Structural bioinformatics and deep learning of metalloproteins: recent advances and applications

Claudia Andreini[1,2], Antonio Rosato[1,2,*]

[1] Consorzio Interuniversitario di Risonanze Magnetiche di Metallo Proteine, Via Luigi Sacconi 6, 50019 Sesto Fiorentino, Italy.

[2] Department of Chemistry and Magnetic Resonance Center (CERM), University of Florence, Via Luigi Sacconi 6, 50019 Sesto Fiorentino, Italy.

***Correspondence:**

Antonio Rosato: rosato@cerm.unifi.it (https://orcid.org/0000-0001-6172-0368)

Via Luigi Sacconi 6

50019, Sesto Fiorentino

Italy

Abstract

All living organisms require some metal ions for their energy production as well as metabolic and biosynthetic processes. Within cells, metal ions are involved in the formation of adducts interact with metabolites and macromolecules (proteins and nucleic acids). The proteins that require binding to one or more metal ions to be able to carry out their physiological function are called metalloproteins. About one third of all protein structures in the Protein Data Bank involve metalloproteins. Over the past few years there has been a tremendous progress in the number of computational tools and techniques making use of 3D structural information to support the investigation of metalloproteins. This trend has been boosted also by the successful applications of neural networks and deep learning approaches in molecular and structural biology at large. In this review, we discuss recent advances in the development and availability of resources dealing with metalloproteins from a structure-based perspective. We start by addressing tools for the prediction of metal-binding sites (MBSs) using structural information on apo-proteins. Then, we provide an overview of methods for and lessons learned from the structural comparison of MBSs in a fold-independent manner. We then move to describing databases of metalloprotein/MBS structures. Finally, we summarize recent DL applications enhancing the functional interpretation of metalloprotein structures.

Keywords

Bioinorganic chemistry; metal-binding; structural biology; zinc; iron; copper; transition metals.

 doi:10.20944/preprints202206.0253.v1

1. Introduction

Living organisms require a variety of metal ions for their optimal functioning [1,2]. The roles of metal ions in cellular and biochemical processes are many, including: the stabilization of the three-dimensional (3D) structure of macromolecules; the direct participation in the catalytic mechanism of enzymes; the transfer of electrons to/from other molecules; the regulation of biological processes. In line with their importance, the concentration of metal ions in the cell is tightly regulated [3,4]. This relies upon the combined action of transport, delivery, storage, detoxification, and efflux machineries. Bacterial pathogens share the same requirements for metal ions as all other organisms https://portlandpress.com/biochemsoctrans/article-abstract/47/1/77/108/The-role-of-metal-ions-in-the-virulence-and?redirectedFrom=fulltext, which they need to acquire from the host organism. Thus, the host can deploy a protective mechanism, called nutritional immunity, which inhibits the growth of pathogens by limiting the availability of crucial metal ions Nutritional immunity: the battle for nutrient metals at the host–pathogen interface | Nature Reviews Microbiology. A similar related strategy can be pursued also through pharmacological treatment https://pubs.acs.org/doi/10.1021/acs.inorgchem.9b01029.

This review will focus on proteins requiring one or more metal ion to be able to carry out their biological function, or for achieving their correct fold. These are known as metalloproteins (MPs). MPs can bind individual metal ions directly into their specific binding sites. In parallel, there is an extensive casuistry of metal-containing cofactors, ranging from polymetallic clusters, which can be homo- (such as iron-sulfur clusters) or hetero-metallic (such as the FeMo cofactor), to organic molecules forming metallic complexes that are then incorporated into the protein (such as cobalamin or protoporphyrin IX). In the PDB, 38% of the entries contain at least one metal ion, [5,6] while it has been estimated that no less than 40% of enzymes require metal ions for their biological function. [7,8] The reactivity and physiological role of metal ions in MPs is largely determined by local protein structure environment through the modulation of how the metal is positioned in the active site, of how it interacts with the substrate and, for redox-active metals, of its reduction potential.[9,10]

Recent years have witnessed a steady growth in the application of bioinformatics methods to the investigation of MPs. In this context, the first area of application has been the prediction of the occurrence of metal-binding sites based only on protein sequence information, as the result of the success of genomics initiatives [11-15]. There are numerous reviews of these methodological developments, also recently published [16-18]. For this reason, sequence-based tools for the prediction of MPs will not be addressed here. A field of application that has received significant attention is 3D-structure-based prediction of the occurrence of metal-sites, which leverages the knowledge about the relative position in space of the amino acids potentially providing donor atoms for metal coordination. Indeed, a significant boost in this kind of methods is being received from the success of AlphaFold and AlphaFold2 in the CASP initiatives, [19-22], which have greatly improved the availability of useful 3D structural models for proteins not yet characterized experimentally [23-25]. With an increase in the number of MP structures available, there is also an increase in the opportunity to apply structural comparison methods to identify functional and/or evolutionary links within and among MP families. The latter is thus another topic of interest for this work. Finally, recent updates regarding databases relevant for the study of MPs will be covered, next to recent applications of deep learning methods (DL) to these systems. Table 1 summarizes the resources and applications mentioned in this contribution; we focused mostly on developments achieved in the past decade, except for examples that constituted important conceptual innovation.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 June 2022 doi:10.20944/preprints202206.0253.v1

Table 1. **Summary of all the resources mentioned in this article.** The resources are listed in the same order as they are discussed in the corresponding sections.

Tool name	Implemented approach	Reference
Template-based methods		
	Identification of cavities with high hydrophobicity contrast	[30]
CHED	Identification of suitable arrangement(s) of triads of the CHED residues based on the distances between candidate donor atom	[31]
IonCom	Integration of four structure-based predictors and a novel sequence-based predictor	[33]
MIB	Docking MBS templates with the fragment transformation method	[37]
ZINCCLUSTER	Detection of known structural patterns	[38]
Predictive algorithm in the GaudiMM modeling suite	Identification of accessible cavities whose center of mass is within 3.5 Å from the β-carbon atoms of three or more CHED residues	[40]
BioMetAll	Identification of cavities followed by their validation against pre-defined geometric patterns of the protein backbone	[41]
	Docking MBS templates with geometric hashing against an ensemble of 11 structural conformations for the query protein, generated with coarse-grained molecular mechanics	[43]
Random forest methods		
Zincbindpredict	Application of a portfolio of predictive models, each optimized to detect a specific type of zinc-binding site. Each type corresponds to a different zinc-binding patterns.	[44]
	Prediction of positions where metal ligands can be introduced, based on protein backbone coordinates, to design artificial MPs	[45]
Structural comparison of metal sites		
MetalS^2	Pairwise metal-centered superposition of MBSs based on a combination of sequence and structural similarity	[46]
MetalS^3	A web server using an optimized version of MetalS^2 to search the MetalPDB database for MBSs structurally similar to the query	[49]
mFASD	A structure-based algorithm to predict which metal populates a MBS based on systematic comparison against a template library	[50]
MeCOM	Pairwise superposition of MBSs based on a combination of site features and the position of the Cα atoms	[51]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 June 2022 doi:10.20944/preprints202206.0253.v1

TopMatch + Sahle	Scoring of pairwise structural superpositions computed by the TopMatch tool, which ignores metal ions, with the sahle function to detect alignments having a good overlap of the MBSs	[54]
Metalloprotein databases		
MetalPDB	MetalPDB collects structural information on all the MBSs present in the Protein Data Bank	[6]
BioLiP	A database collecting structures of protein adducts, including metal-protein complexes	[59]
ZincBind	A database specialized on zinc-binding sites built on biological assemblies	[60]
PyDISH	PyDISH is specialized on the analysis of heme-binding sites in PDB structures	[61]
VirusMED	A database of epitopes, drug binding site and metal binding sites in viral proteins of known 3D structure	[63]
InterMetalDB	A database of MBSs occurring at macromolecular interfaces, built on biological assemblies	[28]
MetLigDB	MetLigDB focuses on the structural and chemical properties of small molecules that bind directly to the metal ion(s) in MP structures	[67]
MeLAD	A database derived from the 3D structures of all metalloenzyme-ligand adducts, which integrates detailed analyses of metal-binding pharmacophores, metalloenzyme structural similarity and ligand chemical similarity	[68]
AI methods applied to metalloproteins		
	Use of conditional variational autoencoders for the automated design of artificial metalloproteins	[72]
	Identification of disease-related mutations through a multichannel convolutional neural network (MCCNN)	[73]
DeepCys	Discrimination of four cysteine different roles, i.e. metal-binding, disulphide formation, sulphenylation and thioether	[75]
MAHOMES	Discrimination of enzymatic and non-enzymatic metals in MPs	[76]
AlphaFill	A database derived from AlphaFold predictions of apo-proteins where holo-structures of MPs have been reconstructed	[81]
bindEmbed21	bindEmbed21 uses a combination of homology-based inference and a convolutional neural network to predict whether a protein residue binds to a metal ion, a nucleic acid, or a small molecule	[87]
mebipred	Sequence-based prediction of MPs using a NN trained with information derived from 3D structures	[89]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 June 2022 doi:10.20944/preprints202206.0253.v1

	Discrimination of physiological and adventitious zinc-binding sites in MPs using a recurrent neural network (RNN)	[90]

2. Structure-based definition of metal binding site (MBS)

In this review we will refer often to the metal binding site (MBS) as a substructure that can be defined in a way such that it is possible to automatically extract these sites from 3D structures deposited in the Protein Data Bank (PDB) [26]. Different research teams have proposed different definitions, which however typically tend to produce comparable results in all downstream analyses of MBSs. Typically, the concept is that of extracting a substructure around the metal ion(s) that represents the macromolecular environment sensed by the metal. This substructure should correspond to the minimal environment determining the function of the metal, i.e., the "minimal functional site" [27].

