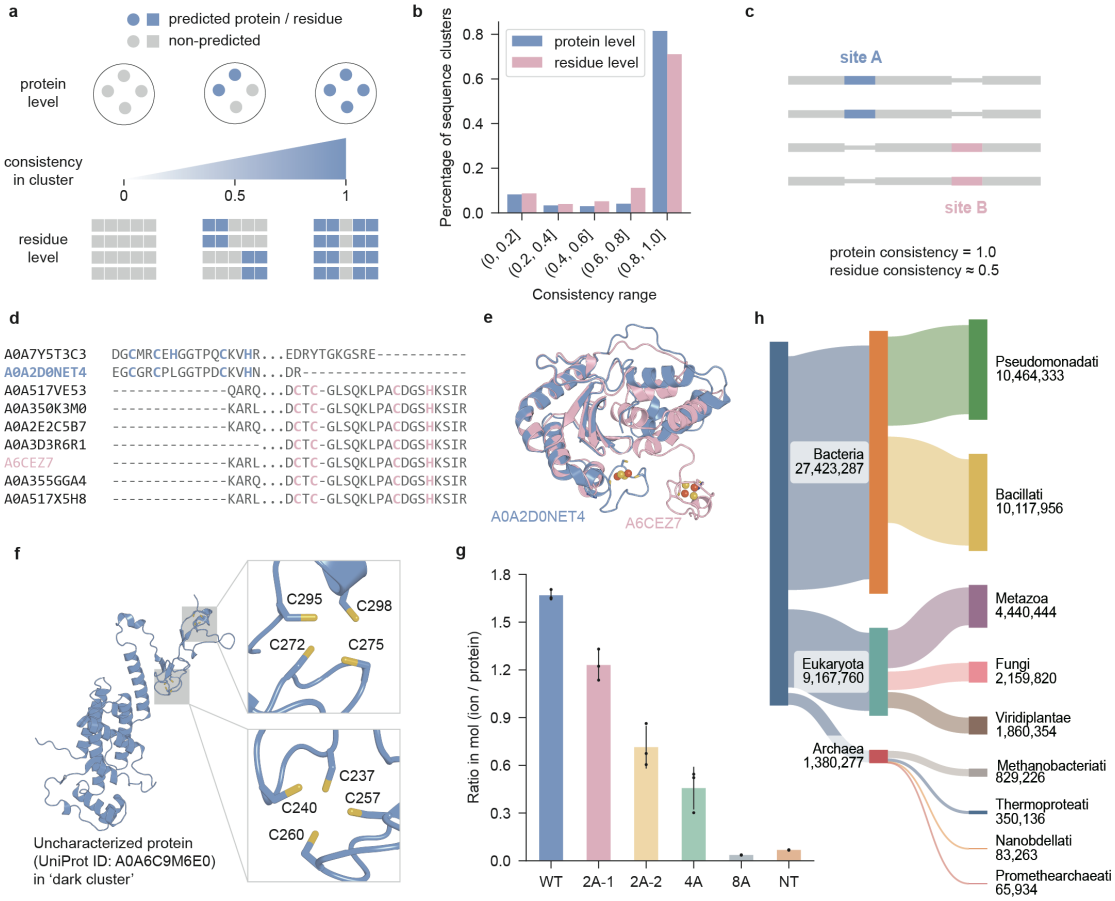

Figure 2: Analysis of sequence clusters in MetalDB. a, Diagram illustrating the predicted metal-binding consistency at the protein or residue level in a sequence cluster. **b**, Distribution of the metal-binding consistency at the protein or residue level for all MetalDB predictions across sequence clusters with at least 5 members. **c**, Diagram illustrating the concept of metal-binding site migration as discovered by the discrepancy between the protein-level and residue-level consistency within a sequence cluster. **d**, Example of metal-binding site migration within the sequence cluster of YdhG-like domain-containing proteins. Sequence alignment shows two subclusters were predicted with N-terminal (e.g., A0A2D0NET4) or C-terminal (e.g., A6CEZ7) metal-binding sites, respectively. **e**, Structural modeling of A0A2D0NET4 (blue cartoon) and A6CEZ7 (pink cartoon) demonstrate the diverged Fe_2S_2-binding sites (yellow and red balls). **f**, Structural modeling of an uncharacterized protein (A0A6C9M6E0) from the 'dark cluster' lacking existing functional annotations. Two zinc-binding sites predicted by MetalNet3 were shown in zoom-in view as sticks. **g**, Experimental validation of zinc-binding activity of A0A6C9M6E0 by ICP-MS. "WT" indicates the wild-type protein, "2A-1" indicates two residues mutation (C237A, C240A), "2A-2" indicates two residues mutation (C272A, C275A), "4A" combines the mutations in "2A-1" and "2A-2", "8A" further combines "4A" with four additional mutations (C257A, C260A, C295A, C298A), "NT" indicates a truncated N-terminal region (1-224) of the protein. **h**, Sankey plot illustrating the taxonomic distribution of MetalDB predictions across three domains of life.

