Hindawi
BioMed Research International
Volume 2022, Article ID 8965712, 19 pages
https://doi.org/10.1155/2022/8965712

Review Article

A Comprehensive Review of Computation-Based Metal-Binding Prediction Approaches at the Residue Level

Nan Ye,[1] Feng Zhou,[2] Xingchen Liang,[2] Haiting Chai,[3] Jianwei Fan,[2] Bo Li,[4] and Jian Zhang[2]

[1] *School of Finance and Economics, Xinyang Agriculture and Forestry University, Xinyang 464000, China*
[2] *School of Computer and Information Technology, Xinyang Normal University, Xinyang 464000, China*
[3] *College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow G12 8QQ, UK*
[4] *College of Electronic Science and Engineering, Jilin University, Changchun 130012, China*

Correspondence should be addressed to Nan Ye; leavesyn@yeah.net

Received 2 January 2022; Accepted 4 March 2022; Published 31 March 2022

Academic Editor: Bing Wang

Clear evidence has shown that metal ions strongly connect and delicately tune the dynamic homeostasis in living bodies. They have been proved to be associated with protein structure, stability, regulation, and function. Even small changes in the concentration of metal ions can shift their effects from natural beneficial functions to harmful. This leads to degenerative diseases, malignant tumors, and cancers. Accurate characterizations and predictions of metalloproteins at the residue level promise informative clues to the investigation of intrinsic mechanisms of protein-metal ion interactions. Compared to biophysical or biochemical wet-lab technologies, computational methods provide open web interfaces of high-resolution databases and high-throughput predictors for efficient investigation of metal-binding residues. This review surveys and details 18 public databases of metal-protein binding. We collect a comprehensive set of 44 computation-based methods and classify them into four categories, namely, learning-, docking-, template-, and meta-based methods. We analyze the benchmark datasets, assessment criteria, feature construction, and algorithms. We also compare several methods on two benchmark testing datasets and include a discussion about currently publicly available predictive tools. Finally, we summarize the challenges and underlying limitations of the current studies and propose several prospective directions concerning the future development of the related databases and methods.

1. Introduction

Metal ions are certain atom compounds that usually form cations that have (a) positive electric charge(s). Metal ions play pivotal roles in protein structure, function, regulation, and stability [1, 2]. Common metal ions include zinc (Zn^{2+}), calcium (Ca^{2+}), magnesium (Mg^{2+}), manganese (Mn^{2+}), iron (Fe^{3+} or Fe^{2+}), copper (Cu^{2+}), cobalt (Co^{2+}), sodium (Na^{+}), potassium (K^{+}), and nickel (Ni^{2+}) ions. Recent estimates have shown that approximately 30%-40% of proteins require one or several metal cofactors to together express biological function [3]. The proportion varies in different types of organisms or tissues. For instance, K^{+} is mostly found inside the cell, while Na^{+} is abundant outside of the cell [4]. Mn^{2+} is found accumulated in leafy green plants [5]. In the human body, Ca^{2+} accounts for approximately 1.5% of total body weight. The bulk of Ca^{2+} is aggregated in bones and teeth [6].

Metal ion binding proteins, i.e., metalloproteins, play critical roles in a biological and chemical process in cellular reactions [7]. Inside the cell, the dynamic homeostasis of the metal ions is strongly connected and delicately tuned [8]. Reinhard et al. claimed that K^{+} and Na^{+} are involved in processing cell signaling, intercellular communication, and maintaining tissue electrolyte balance [9]. A small change in the concentration of metal ions may shift the effects of metal ions from natural beneficial to harmful [10]. A recent study pointed out that metalloproteins are associated with degenerative diseases, including Parkinson's

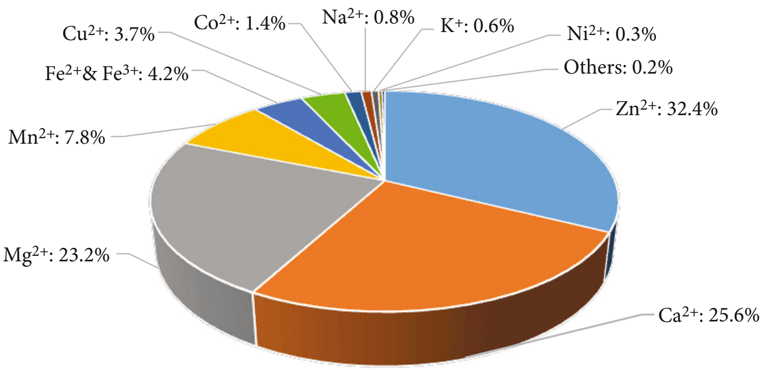


Figure 1: Fraction of top 10 metal-binding interactions that stored in PDB (date: December 20, 2021).

disease and Alzheimer's disease [11]. For instance, α-synuclein (Cu^{2+}-protein complex) constitutes the main component in Lewy bodies in Parkinson's disease [11]. Mn^{2+} and Fe^{3+} are responsible for inducing tangle pathology in Alzheimer's disease [8]. The aging of the brain or the development of diseases is associated with the deregulation of the management of metal ions [10]. Particularly, recent evidence indicates that if different types of metalloproteins interact in a certain salt solution, the potential galvanic erosion may dissolute the compound surface and result in inducing tumor formation [12, 13].

Elucidated protein-metal ion interactions rely in part on the advancement of various accurate characterizations and predictions of metalloproteins at the residue level. The traditional methods that are used to identify metal-binding conformation or binding residues include biophysics- or biochemistry-related wet-lab experiments, such as mass spectrometry [14], X-ray crystallography [15], and surface Plasmon resonance [16]. Since these technologies need expensive instruments, complex procedures, and elaborate labors, they shall benefit from the recent development of computation-aided methods.

We found 12 reviews that focused on the topic of exploring metal-binding residues or proteins in the past decade [7, 10, 11, 17–24]. Mallick et al. shed light on in silico methods including nine predictive tools and discussed the intrinsic mechanisms of metal-protein binding [24]. Thirumoorthy et al. investigated metallothionein isoforms and their role in pathophysiology [17]. They also provided the analysis of how metallothionein impact complex disease scenarios. In [18], the authors focused on structural variability and corresponding mechanisms of polymorphic amyloid oligomers complexed with metal ions. Bal et al. discussed ability constants, dissociation rates, and coordination chemistry of metal-binding residues in albumin [19]. Roohani et al. reviewed the literature related to zinc biochemical and physiological functions, metabolism, and zinc bioavailability in the human body [20]. The authors in [21] summarized the web tools that were proposed to identify metal-binding residues. Liu et al. systematically analyzed the structural features of Zn^{2+}-binding sites and proposed an online predictor [22]. Akcapinar and Sezerman collected and surveyed computational toolboxes designed for the recognition of metal-binding sites or metalloproteins [7]. Quintanar and Kim summarized the research in degenerative diseases related in metal ions [11]. Witkowska and Rowińska-Żyrek overviewed the analytical and biophysical methods utilized for studies on metal-protein interactions [23]. Krzywoszyńska detailed the involvement of metal ions in signaling processes within the cell and its influence in health and disease [10]. Rauer et al. scrutinized computational approaches that are associated with the prediction of protein functional sites and also discussed metal-binding related works [25].

Broadly speaking, these reviews discuss some aspects of the predictive methods. Some of them provide sufficient coverage of databases and predictive models and discuss the challenges and limitations of considered approaches. These reviews bring informative clues for the following researchers in this field. From the pertinence of the research, the prediction of metal-binding can be divided into general and specific approaches. The former recognizes metal-binding residues without considering their types, while, the latter is aimed at identifying one or several specific metal-protein interactions. According to the basic design and scheme, we classify these methods into four categories, namely, learning-, docking-, template-, and meta-based methods.

This review covers a comprehensive set of 44 computation-based methods, and 25 of them were published in the past three years. Specifically, we survey 32 learning-based, 4 docking-based, 6 template-based, and 2 meta-based methods. Depending on whether the structure of a target protein is known or available, we further divide learn-binding methods into the structure- and sequence-based ones. We discuss their benchmark datasets, features, algorithms, and measurements, respectively. We also detail the docking-, template-, and meta-based methods and point out their advantages and limitations.

2. Public Databases for Metal Binding

The development in biochemistry and biophysics leads to a fast increasing number of protein-metal ion binding complexes. Figure 1 draws the top 10 metal-binding annotations in PDB. Our survey reveals that Zn^{2+}, Ca^{2+}, and Mg^{2+} occupy the top three prevalent metal ions. The Zn^{2+} is currently the best-explored and described metal ion [26]. Zn^{2+} participates in many biological processes, such as metabolism, immune system, neurotransmission, hormone secretion, and

Table 1: Summary of recently released database of metal ion binding interactions.

Name	Year	Considered metal ions	Number of sites	Web link	Ref.	Citation	Availability
InterMetalDB	2021	All metal ion binding	6,423	https://intermetaldb.biotech.uni.wroc.pl/	[26]	N/A	Yes
MeLAD	2020	All metal ion binding	N/A	https://melad.ddtmlab.org/	[33]	9	Yes
ZincBindDB	2019	Zn	24,992	https://github.com/samirelanduk/ZincBindDB	[49]	23	Yes
MetalPDB (v2)	2018	All metal ion binding	N/A	http://metalweb.cerm.unifi.it	[34]	90	No
BioLiP	2013	All metal ion binding	146,969	https://zhanggroup.org/BioLiP/	[36]	446	Yes
ZiFDB (v2)	2013	Zn	N/A	http://bindr.gdcb.iastate.edu/ZiFDB	[37]	25	No
MetalPDB (v1)	2013	All metal ion binding	N/A	http://metalweb.cerm.unifi.it	[35]	108	No
BioMe	2012	All metal ion binding	20,307	http://metals.zesoi.fer.hr	[39]	30	No
MetLigDB	2011	Zn, Mn, Fe, Ni, mg, cu, co, Mo	732	http://silver.sejong.ac.kr/MetLigDB	[40]	13	Yes
MIPS	2010	All metal ion binding	N/A	http://dicsoft2.physics.iisc.ernet.in/mips/	[41]	28	Yes
MEDB	2010	All metal ion binding	N/A	http://www.uohyd.ernet.in/anambs/	[42]	14	No
ZiFDB (v1)	2009	Zn	N/A	http://bindr.gdcb.iastate.edu/ZiFDB	[38]	87	No
MetalMine	2009	All metal ion binding	412	http://metalmine.naist.jp	[43]	3	No
Metal-MACiE	2009	All metal ion binding	N/A	https://www.ebi.ac.uk/thornton-srv/databases/Metal_MACiE/home.html	[44]	60	Yes
ZifBASE	2009	Zn	N/A	https://web.iitd.ac.in/~sundar/zifbase/	[45]	35	Yes
MESPEUS	2008	Na, mg, K, ca, Mn, Fe, co, Ni, cu, Zn	34,896	http://eduliss.bch.ed.ac.uk/MESPEUS/	[46]	102	No
MSDsite	2005	All metal ion binding	N/A	http://www.ebi.ac.uk/msd-srv/msdsite	[47]	122	Yes
MDB	2002	All metal ion binding	N/A	http://metallo.scripps.edu/	[48]	276	No

[1] We estimate the availability on December 1st, 10th, and 20th of 2021, respectively.

signaling [27]. According to a rough statistic, approximately 10% of eukaryotic proteins bind Zn^{2+} [28]. Ca^{2+} is mainly aggregated in bones and teeth vertebrates [29]. It helps form solid support structures through biomineralization [6]. Mg^{2+} is usually associated with solvent water molecules, which endow it with a good capability of binding affinity with proteins and movement. The solvation state of Mg^{2+} usually serves as the enzyme in which Mg^{2+} acts as a coenzyme [6].

Besides RCSB PDB (https://www.rcsb.org/) [30], recent years have witnessed several specific databases that collect, categorize, and store these metal-protein interactions. Table 1 summarizes the publication year, considered metal ions, size of the database, web link, citations, and availability for the recently released database. We use citations as a one direct and good way to quantify the impact of these resources within the community [31]. The citation counts were collected from Google Scholar (https://scholar.google.com/) on December 20, 2021.

