Quarterly Reviews of Biophysics

www.cambridge.org/qrb

Computational methods for binding site prediction on macromolecules

Igor Kozlovskii[1,2,3] and Petr Popov[1,2,3]

[1]Constructor Knowledge Labs, Bremen, Germany; [2]School of Science, Constructor University Bremen gGmbH, Bremen, Germany and [3]Tetra D AG, Schaffhausen, Switzerland

Review

Cite this article: Kozlovskii I and Popov P (2025). Computational methods for binding site prediction on macromolecules. *Quarterly Reviews of Biophysics*, **58**, e12, 1–31
https://doi.org/10.1017/S003358352500006X

Received: 05 November 2024
Revised: 22 February 2025
Accepted: 24 February 2025

Keywords:
Bioinformatics; dynamics; function; nucleic acid structure; protein structure

Corresponding author:
Petr Popov;
Email: ppopov@constructor.university

Abstract

Binding sites are key components of biomolecular structures, such as proteins and RNAs, serving as hubs for interactions with other molecules. Identification of the binding sites in macromolecules is essential for structure-based molecular and drug design. However, experimental methods for binding site identification are resource-intensive and time-consuming. In contrast, computational methods enable large-scale binding site identification, structure flexibility analysis, as well as assessment of intermolecular interactions within the binding sites. In this review, we describe recent advances in binding site identification using machine learning methods; we classify the approaches based on the encoding of the macromolecule information about its sequence, structure, template knowledge, geometry, and energetic characteristics. Importantly, we categorize the methods based on the type of the interacting molecule, namely, small molecules, peptides, and ions. Finally, we describe perspectives, limitations, and challenges of the state-of-the-art methods with an emphasis on deep learning-based approaches. These computational approaches aim to advance drug discovery by expanding the druggable genome through the identification of novel binding sites in pharmacological targets and facilitating structure-based hit identification and lead optimization.

Table of contents

Introduction

Proteins are essential for many cellular functions, including enzymatic activity, structural support, transport, and cell signaling (Alberts, 2017). Structurally, they are large macromolecules composed of long chains of amino acids, which fold into unique three-dimensional shapes specific to each protein (Rodwell *et al.*, 2018). Their functional roles are driven by local intermolecular interactions within specific regions called binding sites. Binding sites play a crucial role in drug discovery. They serve as 'hot spots' on pharmacological targets where designed drug-like molecules bind. Identifying novel binding sites expands the 'druggable genome', offering new strategies for therapeutic development and drug discovery (Hopkins and Groom, 2002). Drug-like molecules typically target either the orthosteric binding site, where proteins interact with their natural ligands, or distinct allosteric binding sites, which have garnered special interest. Allosteric sites show higher sequence variability between protein subtypes, enabling the design of more selective drug-like molecules compared to those targeting orthosteric binding sites (Changeux, 2013; Wagner *et al.*, 2016; Lu *et al.*, 2018). Furthermore, the binding sites can be formed by several protein molecules at their interaction interface (Ferré *et al.*, 2014; Wang *et al.*, 2018a), opening another opportunity for proximity-induced drug discovery (Békés *et al.*, 2022; Dewey *et al.*, 2023; Liu and Ciulli, 2023; Tan *et al.*, 2024). While proteins are the most common pharmacological targets, nucleic acids, particularly RNAs, are gaining increasing interest in structure-based drug design (Chen *et al.*, 2024; Tong *et al.*, 2024). RNA plays a vital role in gene regulation and information transfer, making it an appealing target for drug development (Warner *et al.*, 2018). Like proteins, RNA molecules are highly structured and contain binding sites that can be modulated by small molecules (Yu *et al.*, 2020). Both proteins and nucleic acids are flexible macromolecules, adopting multiple conformations throughout their life cycle. Accordingly, binding sites are dynamic properties, influenced by the conformational changes of macromolecules (Laskowski *et al.*, 2009; Changeux and Christopoulos, 2016). A single structure of a macromolecule represents only a single point of the complete conformational space. Therefore, it is possible to overlook binding sites in the static structures (Di Pietro *et al.*, 2017; Sun *et al.*, 2020). A remarkable progress has been made in developing experimental methods for identifying binding sites, including fragment screening, site-directed tethering (Hardy and Wells, 2004; Ludlow *et al.*, 2015), antibody-based techniques (Lawson, 2012), small molecule microarrays (Doyle *et al.*, 2016), hydrogen-deuterium exchange (Chalmers *et al.*, 2006), and site-directed mutagenesis (Gelis *et al.*, 2012). However, experimental methods are often resource-intensive and may yield negative results. In contrast, computational approaches enable large-scale identification of binding sites, exploration of macromolecular flexibility, and the ability to assess how well chemical compounds fit into these sites.

