Supplementary Materials for Manuscript

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;

This file contains:

Supplementary Tables

Table S1. Performance of the MetalNet2 model when the coevolution threshold was tuned. CHED coverage indicates the proportion of CHED residues that appear in coevolution pairs relative to the total number of CHED residues across the sequences. With the coevo_threshold of 0.01, most of CHED pairs in sequence are considered co-evolved, which indicated that the coevolution signals were not strictly required by MetalNet2 for metal-binding site prediction.

coevo_threshold	F1_avg (%)	CHED coverage (%)
0.01	73.19	99.68
0.1	75.37	67.97
0.2	74.37	53.63

Table S2. Performance of different combinations of PLM and classifier. For each combination, shown are precision, recall and F1-score, as well as the time cost for predicting 10000 sequences.

Model name	PLM	classifier	precision (%)	recall (%)	F1-score (%)	Time cost (min)
-	ESM2-150M	Autogluon	84.94	74.00	79.09	5.58
-	ESM2-650M	Autogluon	83.90	74.96	79.18	12.07
-	ProtT5-XL-UniRef50	SVM	86.99	74.26	80.12	7.87
-	ESM2-3B	SVM	87.18	74.15	80.14	17.40
MetalNet3	Ankh-base	SVM	84.92	76.48	80.48	4.69
-	Ankh-large	SVM	85.26	77.08	80.97	8.72
MetalNet2	-	-	82.56	70.02	75.77	10012.07

Table S3. Performance of the metal-type prediction by MetalNet3 on the test dataset. For each metal type, shown are precision, recall and F1-score, as well as number of samples in the test dataset.

metal type	precision (%)	recall (%)	F1-score (%)	Number of samples
ZN	87.05	76.35	81.35	854
CA	85.71	81.25	83.42	576
MG	42.93	47.02	44.89	168
SF4	89.81	88.99	89.40	109
MN	31.93	58.89	41.41	90
FE	31.18	54.72	39.73	53
NI	17.07	17.95	17.50	39
FES	90.32	80.00	84.85	35
CU	87.50	100.0	93.33	28
CO	30.00	30.00	30.00	20
F3S	54.55	66.67	60.00	9
macro average	58.91	63.08	60.53	1981

Table S4. Performance of the metal-type group prediction by MetalNet3 on the test dataset. Metal-type groups include alkaline metals (Mg, Ca), transition metals (Zn, Mn, Fe, Cu, Ni) and Fe-S clusters (Fe_2S_2, Fe_3S_4, Fe_4S_4). For each metal type group, shown are precision, recall and F1-score, as well as number of samples in the test dataset.

metal group type	precision (%)	recall (%)	F1-score (%)	Number of samples
Alkaline metals	90.15	87.37	88.74	744
Transition metals	91.18	93.45	92.30	1084
Fe-S clusters	100.0	97.39	98.68	153
macro average	93.78	92.73	93.24	1981

Table S5. Ratio of the predicted MBPs in MetalDB to the proteome size from representative species from three domains of life. For each domain of Archaea, Bacteria and Eukaryota, three representative species are selected. Shown are name of the species, taxonomy ID, proteome size, the number of predicted MBPs by MetalDB, and the relative ratio.

Taxonomy	Taxonomy ID	Proteome Size	Predicted	Ratio	Domain
Methanococcus jannaschii	243232	1787	457	0.2557	Archaea
Haloferax volcanii	309800	3922	870	0.2218	Archaea
Sulfolobus acidocaldarius	330779	2223	503	0.2263	Archaea
Escherichia coli K-12	83333	4402	818	0.1858	Bacteria
Bacillus subtilis	224308	4271	731	0.1712	Bacteria
Synechocystis sp.	1111708	3508	619	0.1765	Bacteria
Saccharomyces cerevisiae S288C	559292	6065	1040	0.1715	Eukaryota
Oryza sativa Japonica Group	39947	43668	5888	0.1348	Eukaryota
Homo sapiens	9606	20650	4513	0.2186	Eukaryota

Table S6. List of proteins in AFDB that belong to the Apoptosis Harakiri ("HRK") family (PF15196) and their predicted metal-binding sites in MetalDB.

No.	UniProt ID	Predicted residues
1	A0A096N944	C2, C4, C16, C18
2	A0A0D9SC73	C2, C4, C16, C18
3	A0A1U7R468	C2, C4, C16, C18
4	A0A2K5BYA0	C2, C4, C16, C18
5	A0A2K5KZ69	C2, C4, C16, C18
6	A0A2K5PIE7	C2, C4, C16, C18
7	A0A2K6AYL7	C2, C4, C16, C18
8	A0A2K6GL98	C2, C4, C16, C18
9	A0A2U3V4V2	C2, C4, C16, C18
10	A0A2U3VNS6	C2, C4, C16, C18
11	A0A2Y9DQC8	C2, C4, C16, C18
12	A0A2Y9HRX6	C2, C4, C16, C18

13	A0A2Y9KI81	C2, C4, C16, C18
14	A0A2Y9Q8G3	C2, C4, C16, C18
15	A0A2Y9SD47	C2, C4, C16, C18
16	A0A337SQ56	C2, C4, C16, C18
17	A0A340X3L1	C2, C4, C16, C18
18	A0A3M0K0Y0	C2, C4, C13, C16
19	A0A3Q7QF38	C2, C4, C16, C18
20	A0A452DXE5	C2, C4, C16, C18
21	A0A452H537	C2, C4, C14, C16
22	A0A485NQ56	C2, C4, C16, C18
23	A0A4W2CRW4	C2, C4, C16, C18
24	A0A4X2K3J4	C2, C4, C16, C18
25	A0A5E4BIT0	C2, C4, C16, C18
26	A0A5N3XAF9	C2, C4, C16, C18
27	A0A671EZ71	C2, C4, C16, C18
28	A0A674IJJ5	C2, C4, C14, C16
29	A0A6J1W7B0	C14, C16
30	A0A6J1Y4X9	C2, C4, C16, C18
31	A0A6J2MYH0	C2, C4, C16, C18
32	A0A6J3EVK8	C2, C4, C16, C18
33	A0A6P3FFE5	C2, C4, C16, C18
34	A0A6P3G4Z0	C2, C4, C16, C18
35	A0A6P5LYY1	C2, C4, C16, C18
36	A0A6P9AX63	C2, C4, C14, C16
37	A0A7J7S1M4	C2, C4, C16, C18
38	A0A7J7UG34	C2, C4, C16, C18
39	A0A7J8BXL8	C2, C4, C16, C18
40	A0A7J8GZT5	C2, C4, C16, C18
41	A0A811YIY8	C2, C4, C16, C18
42	F7G1G0	C2, C4, C16, C18
43	G1QHD8	C2, C4, C16, C18
44	G1T7W1	C2, C4, C16, C18
45	G3QTJ6	C2, C4, C16, C18
46	G3UH24	C2, C4, C16, C18