Specifically, InterMetalDB collects and presents metal ion binding proteins from RCSB PDB. It uses MMseq2 [32] to cluster the structure chains with the 50% sequence identity. Then, it groups similar binding sites and selects the best-resolution structure as a representative. MeLAD is a metalloenzyme-ligand association database, which contains structural data, metal-binding pharmacophores, and ligand chemical similarity of metalloenzyme-ligand interactions [33]. MetalPDB details the local environment, three-dimensional (3D) structure, secondary structure, and solvent accessibility of the metal ion binding sites [34, 35]. BioLiP is a semimanually curated database, which includes protein-peptide, protein-nucleic acid, and protein-ligand annotations [36]. BioLiP stores and periodically updates all types of metal ion binding information from PDB. ZiFDB is a database that collects information about individual zinc fingers, engineered zinc-finger arrays, and related target sequences [37, 38]. BioMe provides a web interface for biologists to capture coordination numbers, distances, geometry, and percentage of monodentate and bidentate bound aspartic acid and glutamic acid carboxyl groups [39]. MetLigDB is specially designed to select chelating groups or chemical moieties that might be presented in the inhibitor of a metalloprotein [40]. MIPS stores the geometric information, macromolecular function, different chemical behavior of metals, and metalloproteins [41]. MEDB presents quantitative information on metal-binding sites in protein structures and can be used for the identification of trends or patterns in the metal-binding sites [42]. MetalMine automatically collects and categorizes different types of metal-binding sites that derived from the structures of protein-metal-ion complexes [43]. Metal-MACiE gathers all available metalloenzymes and includes structural and functional information of metal ions in the context of the catalytic mechanisms of these metalloenzymes [44]. ZifBASE deposits engineered and natural zinc finger proteins and provides sequences and structural

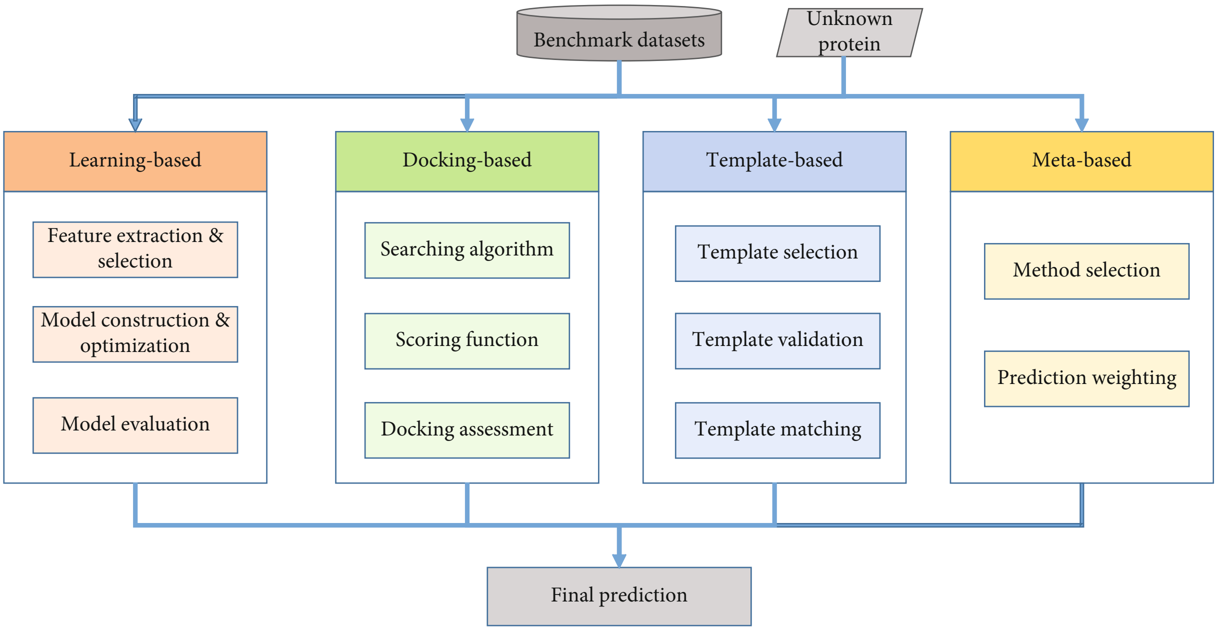


FIGURE 2: The flowchart of computation-based methods for prediction of metal-binding residues.

features and associated potential target sites of these proteins [45]. MESPEUS [46] focuses on the geometry of metal sites in proteins at resolution ≤ 2.5 Å. It provides an open web interface for further identifying and displaying the metal sites. MSDsite deposits computation-based metal-binding geometries and residues [47]. MDB offers quantitative information about metalloproteins [48]. It provides functions to analyze the binding attributes such as metal-ligand bond distances and side-chain torsion angles in metal sites.

We show that twelve source databases are designed for all metal ion binding data. Two databases, namely, MetLigDB and MESPEUS, consider several types of metal ions. There are four specific zinc-binding-related databases. Our survey also reveals that BioLiP is the most favored database, given the fact that its citations are average about 56 (446/8≈56) per year. Moreover, we notice that only half of the databases are available. Thus, we recommend that future databases shall be chronically maintained, periodically updated, and easy expanded.

3. Method Development of Metal-Binding Prediction

Figure 2 illustrates the flowchart of computation-based methods for the prediction of metal-binding residues. Generally, based on the basic design and scheme, these methods can be categorized into four groups. The learning-based methods regard the identification of metal-binding residues as a typical classification problem and attempt to use machine learning or deep learning algorithms to construct prediction models. The docking-based approaches are aimed at finding proper binding conformation as well as the appropriate target binding residues by scanning protein surface. The scoring functions are introduced to assess the selected pockets and quantify the strength of binding affinity. The template-based methods are designed to select the optimal template structures for a given unknown protein. Then, they map and transfer the binding annotations from similar spatial conformation to the target protein. By contrast, the meta-based methods focus on combing the predictions from other methods in order to build more accurate predictors.

3.1. Benchmark Datasets. The sequences and structures of protein-metal ion complexes are available in public databases for the end-users to customize the benchmark datasets. As shown in Table 2, the considered methods use various numbers of sequences/chains, ranging from several dozens to thousands. Besides that, protein complexes with high resolution indicate relatively more comprehensive and accurate annotations of protein-metal ion interactions. According to our survey, 12 out of 23 sequence-based and 6 out of 9 structure-based methods filter the candidate complexes using high resolution with ≤3 Å. Some methods [50–57] remove the sequences/chains whose lengths are less than 50 residues (or 45 residues [58]) since they might be potential segments or peptides. To build an unbiased dataset, it is necessary to remove homologous or redundant proteins. The cutoff threshold which researchers choose varies from minimal 25% to maximal 90%. Generally, a higher identity means a higher chance in local alignments [59]. The literatures in [60, 61] point out that if a pair of proteins have a sequence identity lower than 30%, they have little chance to share the same biological processes. Three tools, namely, BLASTclust [62], PISCES [63], and CD-HIT [64], are mainly used to cluster homologous proteins.

3.2. The Validation and Evaluation Metrics

3.2.1. Cross-Validation and Independent Test. To construct a predictor with high accuracy and decent generalization ability, it is necessary to avoid potential overfitting. In practice, cross-validation and independent test are two popular ways

Table 2: Summary of learning-based methods.

Type	Method[1]	Ref.	Year	Metal ion binding[2]	Dataset[3]	Resolution	Sequence similarity (tool)[4]	Prediction model[5]	Cross-validation	Independent test	Measurements[6]
Sequence-based	Liu et al.	[56]	2020	Zn, Cu, Fe, Co, Mn, Ca, Mg, Na, K	5,340	≤3 Å	30% (CD-HIT)	RF	5-fold	√	SN, SP, ACC, MCC
	MIonSite	[90]	2019	Zn, Ca, Mg, Mn, Fe, Cu, Fe, Co, Na, K, Cd, Ni	7,676	N/A	30% (CD-HIT)	SVM, AdaBoost	5-fold	√	SN, SP, ACC, MCC, AUC
	MPLs-Pred	[91]	2019	General metal ions	1,492	N/A	30% (CD-HIT)	RF	10-fold	√	SN, SP, ACC, MCC
	SXGBsite	[92]	2019	Ca, Zn, Mg, Mn, Fe	4,421	N/A	40% (PISCES)	GBM	5-fold	√	SN, SP, ACC, MCC, AUC
	Wang et al.	[50]	2019	Zn, Cu, Fe, Mn, Ca, Mg, Na, K	5,146	≤3 Å	30% (N/A)	SVM, SMO	5-fold	√	SN, SP, ACC, MCC
	znMachine	[51]	2019	Zn	2,043	≤3 Å	30% (BLASTclust)	SVM, NN	5-fold	√	SN, SP, ACC, MCC, PRE, AUC
	SSWPNN	[93]	2019	Zn	213	≤2.5 Å	70% (N/A)	SVM, NN	5-fold	√	SN, SP, PRE, F1, MCC, ACC
	ZinCaps	[94]	2019	Zn	738	≤3 Å	N/A (N/A)	CN	5-fold	√	SN, SP, ACC, MCC, AUC
	Haberal and Oğul	[65]	2018	General metal ions	2,727	N/A	N/A (N/A)	CNN	5-fold	√	SN, ACC, PRE, F1
	ZincBinder	[87]	2018	Zn	738	≤2.5 Å	30% (PISCES)	SVM	5-fold	√	SN, SP, ACC, MCC, AUC
	EC-RUS	[95]	2017	Ca, Mg, Mn, Fe, Zn	4,421	N/A	40% (PISCES)	WSRC	5-fold	√	SN, SP, ACC, MCC, AUC
	Cao et al.	[52]	2017	Zn, Cu, Fe, Co, Mn, Ca, Mg, K, Na	5,340	≤3 Å	30% (CD-HIT)	SVM	5-fold	√	SN, SP, ACC, MCC
	Kumar	[96]	2017	Cu, Ca, Co, Fe, Mg, Mn, Ni, Zn	3,922	N/A	50% (CD-HIT)	RF	10-fold	√	SN, SP, ACC, MCC
	DeepMBS	[97]	2017	General metal ions	2,727	≤3 Å	N/A (N/A)	CNN	5-fold	√	SN, PRE, F1
	Qiao et al.	[98]	2017	Ca	2,239	N/A	30% (CD-HIT)	SVM	5-fold	√	SN, ACC, PRE, MCC, AUC
	IonCom	[99]	2016	Zn, Cu, Fe, Ca, Mg, Mn, Na, K	1,374	N/A	30% (CD-HIT)	SVM, AdaBoost	5-fold	√	SN, SP, ACC, MCC
	Jiang et al.	[77]	2016	Ca	1,885	≤3 Å	25% (N/A)	SVM	5-fold	√	SN, SP, ACC, MCC
	TargetCom	[53]	2016	Cu, Fe, Zn	1,373	≤3 Å	40% (CD-HIT)	SVM, AdaBoost	5-fold	√	SN, SP, ACC, MCC
	OSML	[100]	2015	Ca, Zn, Mg, Mn, Fe	4,421	N/A	40% (PISCES)	SVM	5-fold	√	SN, SP, ACC, MCC
	TargetS	[101]	2013	Ca, Zn, Mg, Mn, Fe	4,421	N/A	40% (PISCES)	SVM, AdaBoost	5-fold	√	SN, SP, ACC, MCC, AUC
	ETMB-RBF	[102]	2013	General metal ions	55	N/A	20% (BLASTclust)	RBFN	10-fold	√	SN, SP, ACC, MCC
	ZincExplorer	[103]	2013	Zn	392	≤3 Å	N/A (N/A)	SVM	5-fold	√	SN, SP, PRE, MCC, AUPRC

Table 2: Continued.

Type	Method[1]	Ref.	Year	Metal ion binding[2]	Dataset[3]	Resolution	Sequence similarity (tool)[4]	Prediction model[5]	Cross-validation	Independent test	Measurements[6]
Structure-based	Horst et al.	[58]	2010	Ca	635	≤2.1 Å	35% (N/A)	LR	10-fold	√	MCC,AUC,AUPRC
	Nguyen et al.	[104]	2021	Mn, Fe, Co, Ni, Cu, Zn	9,955	≤2.5 Å	90% (N/A)	RF	5-fold	×	ACC
	TMP-MIBS	[54]	2021	General metal ions	427	N/A	40% (CD-HIT)	RF	10-fold	√	SN, SP, ACC, MCC, AUC
	Zincbindpredict	[105]	2021	Zn	N/A	≤ 2 Å	40% (CD-HIT)	RF	5-fold	√	SN, PRE, F1, MCC
	Wang et al.	[81]	2021	Zn, Cu, Fe, Ca, Mg, Mn, Na, K, Co	5,340	≤3 Å	30% (N/A)	MLP,SVM	5-fold	√	SN, SP, ACC, MCC
	DELIA	[80]	2020	Ca, Mn, Mg	3,966	N/A	30% (CD-HIT)	CNN	5-fold	√	SN, PRE, MCC, AUC
	Hu et al.	[57]	2020	Zn, Cu, Fe, Co, Mn, Ca, Mg, Na, K	5,340	≤3 Å	30% (CD-HIT)	GBM	5-fold	√	SN, SP, FPR, ACC, MCC
	MetalExplorer	[79]	2017	Ca, Co, Cu, Fe, Ni, Mg, Mn, Zn	3,192	≤2.5 Å	30% (CD-HIT)	RF	5-fold	√	SN, FPR, PRE, AUC, AURPC
	FINDSITE-metal	[55]	2011	Ca, Co, Cu, Fe, Mg, Mn, Ni, Zn	860	N/A	35% (PISCES)	SVM	2-fold	√	ACC, SPC, PPV
	Zinc identifier	[78]	2011	Zn	1,103	≤2.5 Å	N/A (N/A)	RF	5-fold	√	SN, PRE, SP, FPR, AUC, AUPRC

[1] The name of each method is provided in either the publication or the last name of its first author. [2] General metal ions mean that the related predictor does not differentiate the types of metal ion binding. Otherwise, we list the specific types of metal-binding in detail. [3] The number represents the size of the benchmark dataset. [4] The value reveals the protein similarity threshold in the benchmark dataset. The content in the blanket indicates the tool that is used for clustering proteins. [5] SMO: sequential minimal optimization; SVM: support vector machine; WSRC: weighted sparse representation based classifier; NN: neural network; CN: capsule network; CNN: convolutional neural networks; RF: random forest; GBM: gradient boosting machine; RBFN: radial basis function networks; LR: logistic regression; MLP: multilayer perceptron. [6] SN: sensitivity/recall; SP: specificity; ACC: accuracy; MCC: Matthew's correlation coefficient; PRE: precision; F1: F1-score; AUC: area under the ROC curve; AUPRC: area under the precision recall curve; FPR: false positive rate (FPR = 1-SP).