While there are several articles describing binding site prediction methods (Laurie and Jackson, 2006; Henrich *et al.*, 2010; Leis *et al.*, 2010; Chen *et al.*, 2011; Roche *et al.*, 2015; Simões *et al.*, 2017; Zhao *et al.*, 2020; Liao *et al.*, 2022; Utgés and Barton, 2024), we found several gaps persisting in the literature. Specifically, most of the works focus only on protein–small molecule interactions, neglecting other important binding site interaction types, such as protein–peptide, nucleic acid–small molecules, or protein–ion. Furthermore, while deep learning-based approaches have gained popularity, there has been a limited discussion on their limitations, applicability, and interpretability compared to traditional methods. In this study, we provide a comprehensive review of computational methods for the prediction of binding sites. We eliminate existing gaps in the literature with a unified overview of computational techniques across diverse binding site interaction types. The computational methods are classified based on the types of binding sites they predict: (i) protein–small molecule binding sites (Section Protein–small molecule binding sites); (ii) protein–peptide binding sites (Section Protein–peptide binding sites); (iii) nucleic acid–small molecule binding sites (Section Nucleic acid–small molecule binding sites); and (iv) protein–ion binding sites (Section Protein–ion binding site prediction). For each type of binding site, the corresponding methods are further divided into categories based on the macromolecule input representation (sequence or structure) and algorithm type (template-based, geometric, energetic, machine learning-based, and deep learning-based). If available, we also list the benchmarks and the performance metrics of different methods for each category of binding sites. Finally, we conclude the review with a discussion of current challenges and future perspectives in the field.

Protein–small molecule binding sites

In this section, we describe computational methods for the prediction of small molecule binding sites on proteins. Small molecules are usually defined as molecules with a mass $\leq$ 500 Da (Benet *et al.*, 2016) designed to interact with biological targets to modulate their functions (Southey and Brunavs, 2023). A binding site usually has specific geometry and physicochemical properties, making the corresponding protein region distinguishable from the rest of the protein surface. Thus, methods for predicting binding sites in proteins aim to identify such regions based on the input information (e.g., sequence, structure, or other type of information). This section is organized as follows: first, we divided methods into two large groups based on the representation of input protein as sequence or structure. After that, we further split all structure-based methods into five categories based on the type of algorithm they use: template-based, geometric, energetic, machine learning-based, and deep learning-based methods. Table 1 provides a list of computational method for prediction of small molecule binding sites on proteins along with type of approach they use.

Sequence-based

Sequence-based methods utilize sequence or sequence-driven information about proteins. In this problem setup, the model takes the protein sequence as input and outputs a score for each position in this sequence , indicating whether the amino acid residue at this position interacts with the ligand or not. Figure 1 provides a schematic representation of the overall pipeline of sequence-based methods. Some methods search for similar sequences in a template database of sequences with known binders and map binding information from them (Kauffman and Karypis, 2009; López *et al.*, 2007; Yang *et al.*, 2013). The idea is supported by the study suggesting, that in most cases the function of the unknown protein can be identified from its sequence or structure by homology (Yao *et al.*, 2003). However, for the correct work of template-based methods, there should be proteins with significant sequence identity to a query sequence in a database (Devos and Valencia, 2000; Wilson *et al.*, 2000), as it was shown that proteins with <35–40% identity may not share the same biochemical function (Todd *et al.*, 2001).

https://doi.org/10.1017/S003358352500006X

The majority of sequence-based methods generate descriptors for each position in the input sequence. The feature vector can be composed of multiple different descriptors: evolutionary information comprised of a position-specific scoring matrix (PSSM) or conservation score; tabular physicochemical properties of amino acid residues on specified position: hydrophobicity, polarity, solvation potential, residue interface propensities, net charge, average accessible surface area (ASA), values from the AAindex (Kawashima *et al.*, 2007) database, and others. Note, that one can also utilize structural features predicted from sequence using other tools (including ML ones) to generate more sophisticated feature vectors; for example, solvent accessible solvent area (SASA)(Dor and Zhou, 2007; Garg *et al.*, 2005; Yuan and Huang, 2004; Ahmad *et al.*, 2003; Adamczak *et al.*, 2004; Heffernan *et al.*, 2015), secondary structure information (Faraggi *et al.*, 2012; Yaseen and Li, 2014; Lin *et al.*, 2005; Bondugula and Xu, 2007; Cheng *et al.*, 2007; Pei and Grishin, 2004), dihedral angles (Wood and Hirst, 2005; Dor and Zhou, 2007; Xue *et al.*, 2008; Heffernan *et al.*, 2015; Lyons *et al.*, 2014), etc. The feature vectors are then used as inputs into the machine learning algorithm. In some methods, feature vectors from several consecutive amino acid residues (typically between 7 and 25 residues) are processed together to form a new feature vector (Chen *et al.*, 2014, 2015). One can feed the feature vectors to a classical classification ML model, such as support vector machine (SVM) (Cortes and Vapnik, 1995) or random forest (RF) (Ho, 1995), which outputs probability scores for the amino acid residues to interact with the ligand (Kauffman and Karypis, 2009; Chen *et al.*, 2014, 2015; Yu *et al.*, 2015; Lu *et al.*, 2019). Other methods are based on larger DL models, such as 1D-CNN, GRU, or LSTM, feeding the whole sequence at once (Cui *et al.*, 2019; Lee and Nam, 2022) and also predicting a probability score for each position. Recently, large pre-trained language models have advanced many tasks in the field of natural language processing (NLP). Protein sequences can be viewed as a 'sentence' with amino acid residues as 'words', and approaches similar to ones from NLP can be applied to them. This idea brought the development of several transformer-based (Vaswani *et al.*, 2017) models, such as ESM (Lin *et al.*, 2023), ProtTrans (Elnaggar *et al.*, 2021) or ProteinBert (Brandes *et al.*, 2022). Most of these models utilize BERT-like (Devlin *et al.*, 2018) architectures and were trained on huge databases in a self-supervised manner for the prediction of masked tokens in sequence. It was shown that such protein language models (PLM) can capture structural information, such as secondary structure or residue-residue contacts (Rives *et al.*, 2021). The sequence or amino acid residue embeddings derived from these models can be used as feature vectors in ML or DL models to predict different types of binding sites (Li *et al.*, 2023d).