47	H0Y001	C2, C4, C16, C18
48	H2NIS8	C2, C4, C16, C18
49	H2Q6Y8	C2, C4, C16, C18
50	O00198	C2, C4, C16, C18
51	P62816	C2, C4, C16, C18
52	P62817	C2, C4, C16, C18
53	U3E722	C2, C4, C16, C18
54	A0A484GLG0	Not in MetalDB
55	A0A6I8PDD6	Not in MetalDB
56	W5P3K2	Not in MetalDB

Table S7. List of novel human MBPs predicted in MetalDB with conserved functions.

No.	Taxonomy	UniProt ID	Gene name
1	MYG1 exonuclease	Q9HB07	MYG1
2	Elongation of very long chain fatty acids protein 6	Q9H5J4	ELOVL6
3	Elongation of very long chain fatty acids protein 3	Q9HB03	ELOVL3
4	Protein ARV1	Q9H2C2	ARV1
5	Elongation of very long chain fatty acids protein 1	Q9BW60	ELOVL1
6	Elongation of very long chain fatty acids protein 7	A1L3X0	ELOVL7
7	Elongation of very long chain fatty acids protein 2	Q9NXB9	ELOVL2
8	Elongation of very long chain fatty acids protein 5	Q9NYP7	ELOVL5
9	Protein MAK16 homolog	Q9BXY0	MAK16
10	Ribose-5-phosphate isomerase	P49247	RPIA
11	Elongation of very long chain fatty acids protein 4	Q9GZR5	ELOVL4
12	Post-GPI attachment to proteins factor 3	Q96FM1	PGAP3
13	Putative RNA-binding protein Luc7-like 1	Q9NQ29	LUC7L
14	Putative RNA-binding protein Luc7-like 2	Q9Y383	LUC7L2

15	Mitochondrial tRNA-specific 2-thiouridylase 1	O75648	TRMU
16	Beclin-2	A8MW95	BECN2
17	Luc7-like protein 3	O95232	LUC7L3
18	ERO1-like protein beta	Q86YB8	ERO1B
19	60S ribosomal export protein NMD3	Q96D46	NMD3
20	Dolichol kinase	Q9UPQ8	DOLK
21	RNA pseudouridylate synthase domain-containing protein 2	Q8IZ73	RPUSD2
22	Sodium/hydrogen exchanger 9	Q8IVB4	SLC9A9
23	Sodium/hydrogen exchanger 7	Q96T83	SLC9A7
24	Sodium/hydrogen exchanger 1	P19634	SLC9A1
25	Sodium/hydrogen exchanger 3	P48764	SLC9A3
26	Sodium/hydrogen exchanger 5	Q14940	SLC9A5

Supplementary Figures

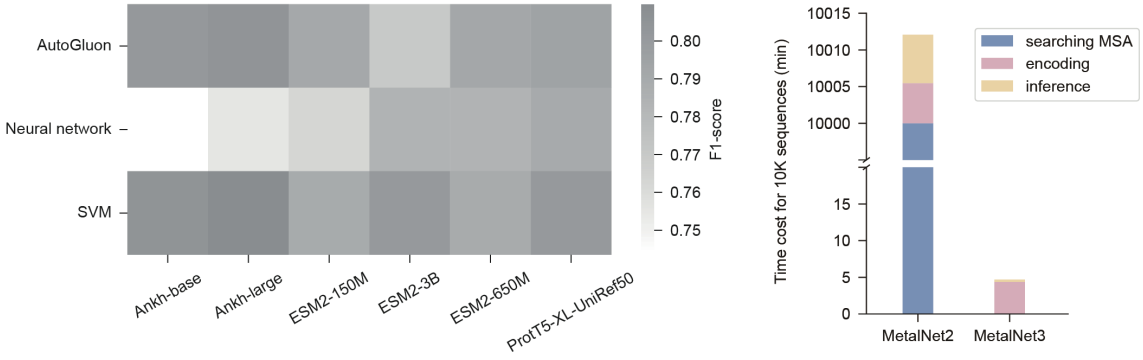

Figure S1. Joint optimization of PLM and classifier. Left, F1-score matrix of different combination of PLM and classifier on test set. Right, comparison of the time cost between MetalNet2 and MetalNet3. Note that the time cost for searching MSA is estimated as one minute per sequence.