In our own work in this field, to build a MBS we implemented a protocol that starts with the identification of the metal ion and its donor atoms, i.e. the atoms that form a coordination bond with the metal (Figure 1). Metal ligands are then defined as the amino acids, nucleotides or other chemical entities (e.g. mono- or poly-atomic anions) that contain at least one donor atom (cyan sticks in Figure 1). The metal ligands provided by protein or nucleic acid residues are called *endogenous* ligands, whereas the metal ligands provided by other chemical entities are called *exogenous* ligands. In proteins, the identity and spacing along the sequence of the amino acid ligands define the metal-binding pattern (MBP) of the metal-binding site. For example, a common MBP in zinc fingers is CX(2)CX(12)HX(2)H, where X denotes any amino acid. To extend the selection to include the environment around the metal and its ligands, we add to the MBS any other residue or chemical species having at least one atom within 5.0 Å from a metal ligand (orange residues in Figure 1). A simpler approach to the definition of the MBS adopted by some authors is simply to include any protein/nucleic acid residue or chemical species having at least one atom that is at a distance lower than an arbitrary threshold from the metal (e.g. [28]). In other words, with the latter definition, a sphere of fixed radius is centered on the metal and the MBS is computed as the ensemble of all the residues or molecules that have at least one atom contained in the sphere.

Usually MBSs do not correspond to continuous stretches of protein sequence. Rather, they are groups of sequence fragments of different lengths, depending on the number and position of the metal ligands, namely the MBP. The fragmented structure of MBSs makes their structural comparison not always possible with standard tools for protein structure superposition, contingent on the specifics of the algorithm used by each tool. Therefore, in some cases *ad hoc* approaches have been developed (Section 4).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 June 2022 doi:10.20944/preprints202206.0253.v1

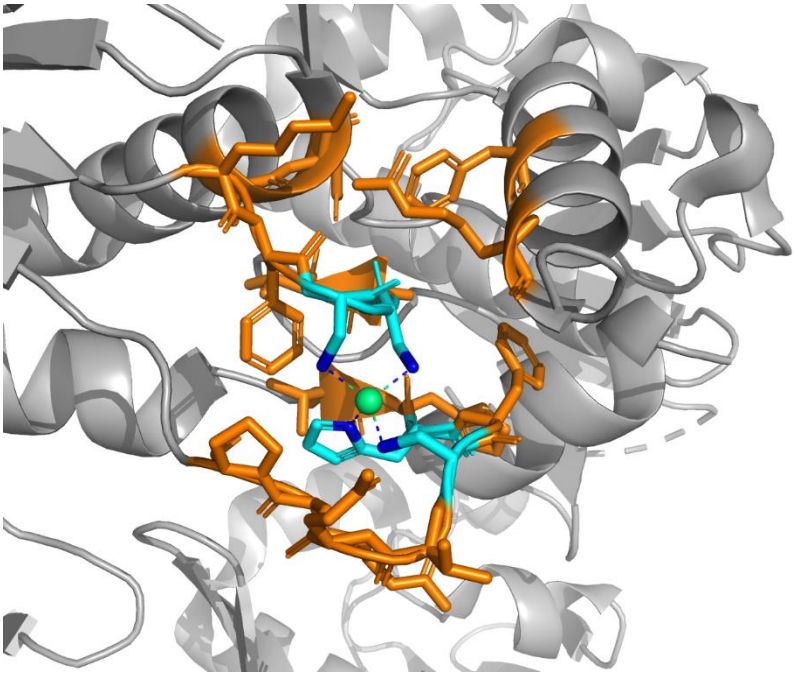

Figure 1. Definition of a MBS according to the MetalPDB protocol. For each metal atom in a given 3D structure, the non-hydrogen non-carbon atoms at a distance smaller than 3.0 Å from the metal ion (green sphere) are identified as its donor atoms (blue atoms), i.e. the atoms that bind directly to the metal. The protein residues or small molecules that contain at least one donor atom are the metal ligands (cyan sticks), and constitute the first coordination sphere of the metal ion. The full MBS is obtained by including any other residue or chemical species having at least one atom within 5.0 Å from a metal ligand (orange sticks). Metal ligands provided by small molecules (e.g. water, ammonia, synthetic inhibitors) or ions (e.g. acetate, hydroxamate) are called *exogenous* ligands.

3. Structure-based prediction of metal sites

3.1 Template-based methods

The first proofs of concept of the feasibility of using 3D structures of apo-proteins (i.e. not containing the metal ion) to identify MBSs date to the 90's. They focused on the analysis of a function measuring the contrast between the hydrophobicity of the metal site itself and the surrounding protein residues (contrast function) [29]. In practice, the contrast function measures how much the outer atoms in a sphere are more hydrophobic than the inner atoms; higher values are associated with spheres centered at MBSs. Thus, predictions were based on the identification of cavities defined by templates of triads of amino acids with suitable relative spatial arrangements and featuring high hydrophobicity contrast [30]. The idea of matching an apo-structure against a set of templates (Figure 2), i.e. pre-arranged spatial distributions of potential metal ligands, is a very logical approach to leverage both the availability of a structural model for the protein of interest and the information stored in the PDB on existing MPs.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 June 2022 doi:10.20944/preprints202206.0253.v1

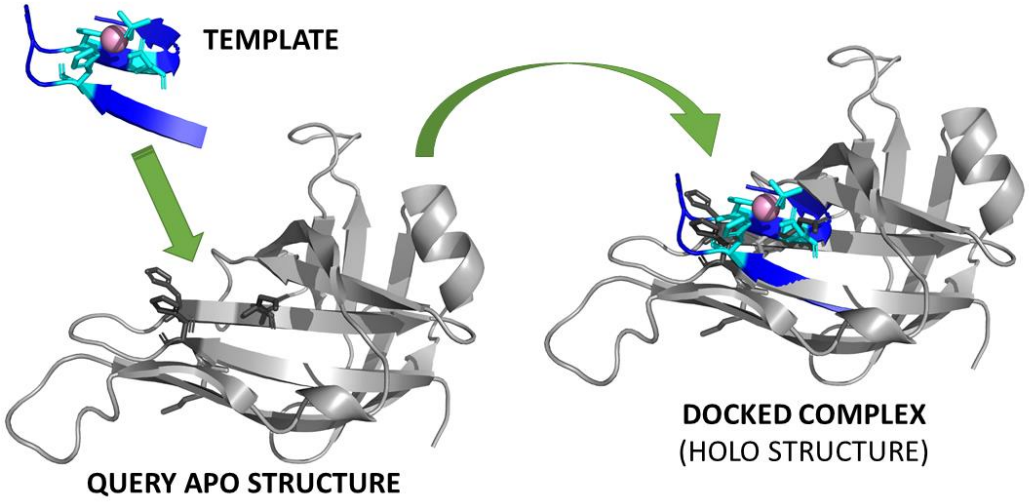

Figure 2. The concept of template-based detection of MBSs. Each template MBS from a suitably designed library, such as all MBSs from the non-redundant PDB, is docked to the query apo-structure. If the docking is successful, the position of the site and the nature of the bound metal are predicted. Often, the docking procedure is guided by first identifying candidate metal ligands in the query structure.