Template-based

The template-based methods operate with a database of protein complexes with known binding sites (Figure 2). Then, for the query protein, they search for similar proteins and retrieve information about binding sites from them.

Some methods rely on a global comparison of a query structure against the template structures, then superimpose known ligands or positions of binding residue position from the identified similar templates (Brylinski and Skolnick, 2008; Wass *et al.*, 2010; Yang *et al.*, 2013; Gao *et al.*, 2016). However, methods that rely on global comparison can miss non-conserved binding sites, as some proteins bind the same molecule at sites with different amino acid patterns (Moodie *et al.*, 1996; Denessiouk *et al.*, 2001). Other template-based

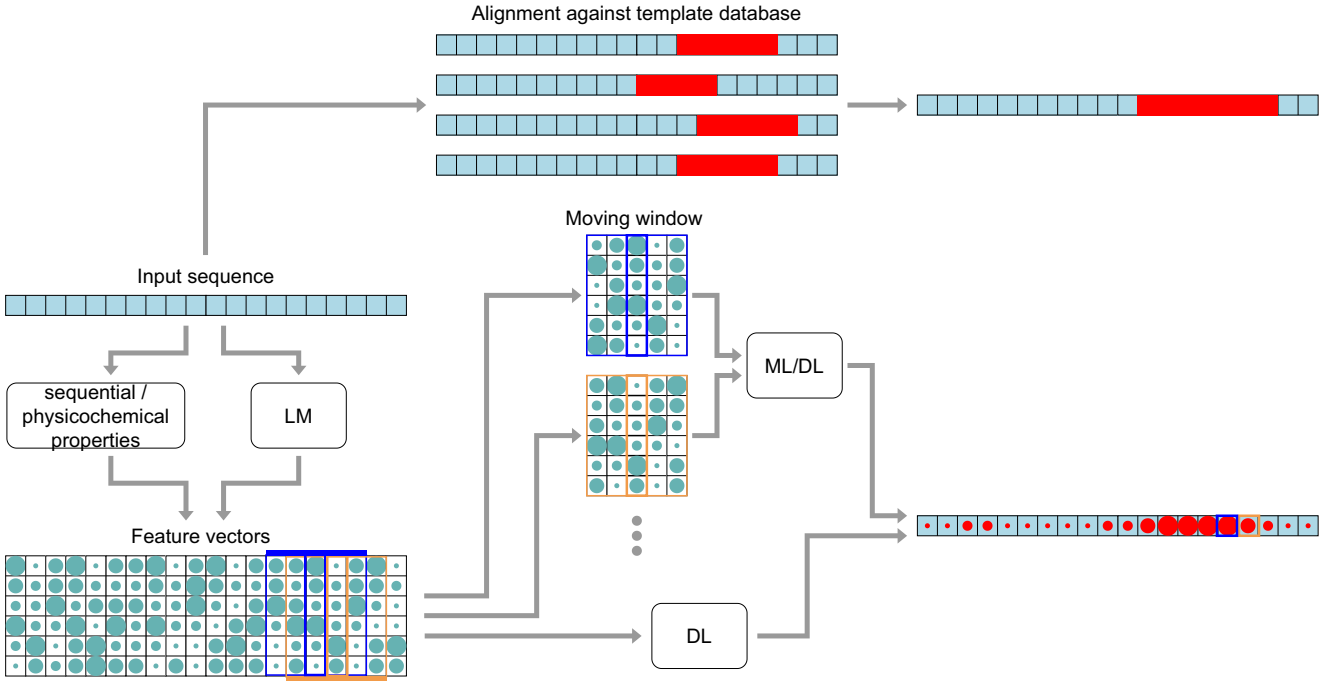

Figure 1. Schematic presentation of the sequence-based methods. The top part demonstrates the pipeline for a template-based approach: the target sequence is aligned against a database of template sequences with known binding residues, and the output binding residues are defined by the consensus score from the alignment. The bottom part demonstrates the pipeline for ML or DL methods. First, the feature vectors (e.g., sequence or physicochemical properties) or the embeddings (e.g., using language models) are calculated. Then, a method uses a moving window across the sequence and feeds feature vectors for each position into an ML or DL model outputting a binding score for each position, or utilizing a larger DL model to get binding scores for each position simultaneously.