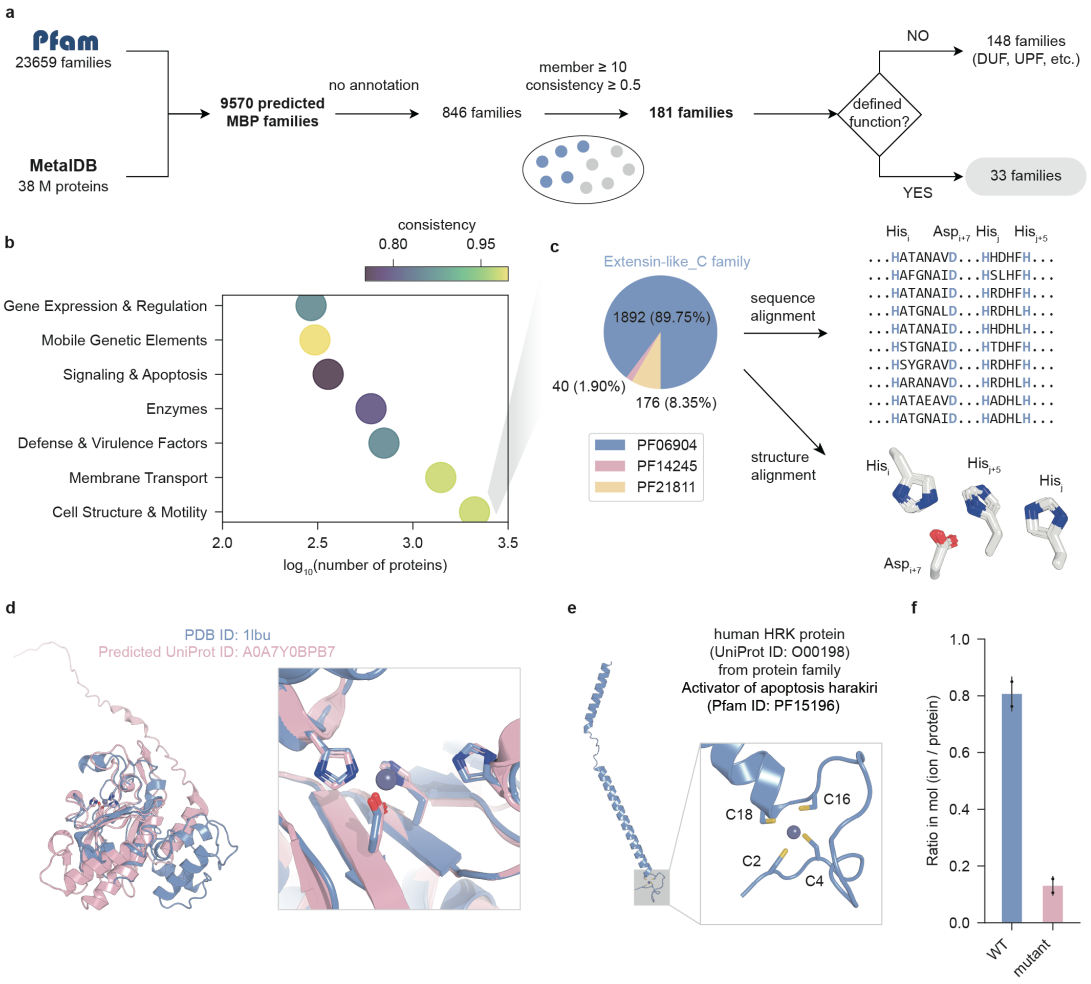

Figure 3: Discovery of Novel Metal-binding Protein Families. a, Schematic workflow of discovering new MBP families by crosschecking MetalDB with Pfam. A total of 181 protein families were predicted as novel MBP families by MetalNet3 with high confidence, including 33 and 148 families with or without defined biological functions, respectively. **b**, Classification of the 33 novel MBP families with defined biological functions into 7 categories. The position and color of each filled circle denotes the number of protein members and the predicted metal-binding consistency at the protein level in each category. **c**. Functional category of "Cell structure & motility" contains three novel MBP families with their family names, number of family members, and relative percentages shown. Ten randomly selected MBPs from the largest Extensin-like_C family (PF06904) are illustrated with the predicted metal-binding residues in sequence and structure alignment. **d**. Structure alignment of the key metal-binding residues predicted in the representative member (A0A7Y0BPB7) of Extensin-like_C family (pink) and those observed in the crystal structure of Zinc D-Ala-D-Ala carboxypeptidase (blue, PDB ID: 1LBU). **e**, AF3 modeling of the human Activator of Apoptosis Harkiri (HRK) protein (O00198) with four predicted metal-binding cysteines (C2, C4, C16 and C18) highlighted. **f**, Experimental validation of metal-binding activity of the human HRK protein by ICP-MS. WT: wild-type; Mutant: all four cysteines mutated to alanines.