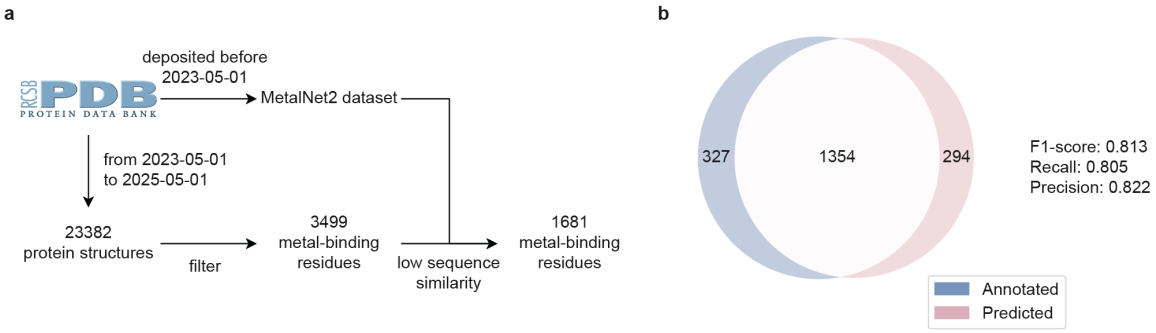

Figure S2. Evaluation of the performance of MetalNet3 on an independent metal-binding data set. **A**, Collection of an independent metal-binding data set by filtering metal-binding residues from structures that were deposited into PDB from 2023-05-01 to 2025-05-01. The filtering computational pipeline is the same as that implemented in MetalNet2. **B**, Venn diagram illustrating the performance of MetalNet3 on the independent metal-binding data set. By calculating the overlap between the predicted metal-binding residues and structurally annotated metal-binding residues, an F1-score of 0.813 was obtained with recall and precision of 0.805 and 0.822, respectively.

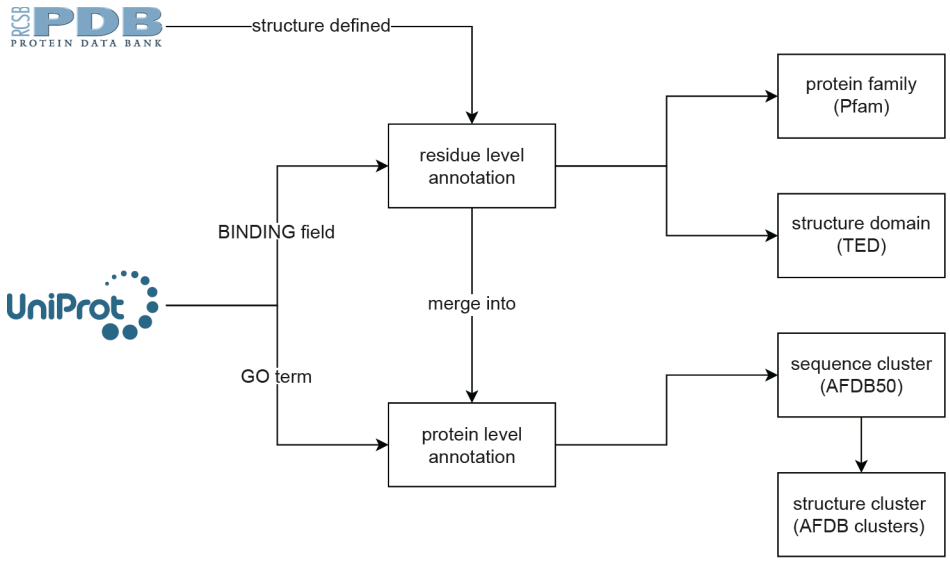

Figure S3. **Workflow of collecting metal-binding annotations for all proteins in AFDB.** Residue-level annotations are generated by combining structural information from the Protein Data Bank (PDB) with residue-specific features from UniProt, including ligand-binding site annotations (BINDING field). These residue-level annotations are further mapped to protein families (Pfam) and structural domains (TED). Protein-level annotations are obtained by merging residue-level information with Gene Ontology (GO) terms from UniProt. The resulting protein annotations are subsequently propagated to sequence clusters (AFDB50) and structure clusters (AFDB clusters).

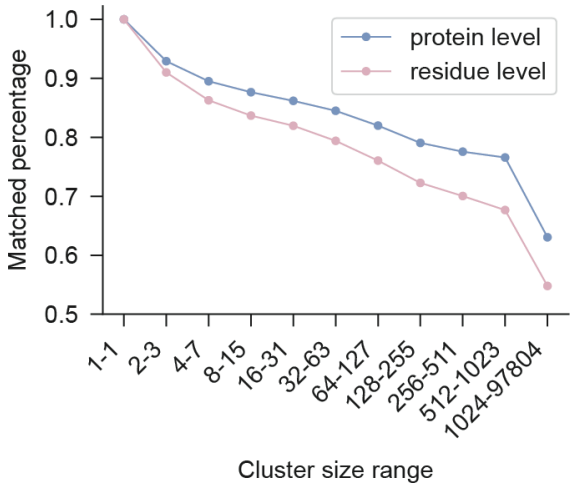

Figure S4. Changes of predicted consistency in MetalDB with respect to the size of sequence clusters. The percentage of matched annotations is shown across different cluster size ranges for protein-level (blue) and residue-level (pink) annotations. Annotation consistency decreases progressively as cluster size increases, indicating greater functional diversity within larger clusters.

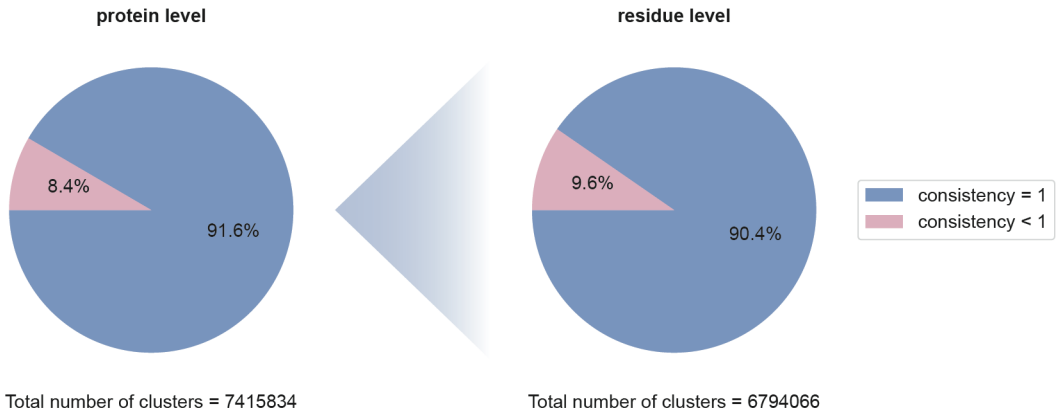

Figure S5. Discrepancy between the protein-level and residue-level consistencies in MetalDB. Pie charts show the distribution of clusters with complete consistency (consistency = 1) and incomplete consistency (consistency < 1) for the protein-level (left) and residue-level (right) predictions. Although 91.6% of sequence clusters exhibit perfect prediction consistency at the protein level, 9.6% of these clusters do not achieve perfect consistency at the residue level.