https://doi.org/10.1017/S003358352500006X

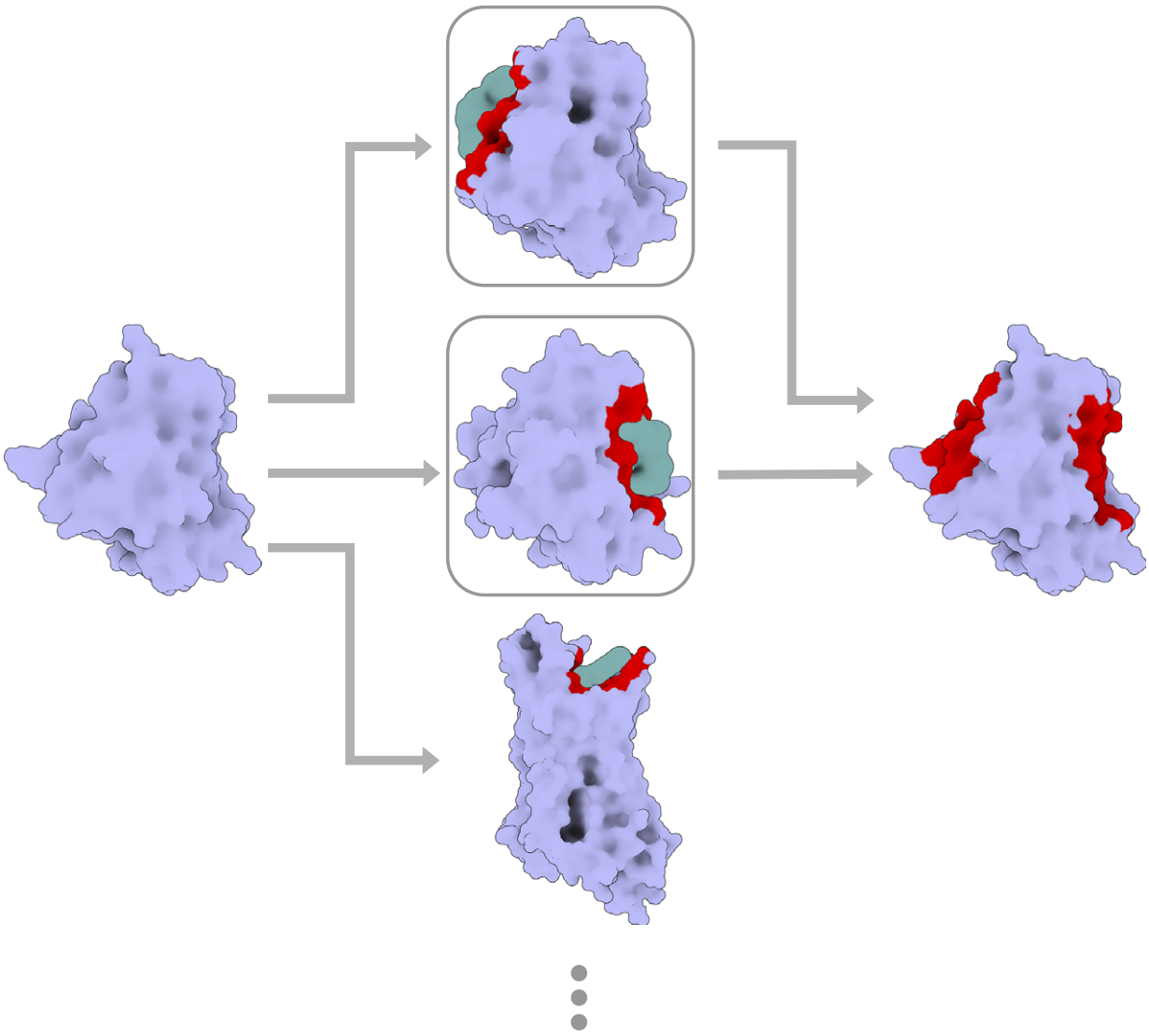

Figure 2. Schematic presentation of the structure template-based methods. In the first stage, the target is screened against a database of template structures with known binding sites. In the second stage, the output prediction is obtained based on the most similar template structures with respect to the target.