(Table 2) to evaluate the proposed models [31]. Specifically, k-fold cross-validation is usually adopted on the training dataset when building the prediction model and optimizing the related parameters [61]. First, the training dataset is equally divided into k parts. The division can be done at residue level or protein level. Next, k-1 subsets are used to train the model, and the last one subset is used for testing. The procedure repeats k times until every subset is been predicted. The performance of the model is usually evaluated by averaging the results of the k repeats.

3.2.2. Performance Measures. According to Table 2, the measures that used to evaluate the performance of the predictors can be divided into binary value-based and propensity score-based ones. The former needs preset thresholds to compute the number of putative binding residues and nonbinding residues. These measures include sensitivity (SN)/recall/true positive rate (TPR), specificity (SP), false positive rate (FPR, FPR = 1-SP) precision (PRE), accuracy (ACC), F1-score (F1), and Matthew's correlation coefficient (MCC). They are defined as follows:

$$\begin{aligned}
\mathrm{SN} = \mathrm{TPR} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}},\\
\mathrm{SP} = 1-\mathrm{FPR} &= \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}},\\
\mathrm{PRE} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}},\\
\mathrm{F1} &= \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}},\\
\mathrm{ACC} &= \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FN}+\mathrm{TN}+\mathrm{FP}},\\
\mathrm{MCC} &= \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FN}\times\mathrm{FP}}{\sqrt{(\mathrm{TP}+\mathrm{FN})\times(\mathrm{TP}+\mathrm{FP})\times(\mathrm{TN}+\mathrm{FP})\times(\mathrm{TN}+\mathrm{FN})}},
\end{aligned} \tag{1}$$

where TP (true positive) indicates the number of correctly recognized metal-binding residues, FP (false positive) means the number of non-metal-binding residues that are incorrectly predicted as metal-binding residues, TN (true negative) stands for the number of correctly predicted non-metal-binding residues, and FN (false negative) is the number of metal-binding residues that are incorrectly predicted as non-metal-binding residues.

The prediction of metal-binding residues is a typical imbalanced classification problem. That is, the number of metal-binding residues is much less than that of the non-metal-binding ones. Therefore, F1-score and MCC are regarded as key criteria since they are featured by assessing the prediction performance for both metal-binding and non-metal ion binding residues.

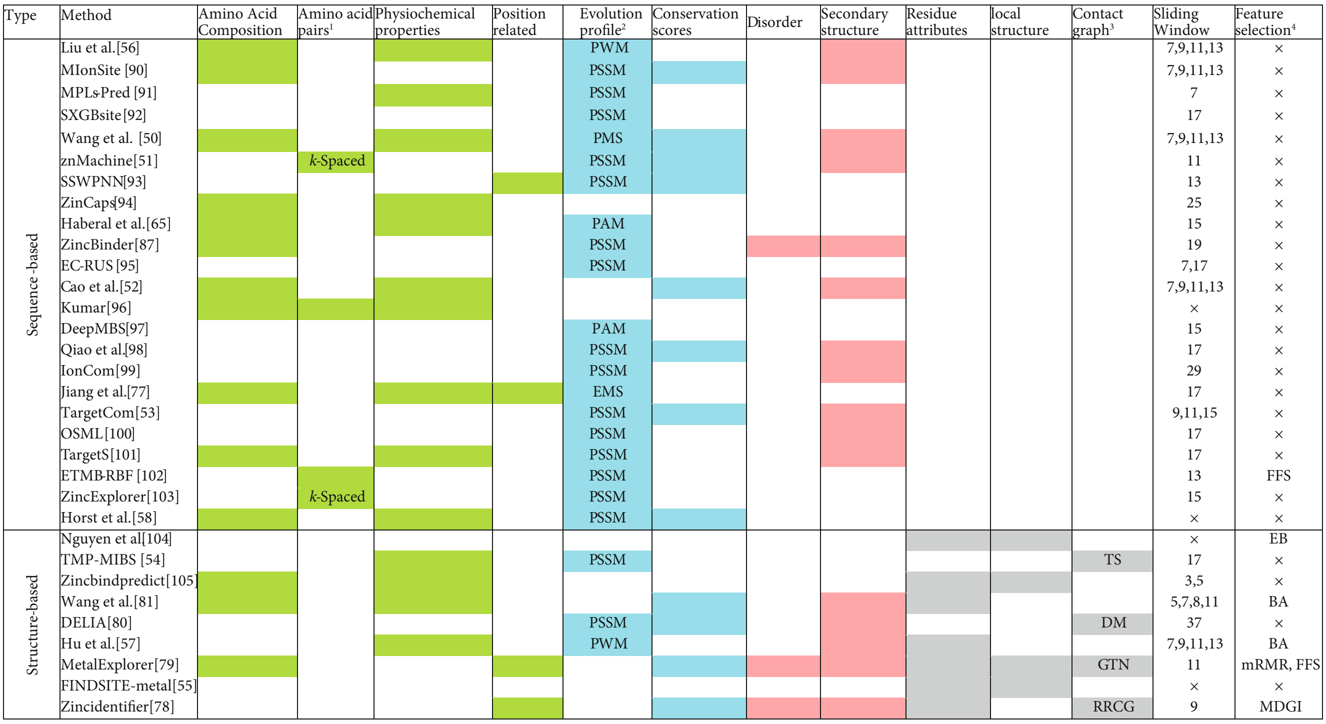
The propensity score-based measures include receiver operating characteristic curve (ROC curve) and precision-recall curve (PR curve). The ROC curve draws the TPR (true positive rate) against the FPR (false positive rate) at various thresholds. The AUC computes the area under the ROC curve and can be used to quantify the ROC curve. The PR curve plots PRE values on the y-axis and recalls values on the x-axis, and the AUPRC estimates the area under the PR curve.

3.3. Learning-Based Methods. Learning-based methods treat the recognition of metal-binding residues as a typical pattern recognition problem. Specifically, the metal-binding residues and nonbinding ones are encoded by using mathematical descriptors, i.e. features. According to the information that used to compute the features, the learning-based methods can be further categorized into sequence-based and structure-based methods. The former only needs simple protein sequences to extract features when encoding the binding residues. These features include sequence directly derived, evolutionary profile-based, and putative structure-based features, while the latter uses both sequence and native structure data to mathematically describe a binding residue. We make a comprehensive literature search and collect 23 sequence- and 9 structure-based methods that were published after the year 2010.

3.3.1. Feature Construction

(1) Sequence Directly Derived Features. We define sequence directly derived features as the ones that are computed from protein primary sequences without using any other information. In Figure 3, 14 out of 32 considered methods consider amino acid composition [50], which quantifies the relative difference in abundance of a given amino acid type [65, 66]. Amino acid pairs, or dipeptides, are based on the observation that amino acid pairs show different propensities in protein structure and function. For instance, pairs of lysine are found present in close spatial vicinity [67]. Moreover, the concept of k-spaced amino acid pairs is introduced in [68]. It calculates the amino acid pairs with k spaces between two residues. Our survey also shows that the majority of studies use physicochemical properties to describe the local environment of the metal-binding residues. The basic physicochemical environment of a metal-protein binding interface is reflected by the specific roles the metal plays in biostructural chemistry and protein function. These properties are crucial since they underpin many of the functional roles of metal ions. These properties include aliphatic [69], sulphur [70], aromatic [71], hydrophobic [72], charge [73], polar [74], positive [73], acidic [75], and hydroxylic [76]. The position-related features mainly consider the influence of the specifically located residues, such as autocross covariance [77] and sequence length [78, 79].

(2) Evolutionary Profile-Based Features. Recent studies [54, 56, 57, 80, 81] pointed out that functional or structural important residues tend to show higher evolutionary conservation. The conserved residues are usually involved in enzyme activity, ligand binding, or protein structural stability [82]. The conserved residues and regions can be identified by multiple sequence alignment [83]. These multiple sequence alignments, also named conservation profiles, include aligning families of homologous sequences and having knowledge of their evolutionary relationships [84]. For



Type	Method	Amino Acid Composition	Amino acid pairs[1]	Physiochemical properties	Position related	Evolution profile[2]	Conservation scores	Disorder	Secondary structure	Residue attributes	local structure	Contact graph[3]	Sliding Window	Feature selection[4]
Sequence -based	Liu et al.[56]	🟩		🟩		🟦 PWM			🟥				7,9,11,13	×
	MIonSite [90]	🟩				🟦 PSSM	🟦		🟥				7,9,11,13	×
	MPLs-Pred [91]			🟩		🟦 PSSM							7	×
	SXGBsite[92]					🟦 PSSM							17	×
	Wang et al. [50]	🟩		🟩		🟦 PMS	🟦		🟥				7,9,11,13	×
	znMachine[51]		🟩 *k*-Spaced			🟦 PSSM	🟦		🟥				11	×
	SSWPNN[93]				🟩	🟦 PSSM	🟦						13	×
	ZinCaps[94]	🟩		🟩									25	×
	Haberal et al.[65]	🟩		🟩		🟦 PAM							15	×
	ZincBinder[87]	🟩				🟦 PSSM		🟥	🟥				19	×
	EC-RUS [95]					🟦 PSSM							7,17	×
	Cao et al.[52]	🟩		🟩			🟦		🟥				7,9,11,13	×
	Kumar[96]	🟩	🟩	🟩									×	×
	DeepMBS[97]					🟦 PAM							15	×
	Qiao et al.[98]					🟦 PSSM	🟦		🟥				17	×
	IonCom[99]					🟦 PSSM			🟥				29	×
	Jiang et al.[77]	🟩		🟩	🟩	🟦 EMS							17	×
	TargetCom[53]					🟦 PSSM	🟦		🟥				9,11,15	×
	OSML[100]					🟦 PSSM			🟥				17	×
	TargetS[101]	🟩		🟩		🟦 PSSM			🟥				17	×
	ETMB-RBF [102]		🟩			🟦 PSSM							13	FFS
	ZincExplorer[103]		🟩 *k*-Spaced			🟦 PSSM							15	×
	Horst et al.[58]	🟩		🟩		🟦 PSSM	🟦						×	×
Structure-based	Nguyen et al[104]									▨	▨		×	EB
	TMP-MIBS [54]			🟩		🟦 PSSM						▨ TS	17	×
	Zincbindpredict[105]	🟩		🟩						▨	▨		3,5	×
	Wang et al.[81]	🟩		🟩			🟦		🟥	▨			5,7,8,11	BA
	DELIA[80]					🟦 PSSM	🟦		🟥			▨ DM	37	×
	Hu et al.[57]			🟩		🟦 PWM			🟥	▨			7,9,11,13	BA
	MetalExplorer[79]	🟩			🟩		🟦	🟥	🟥	▨	▨	▨ GTN	11	mRMR, FFS
	FINDSITE-metal[55]									▨	▨		×	×
	Zincidentifier[78]				🟩		🟦	🟥	🟥	▨		▨ RRCG	9	MDGI

The light green cells indicate sequence directly derived features; The light blue cells stand for profile-based features; The light red cells mean putative structure-based features; The light grey cells are native structure-based features. [1] In the amino acid pairs column, the cells without annotations indicate original amino acid pairs; the cells annotated using 'k-spaced' means k-spaced amino acid pairs. [2] PWM: position weight matrix; PSSM: position specific scoring matrix; PMS: position matrix scoring; PAM: point accepted mutation; EMS: evolutionary matrix scoring. [3] TS: topology structure; DM: distance matrix; GTN: graph theoretic network; RRCG: residue-residue contact graphs. [4] FFS: forward feature selection; EB: experience-based; BA: Boruta algorithm; mRMR: minimum-redundancy maximum-relevancy; MDGI: mean decrease gini index.

Figure 3: Summary of the feature construction and selection for learning-based methods. The light green cells indicate sequence directly derived features. The light blue cells stand for profile-based features. The light red cells mean putative structure-based features. The light grey cells are native structure-based features. [1]In the amino acid pair column, the cells without annotations indicate original amino acid pairs; the cells annotated using "*k*-spaced" means *k*-spaced amino acid pairs. [2]PWM: position weight matrix; PSSM: position specific scoring matrix; PMS: position matrix scoring; PAM: point accepted mutation; EMS: evolutionary matrix scoring. [3]TS: topology structure; DM: distance matrix; GTN: graph theoretic network; RRCG: residue-residue contact graphs. [4]FFS: forward feature selection; EB: experience-based; BA: Boruta algorithm; mRMR: minimum-redundancy maximum-relevancy; MDGI: mean decrease Gini index [50–51, 53–58, 65, 77–81, 87, 90–105].

an unknown protein, although its accurate function is not available, it is expected that we can use its homologous proteins to speculate the function since they share the similar evolutionary profile [85]. Many studies use position-specific scoring matrix (PSSM), which is computed from PSI-BLAST [62], to quantify the evolutionary conservation. PSSM scores the substitution probability of each residue in the protein being substituted by other types of amino acids. Liu et al. [56] and Hu et al. [57] set different weights according to the positions of considered residues within the window and construct position weight matrix (PWM). Wang et al. proposed a customized position matrix scoring (PMS) algorithm, which uses known sequence patterns to describe the composition of amino acids at different positions [50]. Haberal and Oğul introduced a point accepted mutation (PAM) scoring matrix, which measures the rate at which point mutations that substitute one residue for another during evolution [65]. Jiang et al. adopted evolutionary matrix scoring (EMS) algorithm to extract the position conservation of amino acid residues from segments with low dimension feature parameters [77].