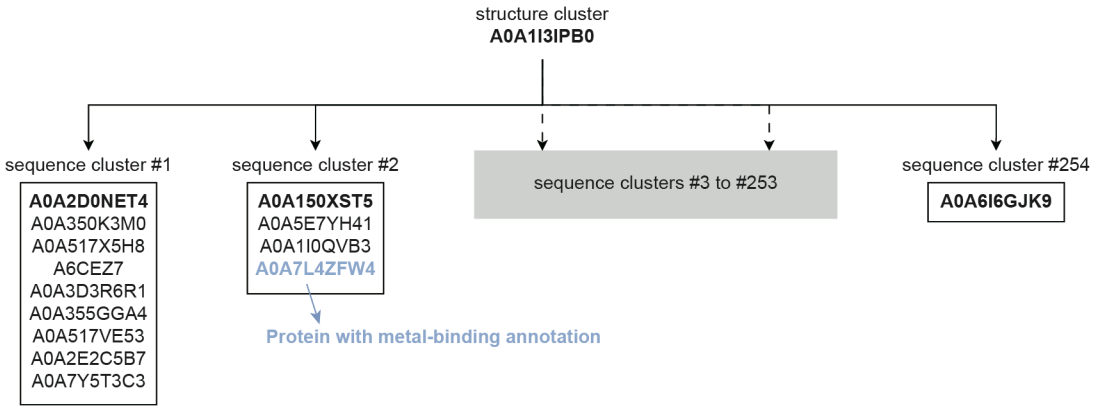

Figure S6. Metal-binding annotation of A0A2D0NET4 revealed by AFDB structural clustering. The top node represents a structural cluster (representative UniProt ID: A0A1I3IPB0) in AFDB that contains 254 sequence clusters. Selected clusters (#1, #2, and #254) are shown as examples, while the remaining 251 clusters are grouped for clarity. The UniProt ID of representative protein in each sequence cluster is highlighted with bold font. The predicted metal-binding site in A0A2D0NET4 (cluster #1) by MetalNet3 can be supported by the existing metal-binding annotation of A0A7L4ZFW4 from the neighboring cluster #2 (blue).

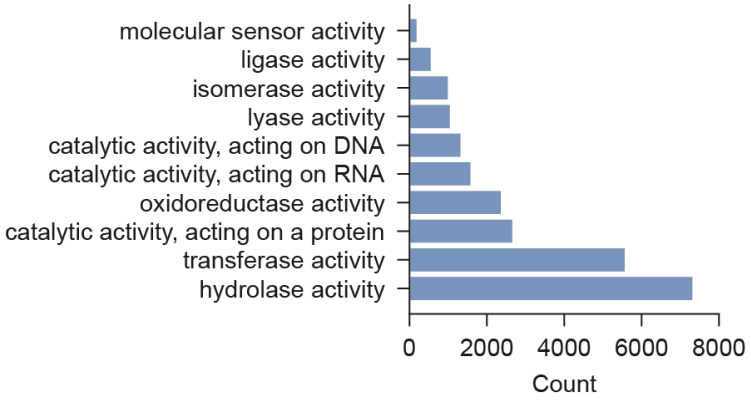

Figure S7. Gene Ontology (GO) analysis of novel metal-binding sequence clusters with high consistency in MetalDB. The top 10 GO terms are shown and 9 of them are involved in catalysis.

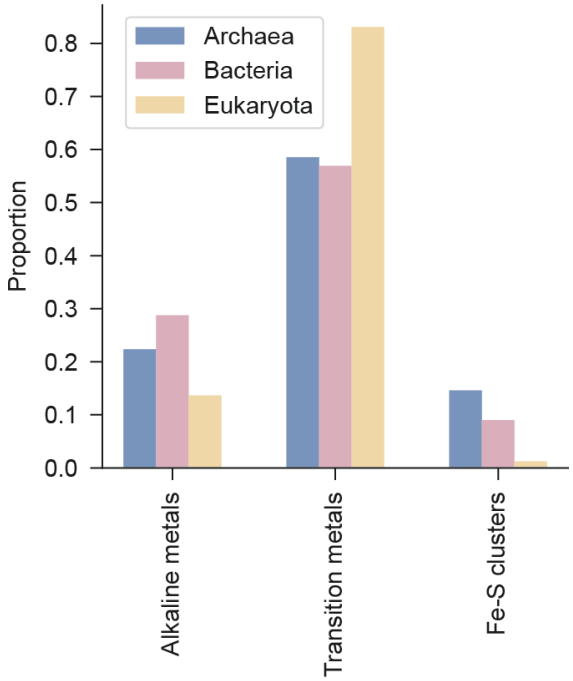

Figure S8. Proportions of predicted metal-binding groups across three domains of life. 200,000 proteins are sampled from each domain. A protein is assigned to a given metal group (alkaline, transition or Fe-S) if it contains at least three residues predicted to bind that metal.