methods incorporate more complicated local comparisons of sub-structures or surface patches (Figure 2). For example, one uses geometric hashing to compare two sets of graph vertices representing the query and template protein structures. These vertices can be centers of 3D cells (Rinaldis *et al.*, 1998), surface residues (Schmitt *et al.*, 2002), surface vertices (Rosen *et al.*, 1998), surface patches (Shulman-Peleg *et al.*, 2004), conserved residues (Roy *et al.*, 2012), or all atoms (Barker and Thornton, 2003; Gold and Jackson, 2006). Some methods utilize the maximum clique detection method (Bron and Kerbosch, 1973) to compare the two sets of residues (Lee and Im, 2013; Viet Hung *et al.*, 2015; Konc and Janežič, 2010) or surfaces (Kinoshita and Nakamura, 2005). Other use sub-graph isomorphism (Ullmann, 1976) for the comparison of two sets of pseudo-atoms representing residue side chains (Spriggs *et al.*, 2003) or calculate root-mean-square-deviation (RMSD) between spatially neighboring sets of residues in the query and the template (Stark *et al.*, 2003). There are also many approaches that use geometric methods to identify pockets in the query protein structure, and then provide a method to compare two binding sites. The comparison methods can be divided into alignment-based or alignment-free. The alignment-based methods calculate alignment for each pair of binding sites to estimate their similarity and are usually computationally demanding. On the other hand, alignment-free methods calculate translation- and rotation-invariant descriptors, which can be compared relatively fast. These methods are much faster than alignment-based approaches, but their results may be difficult to interpret (we refer the reader to this review on binding site comparison methods (Eguida and Rognan, 2022)). Nonetheless, the template-based methods in general have higher interpretability, compared to the ML ones. However, the template-based methods are resource-consuming, as for each query protein one needs to screen the entire database, and the screening time increases as the database grows. They also strongly depend on the database itself – if the database lacks certain type of binding site, the method will not be able to identify such a binding site in a query.

Geometric

Geometric methods identify pockets from the protein shape by analyzing occupancy grids, surfaces, or probes, such as spheres