(3) Putative Structure-Based Features. For an unknown protein, although the accurate function is not available, it is expected that we can use its homologous proteins or template structures to speculate the structure. The secondary structure mainly involves α-helix, β-sheet, and coil, which are fundamental elements of protein tertiary structure [86]. Natively disordered or unstructured regions are proved to be associated with molecular assembly, protein translation, modification, and molecular recognition [78, 79, 87]. Previous studies [79, 87] indicate that disordered regions are strongly correlated with local solvent accessibility areas. Figure 3 reveals that 16 methods introduce secondary structure features and 3 approaches use disorder features, respectively. The secondary structure can be obtained from the primary sequence by using PSIPRED [88]. Putative intrinsic disorder data can be computed by using DISOPRED [89].

(4) Structure-Based Features. The structure-based features include descriptors that are computed from protein 3D structure. These features include solvent exposure, B-factor, spatial cluster properties, and native secondary structure.

Compared with the abovementioned putative structure-based features, the native structure-based features are more accurate since they are directly computed by using residue coordinate data. Besides that, a residue contact network is also considered by some literature. In [79], two residues are defined as being in contact if the distance of their Cα atoms is less than a predefined cutoff distance of 6.5 Å. These features include clustering coefficient, degree, density, distance, topology structure, and graph theoretic network [55, 79, 80].

3.3.2. Sliding Window Optimization and Feature Selection. As shown in Figure 3, many methods adopt a sliding window scheme when they construct different types of features. It is because residues in proteins are influenced by adjacent residues. Besides that, binding residues tend to cluster together. If a central residue is a native-binding residue, its adjacent residues usually have a relatively higher chance to bind the same ligands. Usually, the residues with a long distance away have a lower impact on the central residues when compared with the residues with short distance. Figure 3 summarizes that 19 out of 23 methods use the sliding window scheme. The size of the shortest window is 3 [105], while the size for the longest one is 25 [94]. Some studies [50, 52, 53, 56, 57, 81, 90, 105] use more than one type of window because they consider different types of metal-binding residues. A long window means the introduction of more features.

However, a bigger number of features do not absolutely mean a better prediction performance [106, 107]. The existence of potential "bad" features may interfere with the classifiers and cause unpredictable consequences [108]. The so-called "bad" features include irrelevant and redundant ones. To avoid their terrible influences, it is necessary to perform feature selection before training the model [109]. Figure 3 reveals that 6 out of 32 methods adopt feature selection before training the model. These feature selection approaches include forward feature selection [79, 102], experience-based [104], Boruta algorithm [57, 81], minimum-redundancy maximum-relevancy [79], and mean decrease Gini index [78].

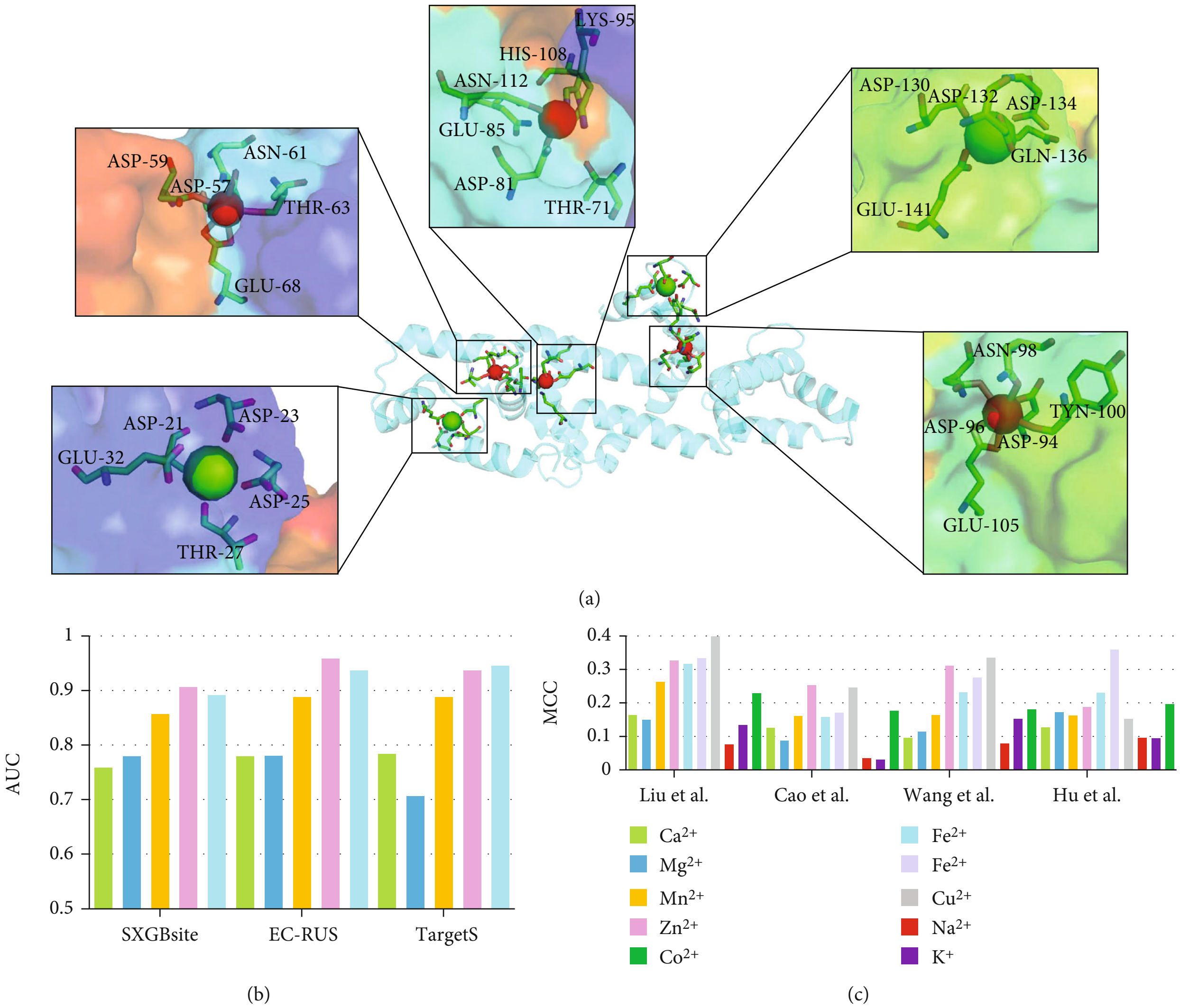
3.3.3. Prediction Algorithms. Learning-based methods use machine-learning or deep-learning-based algorithms to train the model and perform predictions [110]. As shown in Table 2, a variety of algorithms are introduced for solving the problem of correctly recognizing metal-binding residues. Support vector machine (SVM) is a popular machine learning algorithm in bioinformatical research. It is aimed at finding a hyperplane or decision boundary that can segregate a high-dimensional space [111]. Particularly, it uses kernel functions to reduce computation time to avoid strapping into dimension disaster [112]. Sequential minimal optimization (SMO) is an algorithm that is specially used for training support vector machines [113]. The procedure of training large data by SVM usually leads to a complex quadratic programming optimization problem [114]. SMO breaks large programming optimization problems into small ones, which endows SVM a good generalization on large data [113]. The idea of a neural network (NN) comes from the work system of neurons in the biological brain [115]. It learns the correlations between inputs and outputs, making generalizations and build models [116]. The NN algorithm assigns and adjusts different weights for neurons and edges as learning proceeds. The radial basis function network (RBFN) is a variant of the original NN [117]. It adopts radial basis functions as activation functions, which can be used for accelerating learning speed due to their universal approximation [118]. The multilayer perceptron (MLP) algorithm is an improved back propagation NN [119]. It mainly includes three procedures: forward propagation, error evaluation, and error backpropagation [120]. The MLP is featured by its strong generalization and fault tolerance [121]. Therefore, it is proved to be an efficient classification algorithm. The logistic regression (LR) adopts a logistic function to model the probability of an unknown sample being a certain class [122].

Our survey also reveals that the ensemble algorithms are favored by eight studies. The random forest (RF) aggregates the predictions of all the decision trees and performs decisions by most trees [123]. RF can be used for classification, regression, and optimization problems [124]. Adaptive boosting (AdaBoost) is aimed at combining weak learners with strong ones [125, 126]. The key point of AdaBoost is to ensure the diversities of individual learners, which makes it a good generalization ability [90, 127]. The gradient boosting machine (GBM) is another popular ensemble algorithm. During the iterative process, GBM dynamically increases the weight of wrong recognitions and reduces that of the correct ones [128–131]. It should be noted that GBM focuses on the sample residual of the previous iteration instead of the sample itself [132].

Besides machine-learning algorithms, recent studies also use deep-learning methods in this research field. The convolutional neural network (CNN) is one of the most prevalent algorithms that is widely used in bioinformatics [133]. The CNN consists of three main layers, which are the convolutional layer, pooling layer, and fully connected layer [134–136]. Although the CNN is proved to be powerful in dealing with a variety of problems, it performs badly when facing samples with different sizes and orientations [137, 138]. To overcome this shortcoming, the capsule network (CN) is proposed to estimate features of objects by incorporating dynamic routing algorithms [139, 140]. Our review finds two studies use CNN [65, 97] and one uses CN [94].

3.4. Docking-Based Methods. The investigation on the protein-metal complex helps biologists to understand the mechanism of protein-metal interactions. Protein-ligand docking approaches are always based on molecular structure and are used to explore biomolecular interactions and mechanisms [141]. It can be adopted to predict binding conformation as well as the appropriate target binding residues [142, 143].

As shown in Figure 2, the docking-based methods mainly include three steps: searching algorithm, scoring function, and docking assessment [141]. The searching algorithm focuses on creating an optimum number of configurations that properly include the determined binding modes



(a)

(b)

(c)

FIGURE 4: Ribbon and surface model of X-ray structure of Ca^{2+}- and Zn^{2+}-bound calmodulin (PDB: 4HEX) in *Mus musculus*. Red sphere represents bound zinc ion; green one indicates calcium ion; the spatial adjacent residues participating its coordination are shown by the stick model.

[144]. To reduce computation time, it is necessary to make a balance between the computational expense and the searching space. The scoring function includes a series of mathematical functions that quantify the strength of binding affinity [145]. The energy-based scoring functions are always introduced to score the potential interactions between the protein and the corresponding ligands [141]. The frequently used functions include empirical-based, knowledge-based, and consensus-based ones. Finally, the putative docking can be evaluated by using docking accuracy and the correlation between putative and native docking scores [145]. Figure 4 illustrates the structure of a calmodulin (PDB: 4HEX) that is secreted by *Escherichia coli* in *Mus musculus* [146]. Calmodulin is one of the most prevalent EF-hand calcium sensor proteins in eukaryotic cells [147]. It is a highly conserved and soluble protein, which activates enzymes and regulates many cellular functions. 4HEX has three Ca^{2+}-binding and two Zn^{2+}-binding sites. Ca^{2+}-binding causes a change in calmodulin conformation opening both globular domains and exposing hydrophobic surfaces that form binding sites for the target enzymes. Figure 4 shows that these three Ca^{2+} are in the pockets. The binding pockets are half-closed and buried, which substantially limits the capability of Ca^{2+} to escape. Two Zn^{2+}-binding sites are surrounded by a shell of hydrophilic groups that are embedded into a larger shell of hydrophobic groups. The amino acid side chains providing ligands to Zn^{2+} in these structures often form hydrogen bonds with other residues [147].

In [148], He et al. proposed a docking-based predictor named mFASD. It first explored the local biochemical environment of potential functional atoms and then measured the distances between the atoms and bound metal. mFASD also claimed that it can differentiate different types of metal-binding sites. Zhou et al. improved the FEATURE-based calcium model and used the grid scan algorithm to recognize binding sites [149]. GaudiMM [150] adopted a multiobjective genetic algorithm to search metal-binding sites in biological scaffolds. BioMetAll focused on the

Table 3: Summary of docking-based, template-based, and meta-based methods.