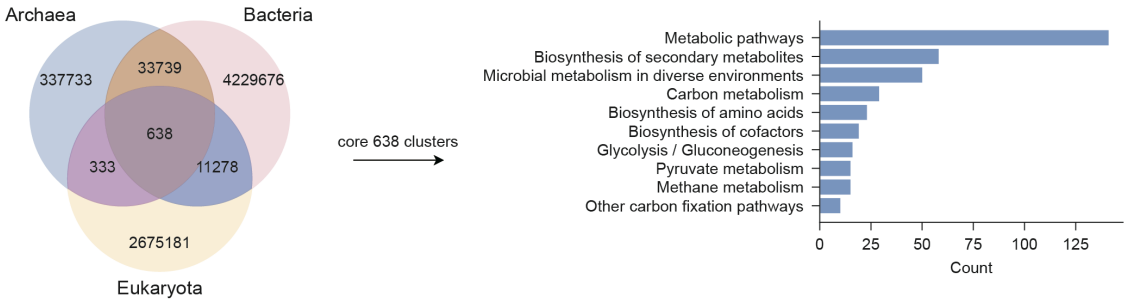

Figure S9. KEGG analysis of the predicted sequence clusters shared across all three domains of life. The overlapping region in the Venn diagram represents 638 universal core clusters shared by all three domains of life. The horizontal bar chart illustrates the functional classification of these 638 core clusters based on KEGG metabolic pathways. "Metabolic pathways" account for the largest proportion of core gene clusters, which are followed by biosynthesis of secondary metabolites and microbial metabolism in diverse environments.

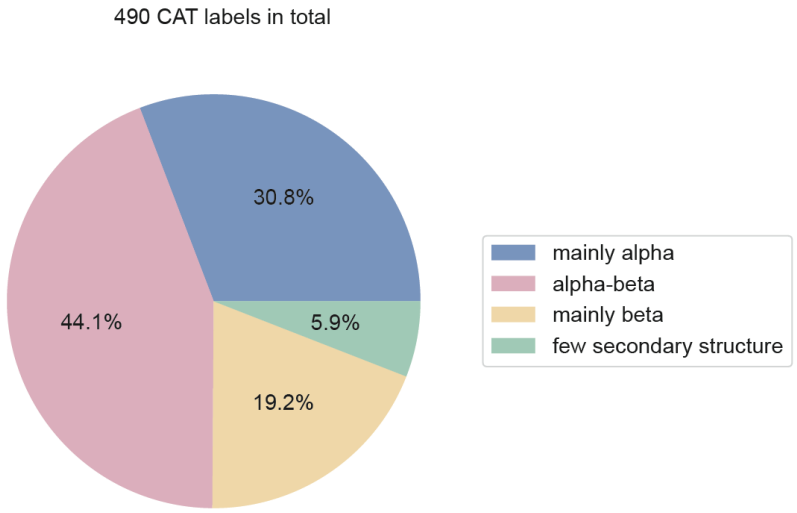

Figure S10. Composition of novel metal-binding domains in terms of structural classification. The four categories represent distinct protein secondary structure types: alpha-beta proteins account for the largest proportion at 44.1%, followed by mainly alpha proteins (30.8%), mainly beta proteins (19.2%), and proteins with few secondary structures (5.9%).

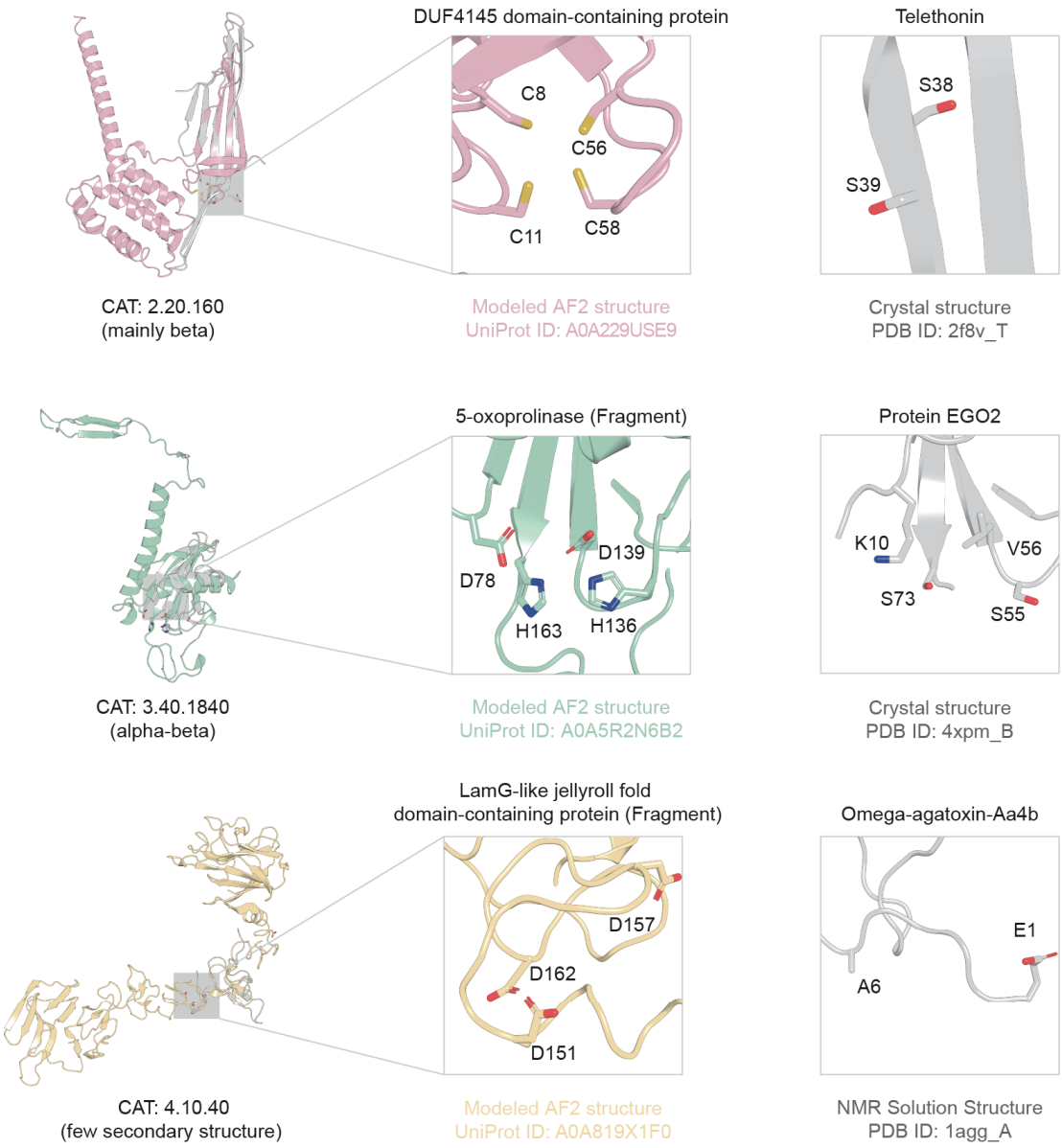

Figure S11. Structural comparison of MetalDB predictions with their corresponding domain neighbors sharing the same CAT label. Three CAT label groups are shown, including CAT 2.20.260 (mainly beta, top), CAT 3.40.1840 (alpha-beta, middle) and CAT 4.10.40 (few secondary structure, bottom). For each CAT label group, the AF2 structure for the MetalDB prediction (colored cartoon) and the PDB structure for its domain neighbor (gray cartoon) are overlaid with the zoom-in view of predicted metal-binding sites shown.