https://doi.org/10.1017/S003358352500006X

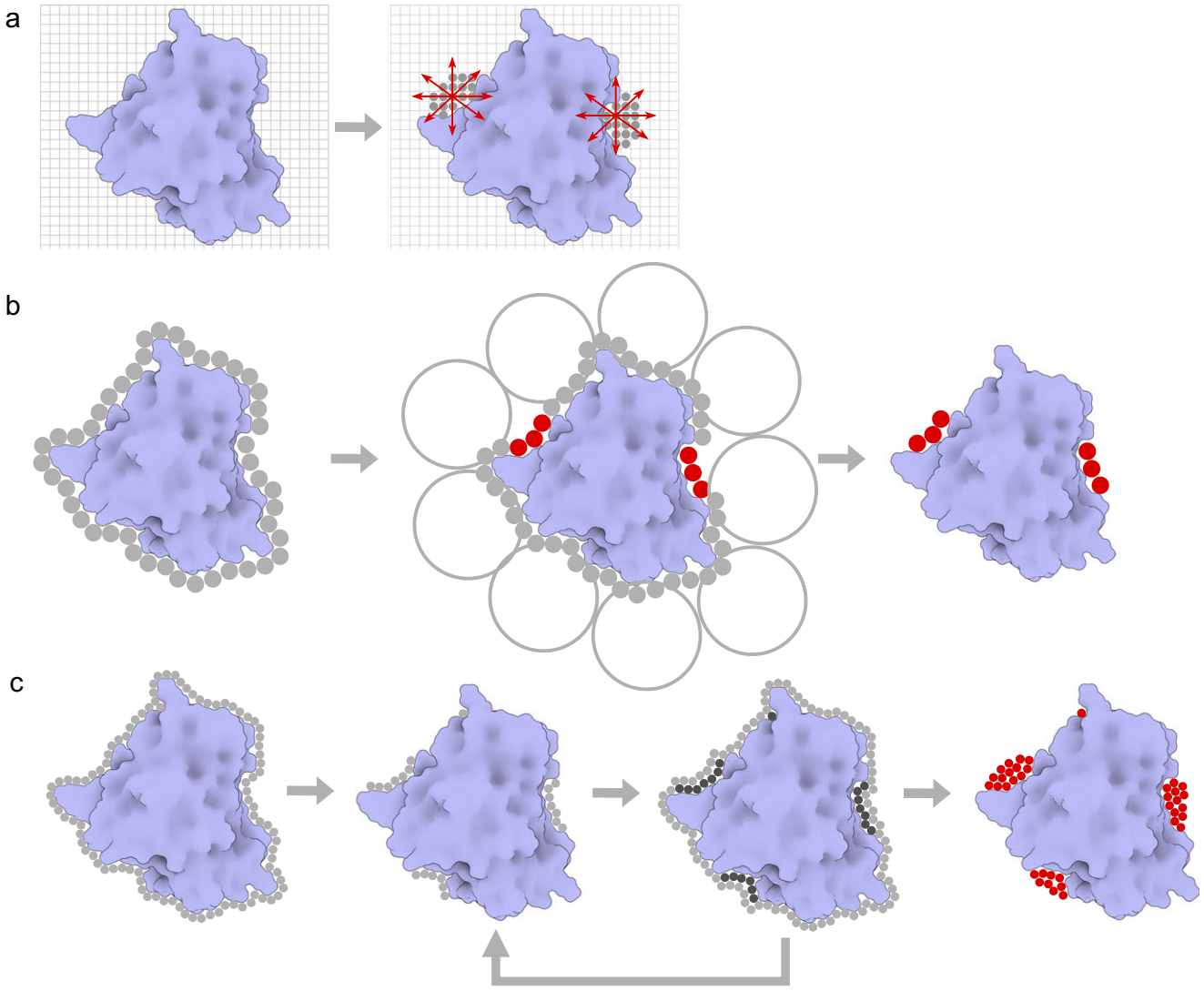

Figure 3. Schematic overview of geometric methods for binding site detection. ***(a)*** Generation of occupancy grid and calculation of the fraction of directions enclosed by the target macromolecule for each empty grid point (used, for example, in POCKET (Levitt and Banaszak, 1992), LIGSITE (Hendlich *et al*., 1997), PocketPocker (Weisel *et al*., 2007), SiteMap (Halgren, 2009), CAVIAR (Marchand *et al*., 2021)). ***(b)*** Rolling of spheres with two different radii around the target macromolecule. The spheres with a larger radius remove the smaller ones. The remaining small spheres are clustered to get final predictions (used, for example, in APROPOS (Peters *et al*., 1996), PHECOM (Kawabata and Go, 2007), (Masuya and Doi, 1995), GHECOM (Kawabata, 2010), and POCASA (Yu *et al*., 2010)). ***(c)*** The addition-removal algorithm, is used in Delaney (1992), Kleywegt and Jones (1994), and Brady and Stouten (2000). Each step consists of adding and removing the surface-exposed points until the convergence. The target macromolecule is represented with a lilac surface, and grid points and probe spheres are shown with circles.

placed around the protein. Figure 3 demonstrates a schematic overview of geometric methods for binding site detection.

SurfNet (Laskowski, 1995) is one of the first geometric algorithms; it generates an occupancy grid for protein atoms and outlines the surface around the occupied voxels to determine the cavities as the binding sites. Many other methods are based on a very similar approach, which generates an occupancy grid and, for each empty grid point, calculates the fraction of directions that are enclosed by protein atoms or surfaces (Levitt and Banaszak, 1992; Hendlich *et al*., 1997; Huang and Schroeder, 2006; Weisel *et al*., 2007; Halgren, 2009; Marchand *et al*., 2021) (see Figure 3*a*). POCKET (Levitt and Banaszak, 1992) casts rays along three directions and determines, if a ray goes through protein–empty points–and then again protein; in this case, the point is considered to be in the pocket. Similarly, LIGSITE (Hendlich *et al*., 1997) casts rays along seven directions (diagonals added), and if the number of intersections is higher than a threshold, the point is a pocket point. PocketPicker (Weisel *et al*., 2007) calculates buriedness for each grid point, scans into 30 directions with rays of length 10Å and width 0.9Å, and counts the number of intersections. SiteMap (Halgren, 2009) calculates the fraction of 110 rays striking the receptor within 8Å distance. SiteMap also calculates multiple descriptors to evaluate the druggability of the detected pocket. Similarly, CAVIAR (Marchand *et al*., 2021) casts rays in 14 directions, selects relevant grid points surrounded by protein atoms, and clusters the grid points forming a binding site. Another common approach involves generating two representations of the protein, corresponding to the spheres probes with two different radii placed around the protein (Peters *et al*., 1996; Kawabata and Go, 2007). More specifically, APROPOS (Peters *et al*., 1996) creates a Delaunay representation of a protein and, then, rolls spheres of two different radii over the structure to remove some of the sides. Shapes removed by small spheres and not removed by large ones are considered pockets (see Figure 3*b*). Similarly, PHECOM