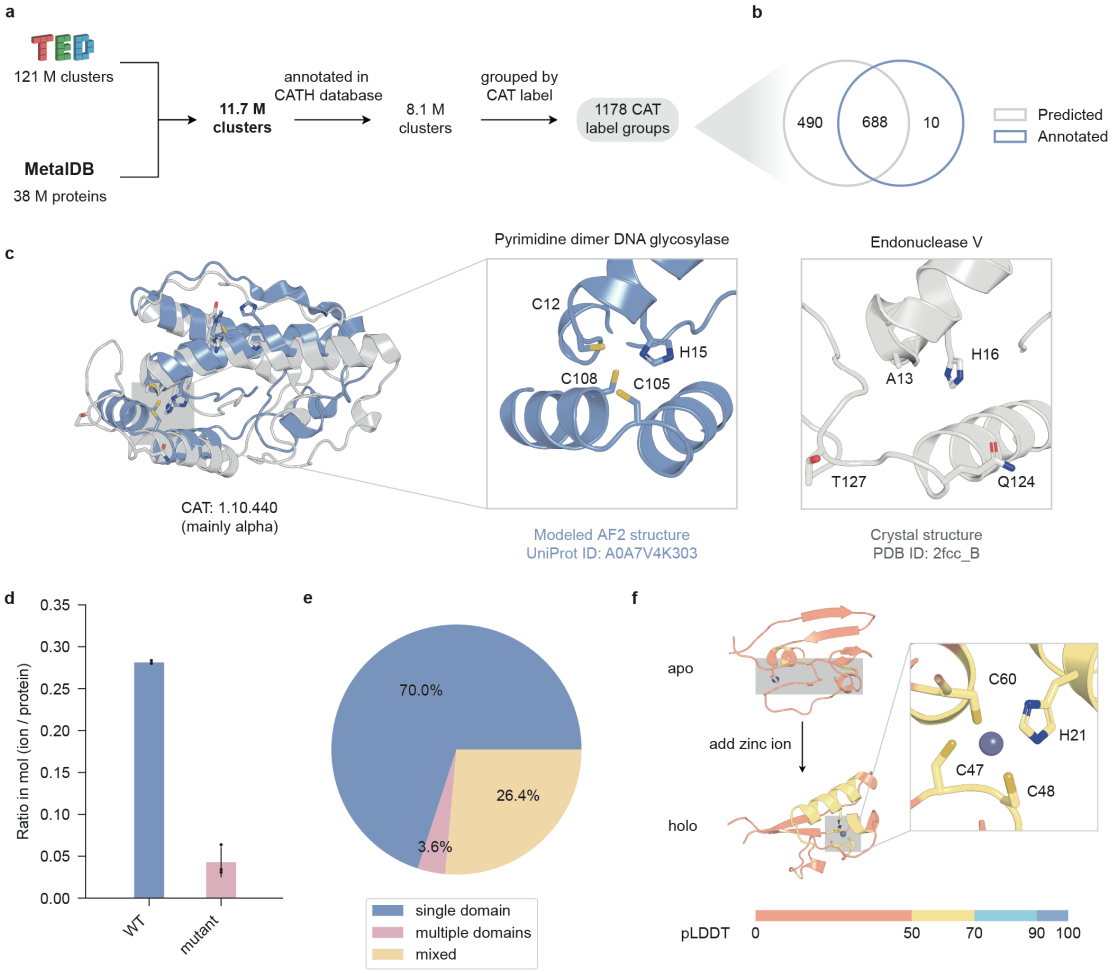

Figure 4: Discovery of Novel Metal-binding Structural Domains. **a**, Schematic workflow of identifying novel metal-binding structure domains by crosschecking MetalDB with "The Encyclopedia of Domains" (TED) database. A total of 8.1M domain clusters from 1178 CAT label groups contain predicted metal-binding sites in MetalDB. **b,** Among the 1178 CAT label groups, 688 and 490 groups are with or without existing metal-binding annotation, respectively. **c**, Structure alignment of the pyrimidine dimer DNA glycosylase from Deltaproteobacteria bacterium (AF2 model, UniProt ID: A0A7V4K303, blue cartoon) and its homologue from Tequatrovirus T4 (Crystal structure, PDB ID: 2fcc, gray cartoon) that share the same CAT label of 1.10.440. The former was predicted by MetalNet3 with a zinc-binding site containing C12, H15, C105, C108, which was lost in the latter due to mutation of these metal-chelating residues. **d**, Experimental validation of the zinc-binding activity of the pyrimidine dimer DNA glycosylase from Deltaproteobacteria bacterium by ICP-MS. WT: wild-type protein; Mutant: four metal-chelating residues mutated to alanines. **e**, Distribution of metal-binding sites in MetalDB in terms of their locations in structural domains. Single domain: sites that are entirely contained within one high/medium-consensus TED structural domain. Multiple domains: sites that are coordinated by residues spanning across two or more distinct domains. Mixed: sites that are partially located outside of the domains defined by TED domains (e.g., residues in low-consensus regions or inter-domain linkers). **f**, Structural modeling of Alpha-conotoxin

VxXXIVA (UniProt ID: J7JU64) without ion (apo) and with zinc ion (holo) by Protenix. The presence of zinc could induce the folding of an otherwise low-confidence region (pLDDT<50) into a relatively stable α-helix structure (50<pLDDT<70) with the defined metal-binding site.