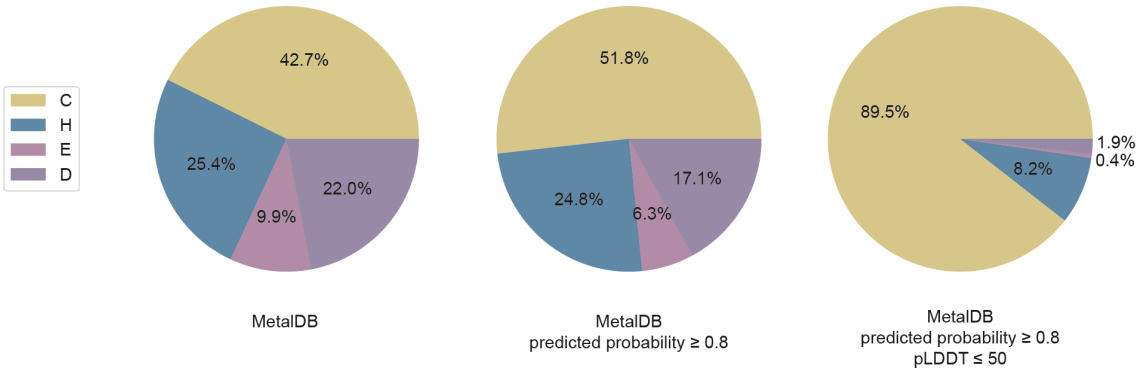

Figure S12. Enrichment of cysteines in metal-binding sites located in structurally flexible regions. CHED residue compositions in metal-binding sites are counted for each group of predictions in MetalDB. Left, all predictions in MetalDB. Middle, predictions in MetalDB with predicted probability of residue no less than 0.8. Right, predictions in MetalDB with predicted probability of residue no less than 0.8 and pLDDT no more than 50.

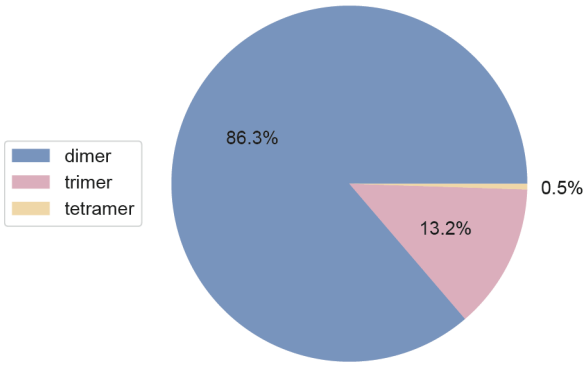

Figure S13. Composition of different types of homomeric proteins compiled by Schweke *et al* (n=386). Dimer is the dominant form, accounting for 86.3% of the total entries, followed by trimer (13.2%), while tetramer only makes up a tiny fraction of 0.5%**.**

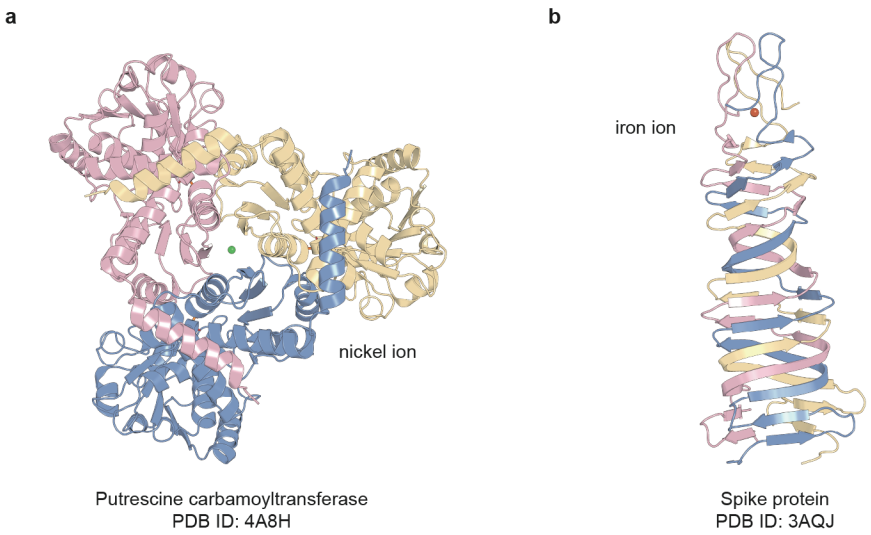

Figure S14. PDB structures of trimeric protein complexes with metal-binding sites at the interface. **a**, Trimeric structure of putrescine carbamoyltransferase (PDB ID: 4A8H) with nickel ion-binding (green ball). **b,** Trimeric structure of spike protein (PDB ID: 3AQJ) with iron ion-binding (brown ball).

```
h.sapiens       ...AIDGADEVDADLNLIKGGGGCLTQEKIV...
b.taurus        ...AIDGADEVDADLNLIKGGGGCLTQEKIV...
m.musculus      ...AIDGADEVDAELNLIKGGGGCLTQEKIV...
x.tropicalis    ...AIDGADEVDSELNLIKGGGGCLAQEKIV...
d.rerio         ...AIDGADEVDTALTLIKGGGGCLTQEKIV...
a.gambiae       ...AIDGADEVDAKMVLIKGGGGCLLQEKIV...
c.elegans       ...CIDGADEVDGQFTCIKGGGGCLAQEKIV...
s.cerevisiae    ...AFDGADEVDENLQLIKGGGACLFQEKLV...
a.thaliana      ...SIDGADEVDPYLNLVKGRGGSLLREKMI...
```

Figure S15. Multiple sequence alignment of RPIA in representative eukaryotic species. Red labeled residues indicate the predicted metal-binding residues, which show high sequence conservation.