In this direction, a very successful implementation was that of the CHED algorithm [31]. CHED focuses on four residue types, namely Cys (C), His (H), Glu (E) and Asp (D) (hence the acronym) and searches the query apo-structure for suitable arrangement(s) of triads of the CHED residues. In these arrangements, the donor atoms are at distances from each other consistent with cut-off values taken from the analysis of the MBSs in the PDB. Some possible structural rearrangements are also taken into account, by looking for alternative conformations of one side chain at a time. The initial hits are then re-examined using two different filters to remove false positive predictions. A relatively similar approach taking additionally into account the volume available to the side chains of the candidate metal ligands was proposed by Goyal and Mande [32].

IonCom [33] has integrated four existing structure-template-based predictors for general ligand binding, also exploiting local similarity to templates, with a novel sequence-based predictor, called IonSeq, to predict binding sites of metal ions as well as of other ions from 3D structural models generated by the i-Tasser modelling tool [34]. Among the template-based predictors used, COACH [35] afforded a particularly relevant contribution. The latter is a consensus method, combining an approach leveraging the 3D alignment of binding-site substructures (TM-SITE), an approach based on sequence profile alignment (S-SITE) and the output of other structure-based predictors using a SVM method. The combined approach achieved a MCC that was 12.5% higher than the individual predictors. A major advantage of IonCom over COACH was due to the integration of complementary sequence-based and template-based

methods. The authors noted that accuracy of IonCom is lower for some metal ions with highly variable MBSs, e.g. alkali metals. This may actually be a general issue, as many tools indeed focus only or mainly on transition metal ions.

The MIB server for the prediction of MBSs and docking of metal ions [36] uses The fragment transformation method [37] to predict the location of MBS for twelve different metals and dock the appropriate metal ion (one among Ca^{2+}, Cu^{2+}, Fe^{3+}, Mg^{2+}, Mn^{2+}, Zn^{2+}, Cd^{2+}, Fe^{2+}, Ni^{2+}, Hg^{2+}, Co^{2+}, Cu^{+}) into the query 3D structure. To generate the library of templates, the structures of MPs containing at least one relevant ion were collected from the PDB and filtered at the 30% sequence identity level. The MBS templates were then extracted by selecting all residues with at least one heavy atom within 3.5 Å of the metal ion. Then, by applying the fragment transformation algorithm, clusters of residues in the query apo-structure are matched to the templates of the library. Each identified cluster is scored according to its sequence and structural similarity to the template. In practice, this procedure identifies the best superposition (taking into account also sequence similarity) of each template to the query structure and then ranks the results. Because each template contains also the metal ion, the predicted position of the latter within the structure is readily available.

ZINCCLUSTER focuses on the prediction of zinc-binding sites [38], based on the occurrence within the query protein of the structural patterns detected from the analysis of MBSs present in MPs of known 3D structure.

The implementation of a predictive algorithm in the GaudiMM modeling suite [39] did not explicitly make use of templates. Instead, the authors implemented a method where groups of protein residues potentially able to coordinate a metal ion are discovered in the query structure by seeking suitable donor atoms within a user-defined distance (3.5 Å in the paper) from the metal ion [40]. When the input is an apo-protein, potential MBSs are initially found by probing the structure for accessible pockets whose center of mass is within 3.5 Å from the β-carbon atoms of three or more amino acids of the CHED group (as in the CHED method described above). The identification of donors is then performed from the pocket center. After defining the candidate metal position and the surrounding donor atoms, a series of geometrical aspects are calculated to validate the site prediction. Note that, in practice, this procedure requires that a minimum of three donor atoms are present in the site.

BioMetAll expands upon the concept incorporated as part of the GaudiMM suite (see preceding paragraph), by making the assumption that the geometric patterns of the protein backbone permit the identification of preorganized MBSs [41]. The structural analysis of the conformation of the backbone instead of the side chains, should make the predictions less dependent on the high quality of the structure and also on the metal-induced reorganization of the site, which often does not greatly affect the protein backbone [42]. The BioMetAll algorithm starts by embedding the apo protein structure in a grid of virtual metal probes. This is done by retaining only the coordinates of the Cα, Cβ, C' and backbone O atoms. These atoms are embedded in a spherical grid of equidistributed probes, with an extra thickness of 8 Å to account for the atoms of the residues at the protein surface. Probes at less than 1.0 Å from any protein atom are removed. For every remaining probe, BioMetAll evaluates which protein residues surrounding it meet the geometric parameters defined by the authors through a statistical analysis of the sites in the MetalPDB database [6]. Restrictions on the minimum number and type of the metal ligands are applied to finally produce a list of valid probes along with their potential ligands. With this procedure, each MBS

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 June 2022 doi:10.20944/preprints202206.0253.v1

can be associated to a number of metal probes; also, any protein will be associated with several predicted MBSs. The authors observed that there was a good correlation between the number of probes associated with a predicted MBS and the likelihood that the prediction was correct: in 75% of the cases the most populated solution overlapped with the experimental site.

The flexibility of MBS has been taken explicitly into account for the predictive method described in [43]. In this work, coarse grained molecular mechanics was applied to produce meaningful ensembles of 11 structural conformations for each query protein. The ensemble represents the conformational space available to the protein, based on its input structure and the force field used, thus contributing to overcome the problem of metal-induced rearrangements at the site. Then, recognition of MBS templates from a predefined library is carried out using geometric hashing. Geometry hashing has been chosen because it speeds up the comparison, allowing the software to deal with more structures for a single query. On the other hand, it provides limitations to the minimum number of residues in the MBS, which, in the current implementation, should be at least four. The method includes in the evaluation of the template matches only donor atoms from the amino acid sidechains, in order to simplify the identification of candidate sites in the query structure.

3.2 Random forest methods

Two recent predictors exploited a random forest algorithm. In this type of approaches, a computational model is trained over a set of features (i.e. specific properties) extracted from a large number of positive and negative examples of MBSs. The optimized model is then used without further modifications to classify novel structures. The Zincbindpredict tool employs a random forest classifier that has been trained to predict entire zinc binding sites, as opposed to individual zinc binding residues [44]. In practice, this tool employs a portfolio of predictive models, each optimized to detect a specific type of zinc-binding site, where the different types, called families of sites in the article, correspond to different metal-binding patterns, e.g. C2H2, H3, etc. The features of each family of sites included sequence-derived properties (inter-residue distance, average hydrophobicity and average number of charges around each residue, both computed over three different windows) and structure-derived properties (various combinations of Cα-Cα as well as Cβ-Cβ distances within the MBS, plus the hydrophobic contrast function) [44]. To collect a dataset of negative samples, an arrangement of residues matching the pattern of the family in question was taken from a randomly chosen non-zinc-binding PDB structure, and the corresponding feature vector created. Only residue combinations where the Cα-Cα distances were all below 30 Å were taken into account, in order to exclude physically sites from the negative dataset. The query structure is thus processed in order to identify the combinations of residues matching the different families of sites included in Zincbindpredict; for each potential site, the feature vector generated from the query sequence and structure is fed to the classifier of the corresponding family to evaluate the likelihood that it is a true site.

Although not aimed at the identification of MBSs in apo-protein structures, another tool employing a random forest classifier has been developed to analyze backbone protein structures to identify suitable positions to introduce metal-binding residues in order to engineer MBSs in proteins (i.e. to artificially design a MP given a protein scaffold of known 3D structure) [45]. In practice, the training set contains features that are based only on the coordinates of the backbone atoms, whereas all side chain

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 June 2022 doi:10.20944/preprints202206.0253.v1

atoms are removed. The position of the Cβ is recalculated from the backbone coordinates to permit the computation of geometric parameters that are fed into the random forest classifiers. Neglecting the position of the side chain atoms makes it possible to predict the occurrence of potential MBSs independently of the identity of amino acids in the current structure, thereby opening the way towards creating engineered binding sites by introducing the appropriate side chains in the positions highlighted by the tool [45].

4. Structural comparison of metal sites

The macromolecular context surrounding the MBS is a major determinant of its chemico-physical properties and, consequently, of its reactivity. The structural comparison of MBSs therefore is a tool enabling functional analysis of MP families as well as of evolutionary unrelated proteins that harbor similar MBSs. Structural comparison of MBSs is typically performed regardless of the similarity of the global protein fold, in order to focus only on the evolution of the MBSs (Figure 3).