https://doi.org/10.1017/S003358352500006X

(Kawabata and Go, 2007) and (Masuya and Doi, 1995) roll spheres around protein atoms, and GHECOM (Kawabata, 2010) and POCASA (Yu *et al.*, 2010) place spheres on a 3D grid. Small spheres are removed, if they intersect with large ones; after that, clusters of small spheres are considered as pockets. In a similar approach (Kim *et al.*, 2008), one generates inner and outer surface meshes through Voronoi diagrams with different probe radii; the pocket is then defined as the cavity between inner and outer meshes. Note, that multiple methods (Delaney, 1992; Kleywegt and Jones, 1994; Brady and Stouten, 2000) comprise an addition-removal iterative process, where at each step a buffer around the protein is added and some of the points are removed again until the pocket is identified or there is no change after the iteration (see Figure 3*c*). One can use grid-based approaches, where the flood fill algorithm is performed after each addition until the pocket points become enclosed and cannot be removed (Delaney, 1992; Kleywegt and Jones, 1994). DoGSite (Volkamer *et al.*, 2010) generates a 3D occupancy grid for the protein, then repeatedly applies Gaussian filters to remove points with values exceeding a specified threshold; the remaining grid points are clustered to form the binding site. Other heuristics can also be applied; for example, LISE (Xie and Hwang, 2012) generates sets of triangle motifs with assigned scores from protein atoms. Then, the method generates a 3D grid around the protein, and for each empty point, the sum of scores from triangles whose centers lie inside the voxel is assigned. In the next step, for each empty point, the score is recalculated as a sum of scores of all empty points within a sphere of radius 11Å. Finally, top-score points are selected as final pocket centers. Another example is PASS (Brady and Stouten, 2000), which adds spheres around triplets of protein atoms and filters them until no spheres can be added. There are many other types of geometric approaches that treat protein structures as 3D object and apply geometry-based algorithms to detect binding sites. CAST (Binkowski *et al.*, 2003; Liang *et al.*, 1998) creates a Delaunay representation of a protein and then applies a flow theory to determine pockets. In Del Carpio *et al.* (1993), the authors utilized an iterative process in which they first calculate a protein center, identify the closest surface atom, and flag all surface atoms in sight from this atom. Further, the next closest unflagged surface atom is selected and the process repeats until all atoms are flagged. In Coleman and Sharp (2006), the method calculates the surface and the convex hull, which is defined as the smallest convex polyhedron containing all the surface points. For surface points, it calculates 'travel depth', as the minimal distance from a point to the convex hull, and determines pockets as points with higher 'travel depth'. In Bock *et al.* (2007), the method generates a protein surface, and surface points calculates 'spin-images' from classical computer vision algorithms, from which the largest spheres that can be placed on a particular surface point without intersection with other surfaces are defined. Then the method clusters large spheres and outputs them as predicted pockets. MSPocket (Zhu and Pisabarro, 2011) generates a surface and converts it to a graph, where two surface points are considered adjacent if moving these points along their normals makes them closer to each other. Then this graph is pruned, and surface points in the left subgraphs represent final pockets. CurPocket (Liu *et al.*, 2020b) generates a solvent-accessible surface for a protein, calculates curvature at each point, and then clusters points with high curvature. In Xie and Bourne (2007), the method constructs a Delaunay tessellation of protein Cα atoms. Then it removes too long edges and determines edges for protein boundaries. And, finally, it calculates surface directions and geometric potentials, from which the binding site is predicted. Fpocket (Le Guilloux *et al.*, 2009) is the most widely used geometric method for binding pocket detection. It operates via alpha spheres. An alpha sphere is a sphere that contacts with four atoms and does not contain atoms inside. Intuitively, small spheres should lie inside the protein, large spheres are outside, and cavities should correspond to spheres of intermediate radii. So, the algorithm consists of the following steps: (i) detection of alpha spheres via Voronoi tessellation; (ii) filtering out too small and too large spheres; (iii) clustering alpha spheres; and (iv) calculation of additional descriptors and pocket re-ranking. It is worth mentioning, that there are other geometric methods that rely on the previously mentioned assumption, that residues in protein functional sites are more conserved. These methods map conservation scores of residues onto surface points of respective residues, and cluster the most conserved points in space to get binding sites (Glaser *et al.*, 2006; Pupko *et al.*, 2002; Armon *et al.*, 2001; Nimrod *et al.*, 2008; Panchenko *et al.*, 2004; Capra *et al.*, 2009).

Geometry-based methods are usually faster than other methods, but they often have lower accuracy due to the lack of information about the physicochemical and energetic properties of a protein structure.

Energetic

Most energetic methods operate with atom probes placed in a 3D grid around the protein and determine low-energy clusters (see Figure 4). In Goodford (1985), the authors proposed the first probe-based method, searching for energetically favorable positions on 3D maps for three types of probes: water probe, amino group NH_3^+, and methyl group CH_3. For this, they used three-term energy functions including Lennard-Jones, electrostatic, and hydrogen-bond potentials. In Ruppert *et al.* (1997), the authors used another three types of probes (hydrogen atom for hydrophobic interaction, NH for hydrogen bond donor, and C=O for hydrogen bond acceptor probe) to obtain clusters of the lowest-energy points on the protein surface. DrugSite (An *et al.*, 2004), Q-SiteFinder (Laurie and Jackson, 2005), and PocketFinder (An *et al.*, 2005) identify binding pockets via calculation of potential energy maps with aliphatic carbon probe using Lennard-Jones potential with parameters from ECEPP/3 (Nemethy *et al.*, 1992) or GRID (Wade *et al.*, 1993) force fields. SiteHound (Ghersi and Sanchez, 2009; Hernandez *et al.*, 2009) creates maps of potential energies for six probes (methyl, phosphate oxygen, hydroxyl oxygen, peptide nitrogen, water, and carbon), where potential energy is calculated as a sum of van der Waals and electrostatic interactions with parameters from GROMOS (Van Der Spoel *et al.*, 2005) force field. FTSite (Ngan *et al.*, 2012; Brenke *et al.*, 2009) places 16 small molecular probes (ethanol, isopropanol, isobutanol, acetone, acetaldehyde, dimethyl ether, cyclohexane, ethane, acetonitrile, urea, methylamine, phenol, benzaldehyde, benzene, acetamide, and N,N-dimethylformamide) on a dense grid around the protein, optimizes their positions with extended energy expression using CHARMM force field (Brooks *et al.*, 1983) and obtains low-energy clusters of probes. AutoSite (Ravindranath and Sanner, 2016) calculates affinity maps for hydrophobic (carbon) and hydrophilic (oxygen, hydrogen) probes using the van der Waals interaction term from AutoDock energy function (Morris *et al.*, 2009; Huey *et al.*, 2007). SuperStar (Verdonk *et al.*, 2001) uses a slightly modified approach. This method places four different probes, NH_3^+ nitrogen atom, carbonyl oxygen atom, hydroxyl oxygen atom, and methyl carbon atom, into the grid, and converts these maps according to distributions of densities observed in a database of crystallographic structures. In Tsujikawa *et al.* (2016), the atom probe approach in addition takes into account the conservation of amino acid residues.