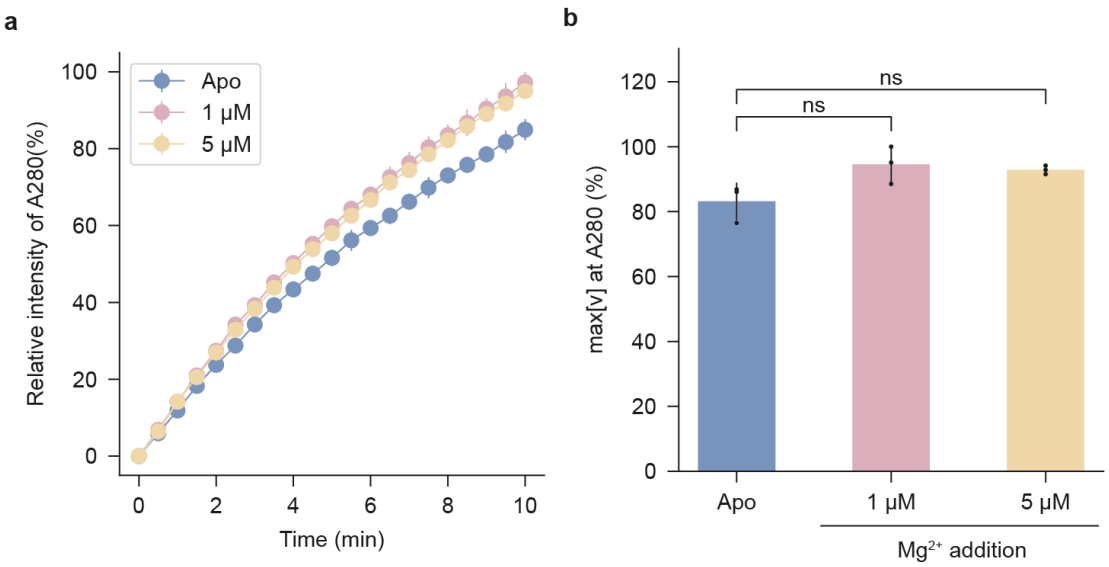

Figure S16. Evaluation of RPIA activity under Mg^{2+} treatment. **a**, Enzymatic activity of human RPIA with increasing concentrations of Mg^{2+}. Reaction progress was monitored by the absorbance at 280 nm. Supplementation with Mg^{2+} (1 μM or 5 μM) did not change RPIA's activity significantly compared to the Apo enzyme (with EDTA treatment). **b**, Maximum reaction rates derived from the kinetic traces shown in (a). Mg^{2+} supplementation did not significantly affect the enzymatic activity of apo-RPIA. Statistical significance was determined using ANOVA followed by Holm-adjusted pairwise comparisons; ns, not significant.

Supplementary Methods

Optimization of the single-sequence prediction model. The optimization framework focused on two primary components: protein language model (PLM) encoding and downstream classifier architecture. Performance was evaluated based on the F1-score, reported as the average across three independent validation runs. Inference benchmarks were conducted on a single NVIDIA L40 GPU with a batch size limit of 16,384 tokens (reduced to 4,096 for ESM2-3B to accommodate memory constraints).

All PLMs utilized in this study are based on the Transformer architecture and were implemented via a unified interface using the transformers library. The evaluated models included: ESM2 series[1] (ESM2-150M, ESM2-650M, ESM2-3B), Ankh series[2] (Ankh-base, Ankh-large), ProtTrans[3] (ProtT5-XL-UniRef50). For classifier architectures, three classifier types were evaluated on top of the PLM embeddings: (i) Support Vector Machine (SVM). An RBF kernel was utilized with GPU acceleration provided by the cuml library[4]. The regularization parameter C was optimized on a logarithmic scale (0.01 to 100) using Optuna. (ii) Ensemble Baseline. AutoGluon[5] was used without manual modification to serve as an automated ensemble baseline. (iii) Multi-Layer Perceptron (MLP). Hyperparameter management was handled via hydra. Optimization was conducted sequentially in the following order: learning rate, depth (number of hidden layers), sampling strategy (weighted vs. standard), loss function weighting (class weights and focal loss gamma), hidden dimensionality, dropout rate, and batch size. Notably, the extensive hyperparameter search for the MLP was performed exclusively using the ESM2-650M encoding.

Identification of high-confidence metal-binding sites to construct MetalDB. High-confidence metal-binding sites were defined based on spatial proximity and graph connectivity. Two residues were considered spatially adjacent if their C-beta atoms were within 15 Å. Consequently, an initial metal-binding site was defined as a clique ("a fully connected graph") of no less than 3 mutually adjacent residues. In the context of homomers, valid sites were restricted to those formed at the interfaces of at least three chains. We then grouped these residues into site-level predictions using a clique-merging strategy, in which a graph where cliques sharing common residues were iteratively merged. The final binding sites corresponded to the connected components of this graph. After this filtering, MetalDB with 38 M proteins and 46 M sites is constructed. The implementation utilized the find_cliques and connected_components modules in the networkx Python library[5].

Metal-binding annotation in AFDB. Metal-binding annotations were integrated from UniProt and PDB. Site-level annotation protocols generally followed those described in MetalNet2[6], with two key modifications: (1) the inclusion of all entries from both Swiss-Prot and TrEMBL, and (2) the use of PDB-provided ID mappings to link PDB chains with UniProt accessions. Protein-level annotations were derived from UniProt Gene Ontology (GO) terms. Proteins associated with "metal ion binding" (GO:0046872), "metal cluster binding" (GO:0051540), or any of their descendant terms were classified as metal-binding.