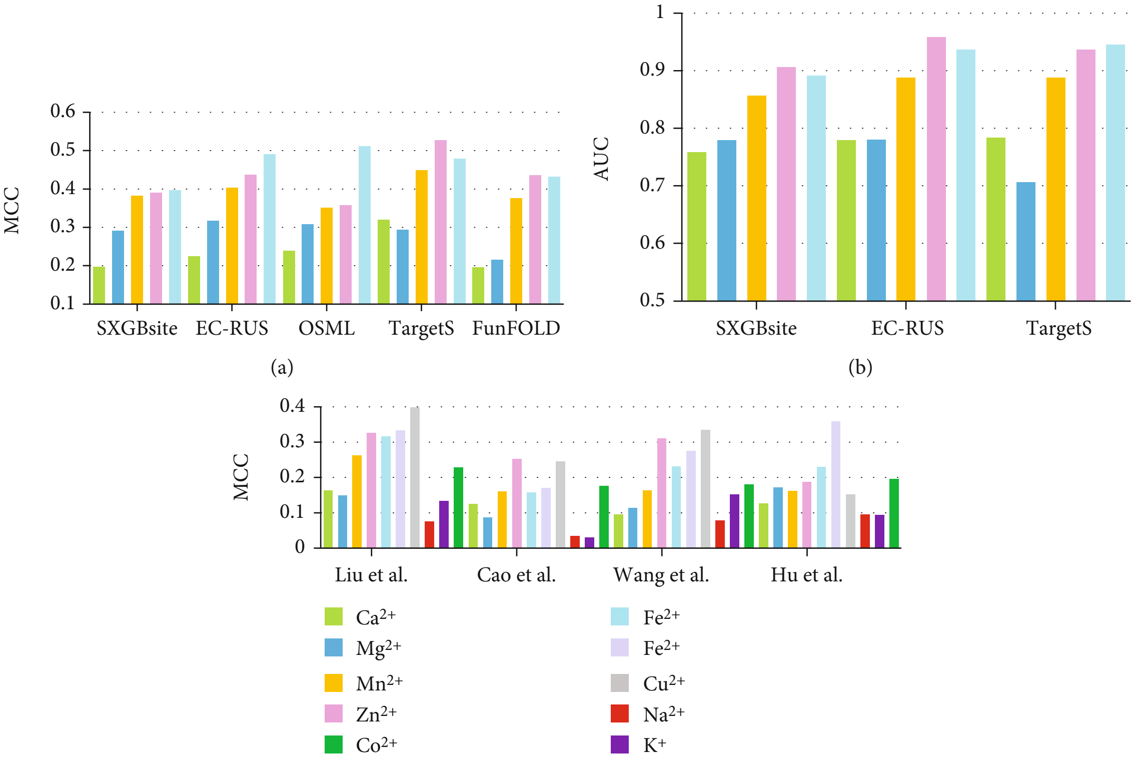
Type	Method	Year	Notes
Docking-based	mFASD [148]	2015	Capture the characteristics of metal-binding sites and discriminate most types of these sites
	Zhou et al. [149]	2015	Use a FEATURE-based calcium model and convert high scoring regions into specific site predictions
	GaudiMM [150]	2019	Find poses that satisfy metal-derived geometrical rules and use post optimizations
	BioMetAll [151]	2020	Predict metal-binding sites with particular motifs, determine transient sites in structures, and predict potential mutations to generate convenient sites
Template-based	Deng et al. [152]	2006	Use a graph theory algorithm to find oxygen clusters of the protein (high potential for calcium binding)
	Goyal et al. [153]	2008	Describe generation of 3D-structural motifs for metal-binding sites from the known metalloproteins
	Levy et al. [154]	2009	Analyze whether structural models based on remote homology are effective in predicting 3D metal binding sites
	FunFOLD [155]	2011	Use an automated method for ligand clustering and identification of binding residues
	FunFOLDQA [156]	2012	Use a fully automated agglomerative clustering approach for both ligand identification and residue selection
	FunFOLD2 [157]	2013	Propose a method that include protein-ligand binding prediction and quality assessment protocol
Meta-based	Li et al. [158]	2017	Integrate the results of ZincExplorer [103], zincFinder [159], and zincPred [160]
	IBayes_Zinc [161]	2019	Adopt Bayesian method and combine the predictions from ZincExplorer [103], zincFinder [159], and zincPred [160]

conformation of the potential metal-binding site, associated with the geometric organization of the protein backbone [151]. It was also proved to have good performance on the applications including the modulation and mutation of the metal-binding residues. Table 3 summarizes the key notes of the abovementioned 4 docking-based methods.

3.5. Template-Based Methods. It is well known that protein structure determines function, and similar interface conformation indicates similar bound regions [162]. The template-based methods are based on the abovementioned hypothesis. Therefore, the most important thing for template-based methods is to find and validate proper structural templates. The fold recognition algorithms, which quantify the best matches from candidate templates, are commonly used to select the optimal template structures [163]. Next, the selected templates are used to map onto the target protein given the alignments with the template structures [164].

As shown in Table 3, Goyal and Mande analyzed the metal-binding sites by using structure templates and designing 3D motifs for several types of metal-binding interactions [153]. In [154], the authors analyzed whether structural models based on remote homology are effective in recognizing structural metal-binding residues based on simple protein primary sequences. Deng et al. applied a graph theory algorithm to identify, predict, and analyze calcium-binding residues [152]. However, it should be noted that this strategy produces good prediction performance when a decent complex is available as a template. If the template structure information is not available, this strategy might have poor predictions [164]. The FunFOLD was an automatic method that uses protein structure superposition of distantly related templates to a modelled protein for the clustering of ligands and prediction of metal binding residues [155]. The FunFOLDQA [156] approach determined the reliability of our FunFOLD [155] by assigning the quality assessment scores. FunFOLD2 was a web server that integrated cutting edge function and putative 3D structures to identify metal-binding residues [157].

3.6. Meta-Based Methods. The meta-based methods use a meta-learning strategy from fewer samples than traditional machine learning models. Since meta-based methods can only use limited data, they must ensure that the data is featured with high accuracy. As a result, a meta-based approach always directly combines the predictions of other methods. It uses weights or voting strategy on the available propensity scores or binary values. Thus, the meta-based method promises a robust accurate prediction on the metal-binding residues. In [158], Li et al. collected the predictions from ZincExplorer [103], ZincFinder [159], and ZincPred [160] (Table 3). Then, they built a linear regression model and optimized corresponding parameters on the training dataset. They claimed that the meta-model, which was named meta-zincPrediction, improves the AUPRC by about 2%~8%. IBayes_Zinc [161] was another meta-based predictor for the identification of zinc-binding residues (Table 3). It firstly computed the predictions of zinc-binding probabilities from ZincExplorer [103], ZincFinder [159], and ZincPred [160]. Next, IBayes_Zinc processed the missing attribute values and adopts Bayesian theory [165] to construct a meta-based model. The performance on the independent dataset proved that the MCC value of IBayes_Zinc was about 5~13% higher than the considered three predictors.



(a)

(b)

(c)

Figure 5: Comparative assessment of several predictors on two benchmark dataset. (a) and (c) indicate the MCC bar charts for considered methods on different metal ion binding residues on Yu et al.'s and Cao et al.'s testing datasets, respectively. (b) draws the AUC values of three predictors on corresponding metal ion binding residues.

3.7. Prediction Results. This review surveys 44 computation-based methods. It is necessary to make a consensus comparison for these methods. However, since there is no standard benchmark dataset and some methods are not currently available, we use two datasets that are used by some methods to perform the evaluations. The first dataset is compiled by Yu et al. in [101], which includes five types of metal ions binding annotations. The second dataset is obtained from [52], consisting of ten types of metal ions binding annotations.

Figure 5 illustrates the predictive performance on two benchmark test datasets. Details are provided in Table S1 and Table S2 in Supplementary Materials, respectively; the corresponding results are sourced from [56, 57, 81, 92]. We notice that the predictors show relatively big differences in recognizing various types of metal-binding residues. On Yu et al.'s dataset [101], TargetS shows the best results in predicting Ca^{2+}-, Zn^{2+}-, and Mn^{2+}-binding residues; EC-RUS [95] performs best in recognizing Mg^{2+}-binding residues; OSML [100] achieves the highest MCC on Fe^{3+}-binding predictions. Besides that, Figure 5(a) indicates that all five methods show a decent performance on recognizing Fe^{3+}-binding residues (MCC values close or higher than 0.4), compared with MCC close or less than 0.2 on Ca^{2+} binding residues. Figure 5(b) draws the bars of AUC values for SXGBsite [92], EC-RUS [95], and TargetS [101], respectively. These three predictors all achieve high AUC scores (close or higher than 0.9) on Zn^{2+}- and Fe^{3+}-binding residues. Figure 5(c) summarizes the results of ten metal ions binding residues on Cao et al.'s dataset [52]. Among these predictors, Liu et al. [56] performs the best on Zn^{2+}, Fe^{3+}, and Cu^{2+}, compared to [81], Wang et al. shows best on Zn^{2+} and Cu^{2+}, and Hu et al. [57] achieves the highest on Fe^{2+}. Interestingly, the binding residues associated with relatively inactive metal (Zn^{2+}, Fe^{3+}, and Cu^{2+}) ions show relatively better results compared to that of the active metal ions (Na^{+} and K^{+}). Particularly, four methods all give better results on Fe^{3+}-binding residues than that on Fe^{2+}-binding residues, which keep consistent with our observations as mentioned above.

3.8. Publicly Available Tools. The publicly available standalone software or web server that implements the proposed approach provides convenience for biologists and researchers [79, 105, 122]. These tools help the community to repeat the results and build a platform for easy understanding and improvement. Table 4 summarizes the public availability of implementations for the considered methods. These 28 predictors are implemented as standalone software or web servers. Among these predictive tools, 16 (or 57%) of them are currently publicly available. Standalone software requires the biologists to build the same running environment. By contrast, the web server provides the most convenient since the users only need to submit their queries via

Table 4: A breakdown of predictive tools of metal-binding residues.

Method	Year	Platform[1]	Web link	Availability[2]
TMP-MIBS [54]	2021	SS	https://github.com/QuJing785464/TMP_MIBS	Yes
Wang et al. [50]	2021	WS	http://39.104.77.103:8081/lsb/HomePage/HomePage.html	No
Zincbindpredict [105]	2021	WS	https://zincbind.bioinf.org.uk/predict/	No
DELIA [80]	2020	WS	http://www.csbio.sjtu.edu.cn/bioinf/delia/	Yes
BioMetAll [151]	2020	SS	https://github.com/insilichem/biometall	Yes
MPLs-Pred [91]	2019	WS	http://icdtools.nenu.edu.cn/	Yes
SXGBsite [92]	2019	SS	https://github.com/Lightness7/SXGBsite	Yes
MIonSite [90]	2019	SS	https://github.com/LiangQiaoGu/MIonSite.git	Yes
znMachine [51]	2019	WS&SS	http://bioinformatics.fzu.edu.cn/znMachine.html	No
ZinCaps [94]	2019	SS	https://github.com/clemEssien/ActiveSitePrediction	Yes
EC-RUS [95]	2017	SS	https://github.com/6gbluewind/protein_ligand_binding_site	Yes
MetalExplorer [79]	2017	WS	http://metalexplorer.erc.monash.edu.au/	No
Cao et al. [52]	2017	WS	http://60.31.198.140:8081/metal/HomePage/HomePage.html	No
ZincBinder [103]	2017	WS&SS	http://proteininformatics.org/mkumar/znbinder/	Yes
SSWPNN [93]	2017	SS	http://net.jitsec.cn:88/UploadedImages/SSWPNN.rar	Yes
Jiang et al. [77]	2016	WS	http://202.207.29.245/	No
TargetCom [53]	2016	SS	http://dase.ecnu.edu.cn/qwdong/TargetCom/TargetCom_standalone.tar.gz	No
OSML [100]	2015	WS	http://www.csbio.sjtu.edu.cn/OSML/	Yes
mFASD [148]	2015	SS	http://staff.ustc.edu.cn/liangzhi/mfasd/	Yes
FunFOLD2 [157]	2013	WS	http://www.reading.ac.uk/bioinf/FunFOLD/FunFOLD_form_2_0.html	Yes
ZincExplorer [103]	2013	WS	http://protein.cau.edu.cn/ZincExplorer	No
TargetS [101]	2013	WS	http://www.csbio.sjtu.edu.cn/TargetS/	Yes
FunFOLDQA [156]	2012	SS	http://www.reading.ac.uk/bioinf/downloads/	Yes
Zincidentifier [78]	2012	WS	http://protein.cau.edu.cn/zincidentifier/	No
FINDSITE-metal [55]	2011	WS	http://cssb.biology.gatech.edu/findsite-metal/	No
FunFOLD [155]	2011	WS&SS	http://www.reading.ac.uk/bioinf/FunFOLD/	Yes
Goyal et al. [153]	2008	WS	http://sunserver.cdfd.org.in:8080/protease/PAR_3D/index.html	No
Deng et al. [152]	2006	SS	http://chemistry.gsu.edu/faculty/Yang/GG.htm	No

[1]WS: web server; SS: standalone software. [2]The availability was estimated on Dec 1st, 10th, and 20th of 2021, respectively.

the browser, and the server helps to do the computations. Three methods, namely, znMachine [51], ZincBinder [103], and FunFOLD [155], provide both web server and standalone software. TMP_MIBS [54] is designed to predict general metal-binding residues and deployed using the Python language. DELIA [80] requires PDB-formatted 3D coordinates input and produces both binary prediction and putative probability of a residue being potential specific metal-binding. BioMetAll uses a docking-based strategy to scan specific motifs, putative mutations, and binding residues. Another available docking-based method is mFASD, which distinguishes different types of metal-binding sites according to the interaction distances. MPLs-Pred [91], SXGBsite [92], MIonSite [90], OSML [100], EC-RUS [95], and TargetS [101] are all sequence-based predictive tools, which accepts FASTA-formatted input and produced the results of putative metal-binding residues. ZinCaps [94], SSWPNN [93], and ZincBinder [103] are specially designed for the identification of zinc-binding residues. FunFOLD [155], FunFOLDQA [156], and FunFOLD2 [157] are a series of template-based methods.