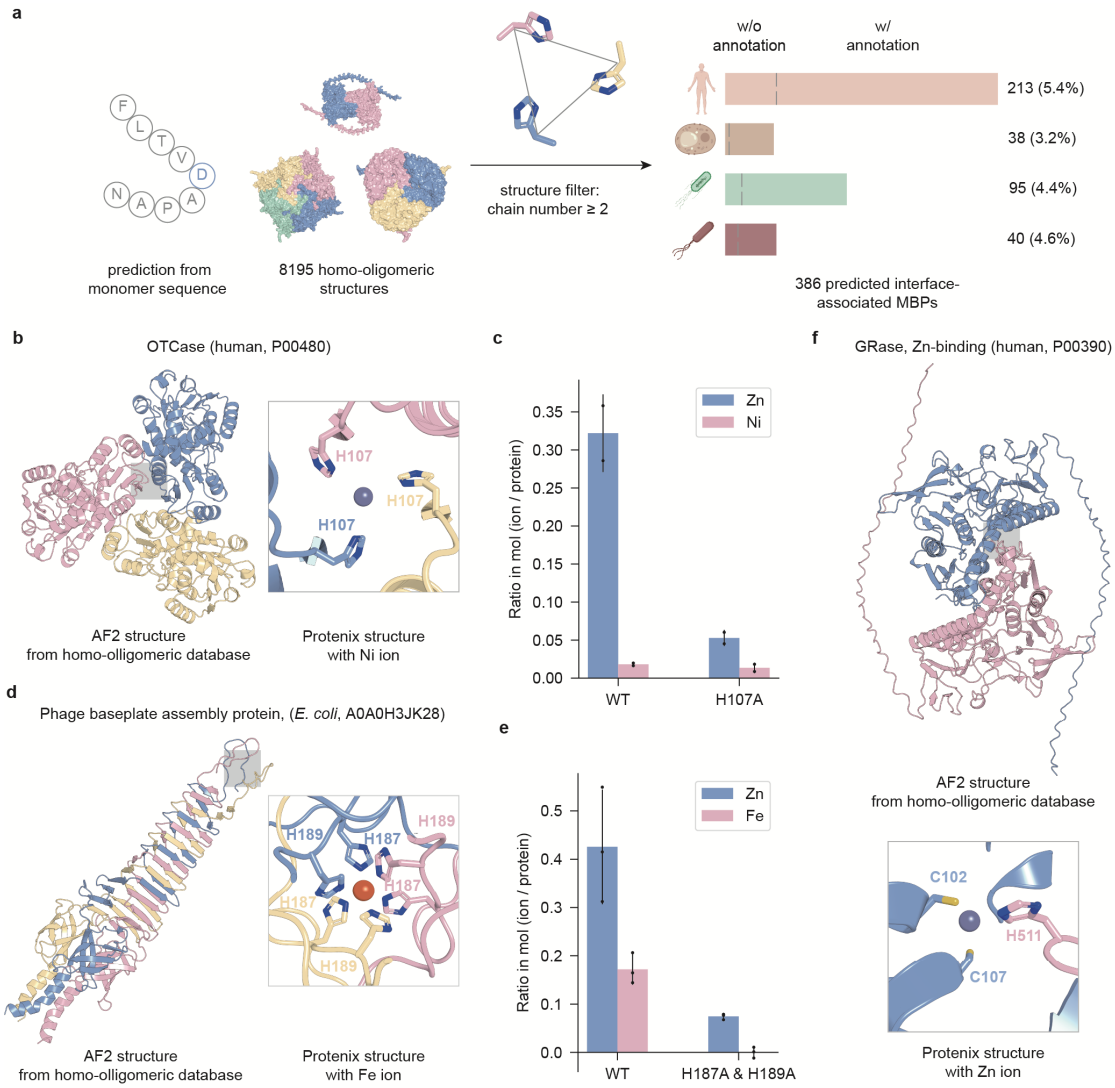

Figure 5: Prediction of metal-binding sites at homomeric protein interfaces. a, Schematic workflow of identifying metal-binding sites at the homomeric protein interface by crosschecking MetalDB with the database compiled by Schweke *et al* that contain 8195 homo-oligomeric structures from four species including *H. sapiens*, *S. cerevisiae*, *E. coli*, *P. furiosus*. For each species, the number of predicted metal-binding proteins at the homomeric protein interfaces is shown and the percentage relative to the total number of homologous oligomers provided by the database is shown in parenthesis. The dashed line separates predictions with (right) or without metal-binding annotation (left). **b,** Trimeric structure of the human OTCase with the modeled nickel-binding site. **c**, Experimental validation of the zinc-binding activity of the human OTCase over nickel by ICP-MS. **d**, Trimeric structure of *E. coli* phage baseplate assembly protein with the modeled iron-binding site. **e,** Experimental validation of the zinc-binding activity of *E. coli* phage baseplate assembly protein over iron by ICP-MS. **f**, Dimeric structure of the human glutathione reductase with the modeled zinc-binding site. In **b,d,f**, the overview structure was obtained from the homo-oligomeric database by Schweke *et al*, whereas the local structure containing the metal ion was modeled using Proteinx.