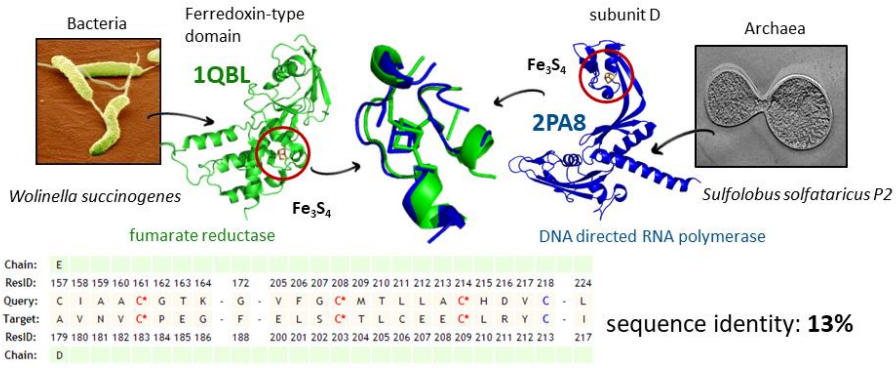

Figure 3. Detection of relationships between MPs with divergent sequences. MBS-driven structural superposition (in this example computed with the MetalS^2 [46] tool) permits the detection of similarities between proteins having no homology or having experienced extensive sequence divergence.

The MetalS^2 tool (http://metalweb.cerm.unifi.it/tools/metals2/) takes as input the structures of two MBSs (as defined in the MetalPDB database) and superimposes them regardless of the protein fold [46]. The very first step of MetalS^2 is to identify the geometric center of the metal ions (in order to be able to handle also polymetallic sites) in each site and then overlap them. Each MBS is then decomposed into triangular units defined by the geometric center of the metals and two donor atoms and proceeds to systematically superimpose all possible unit pairs from the two sites, always keeping the vertices corresponding to the metal positions coincident. All the poses generated are ranked according to the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 June 2022 doi:10.20944/preprints202206.0253.v1

MetalS^2 scoring function, which takes into account both sequence and structure similarity for the set of relationships defined by each pose. The best ranking poses are optimized by minimizing the RMSD of the Cα and Cβ pairs of the two MBSs and re-ranked to provide the final best-scoring superposition [46-48]. We subsequently implemented an optimized version of MetalS^2 to allow users to search the entire MetalPDB database for MBSs that are structurally similar to the site of a MP structure of interest, either taken from the PDB or input by the user. The latter tool, called MetalS^3, is available as a web server at http://metalweb.cerm.unifi.it/tools/metals3/ [49].

mFASD is a structure-based algorithm that predicts which metal is most likely to populate a MBS [50]. In this tool, the MBS is a represented as a group of functional atoms (functional atoms set, FAS). The local chemical environment of each atom in the FAS is described by integrating information on its chemical properties and the chemical properties of its neighboring atoms. This allowed the authors to define a similarity measure between pairs of FAS atoms. In turn, this enabled the pairwise comparison of FASs, by all-versus-all comparison between the atoms in the two sets. A predefined threshold was introduced by analyzing the ROC curve computed for the case of identifying zinc(II) sites in an ensemble of all MPs binding one metal among Cu, Fe, Mg, Mn, Zn or Ca, derived from PDBSelect25. mFASD uses this threshold to assign pairs of FASs to same metal-binding type. Finally, the authors created a reference dataset of FASs, i.e. MBS templates, against which a query MBS could be scanned in order to assign its metal-binding type.

The MeCOM tool also aims at the superimposition of metal sites [51]. The algorithm starts by identifying the metal ion and its protein ligands in each MBS. Then, a metal-centered 40 Å grid is used to create a set of solvent-accessible lattice points, within which the active sites of the MPs are identified. Multiple metal ions within 5 Å from each other are treated as a single cluster when identifying the active site. MeCOM uses the atoms at the surface of the active site to assign specific features to the sites; these features include pharmacophore properties (e.g. presence of hydrogen bond donor/acceptors) as well as metal coordination and cofactor properties. A quaternion approach is finally deployed to superimpose the two MBSs based on the comparison of their features and the position of the Cα atoms. MeCOM automatically detects and builds MP active sites; this was tested on a dataset of 4223 structures with a resulting accuracy of 95.5%. Furthermore, a PyMOL plug-in has been made available to view and analyze MeCOM comparison results.

In a recent publication the already available TopMatch tool for structure superposition [52,53] has been used in conjunction with a novel scoring function, called sahle [54]. In this work, MBSs are extracted from the MP structures as spheres with a 15 Å radius centered on the metal. The aim of the sahle function, which depends on the length and sequence similarity of the aligned MBSs but not on the RMSD, is to identify functional relationships between the structurally aligned (with TopMatch) MBSs. As TopMatch does not explicitly take into account the metal ions to build the superposition, the authors optimized the parameters in the sahle function to make it capable of detecting structural alignments having a short metal-metal distance between the two superposed MBSs. With the optimized sahle function, the authors could detect structural similarity of the MBSs in evolutionarily distant MP families and identified six clusters of ancient metal-binding motifs [54]. Previously, a similar concept was implemented using Pymol as the structure alignment tool and a scoring function based solely on structural similarity parameters [55]. In the latter study, the aim was to define clusters of structurally similar MBSs, also extracted using a 15-Å sphere, and subsequently link them if members of different clusters co-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 June 2022 doi:10.20944/preprints202206.0253.v1

occurred in the same structure at a distance compatible with an electron-transfer interaction. In this way, a spatial adjacency network (SPAN) was built, based on the structural proximity of MBSs in electron transfer (ET) chains. The network provided evidence for the existence of four ancient folds that recurred frequently in ET chains.

5. Metalloprotein databases

Numerous databases exist addressing MPs in general or some specific aspects of their chemistry and biology [16]. As with many other databases in biology [56-58], a recurring issue is the obsolescence of their contents, even when the sites are still reachable. In this section, we describe some recently created or updated resources on MPs, whose contents are still current and accessible as of June 15, 2022.

MetalPDB [5,6] is available at https://metalpdb.cerm.unifi.it. MetalPDB collects structural information on all the MBSs present in the PDB, and links them to other biological resources such as protein domain databases. By construction of the MetalPDB database, the MBS is the ensemble of the atoms in the metal ligands and any other atom belonging to a chemical species within 5 Å from a ligand. The web site provides extensive statistical analyses on the databases contents, to facilitate a deeper comprehension of the diversity of the biochemistry of metals. MBSs are grouped into sets of equivalent and equistructural sites, which correspond to sites at a corresponding position within a given protein fold that are populated respectively by the same or different metal ions. These groups are linked to apo-protein structures with the same fold, which potentially are missing the metal cofactor.

BioLiP (https://zhanggroup.org/BioLiP/) is a semi-manually curated database of molecular adducts involving proteins [59]. The structure contents in BioLIP are harvested from the PDB and cross-referenced with the literature as well as with databases on biological function. Adducts between proteins and metal ions, i.e. MPs, are collected in a specific section of BioLiP.

ZincBind (https://zincbind.net/) is a database specialized on zinc-binding sites [60], where the sites have been built taking into account the biological assemblies rather than the asymmetric units deposited in the PDB. This is quite relevant for all MBSs at the interface between copies of a chain, where taking into consideration only the contribution of a single chain from the asymmetric unit is not biologically correct. ZincBind automatically discards zinc sites that are detected because of crystallization conditions, identified as zinc ions that have less than two protein ligands with three donor atoms. Furthermore, a 90% sequence identity filter has been applied to remove redundancy, except when the sites differ in the number, order or type of protein ligands. The software used to generate the database contents is open source.

Another specialized database is PyDISH (https://pydish.bio.info.hiroshima-cu.ac.jp/), which focuses on the analysis of heme-binding sites in PDB structures [61]. PyDISH focuses on the coordination of the heme iron (axial ligands), on the occurrence of the different heme types and on the distortions of the heme porphyrin. Statistical analyses can be obtained from the web site. Normal-coordinate structural decomposition [62] has been applied to define the porphyrin distortion as displacements from its equilibrium planar structure having D_{4h} symmetry.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 June 2022 doi:10.20944/preprints202206.0253.v1

VirusMED (https://virusmed.biocloud.top/) is a database of epitopes, drug binding site and metal binding sites in viral proteins of known 3D structure [63]. For metal binding sites, this database provides information on the coordination bonds between the protein to the metal ion(s). The enzymatic classification number (EC number) of each polypeptide chain coordinating the metal ion is included in the annotation, along with the taxonomic classification of the virus. A unique feature of VirusMED is that the quality of each site is evaluated using state-of-the-art methods for the validation of metal sites in crystallographic structures [64,65].