4. Conclusions and Future Perspectives

This review summarizes the public database of metal ions binding interactions, discusses the architectures of computation-based methods for identifying binding residues, and comparatively evaluates four types of methods. Based on the observations made in this work, we propose a few recommendations for future research in this field:

First, the researchers should maintain and update the database regularly. This will significantly improve effectiveness and completeness for these databases and provide convenience for the computation-based methods, which depend on the accurate internal database. We expect a high-quality metal ion binding-related database with an advanced searching engine, high-speed download service, complete annotation information, etc. Particularly, a decent database should be designed to open for easy expanding and improvement. *Second*, standard benchmark datasets that related to general or ligand-specific metal-binding residues should be periodically compiled and made available. This will ensure consistent evaluation and comparative analysis of the performance of the existing and novel methods. *Third*, these predictors are expected to use delicate architectures and powerful algorithms. Since the differences between different types are quite small, the novel predictors shall not only correctly identify metal-binding residues but also distinguish different types of metal ions. *Fourth*, the authors of the metal-binding predictors are suggested to make their approaches publicly available, preferably as both webservers and standalone software. Particularly, high-throughput predictors promise a wide application among the research community since they can be used to perform large-scale computations, such as proteome-level predictions.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Nos. 62002307 and 61802329), by the Innovation Team Support Plan of University Science and Technology of Henan Province (No. 19IRTSTHN014), by the Key Scientific Research Project of Colleges and Universities in Henan Province (Grant No. 22A170019), by the Project of Science and Technology Department of Henan Province (No. 212102210392), and by Nanhu Scholars Program for Young Scholars of Xinyang Normal University.

Supplementary Materials

Table S1: comparative assessment of considered methods on Yu et al.'s independent testing dataset. Table S2: comparative assessment of considered methods on Cao et al.'s independent testing dataset. *(Supplementary Materials)*

References

[1] K. H. Thompson and C. Orvig, "Boon and bane of metal ions in medicine," *Science*, vol. 300, no. 5621, pp. 936–939, 2003.

[2] A. T. Aron, D. Petras, R. Schmid et al., "Native mass spectrometry-based metabolomics identifies metal-binding compounds," *Nature Chemistry*, vol. 14, no. 1, pp. 100–109, 2022.

[3] C. Andreini, I. Bertini, and A. Rosato, "Metalloproteomes: a bioinformatic approach," *Accounts of Chemical Research*, vol. 42, no. 10, pp. 1471–1479, 2009.

[4] M. J. V. Clausen and H. Poulsen, "Sodium/potassium homeostasis in the cell," in *Metallomics and the Cell*, pp. 41–67, Springer, 2013.

[5] H. Lambers, P. E. Hayes, E. Laliberte, R. S. Oliveira, and B. L. Turner, "Leaf manganese accumulation and phosphorus-acquisition efficiency," *Trends in Plant Science*, vol. 20, no. 2, pp. 83–90, 2015.

[6] W. Maret and A. Wedd, *Binding, Transport and Storage of Metal Ions in Biological Cells*, Royal Society of Chemistry, 2014.

[7] G. B. Akcapinar and O. U. Sezerman, "Computational approaches forde novodesign and redesign of metal-binding sites on proteins," *Bioscience Reports*, vol. 37, no. 2, 2017.

[8] K. P. Kepp, "Alzheimer's disease: how metal ions define β-amyloid function," *Coordination Chemistry Reviews*, vol. 351, pp. 127–159, 2017.

[9] L. Reinhard, H. Tidow, M. J. Clausen, and P. Nissen, "Na+, K+-ATPase as a docking station: protein–protein complexes of the Na+, K+-ATPase," *Cellular and Molecular Life Sciences*, vol. 70, no. 2, pp. 205–222, 2013.

[10] K. Krzywoszyńska, D. Witkowska, J. Świątek-Kozłowska, A. Szebesczyk, and H. Kozłowski, "General aspects of metal ions as signaling agents in health and disease," *Biomolecules*, vol. 10, no. 10, p. 1417, 2020.

[11] L. Quintanar and M. H. Lim, *Metal ions and degenerative diseases*, Springer, 2019.

[12] E. J. Tokar, L. Benbrahim-Tallaa, and M. P. Waalkes, "14 Metal ions in human cancer development," *Metal Ions in Toxicology: Effects, Interactions, Interdependencies*, vol. 8, p. 375, 2015.

[13] J. Zhang, "Current advances in drug design and cancer research," *Current Topics in Medicinal Chemistry*, vol. 21, no. 15, pp. 1307–1309, 2021.

[14] N. Potier, H. Rogniaux, G. Chevreux, and A. Van Dorsselaer, "Ligand-Metal Ion Binding to Proteins: Investigation by ESI Mass Spectrometry," *Methods in Enzymology*, vol. 402, pp. 361–389, 2005.

[15] K. B. Handing, E. Niedzialkowska, I. G. Shabalin, M. L. Kuhn, H. Zheng, and W. Minor, "Characterizing metal-binding sites in proteins with X-ray crystallography," *Nature Protocols*, vol. 13, no. 5, pp. 1062–1090, 2018.

[16] D. Shen, X. Xu, H. Wu et al., "Metal ion binding to anticoagulation factor II from the venom of Agkistrodon acutus: stabilization of the structure and regulation of the binding affinity to activated coagulation factor X," *Journal of Biological Inorganic Chemistry*, vol. 16, no. 4, pp. 523–537, 2011.

[17] N. Thirumoorthy, A. S. Sunder, K. M. Kumar, G. Ganesh, and M. Chatterjee, "A review of metallothionein isoforms and their role in pathophysiology," *World Journal of Surgical Oncology*, vol. 9, no. 1, pp. 1–7, 2011.

[18] Y. Miller, B. Ma, and R. Nussinov, "Metal binding sites in amyloid oligomers: complexes and mechanisms," *Coordination Chemistry Reviews*, vol. 256, no. 19-20, pp. 2245–2252, 2012.

[19] W. Bal, M. Sokołowska, E. Kurowska, and P. Faller, "Binding of transition metal ions to albumin: sites, affinities and rates," *Biochimica et Biophysica Acta (BBA)-General Subjects*, vol. 1830, no. 12, pp. 5444–5455, 2013.

[20] N. Roohani, R. Hurrell, R. Kelishadi, and R. Schulin, "Zinc and its importance for human health: an integrative review," *Journal of research in medical sciences: the official journal of Isfahan University of Medical Sciences*, vol. 18, no. 2, pp. 144–157, 2013.

[21] V. Sobolev and M. Edelman, "Web tools for predicting metal binding sites in proteins," *Israel Journal of Chemistry*, vol. 53, no. 3-4, pp. 166–172, 2013.

[22] Z. Liu, Y. Wang, C. Zhou, Y. Xue, W. Zhao, and H. Liu, "Computationally characterizing and comprehensive analysis of zinc-binding sites in proteins," *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, vol. 1844, no. 1, pp. 171–180, 2014.

[23] D. Witkowska and M. Rowińska-Żyrek, "Biophysical approaches for the study of metal-protein interactions," *Journal of Inorganic Biochemistry*, vol. 199, article 110783, 2019.

[24] M. Mallick, A. Sharan Vidyarthi, and Shankaracharya, "Tools for predicting metal binding sites in protein: a review," *Current Bioinformatics*, vol. 6, no. 4, pp. 444–449, 2011.

[25] C. Rauer, N. Sen, V. P. Waman, M. Abbasian, and C. A. Orengo, "Computational approaches to predict protein functional families and functional sites," *Current Opinion in Structural Biology*, vol. 70, pp. 108–122, 2021.

[26] J. B. Tran and A. Krężel, "InterMetalDB: a database and browser of intermolecular metal binding sites in macromolecules with structural information," *Journal of Proteome Research*, vol. 20, no. 4, pp. 1889–1901, 2021.

[27] A. J. R. Cabrera, "Zinc, aging, and immunosenescence: an overview," *Pathobiology of Aging & Age-related Diseases*, vol. 5, no. 1, p. 25592, 2015.

[28] C. C. Staats, L. Kmetzsch, A. Schrank, and M. H. Vainstein, "Fungal zinc metabolism and its connections to virulence," *Frontiers in Cellular and Infection Microbiology*, vol. 3, p. 65, 2013.

[29] M. J. Glimcher, "Bone: nature of the calcium phosphate crystals and cellular, structural, and physical chemical mechanisms in their formation," *Reviews in Mineralogy and Geochemistry*, vol. 64, no. 1, pp. 223–282, 2006.

[30] H. M. Berman, J. Westbrook, Z. Feng et al., "The protein data bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000.

[31] N. Ye, "Survey of in-silico prediction of anticancer peptides," *Current Topics in Medicinal Chemistry*, vol. 21, no. 15, pp. 1310–1318, 2021.

[32] M. Hauser, M. Steinegger, and J. Söding, "MMseqs software suite for fast and deep clustering and searching of large protein sequence sets," *Bioinformatics*, vol. 32, no. 9, pp. 1323–1330, 2016.

[33] G. Li, Y. Su, Y.-H. Yan et al., "MeLAD: an integrated resource for metalloenzyme-ligand associations," *Bioinformatics*, vol. 36, no. 3, pp. 904–909, 2020.

[34] V. Putignano, A. Rosato, L. Banci, and C. Andreini, "MetalPDB in 2018: a database of metal sites in biological macromolecular structures," *Nucleic Acids Research*, vol. 46, no. D1, pp. D459–D464, 2018.

[35] C. Andreini, G. Cavallaro, S. Lorenzini, and A. Rosato, "MetalPDB: a database of metal sites in biological macromolecular structures," *Nucleic Acids Research*, vol. 41, no. D1, pp. D312–D319, 2012.

[36] J. Yang, A. Roy, and Y. Zhang, "BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions," *Nucleic Acids Research*, vol. 41, no. D1, pp. D1096–D1103, 2012.

[37] F. Fu and D. F. Voytas, "Zinc finger database (ZiFDB) v2. 0: a comprehensive database of C2H2 zinc fingers and engineered zinc finger arrays," *Nucleic Acids Research*, vol. 41, no. D1, pp. D452–D455, 2012.

[38] F. Fu, J. D. Sander, M. Maeder et al., "Zinc finger database (ZiFDB): a repository for information on C2H2 zinc fingers and engineered zinc-finger arrays," *Nucleic Acids Research*, vol. 37, no. Database, pp. D279–D283, 2009.

[39] A. Tus, A. Rakipović, G. Peretin, S. Tomić, and M. Šikić, "BioMe: biologically relevant metals," *Nucleic Acids Research*, vol. 40, no. W1, pp. W352–W357, 2012.

[40] H. Choi, H. Kang, and H. Park, "MetLigDB: a web-based database for the identification of chemical groups to design metalloprotein inhibitors," *Journal of Applied Crystallography*, vol. 44, no. 4, pp. 878–881, 2011.

[41] K. Hemavathi, M. Kalaivani, A. Udayakumar, G. Sowmiya, J. Jeyakanthan, and K. Sekar, "MIPS: metal interactions in protein structures," *Journal of Applied Crystallography*, vol. 43, no. 1, pp. 196–199, 2010.

[42] B. Kumar Kuntal, P. Aparoy, and P. Reddanna, "Development of tools and database for analysis of metal binding sites in protein," *Protein and Peptide Letters*, vol. 17, no. 6, pp. 765–773, 2010.

[43] K. Nakamura, A. Hirai, M. Altaf-Ul-Amin, and H. Takahashi, "MetalMine: a database of functional metal-binding sites in proteins," *Plant Biotechnology*, vol. 26, no. 5, pp. 517–521, 2009.

[44] C. Andreini, I. Bertini, G. Cavallaro, G. L. Holliday, and J. M. Thornton, "Metal-MACiE: a database of metals involved in biological catalysis," *Bioinformatics*, vol. 25, no. 16, pp. 2088-2089, 2009.

[45] M. Jayakanthan, J. Muthukumaran, S. Chandrasekar, K. Chawla, A. Punetha, and D. Sundar, "ZifBASE: a database of zinc finger proteins and associated resources," *BMC Genomics*, vol. 10, no. 1, pp. 1–7, 2009.

[46] K. Hsin, Y. Sheng, M. Harding, P. Taylor, and M. Walkinshaw, "MESPEUS: a database of the geometry of metal sites in proteins," *Journal of Applied Crystallography*, vol. 41, no. 5, pp. 963–968, 2008.

[47] A. Golovin, D. Dimitropoulos, T. Oldfield, A. Rachedi, and K. Henrick, "MSDsite: a database search and retrieval system for the analysis and viewing of bound ligands and active sites," *Proteins: Structure, Function, and Bioinformatics*, vol. 58, no. 1, pp. 190–199, 2005.

[48] J. M. Castagnetto, S. W. Hennessy, V. A. Roberts, E. D. Getzoff, J. A. Tainer, and M. E. Pique, "MDB: the metalloprotein database and browser at the Scripps Research Institute," *Nucleic Acids Research*, vol. 30, no. 1, pp. 379–382, 2002.

[49] S. M. Ireland and A. C. Martin, "ZincBind—the database of zinc binding sites," *Database*, vol. 2019, 2019.

[50] S. Wang, X. Hu, Z. Feng et al., "Recognizing ion ligand binding sites by SMO algorithm," *BMC Molecular and Cell Biology*, vol. 20, no. S3, pp. 53–59, 2019.