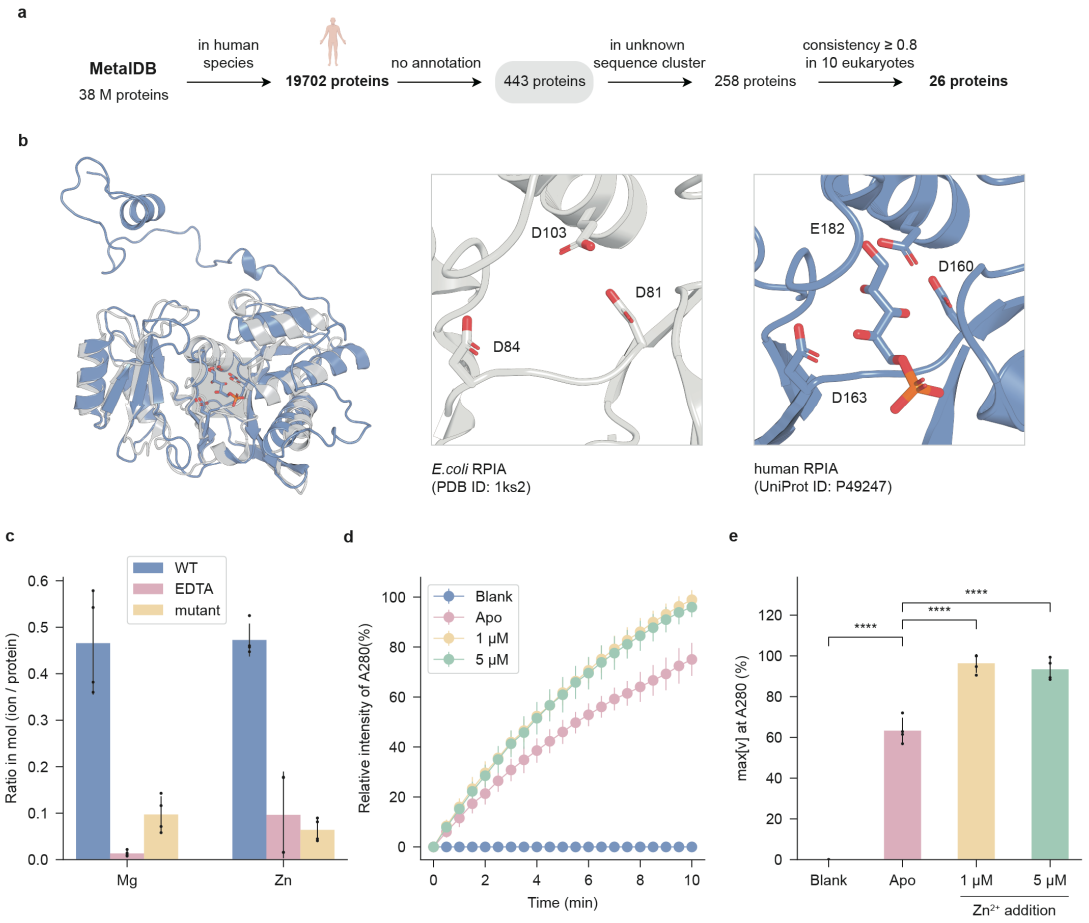

Figure 6: Functional characterization of RPIA as a novel metal-binding enzyme. **a**, Workflow for identifying novel metal-binding enzymes in the human proteome. 443 human proteins without existing metal-binding annotations were firstly extracted from MetalDB, and after filtering by membership in unknown sequence clusters and by requiring a residue level consistency higher than 0.8 among at least 10 eukaryotic homologs, a final list of 26 candidate proteins was obtained. **b**, Structural comparison of the predicted metal-binding site in human ribose-5-phosphate isomerase A (RPIA) with its bacterial homolog. Left, structure overlay of *E. coli* RPIA (grey, PDB ID: 1KS2) and human RPIA complex (blue, UniProt ID: P49247, modeled by Proteinx). Middle, zoom-in view of the catalytic site of *E. coli* RPIA including D81, D84, and D103 (sticks). Right, the corresponding region in human RPIA with ribose-5-phosphate as the substrate. The conserved substrate-binding residues (D160, D163, and E182) are also predicted as metal-binding residues by MetalNet3. **c**, Experimental validation of the zinc-binding activity and magnesium-binding of human RPIA by ICP-MS. WT: wild-type protein. Mutant: three metal-chelating residues mutated to alanines. EDTA: wide-type protein treated with EDTA to remove metals. **d**, Enzymatic activity of human RPIA with increasing concentrations of Zn^{2+}. Reaction progress was monitored by the absorbance at 280 nm. Supplementation with Zn^{2+} (1 μM or 5 μM) enhanced RPIA's activity compared to the Apo enzyme (with EDTA treatment). **e**, Maximum reaction rates derived from the kinetic traces shown in (**d**). Zn^{2+} supplementation significantly

increased the enzymatic activity of apo-RPIA. Statistical significance was determined using ANOVA followed by Holm-adjusted pairwise comparisons; ****, $P < 0.0001$.