InterMetalDB (https://intermetaldb.biotech.uni.wroc.pl/) has a focus on MBSs occurring at intermolecular interfaces [28]. As mentioned for ZincBind, a stringent requisite to investigate this class of sites is to reconstruct the biological assembly from the asymmetric unit [66]. For the construction of InterMetalDB, intermolecular MBSs were identified by detecting metal ions having donor atoms (within a 3 Å threshold) from at least two protein or nucleotide residues belonging to different macromolecular chains. This criterion explicitly excludes non-macromolecular ligands from the definition. The redundancy of the database contents was reduced by using both a protein sequence filter (at 50% level) and a clustering approach to identify unique MBSs in structures harboring multiple sites. The analysis of InterMetalDB permitted the identification of metal preferences in interfacial sites as well as of the corresponding macromolecular environments [28].

MetLigDB (http://silver.sejong.ac.kr/MetLigDB/home.html) focuses on the structural and chemical properties of small molecules that bind directly to the metal ion(s) present in MPs [67]. MetLigDB entries were derived from the analysis of ligand-containing PDB structures. In addition to the structural view of each metal site containing an organic ligand, derived from the relevant PDB entry, the web pages of this database provide information on the binding affinity of the inhibitor for the target MP. This resource is mainly intended to support researchers in the development of novel metal-targeted inhibitors by looking at previously released molecules.

A related, more recent database is MeLAD (https://melad.ddtmlab.org/), which is also derived by extracting from the PDB database all the 3D structures of metalloenzyme-ligand adducts [68]. MeLAD extends the overview introduced by MetLigDB by integrating the structural view with detailed analyses of the properties of these systems, including metal-binding pharmacophores, metalloenzyme structural similarity and ligand chemical similarity. For example, MeLAD divides organic metal ligands into monodentate, bidentate and tridentate chemotypes, which are then linked to different metal ions and coordination modes. The analysis of the chemical similarity between ligands allowed MeLAD to identify groups of ligands harboring common metal-binding moieties. In turn, these associations are leveraged to cluster the metalloenzymes whose metal sites interact with the ligands in a given group. Besides their relevance to the development of novel metal-targeted inhibitors, the contents and underlying ideas of MeLAD provide hints to understand ligand selectivity in the context of the entire metalloproteome.

A missing database in the field of metal-based medicinal chemistry is one on metallodrugs (e.g., metallodrug-DB) [69]. Metal-containing complexes formed by small organic molecules are used as effective pharmaceuticals in a variety of contexts, from cancer treatment to antimicrobial and diagnostic agents. There are a number of subtleties associated with metallodrugs, starting with the cytotoxicity of free metal ions, which require a proper understanding of the molecular basis of their action mechanisms [70]. A metallodrug database could address not only metallodrugs already approved for clinical use or

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 June 2022 doi:10.20944/preprints202206.0253.v1

under clinical development, but also harvest information on metal-based compounds tested in relevant biochemical and cellular assays.

6. AI methods applied to metalloproteins

AI and DL methods have gained great popularity in the investigation of the 3Dstructure and reactivity of proteins, and the field of bioinformatics studies of MPs makes no exception [71]. However, the application of DL to MP structures is relatively recent, in spite of the extensive information available on these systems and of their biological relevance. In line with the rest of this contribution, here we do not address AI methods for sequence-based detection of metal-binding sites.

A pioneering application of DL to MPs is the use of conditional variational autoencoders for the insertion of metal-binding sites in non-metal binding proteins without human input [72]. The developed methodology was able to design protein sequences that matched specified attributes, such as metal-binding. The performance of the DL method was evaluated in comparison to a predictor based on hidden Markov models (HMMs), by estimating the stability of the predicted novel metal-binding structures. This analysis showed that the former approach could generate substantially more stable structures.

At the functional level, a relevant application is the investigation of disease-related mutations through a multichannel convolutional neural network (MCCNN) [73]. The MCCNN was trained using spatial and sequential features for each selected MBS (including both positive and negative examples, i.e. sites with and without known disease-associated missense mutations). The positive examples in the training set included 1256 disease-associated mutations related to ten metals, identified by integrating the information contained in clinical and human genetics databases with the MBSs of MetalPDB. The selected features input to the network included the occurrence of aliphatic and aromatic carbon atoms, the presence of hydrogen bond donors and acceptors, computed interaction energies with the MBS, the physicochemical properties of the aminoacids in the MBS, and data about the mutation. The trained MCCNN can predict disease-associated mutations in both the first and second sphere of MBSs [74] with a very satisfactory performance.

The DeepCys tool uses a NN to predict the probabilities of four cysteine different roles, i.e. metal-binding, disulphide formation, sulphenylation and thioether [75]. The most probable function of each cysteine in the input structure can then be assessed. In particular, the network learned how to identify metal-binding cysteines thanks to the inclusion in the training dataset of PDB structures binding Zn^{2+}, Cu^{2+}, Cd^{2+}, Fe^{2+}/Fe^{3+} or Hg^{2} ions. The input features to train the NN included descriptors of the cysteine microenvironment, secondary structure, protein family, the hydrophilicity around each cysteine as defined by the protein residues in contact with it, and the occurrence of specific patterns (e.g. CC, CSC, CXXC, …). The accuracy of DeepCys for its four different predictions ranged between 75% (thioether) and 87% (disuplhide).

MAHOMES is a recently developed approach aiming at distinguishing between enzymatic and non-enzymatic metals in MPs [76]. In this work, the authors applied fourteen different machine learning methods, including a neural network approach. These fourteen algorithms were trained on the same data and the Matthews correlation coefficient (MCC) was the selection criterion to identify the best performing approach. The MCC is a performance measure that is not particularly sensitive to imbalances in the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 June 2022 doi:10.20944/preprints202206.0253.v1

training set, as non-enzymatic data were about three times the enzymatic data in the input dataset. The best performing method was an extra trees algorithm, with which MAHOMES achieved a 94.2% accuracy on a validation dataset composed by enzyme structures deposited in 2018 or later. The input features used in MAHOMES included Rosetta energy terms, information on the MBS geometry, a description of the residues defining the MBS, electrostatic energy terms, and coordination geometry information, for a total of 391 features. Information on sequence conservation or secondary structure was not included by design in order to avoid potential biases towards specific folds. A further interesting outcome of this work was the analysis of which features were more important to discriminate the two categories of sites. The Rosetta energy summed over the spherical volume of the MBS was the most distinctive feature. The other most important features were based on the number and volume of side chain and backbone atoms lining the MBS, showing that enzymatic MBSs had larger volumes and involved a larger number of residues.

An "indirect" use of artificial intelligence in the study of MPs is the exploitation of AlphaFold [19,77] or RoseTTAFold [20] predictions to model or predict the occurrence of MBSs. In fact, the structural models in the AlphaFold database do not contain chemical entities other than natural aminoacids, even when the presence of the cofactor would be required to achieve the proper folding of the polypeptide chain, and also do not take into account quaternary structure [78,79]. The latter issue is already addressed by RoseTTAFold and is also being tackled by a novel version of AlphaFold, called AlphaFoldMultimer [80]. The AlphaFill database [81] aims at filling the gap regarding cofactor binding to the models in the AlphaFold database by docking small molecules and ions that have been observed in complex with homologous proteins in experimental structures from the PDB-REDO [82,83] repository. In practice, the AlphaFill algorithm uses BLAST [84] to identify close homologs of each AlphaFold model among the PDB-REDO structures that contain a metal ion (or another common cofactor, excluding crystallization agents or metals typically used in heavy-metal derivatives to help phasing). The residues surrounding the cofactor (i.e. the MBSs) in the BLAST hits are used for a local structural alignment of each PDB-REDO structure to the AlphaFold model, thereby allowing placing the cofactor within the latter model. The resulting holo-structures are vailable from the AlphaFill interface (https://alphafill.eu/).

In a related work, a thorough search of zinc and iron-sulfur binding sites was performed on all AlphaFold models [85]. The results hinted at the occurrence of a large variety of novel sites that could be predicted thanks to the availability of the 3D models. The protocol starts with identifying the coordinates of all sidechain or backbone atoms (e.g. the Sγ of cysteine or the Nδ1 and Nε2 of histidine). These potential donor atoms are then clustered using a single-linkage clustering algorithm with a distance threshold of 8 Å. Each cluster is then used to identify the possible superpositions of the donor atoms of a template MBS, with an approach analogous to template-based docking. All possible permutations of donor atoms are evaluated, and only those featuring a RMSD lower than 0.5 Å between the donor atoms of the template and of the AlphaFold model are retained to be checked for steric clashes between the cofactor and the protein atoms. After rejecting all permutations with poor RMSDs or steric clashes, the permutation with the lowest RMSD is retained for the current cluster of potential donor atoms. Twelve different template MBSs were analyzed in the work, six for iron-sulfur clusters and six for different zinc(II) sites, containing a single ion with three or four donor atoms. In practice, by looking at whether the backbone and sidechain atoms in each AlphaFold model were pre-organized to allow one of the twelve template MBSs to be docked to the protein with a low RMSD and no clashes, the authors predicted as many as 13,139 binding sites in 7,490 unique proteins with no known structural homologs [85]. Intriguingly, this repertoire might

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 June 2022 doi:10.20944/preprints202206.0253.v1

be even larger if one takes into account that proteins can populate different conformational states, while AlphaFold predicts only a single state [86].