[51] R. Yan, X. Wang, Y. Tian, J. Xu, X. Xu, and J. Lin, "Prediction of zinc-binding sites using multiple sequence profiles and machine learning methods," *Molecular omics*, vol. 15, no. 3, pp. 205–215, 2019.

[52] X. Cao, X. Hu, X. Zhang et al., "Identification of metal ion binding sites based on amino acid sequences," *PLoS One*, vol. 12, no. 8, article e0183756, 2017.

[53] X. Hu, K. Wang, and Q. Dong, "Protein ligand-specific binding residue predictions by an ensemble classifier," *BMC Bioinformatics*, vol. 17, no. 1, pp. 1–12, 2016.

[54] J. Qu, S. S. Yin, and H. Wang, "Prediction of metal ion binding sites of transmembrane proteins," *Computational and Mathematical Methods in Medicine*, vol. 2021, 11 pages, 2021.

[55] M. Brylinski and J. Skolnick, "FINDSITE-metal: integrating evolutionary information and machine learning for structure-based metal-binding site prediction at the proteome level," *Proteins: Structure, Function, and Bioinformatics*, vol. 79, no. 3, pp. 735–751, 2011.

[56] L. Liu, X. Hu, Z. Feng, S. Wang, K. Sun, and S. Xu, "Recognizing ion ligand–binding residues by random forest algorithm based on optimized dihedral angle," *Frontiers in Bioengineering and Biotechnology*, vol. 8, p. 493, 2020.

[57] X. Hu, Z. Feng, X. Zhang, L. Liu, and S. Wang, "The identification of metal ion ligand-binding residues by adding the reclassified relative solvent accessibility," *Frontiers in Genetics*, vol. 11, p. 214, 2020.

[58] J. A. Horst and R. Samudrala, "A protein sequence metafunctional signature for calcium binding residue prediction," *Pattern Recognition Letters*, vol. 31, no. 14, pp. 2103–2112, 2010.

[59] W. R. Pearson, "An introduction to sequence similarity ("homology") searching," *Current Protocols in Bioinformatics*, vol. 42, no. 1, 2013.

[60] T. Joshi and D. Xu, "Quantitative assessment of relationship between sequence similarity and function similarity," *BMC Genomics*, vol. 8, no. 1, 2007.

[61] J. Zhang, H. Chai, G. Yang, and Z. Ma, "Prediction of bioluminescent proteins by using sequence-derived features and lineage-specific scheme," *BMC Bioinformatics*, vol. 18, no. 1, p. 294, 2017.

[62] S. F. Altschul, T. L. Madden, A. A. Schäffer et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[63] G. Wang and R. L. Dunbrack Jr., "PISCES: a protein sequence culling server," *Bioinformatics*, vol. 19, no. 12, pp. 1589–1591, 2003.

[64] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, pp. 1658-1659, 2006.

[65] I. Haberal and H. Oğul, "Prediction of protein metal binding sites using deep neural networks," *Molecular Informatics*, vol. 38, no. 7, article 1800169, 2019.

[66] J. Zhang, X. Liang, F. Zhou, B. Li, and Y. Li, "TYLER, a fast method that accurately predicts cyclin-dependent proteins by using computation-based motifs and sequence-derived features," *Mathematical Biosciences and Engineering*, vol. 18, no. 5, pp. 6410–6429, 2021.

[67] L. Homchaudhuri and R. Swaminathan, "Near ultraviolet absorption arising from lysine residues in close proximity: a probe to monitor protein unfolding and aggregation in lysine-rich proteins," *Bulletin of the Chemical Society of Japan*, vol. 77, no. 4, pp. 765–769, 2004.

[68] K. Chen, L. A. Kurgan, and J. Ruan, "Prediction of flexible/rigid regions from protein sequences using k-spaced amino acid pairs," *BMC Structural Biology*, vol. 7, no. 1, pp. 1–13, 2007.

[69] D. Avrahami, Z. Oren, and Y. Shai, "Effect of multiple aliphatic amino acids substitutions on the structure, function, and mode of action of diastereomeric membrane active peptides," *Biochemistry*, vol. 40, no. 42, pp. 12591–12603, 2001.

[70] M. E. Suliman, P. Bárány, J. C. Divino Filho et al., "Influence of nutritional status on plasma and erythrocyte sulphur amino acids, sulph-hydryls, and inorganic sulphate in end-stage renal disease," *Nephrology, Dialysis, Transplantation*, vol. 17, no. 6, pp. 1050–1056, 2002.

[71] S. Scheiner, A. Tapas Kar, and J. Pattanayak, "Comparison of various types of hydrogen bonds involving aromatic amino acids," *Journal of the American Chemical Society*, vol. 124, no. 44, pp. 13257–13264, 2002.

[72] C. Strub, C. Alies, A. Lougarre, C. Ladurantie, J. Czaplicki, and D. Fournier, "Mutation of exposed hydrophobic amino acids to arginine to increase protein stability," *BMC Biochemistry*, vol. 5, no. 1, pp. 9–9, 2004.

[73] T. S. Heard and H. Weiner, "A regional net charge and structural compensation model to explain how negatively charged amino acids can be accepted within a mitochondrial leader sequence," *Journal of Biological Chemistry*, vol. 273, no. 45, pp. 29389–29393, 1998.

[74] S. Kamtekar, J. M. Schiffer, H. Xiong, J. M. Babik, and M. H. Hecht, "ChemInform Abstract: Protein design by binary patterning of polar and nonpolar amino acids," *Science*, vol. 25, no. 10, 1994.

[75] G. H. Goodwin, C. Sanders, and E. W. Johns, "A new group of chromatin-associated proteins with a high content of acidic and basic amino acids," *FEBS Journal*, vol. 38, no. 1, pp. 14–19, 1973.

[76] J. Zhang, Z. Ma, and L. Kurgan, "Comprehensive review and empirical analysis of hallmarks of DNA-, RNA- and protein-binding residues in protein chains," *Briefings in Bioinformatics*, vol. 20, no. 4, pp. 1250–1268, 2019.

[77] Z. Jiang, X. Hu, G. Geriletu, H. Xing, and X. Cao, "Identification of Ca2+-binding residues of a protein from its primary sequence," *Genetics and Molecular Research*, vol. 15, no. 2, 2016.

[78] C. Zheng, M. Wang, K. Takemoto, T. Akutsu, Z. Zhang, and J. Song, "An integrative computational framework based on a two-step random forest algorithm improves prediction of zinc-binding sites in proteins," *PLoS One*, vol. 7, no. 11, article e49716, 2012.

[79] J. Song, C. Li, C. Zheng, J. Revote, Z. Zhang, and G. I. Webb, "MetalExplorer, a bioinformatics tool for the improved prediction of eight types of metal-binding sites using a random forest algorithm with Two- Step feature selection," *Current Bioinformatics*, vol. 12, no. 6, pp. 480–489, 2017.

[80] C.-Q. Xia, X. Pan, and H.-B. Shen, "Protein–ligand binding residue prediction enhancement through hybrid deep heterogeneous learning of sequence and structure data," *Bioinformatics*, vol. 36, no. 10, pp. 3018–3027, 2020.

[81] S. Wang, X. Hu, Z. Feng, L. Liu, K. Sun, and S. Xu, "Recognition of ion ligand binding sites based on amino acid features with the fusion of energy, physicochemical and structural features," *Current Pharmaceutical Design*, vol. 27, no. 8, pp. 1093–1102, 2021.

[82] Y. Liu and I. Bahar, "Sequence evolution correlates with structural dynamics," *Molecular Biology and Evolution*, vol. 29, no. 9, pp. 2253–2263, 2012.

[83] H. Chai and J. Zhang, "Identification of mammalian enzymatic proteins based on sequence-derived features and species-specific scheme," *IEEE Access*, vol. 6, pp. 8452–8458, 2018.

[84] G. Celniker, G. Nimrod, H. Ashkenazy et al., "ConSurf: using evolutionary data to raise testable hypotheses about protein function," *Israel Journal of Chemistry*, vol. 53, no. 3-4, pp. 199–206, 2013.

[85] S. Chowdhury, J. Zhang, and L. Kurgan, "In silico prediction and validation of novel RNA binding proteins and residues in the human proteome," *Proteomics*, vol. 18, no. 21-22, article 1800064, 2018.

[86] J. Zhang, H. Chai, S. Guo, H. Guo, and Y. Li, "High-throughput identification of mammalian secreted proteins using species-specific scheme and application to human proteome," *Molecules*, vol. 23, no. 6, p. 1448, 2018.

[87] A. Srivastava and M. Kumar, "Prediction of zinc binding sites in proteins using sequence derived information," *Journal of Biomolecular Structure and Dynamics*, vol. 36, no. 16, pp. 4413–4423, 2018.

[88] L. J. McGuffin, K. Bryson, and D. T. Jones, "The PSIPRED protein structure prediction server," *Bioinformatics*, vol. 16, no. 4, pp. 404-405, 2000.

[89] J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton, and D. T. Jones, "Prediction and functional analysis of native disorder in proteins from the three kingdoms of life," *Journal of Molecular Biology*, vol. 337, no. 3, pp. 635–645, 2004.

[90] L. Qiao and D. Xie, "MIonSite: ligand-specific prediction of metal ion-binding sites via enhanced AdaBoost algorithm with protein sequence information," *Analytical Biochemistry*, vol. 566, pp. 75–88, 2019.

[91] C. Lu, Z. Liu, E. Zhang, F. He, Z. Ma, and H. Wang, "MPLs-Pred: predicting membrane protein-ligand binding sites using hybrid sequence-based features and ligand-specific models," *International Journal of Molecular Sciences*, vol. 20, no. 13, p. 3120, 2019.

[92] Z. Zhao, Y. Xu, and Y. Zhao, "SXGBsite: prediction of protein–ligand binding sites using sequence information and extreme gradient boosting," *Genes*, vol. 10, no. 12, p. 965, 2019.

[93] H. Li, D. Pi, C. Chen, and H. Li, "A novel prediction method for zinc-binding sites in proteins by an ensemble of SVM and sample-weighted probabilistic neural network," *IEEE Access*, vol. 7, pp. 186147–186157, 2019.

[94] C. Essien, D. Wang, and D. Xu, "Capsule Network for Predicting Zinc Binding Sites in Metalloproteins," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 2337–2341, San Diego, CA, USA, 2019.

[95] Y. Ding, J. Tang, and F. Guo, "Identification of protein–ligand binding sites by sequence information and ensemble classifier," *Journal of Chemical Information and Modeling*, vol. 57, no. 12, pp. 3149–3161, 2017.

[96] S. Kumar, "Prediction of metal ion binding sites in proteins from amino acid sequences by using simplified amino acid alphabets and random forest model," *Genomics & informatics*, vol. 15, no. 4, pp. 162–169, 2017.

[97] İ. Haberal and H. Oğul, "Deepmbs: prediction of protein metal binding-site using deep learning networks," in *2017 Fourth International Conference on Mathematics and Computers in Sciences and in Industry (MCSI)*, pp. 1–25, Corfu, Greece, 2017.

[98] L. Qiao and D. Xie, "Sequence-based protein-Ca^{2+} binding site prediction using SVM classifier ensemble with random under-sampling," in *2017 International Conference on Progress in Informatics and Computing (PIC)*, pp. 86–90, Nanjing, China, 2017.

[99] X. Hu, Q. Dong, J. Yang, and Y. Zhang, "Recognizing metal and acid radical ion-binding sites by integrating ab initio modeling with template-based transferals," *Bioinformatics*, vol. 32, no. 21, pp. 3260–3269, 2016.

[100] D.-J. Yu, J. Hu, Q.-M. Li, Z.-M. Tang, J.-Y. Yang, and H.-B. Shen, "Constructing query-driven dynamic machine learning model with application to protein-ligand binding sites prediction," *IEEE Transactions on Nanobioscience*, vol. 14, no. 1, pp. 45–58, 2015.

[101] D.-J. Yu, J. Hu, J. Yang, H.-B. Shen, J. Tang, and J.-Y. Yang, "Designing template-free predictor for targeting protein-ligand binding sites with classifier ensemble and spatial clustering," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 10, no. 4, pp. 994–1008, 2013.

[102] Y.-Y. Ou, S.-A. Chen, and S.-C. Wu, "ETMB-RBF: discrimination of metal-binding sites in electron transporters based on RBF networks with PSSM profiles and significant amino acid pairs," *PLoS One*, vol. 8, no. 2, article e46572, 2013.

[103] Z. Chen, Y. Wang, Y.-F. Zhai, J. Song, and Z. Zhang, "ZincExplorer: an accurate hybrid method to improve the prediction of zinc-binding sites from protein sequences," *Molecular Bio Systems*, vol. 9, no. 9, pp. 2213–2222, 2013.

[104] H. Nguyen and J. Kleingardner, "Identifying metal binding amino acids based on backbone geometries as a tool for metalloprotein engineering," *Protein Science*, vol. 30, no. 6, pp. 1247–1257, 2021.

[105] S. M. Ireland and A. C. Martin, "Zincbindpredict—prediction of zinc binding sites in proteins," *Molecules*, vol. 26, no. 4, p. 966, 2021.

[106] J. Zhang, Y. Zhang, Y. Li, S. Guo, and G. Yang, "Identification of cancer biomarkers in human body fluids by using enhanced physicochemical-incorporated evolutionary conservation scheme," *Current Topics in Medicinal Chemistry*, vol. 20, no. 21, pp. 1888–1897, 2020.