Additionally, proteins possessing site-level annotations were incorporated into the protein-level dataset.

Site-level annotations were subsequently mapped to Pfam families and TED domains[7]. Specifically, if a metal-binding residue was located within a Pfam region or TED domain, the corresponding family or domain was designated as metal-binding. All TED domains, including redundant entries, were mapped to their respective cluster representatives in TED100. For AFDB50 sequence clusters, annotation was assigned if: (1) at least one member protein possessed a known annotation, or (2) the parent structural cluster (AFDB-cluster) contained at least one other annotated sequence cluster. Ultimately, these annotations were propagated across Pfam families, TED100 clusters, and AFDB50 clusters to generate the comprehensive AFDB annotation set. Groups entirely devoid of constituent proteins with metal-binding annotations were defined as non-annotated.

Prediction consistency. Prediction consistency was evaluated for each sequence cluster. At the protein level, consistency was calculated as the proportion of members predicted to be positive within the cluster. To quantify residue-level consistency, sequences within each cluster (size n) were aligned using Clustal Omega[8]. Predicted binding residues were mapped to the MSA coordinates. We then identified the total number of unique alignment columns containing at least one predicted metal-binding residue (denoted as m; representing the union of predicted sites). The consistency score was calculated by summing the number of predicted residues for each individual sequence (k_i) and dividing this sum by m, $\frac{1}{n}\sum_{i=1}^{n}\frac{k_i}{m}$.

Construction, expression, and purification of potential metal-binding proteins. Human cDNAs of candidates were obtained from the National Center for Protein Sciences (Beijing) – Protein and Gene Platform at Peking University. Prokaryotic genes were synthesized by Beijing Hitrobio Biotechnology Co., Ltd. The coding sequences were cloned into the pGEX-6P-1 expression vector to generate GST-tagged constructs for prokaryotic expression. Proteins were expressed in *E. coli* BL21 (DE3) cells by being cultured in LB medium (containing 5 g of yeast extract, 10 g of NaCl and 10 g of tryptone per 1 liter) with 100 μg/mL ampicillin at 37°C until the OD at 600 nm reached between 0.6 and 0.8. Protein expression was induced by the addition of IPTG to a final concentration of 0.5 mM. After induction of expression at 16°C for 20 h, bacterial cells were harvested by centrifuging at 4,000 × g for 30 min and lysed in PBS buffer by sonication. The supernatant was collected after centrifugation at 12,000 × g for 30 min and incubated with glutathione affinity resin (Smart-Lifesciences, SA008100) for 1 h 4°C. After extensive washing with PBS to remove nonspecifically bound proteins, the target proteins were eluted with buffer containing 20 mM reduced glutathione. The eluted protein was concentrated by ultrafiltration centrifugation (Millipore, UFC900324, MWCO = 3 kDa) and further purified by size-exclusion chromatography on a Superdex 75 10/300 GL (GE

Healthcare) equilibrated with PBS. Protein concentrations were measured by NanoDrop 2000 (Thermo Fisher Scientific) based on its calculated molecular weights and extinction coefficients of the purified proteins.

ICP-MS Analysis. For each sample, at least 50 μg of protein was used, and the initial sample volume was kept below 50 μL. Samples were digested by adding 200 μL of nitric acid (MOS grade) and 20 μL of hydrogen peroxide (MOS grade), followed by incubation at 70 °C for 1–5 h. When the sample volume exceeded 50 μL, the volumes of nitric acid and hydrogen peroxide were increased proportionally to maintain the same digestion reagent-to-sample ratio. The samples were then diluted with ddH_2O, and the metal contents were analyzed by inductively coupled plasma mass spectrometry (ICP-MS) using an Elan DRC-e (PerkinElmer) spectrometer.

RPIA activity analysis. The enzymatic activity of RPIA was assessed as previously described[9]. Briefly, EDTA-pretreated apo-RPIA (final concentration, 50 nM) was incubated with the indicated concentrations of $MgCl_2$ or $ZnCl_2$ in 100 μL of reaction buffer (50 mM Tris-HCl, pH 7.4) at room temperature for 30 min to allow metal ion reconstitution. In parallel, the substrate D-ribose 5-phosphate (MedChemExpress, HY-W009371) was prepared at a final concentration of 20 mM in an additional 100 μL of the same reaction buffer. The enzyme and substrate solutions were then rapidly combined in 96-well UV-transparent microplates (Greiner, GN655801), and the reaction was immediately monitored by recording the absorbance at 280 nm in kinetic mode using a microplate reader.

Supplementary References

1. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
2. Elnaggar, A. *et al.* Ankh: Optimized Protein Language Model Unlocks General-Purpose Modelling. *arXiv preprint arXiv:2301.06568* (2023).
3. Elnaggar, A. *et al.* ProtTrans: Towards Cracking the Language of Lifes Code Through Self-Supervised Deep Learning and High Performance Computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1–1 (2021) doi:10.1109/TPAMI.2021.3095381.
4. Raschka, S., Patterson, J. & Nolet, C. Machine Learning in Python: Main developments and technology trends in data science, machine learning, and artificial intelligence. *arXiv preprint arXiv:2002.04803* (2020).
5. Hagberg, A. A., Schult, D. A. & Swart, P. J. Exploring Network Structure, Dynamics, and Function using NetworkX. in *Proceedings of the 7th Python in Science Conference* (eds Varoquaux, G., Vaught, T. & Millman, J.) 11–15 (Pasadena, CA USA, 2008).
6. Zhang, F. *et al.* MetalNet2: an enhanced server for predicting metal-binding sites in proteomes. *National Science Review* **11**, nwae391 (2024).
7. Lau, A. M. *et al.* Exploring structural diversity across the protein universe with The Encyclopedia of Domains. *Science* **386**, eadq4946 (2024).
8. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* **7**, MSB201175 (2011).

9. 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).