The bindEmbed21 approach combines homology-based inference with DL to predict whether a protein residue binds to a metal ion, a nucleic acid, or a small molecule [87]. The DL component used protein embeddings as input to a two-layer convolutional neural network (CNN). Protein embeddings consist of fixed-length vector representations for each residue in a sequence, based on the distribution of sequences in an unlabeled set (i.e. a sequence database of proteins without experimental characterization whatsoever). In practice, this type of representation embeds each protein sequence in a vector space and allows the CNN to learn the constraints of protein sequence [88]. The advantage of this approach is that it does not require knowledge of protein structures, which is scarcer than knowledge of protein sequences, nor expert-selected features, which may require prior information on the chemicophysical properties relevant for the problem of interest, nor evolutionary information derived from multiple sequence alignments (MSAs), which are computationally cumbersome. The overall performance of *bindEmbed21* was close to that of specialized zinc-binding prediction methods, including the Zincbindpredict tool described in Section 3.2.

Mebipred is a DL tool to predict whether a protein is a MP, based on sequence information alone [89]. It is relevant for this review because it has been trained using information derived from 3D structures. This tool exploits a feed-forward multi-layer perceptron, a specific type of NN. The training data of mebipred were built on the MP structures available from the PDB, cluster at 70% sequence identity. For each representative structure, the input features were the amino acid composition, the physicochemical properties of the amino acids in the sequence and the frequency of occurrence of pentameric residue sequences within 5 Å from the metal ion. When analyzing a new query sequence, the latter is decomposed into pentamers with a sliding window of one position and the structure-derived feature is calculated as the sum of the occurrences of the pentamers in the PDB dictionary. In practice, mebipred looks in the sequence for pentamers that have been detected previously in the entire PDB database as being within 5 Å of the metal in a MBS of known structure. Based on the pentamers detected in the query the chemical identity of the bound metal can be predicted as well, based on the metal content of the 3D structures where these pentamers occurred. Mebipred can identify MPs with a 80% accuracy, and define the chemical identity of 11 different metals, for which a sufficient number of PDB structures was available.

In a very recent application, our own research team developed a DL classifier that can discriminate physiological and adventitious zinc-binding sites in the 3D structures of MPs with an accuracy of about 90% [90]. To develop the tool, we trained a recurrent neural network (RNN) using a dataset of 1944 physiological and 3352 adventitious zinc-binding sites extracted from MetalPDB and manually annotated. In order to compensate for the imbalance with respect to adventitious sites, the weight of the physiological sites in the cost function of the RNN was scaled up by 1.7. In addition to zinc-binding sites, the same DL classifier (i.e. without further training) could discriminate sites non-heme mononuclear iron sites with an accuracy close to 80%. This indicates that the rules learnt on zinc sites have general relevance, at least for simple transition metal ions. By systematically evaluating the importance of the various features input to the DL classifier, it appeared that MBSs involving 20 protein residues or more (defined according to the protocol of Figure 1) are quite likely to be physiological. The same holds for sites

with four metal ligands or more provided by the protein chain. Furthermore, it was observed that metal ligands in physiological MBSs tend to be buried, as judged from their relatively low solvent accessibility.

7. Concluding remarks

The extensive information available on MP structures has enabled the development of a multitude of applications for a deeper understanding of the biochemistry of MPs (Table 1). The tools for the prediction of the occurrence of MBSs in apo-structures and in structural models lacking their metal cofactor allow researchers to obtain a complete view of the occurrence of MPs in different organisms. Systematic structural comparison of these MBSs then results in the identification of distant evolutionary relationships, which would go unnoticed with other methods, or highlight cases of evolutionary convergence. A blooming sector is the application of DL methods to MPs, which is providing unprecedented insight into structure-function relationship in these systems. This whole plethora of computational advances is supported by public databases, derived from the PDB and integrating specialized functional information together with systematic analyses for selected aspects of the biochemistry of MPs. We are fully confident that this growth trend will reinforce in the next years, leading to an unprecedented level of comprehension of the role of essential metal ions in living organisms.

References

1. Foster, A.W.; Young, T.R.; Chivers, P.T.; Robinson, N.J. Protein metalation in biology. *Current Opinion in Chemical Biology* **2022**, *66*, 102095.
2. Smethurst, D.G.J.; Shcherbik, N. Interchangeable utilization of metals: New perspectives on the impacts of metal ions employed in ancient and extant biomolecules. *Journal of Biological Chemistry* **2021**, *297*, 101374.
3. Chandrangsu, P.; Rensing, C.; Helmann, J.D. Metal homeostasis and resistance in bacteria. *Nature Reviews Microbiology* **2017**, *15*, 338-350.
4. Young, T.R.; Martini, M.A.; Foster, A.W.; Glasfeld, A.; Osman, D.; Morton, R.J.; Deery, E.; Warren, M.J.; Robinson, N.J. Calculating metalation in cells reveals CobW acquires CoII for vitamin B12 biosynthesis while related proteins prefer ZnII. *Nat. Commun.* **2021**, *12*.
5. Andreini, C.; Cavallaro, G.; Lorenzini, S.; Rosato, A. MetalPDB: a database of metal sites in biological macromolecular structures. *Nucleic Acids Res.* **2013**, *41*, D312-D319.
6. Putignano, V.; Rosato, A.; Banci, L.; Andreini, C. MetalPDB in 2018: a database of metal sites in biological macromolecular structures. *Nucleic Acids Res.* **2018**, *46*, D459-d464.
7. Andreini, C.; Bertini, I.; Cavallaro, G.; Holliday, G.L.; Thornton, J.M. Metal-MACiE: a database of metals involved in biological catalysis. *Bioinformatics* **2009**, *25*, 2088-2089.
8. Waldron, K.J.; Rutherford, J.C.; Ford, D.; Robinson, N.J. Metalloproteins and metal sensing. *Nature* **2009**, *460*, 823-830.
9. Valasatava, Y.; Rosato, A.; Furnham, N.; Thornton, J.M.; Andreini, C. To what extent do structural changes in catalytic metal sites affect enzyme function? *J. Inorg. Biochem* **2018**, *179*, 40-53.
10. Ben-David, M.; Soskine, M.; Dubovetskyi, A.; Cherukuri, K.-P.; Dym, O.; Sussman, J.L.; Liao, Q.; Szeler, K.; Kamerlin, S.C.L.; Tawfik, D.S. Enzyme Evolution: An Epistatic Ratchet versus a Smooth Reversible Transition. *Mol. Biol. Evol.* **2019**, *37*, 1133-1147.