[107] M. J. Iqbal, I. Faye, B. B. Samir, and A. Md Said, "Efficient feature selection and classification of protein sequence data in bioinformatics," *The Scientific World Journal*, vol. 2014, 12 pages, 2014.

[108] L. Wang, "Feature selection in bioinformatics," in *Independent Component Analyses, Compressive Sampling, Wavelets, Neural Net, Biosystems, and Nanoengineering X*, p. 8401, International Society for Optics and Photonics, 2012.

[109] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[110] K. A. Shastry, H. A. Sanjay, and H. Sanjay, "Machine learning for bioinformatics," in *Statistical Modelling and Machine Learning Principles for Bioinformatics Techniques, Tools, and Applications*, pp. 25–39, Springer, 2020.

[111] D. A. Pisner and D. M. Schnyer, "Support vector machine," in *Machine Learning*, pp. 101–121, Elsevier, 2020.

[112] L. Liu, B. Shen, X. Wang, and X. Wang, "Research on kernel function of support vector machine," in *Advanced Technologies, Embedded and Multimedia for Human-Centric Computing*, pp. 827–834, Springer, 2014.

[113] J. Platt, "Sequential minimal optimization: a fast algorithm for training support vector machines," Microsoft Research Technical Report, 1998.

[114] X. Huang, L. Shi, and J. A. Suykens, "Sequential minimal optimization for SVM with pinball loss," *Neurocomputing*, vol. 149, pp. 1596–1603, 2015.

[115] L. J. Lancashire, C. Lemetre, and G. R. Ball, "An introduction to artificial neural networks in bioinformatics-application to complex microarray and mass spectrometry datasets in cancer studies," *Briefings in bioinformatics*, vol. 10, no. 3, pp. 315–329, 2009.

[116] B.-H. Kim, K. Yu, and P. C. Lee, “Cancer classification of single-cell gene expression data by neural network,” *Bioinformatics*, vol. 36, no. 5, pp. 1360–1366, 2020.

[117] T.-Y. Lee, S.-A. Chen, H.-Y. Hung, and Y.-Y. Ou, “Incorporating distant sequence features and radial basis function networks to identify ubiquitin conjugation sites,” *PLoS One*, vol. 6, no. 3, article e17331, 2011.

[118] Z. Zainuddin and M. Kumar, “Radial basis function neural networks in protein sequence classification,” *Malaysian Journal of Mathematical Sciences*, vol. 2, no. 2, pp. 195–204, 2008.

[119] H. Taud and J. Mas, “Multilayer perceptron (MLP),” in *Geomatic Approaches for Modeling Land Change Scenarios*, pp. 451–455, Springer, 2018.

[120] H. Ramchoun, M. A. J. Idrissi, Y. Ghanou, and M. Ettaouil, “Multilayer perceptron: architecture optimization and training,” *The International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 4, no. 1, pp. 26–30, 2016.

[121] I. Lorencin, N. Anđelić, J. Španjol, and Z. Car, “Using multilayer perceptron with Laplacian edge detector for bladder cancer diagnosis,” *Artificial Intelligence in Medicine*, vol. 102, article 101746, 2020.

[122] J. Zhang, Y. Zhang, and Z. Ma, “In silico prediction of human secretory proteins in plasma based on discrete firefly optimization and application to cancer biomarkers identification,” *Frontiers in Genetics*, vol. 10, p. 542, 2019.

[123] M. Schonlau and R. Y. Zou, “The random forest algorithm for statistical learning,” *The Stata Journal*, vol. 20, no. 1, pp. 3–29, 2020.

[124] Y. Qi, “Random forest for bioinformatics,” in *Ensemble Machine Learning*, pp. 307–323, Springer, 2012.

[125] J. Zhang, H. Chai, B. Gao, G. Yang, and Z. Ma, “HEMEsPred: structure-based ligand-specific heme binding residues prediction by using fast-adaptive ensemble learning scheme,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 15, no. 1, pp. 147–156, 2018.

[126] H. Lu, H. Gao, M. Ye, and X. Wang, “A hybrid ensemble algorithm combining Ada Boost and genetic algorithm for cancer classification with gene expression data,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, 2021.

[127] P. Chen and C. Pan, “Diabetes classification model based on boosting algorithms,” *BMC Bioinformatics*, vol. 19, no. 1, pp. 1–9, 2018.

[128] C. Yan, F.-X. Wu, J. Wang, and G. Duan, “PESM: predicting the essentiality of miRNAs based on gradient boosting machines and sequences,” *BMC Bioinformatics*, vol. 21, no. 1, pp. 1–9, 2020.

[129] R. Rawi, R. Mall, K. Kunji, C.-H. Shen, P. D. Kwong, and G.-Y. Chuang, “PaRSnIP: sequence-based protein solubility prediction using gradient boosting machine,” *Bioinformatics*, vol. 34, no. 7, pp. 1092–1098, 2018.

[130] L. Deng, J. Pan, X. Xu, W. Yang, C. Liu, and H. Liu, “PDRLGB: precise DNA-binding residue prediction using a light gradient boosting machine,” *BMC Bioinformatics*, vol. 19, no. S19, pp. 135–145, 2018.

[131] X. Zhao, J. Zhang, Q. Ning, P. Sun, Z. Ma, and M. Yin, “Identification of protein pupylation sites using bi-profile Bayes feature extraction and ensemble learning,” *Mathematical Problems in Engineering*, vol. 2013, Article ID 283129, 7 pages, 2013.

[132] V. K. Ayyadevara, “Gradient boosting machine,” in *Pro Machine Learning Algorithms*, pp. 117–134, Springer, 2018.

[133] X. Li, S. Li, Y. Wang, S. Zhang, and K.-C. Wong, “Identification of pan-cancer Ras pathway activation with deep learning,” *Briefings in Bioinformatics*, vol. 22, no. 4, p. bbaa 258, 2021.

[134] S. Albawi, T. A. Mohammed, and S. Al-Zawi, “Understanding of a convolutional neural network,” in *2017 International Conference on Engineering and Technology (ICET)*, pp. 1–6, Antalya, Turkey, 2017.

[135] A. Ghosh, A. Sufian, F. Sultana, A. Chakrabarti, and D. De, “Fundamental concepts of convolutional neural network,” in *Recent Trends and Advances in Artificial Intelligence and Internet of Things*, pp. 519–567, Springer, 2020.

[136] Y. Yang, Z. Hou, Z. Ma, X. Li, and K.-C. Wong, “iCircRBP-DHN: identification of circRNA-RBP interaction sites using deep hierarchical network,” *Briefings in Bioinformatics*, vol. 22, no. 4, article bbaa 274, 2021.

[137] Q. Zhang, M. Zhang, T. Chen, Z. Sun, Y. Ma, and B. Yu, “Recent advances in convolutional neural network acceleration,” *Neurocomputing*, vol. 323, pp. 37–51, 2019.

[138] Y. Wang, Y. Yang, Z. Ma, K.-C. Wong, and X. Li, “EDCNN: identification of genome-wide RNA-binding proteins using evolutionary deep convolutional neural network,” *Bioinformatics*, vol. 38, no. 3, pp. 678–686, 2022.

[139] D. Wang, Y. Liang, and D. Xu, “Capsule network for protein post-translational modification site prediction,” *Bioinformatics*, vol. 35, no. 14, pp. 2386–2394, 2019.

[140] X. Li, S. Li, L. Huang, S. Zhang, and K.-c. Wong, “High-throughput single-cell RNA-seq data imputation and characterization with surrogate-assisted automated deep learning,” *Briefings in Bioinformatics*, vol. 23, no. 1, article bbab 368, 2022.

[141] X. Du, Y. Li, Y.-L. Xia et al., “Insights into protein–ligand interactions: mechanisms, models, and methods,” *International Journal of Molecular Sciences*, vol. 17, no. 2, p. 144, 2016.

[142] D. Salha, M. Andaç, and A. Denizli, “Molecular docking of metal ion immobilized ligands to proteins in affinity chromatography,” *Journal of Molecular Recognition*, vol. 34, no. 2, p. e 2875, 2021.

[143] R. D. Smith, A. L. Engdahl, J. B. Dunbar Jr., and H. A. Carlson, “Biophysical limits of protein–ligand binding,” *Journal of Chemical Information and Modeling*, vol. 52, no. 8, pp. 2098–2106, 2012.

[144] S.-Y. Huang and X. Zou, “Advances and challenges in protein-ligand docking,” *International Journal of Molecular Sciences*, vol. 11, no. 8, pp. 3016–3034, 2010.

[145] S.-Y. Huang, S. Z. Grinter, and X. Zou, “Scoring functions and their evaluation methods for protein–ligand docking: recent advances and future directions,” *Physical Chemistry Chemical Physics*, vol. 12, no. 40, pp. 12899–12908, 2010.

[146] V. Kumar, V. P. R. Chichili, X. Tang, and J. Sivaraman, “A novel trans conformation of ligand-free calmodulin,” *PLoS One*, vol. 8, no. 1, article e54834, 2013.

[147] E. A. Permyakov, “Metal Binding Proteins,” *Encyclopedia*, vol. 1, no. 1, pp. 261–292, 2021.

[148] W. He, Z. Liang, M. Teng, and L. Niu, “mFASD: a structure-based algorithm for discriminating different types of metal-binding sites,” *Bioinformatics*, vol. 31, no. 12, pp. 1938–1944, 2015.

[149] W. Zhou, G. W. Tang, and R. B. Altman, "High resolution prediction of calcium-binding sites in 3D protein structures using FEATURE," *Journal of Chemical Information and Modeling*, vol. 55, no. 8, pp. 1663–1672, 2015.

[150] G. Sciortino, E. Garribba, J. Rodriguez-Guerra Pedregal, and J.-D. Maréchal, "Simple coordination geometry descriptors allow to accurately predict metal-binding sites in proteins," *ACS Omega*, vol. 4, no. 2, pp. 3726–3731, 2019.

[151] J.-E. Sánchez-Aparicio, L. Tiessler-Sala, L. Velasco-Carneros, L. Roldán-Martín, G. Sciortino, and J.-D. Maréchal, "Bio met all: identifying metal-binding sites in proteins from backbone preorganization," *Journal of Chemical Information and Modeling*, vol. 61, no. 1, pp. 311–323, 2021.

[152] H. Deng, G. Chen, W. Yang, and J. J. Yang, "Predicting calcium-binding sites in proteins—a graph theory and geometry approach," *Proteins: Structure, Function, and Bioinformatics*, vol. 64, no. 1, pp. 34–42, 2006.

[153] K. Goyal and S. C. Mande, "Exploiting 3D structural templates for detection of metal-binding sites in protein structures," *Proteins: Structure, Function, and Bioinformatics*, vol. 70, no. 4, pp. 1206–1218, 2008.

[154] R. Levy, M. Edelman, and V. Sobolev, "Prediction of 3D metal binding sites from translated gene sequences based on remote-homology templates," *Proteins: Structure, Function, and Bioinformatics*, vol. 76, no. 2, pp. 365–374, 2009.

[155] D. B. Roche, S. J. Tetchner, and L. J. McGuffin, "FunFOLD: an improved automated method for the prediction of ligand binding residues using 3D models of proteins," *BMC Bioinformatics*, vol. 12, no. 1, pp. 1–20, 2011.

[156] D. B. Roche, M. T. Buenavista, and L. J. McGuffin, "FunFOLDQA: a quality assessment tool for protein-ligand binding site residue predictions," *PLoS One*, vol. 7, no. 5, article e38219, 2012.

[157] D. B. Roche, M. T. Buenavista, and L. J. McGuffin, "The FunFOLD2 server for the prediction of protein–ligand interactions," *Nucleic Acids Research*, vol. 41, no. W1, pp. W303–W307, 2013.

[158] H. Li, D. Pi, Y. Wu, and C. Chen, "Integrative method based on linear regression for the prediction of zinc-binding sites in proteins," *IEEE Access*, vol. 5, pp. 14647–14656, 2017.

[159] A. Passerini, C. Andreini, S. Menchetti, A. Rosato, and P. Frasconi, "Predicting zinc binding at the proteome level," *BMC Bioinformatics*, vol. 8, no. 1, pp. 1–13, 2007.

[160] N. Shu, T. Zhou, and S. Hovmöller, "Prediction of zinc-binding sites in proteins from sequence," *Bioinformatics*, vol. 24, no. 6, pp. 775–782, 2008.

[161] H. Li, D. Pi, and C. Chen, "An improved prediction model for zinc-binding sites in proteins based on Bayesian method," *Tehnički vjesnik*, vol. 26, no. 5, pp. 1422–1426, 2019.

[162] M. Ohue, Y. Matsuzaki, T. Shimoda, T. Ishida, and Y. Akiyama, "Highly precise protein-protein interaction prediction based on consensus between template-based and de novo docking methods," *BMC Proceedings*, vol. 7, 2013.

[163] H. Deng, Y. Jia, and Y. Zhang, "Protein structure prediction," *International Journal of Modern Physics B*, vol. 32, no. 18, article 1840009, 2018.

[164] A. Fiser, "Template-based protein structure modeling," *Computational Biology*, vol. 673, pp. 73–94, 2010.

[165] E. Karni, "Foundations of Bayesian theory," *Journal of Economic Theory*, vol. 132, no. 1, pp. 167–188, 2007.