11. Ridge, P.G.; Zhang, Y.; Gladyshev, V.N. Comparative genomic analyses of copper transporters and cuproproteomes reveal evolutionary dynamics of copper utilization and its link to oxygen. *Plos ONE* **2008**, *3*, e1378.
12. Zhang, Y.; Gladyshev, V.N. Comparative Genomics of Trace Elements: Emerging Dynamic View of Trace Element Utilization and Function. *Chem. Rev.* **2009**, *109*, 4828-4861.
13. Andreini, C.; Bertini, I.; Rosato, A. A hint to search for metalloproteins in gene banks. *Bioinformatics* **2004**, *20*, 1373-1380.
14. Andreini, C.; Banci, L.; Bertini, I.; Rosato, A. Zinc through the three domains of life. *J. Proteome Res* **2006**, *5*, 3173-3178.
15. Andreini, C.; Banci, L.; Bertini, I.; Elmi, S.; Rosato, A. Non-heme iron through the three domains of life. *Proteins: Struct., Funct., Bioinf.* **2007**, *67*, 317-324.
16. Zhang, Y.; Zheng, J. Bioinformatics of Metalloproteins and Metalloproteomes. *Molecules* **2020**, *25*.
17. Zeng, X.; Cheng, Y.; Wang, C. Global Mapping of Metalloproteomes. *Biochemistry* **2021**, *60*, 3507-3514.
18. Grosjean, N.; Blaby-Haas, C.E. Leveraging computational genomics to understand the molecular basis of metal homeostasis. *New Phytol* **2020**, *228*, 1472-1489.
19. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A., et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583-589.
20. Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G.R.; Wang, J.; Cong, Q.; Kinch, L.N.; Schaeffer, R.D., et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, *373*, 871-876.
21. AlQuraishi, M. AlphaFold at CASP13. *Bioinformatics* **2019**, *35*, 4862-4865.
22. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A., et al. Applying and improving AlphaFold at CASP14. *Proteins: Struct., Funct., Bioinf.* **2021**, *89*, 1711-1721.
23. Jones, D.T.; Thornton, J.M. The impact of AlphaFold2 one year on. *Nat Methods* **2022**, *19*, 15-20.
24. Laine, E.; Eismann, S.; Elofsson, A.; Grudinin, S. Protein sequence-to-structure learning: Is this the end(-to-end revolution)? *Proteins: Struct., Funct., Bioinf.* **2021**, *89*, 1770-1786.
25. Masrati, G.; Landau, M.; Ben-Tal, N.; Lupas, A.; Kosloff, M.; Kosinski, J. Integrative Structural Biology in the Era of Accurate Structure Prediction. *Journal of Molecular Biology* **2021**, *433*, 167127.
26. consortium, w. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.* **2019**, *47*, D520-D528.
27. Andreini, C.; Bertini, I.; Cavallaro, G. Minimal functional sites allow a classification of zinc sites in proteins. *Plos ONE* **2011**, *10*, e26325.
28. Tran, J.B.; Krężel, A. InterMetalDB: A Database and Browser of Intermolecular Metal Binding Sites in Macromolecules with Structural Information. *Journal of Proteome Research* **2021**, *20*, 1889-1901.
29. Yamashita, M.M.; Wesson, L.; Eisenman, G.; Eisenberg, D. Where metal ions bind in proteins. *Proc. Natl. Acad. Sci. USA* **1990**, *87*, 5648-5652.
30. Gregory, D.S.; Martin, A.C.; Cheetham, J.C.; Rees, A.R. The prediction and characterization of metal binding sites in proteins. *Protein Eng* **1993**, *6*, 29-35.
31. Babor, M.; Gerzon, S.; Raveh, B.; Sobolev, V.; Edelman, M. Prediction of transition metal-binding sites from apo protein structures. *Proteins: Struct., Funct., Bioinf.* **2008**, *70*, 208-217.
32. Goyal, K.; Mande, S.C. Exploiting 3D structural templates for detection of metal-binding sites in protein structures. *Proteins: Struct., Funct., Bioinf.* **2008**, *70*, 1206-1218.

33. Hu, X.; Dong, Q.; Yang, J.; Zhang, Y. Recognizing metal and acid radical ion-binding sites by integrating ab initio modeling with template-based transferals. *Bioinformatics* **2016**, *32*, 3260-3269.
34. Yang, J.; Yan, R.; Roy, A.; Xu, D.; Poisson, J.; Zhang, Y. The I-TASSER Suite: protein structure and function prediction. *Nat. Meth,* **2015**, *12*, 7-8.
35. Yang, J.; Roy, A.; Zhang, Y. Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics* **2013**, *29*, 2588-2595.
36. Lin, Y.F.; Cheng, C.W.; Shih, C.S.; Hwang, J.K.; Yu, C.S.; Lu, C.H. MIB: Metal Ion-Binding Site Prediction and Docking Server. *J Chem Inf Model* **2016**, *56*, 2287-2291.
37. Lu, C.H.; Lin, Y.S.; Chen, Y.C.; Yu, C.S.; Chang, S.Y.; Hwang, J.K. The fragment transformation method to detect the protein structural motifs. *Proteins* **2006**, *63*, 636-643.
38. Ajitha, M.; Sundar, K.; Arul Mugilan, S.; Arumugam, S. Development of METAL-ACTIVE SITE and ZINCCLUSTER tool to predict active site pockets. *Proteins* **2018**, *86*, 322-331.
39. Rodríguez-Guerra Pedregal, J.; Sciortino, G.; Guasp, J.; Municoy, M.; Maréchal, J.D. GaudiMM: A modular multi-objective platform for molecular modeling. *J Comput Chem* **2017**, *38*, 2118-2126.
40. Sciortino, G.; Garribba, E.; Rodríguez-Guerra Pedregal, J.; Maréchal, J.D. Simple Coordination Geometry Descriptors Allow to Accurately Predict Metal-Binding Sites in Proteins. *Acs Omega* **2019**, *4*, 3726-3731.
41. Sánchez-Aparicio, J.-E.; Tiessler-Sala, L.; Velasco-Carneros, L.; Roldán-Martín, L.; Sciortino, G.; Maréchal, J.-D. BioMetAll: Identifying Metal-Binding Sites in Proteins from Backbone Preorganization. *J. Chem. Inf. Model.* **2021**, *61*, 311-323.
42. Babor, M.; Greenblatt, H.M.; Edelman, M.; Sobolev, V. Flexibility of metal binding sites in proteins on a database scale. *Proteins* **2005**, *59*, 221-230.
43. Garg, A.; Pal, D. Inferring metal binding sites in flexible regions of proteins. *Proteins: Struct., Funct., Bioinf.* **2021**, *89*, 1125-1133.
44. Ireland, S.M.; Martin, A.C.R. Zincbindpredict—Prediction of Zinc Binding Sites in Proteins. *Molecules* **2021**, *26*.
45. Nguyen, H.; Kleingardner, J. Identifying metal binding amino acids based on backbone geometries as a tool for metalloprotein engineering. *Protein Sci.* **2021**, *30*, 1247-1257.
46. Andreini, C.; Cavallaro, G.; Rosato, A.; Valasatava, Y. MetalS^2: a tool for the structural alignment of minimal functional sites in metal-binding proteins and nucleic acids. *J. Chem. Inf. Model* **2013**, *53*, 3064-3075.
47. Valasatava, Y.; Andreini, C.; Rosato, A. Hidden relationship between metalloproteins unveiled by structural comparison of their metal sites. *Scientific Reports* **2015**, *5*, 9486.
48. Rosato, A.; Valasatava, Y.; Andreini, C. Minimal functional sites in metalloproteins and their usage in strucutral bioinformatics. *Int. J. Mol. Sci* **2016**, *17*, 671.
49. Valasatava, Y.; Rosato, A.; Cavallaro, G.; Andreini, C. MetalS^3, a database-mining tool for the identification of structurally similar metal sites. *J. Biol. Inorg. Chem* **2014**, *19*, 937-945.
50. He, W.; Liang, Z.; Teng, M.; Niu, L. mFASD: a structure-based algorithm for discriminating different types of metal-binding sites. *Bioinformatics* **2015**, *31*, 1938-1944.
51. Li, G.; Dai, Q.-Q.; Li, G.-B. MeCOM: A Method for Comparing Three-Dimensional Metalloenzyme Active Sites. *J. Chem. Inf. Model.* **2022**, *62*, 730-739.
52. Sippl, M.J.; Wiederstein, M. Detection of spatial correlations in protein structures and molecular complexes. *Structure* **2012**, *20*, 718-728.
53. Wiederstein, M.; Sippl, M.J. TopMatch-web: pairwise matching of large assemblies of protein and nucleic acid chains in 3D. *Nucleic Acids Res.* **2020**, *48*, W31-w35.

54. Bromberg, Y.; Aptekmann, A.A.; Mahlich, Y.; Cook, L.; Senn, S.; Miller, M.; Nanda, V.; Ferreiro, D.U.; Falkowski, P.G. Quantifying structural relationships of metal-binding sites suggests origins of biological electron transfer. *Science advances* **2022**, *8*, eabj3984.
55. Raanan, H.; Pike, D.H.; Moore, E.K.; Falkowski, P.G.; Nanda, V. Modular origins of biological electron transfer chains. *Proc Natl Acad Sci U S A* **2018**, *115*, 1280-1285.
56. Attwood, T.K.; Agit, B.; Ellis, L.B.M. Longevity of Biological Databases. *EMBnet.journal; Vol 21* **2015**, 10.14806/ej.21.0.803.
57. Wren, J.D.; Georgescu, C.; Giles, C.B.; Hennessey, J. Use it or lose it: citations predict the continued online availability of published bioinformatics resources. *Nucleic Acids Res.* **2017**, *45*, 3627-3633.
58. Imker, H.J. 25 Years of Molecular Biology Databases: A Study of Proliferation, Impact, and Maintenance. *Frontiers in Research Metrics and Analytics* **2018**, *3*.
59. Yang, J.; Roy, A.; Zhang, Y. BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic Acids Res.* **2012**, *41*, D1096-D1103.
60. Ireland, S.M.; Martin, A.C.R. ZincBind-the database of zinc binding sites. *Database (Oxford)* **2019**, *2019*.
61. Kondo, H.X.; Kanematsu, Y.; Masumoto, G.; Takano, Y. PyDISH: database and analysis tools for heme porphyrin distortion in heme proteins. *Database* **2020**, 10.1093/database/baaa066, baaa066.
62. Jentzen, W.; Song, X.-Z.; Shelnutt, J.A. Structural Characterization of Synthetic and Protein-Bound Porphyrins in Terms of the Lowest-Frequency Normal Coordinates of the Macrocycle. *The Journal of Physical Chemistry B* **1997**, *101*, 1684-1699.
63. Zhang, H.; Chen, P.; Ma, H.; Woinska, M.; Liu, D.; Cooper, D.R.; Peng, G.; Peng, Y.; Deng, L.; Minor, W., et al. virusMED: an atlas of hotspots of viral proteins. *IUCrJ* **2021**, *8*, 931-942.
64. Zheng, H.; Shabalin, I.G.; Handing, K.B.; Bujnicki, J.M.; Minor, W. Magnesium-binding architectures in RNA crystal structures: validation, binding preferences, classification and motif detection. *Nucleic Acids Res.* **2015**, *43*, 3789-3801.
65. Zheng, H.; Cooper, D.R.; Porebski, P.J.; Shabalin, I.G.; Handing, K.B.; Minor, W. CheckMyMetal: a macromolecular metal-binding validation tool. *Acta Crystallogr., Sect. D* **2017**, *73*, 223-233.
66. Laitaoja, M.; Valjakka, J.; Janis, J. Zinc coordination spheres in protein structures. *Inorg. Chem* **2013**, *52*, 10983-10991.
67. Choi, H.; Kang, H.; Park, H. MetLigDB: a web-based database for the identification of chemical groups to design metalloprotein inhibitors. *Journal of Applied Crystallography* **2011**, *44*, 878-881.
68. Li, G.; Su, Y.; Yan, Y.H.; Peng, J.Y.; Dai, Q.Q.; Ning, X.L.; Zhu, C.L.; Fu, C.; McDonough, M.A.; Schofield, C.J., et al. MeLAD: an integrated resource for metalloenzyme-ligand associations. *Bioinformatics* **2020**, *36*, 904-909.
69. Medina-Franco, J.L.; López-López, E.; Andrade, E.; Ruiz-Azuara, L.; Frei, A.; Guan, D.; Zuegg, J.; Blaskovich, M.A.T. Bridging informatics and medicinal inorganic chemistry: Toward a database of metallodrugs and metallodrug candidates. *Drug Discovery Today* **2022**, https://doi.org/10.1016/j.drudis.2022.02.021.
70. Anthony, E.J.; Bolitho, E.M.; Bridgewater, H.E.; Carter, O.W.L.; Donnelly, J.M.; Imberti, C.; Lant, E.C.; Lermyte, F.; Needham, R.J.; Palau, M., et al. Metallodrugs are unique: opportunities and challenges of discovery and development. *Chemical Science* **2020**, *11*, 12888-12917.
71. Yu, Y.; Wang, R.; Teo, R.D. Machine Learning Approaches for Metalloproteins. *Molecules* **2022**, *27*, 1277.
72. Greener, J.G.; Moffat, L.; Jones, D.T. Design of metalloproteins and novel protein folds using variational autoencoders. *Sci Rep* **2018**, *8*, 16189.

73. Koohi-Moghadam, M.; Wang, H.; Wang, Y.; Yang, X.; Li, H.; Wang, J.; Sun, H. Predicting disease-associated mutation of metal-binding sites in proteins using a deep learning approach. *Nature Machine Intelligence* **2019**, *1*, 561-567.
74. Levy, R.; Sobolev, V.; Edelman, M. First- and second-shell metal binding residues in human proteins are disproportionately associated with disease-related SNPs. *Hum Mutat* **2011**, *32*, 1309-1318.
75. Nallapareddy, V.; Bogam, S.; Devarakonda, H.; Paliwal, S.; Bandyopadhyay, D. DeepCys: Structure-based multiple cysteine function prediction method trained on deep neural network: Case study on domains of unknown functions belonging to COX2 domains. *Proteins* **2021**, *89*, 745-761.
76. Feehan, R.; Franklin, M.W.; Slusky, J.S.G. Machine learning differentiates enzymatic and non-enzymatic metals in proteins. *Nat. Commun.* **2021**, *12*, 3712.
77. Varadi, M.; Anyango, S.; Deshpande, M.; Nair, S.; Natassia, C.; Yordanova, G.; Yuan, D.; Stroe, O.; Wood, G.; Laydon, A., et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **2022**, *50*, D439-d444.
78. Perrakis, A.; Sixma, T.K. AI revolutions in biology: The joys and perils of AlphaFold. *EMBO Rep* **2021**, *22*, e54046.
79. Thornton, J.M.; Laskowski, R.A.; Borkakoti, N. AlphaFold heralds a data-driven revolution in biology and medicine. *Nat Med* **2021**, *27*, 1666-1669.
80. Evans, R.; O'Neill, M.; Pritzel, A.; Antropova, N.; Senior, A.; Green, T.; Žídek, A.; Bates, R.; Blackwell, S.; Yim, J., et al. Protein complex prediction with AlphaFold-Multimer. *bioRxiv : the preprint server for biology* **2022**, 10.1101/2021.10.04.463034, 2021.2010.2004.463034.
81. Hekkelman, M.L.; de Vries, I.; Joosten, R.P.; Perrakis, A. AlphaFill: enriching the AlphaFold models with ligands and co-factors. *bioRxiv : the preprint server for biology* **2021**, 10.1101/2021.11.26.470110, 2021.2011.2026.470110.
82. van Beusekom, B.; Touw, W.G.; Tatineni, M.; Somani, S.; Rajagopal, G.; Luo, J.; Gilliland, G.L.; Perrakis, A.; Joosten, R.P. Homology-based hydrogen bond information improves crystallographic structures in the PDB. *Protein Sci.* **2018**, *27*, 798-808.
83. Joosten, R.P.; Salzemann, J.; Bloch, V.; Stockinger, H.; Berglund, A.-C.; Blanchet, C.; Bongcam-Rudloff, E.; Combet, C.; Da Costa, A.L.; Deleage, G., et al. PDB_REDO: automated re-refinement of X-ray structure models in the PDB. *Journal of Applied Crystallography* **2009**, *42*, 376-384.
84. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol* **1990**, *215*, 403-410.
85. Wehrspan, Z.J.; McDonnell, R.T.; Elcock, A.H. Identification of Iron-Sulfur (Fe-S) Cluster and Zinc (Zn) Binding Sites Within Proteomes Predicted by DeepMind's AlphaFold2 Program Dramatically Expands the Metalloproteome. *J Mol Biol* **2022**, *434*, 167377.
86. Golinelli-Pimpaneau, B. Prediction of the Iron–Sulfur Binding Sites in Proteins Using the Highly Accurate Three-Dimensional Models Calculated by AlphaFold and RoseTTAFold. *Inorganics* **2022**, *10*, 2.
87. Littmann, M.; Heinzinger, M.; Dallago, C.; Weissenow, K.; Rost, B. Protein embeddings and deep learning predict binding residues for various ligand classes. *Scientific Reports* **2021**, *11*, 23916.
88. Yang, K.K.; Wu, Z.; Bedbrook, C.N.; Arnold, F.H. Learned protein embeddings for machine learning. *Bioinformatics* **2018**, *34*, 2642-2648.
89. Aptekmann, A.A.; Buongiorno, J.; Giovannelli, D.; Glamoclija, M.; Ferreiro, D.U.; Bromberg, Y. mebipred: identifying metal-binding potential in protein sequence. *Bioinformatics* **2022**, 10.1093/bioinformatics/btac358.

90. Laveglia, V.; Giachetti, A.; Sala, D.; Andreini, C.; Rosato, A. Learning to Identify Physiological and Adventitious Metal-Binding Sites in the Three-Dimensional Structures of Proteins by Following the Hints of a Deep Neural Network. *J. Chem. Inf. Model.* **2022**, 10.1021/acs.jcim.2c00522.

Funding

The authors thank the University of Florence and C.I.R.M.M.P. for support.

Author Contributions

All authors have contributed to the conceptualization, writing and editing of this article. All authors have read and agreed to the final version of the manuscript.

Conflicts of interest

The authors have no conflicts or competing interests to declare.