https://doi.org/10.1017/S003358352500006X

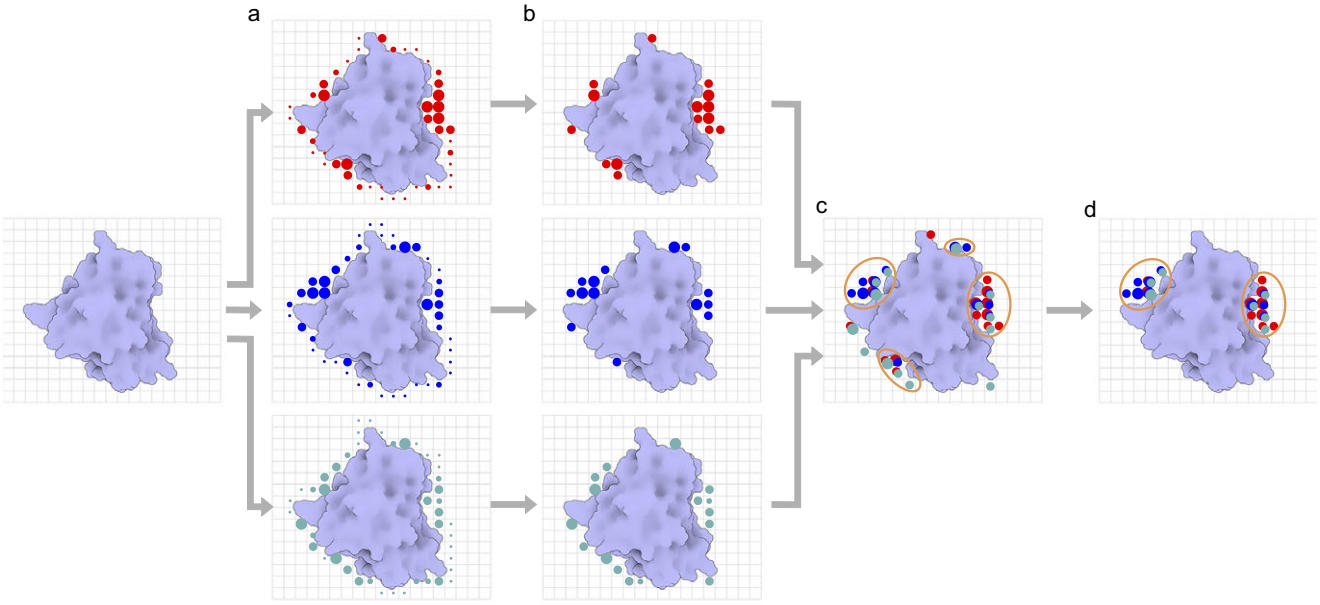

Figure 4. Schematic presentation of the energy probe-based methods. **(*a*)** Different probes (shown as red, blue, and green circles) are placed on a 3D grid around the target macromolecule (shown as a lilac surface) and their interaction energies with the target's atoms are calculated. **(*b*)** The probes corresponding to the high-energy values are filtered out. **(*c*)** The remaining probes are clustered. **(*d*)** The filtering procedure is applied to remove non-relevant clusters.

The method places a carbon atom probe, calculates van der Waals energies for them, and, then, weights interaction energy by conservation scores of nearby amino acid residues. Another class of energy-based methods runs MD simulations and retrieves information about binders from trajectory analysis. OMD (Bhinge *et al.*, 2004) runs short MD simulations of protein in water and then determines binding pockets as volumes, where the RMSD of solvent molecules within the trajectory is low. SILCS (Faller *et al.*, 2015) runs an MD simulation of protein with water solvent and multiple small molecule fragments. It defines what protein regions are more likely to be occupied by which small molecule types. PlayMolecule CrypticScout (Martinez-Rosell *et al.*, 2020) runs mixed-solvent MD with benzene molecules and defines binding hotspots as regions with high occupancy of benzene molecules or regions with low RMSD for these molecules within the trajectory. Another approach is utilized in MDPA (Gu *et al.*, 2022). It is based on the assumption, that ligand binding occurs in regions with higher conformational dynamics (Ming and Wall, 2006). To calculate the external interaction of proteins with test points, the method treats proteins as elastic network structures and simulates interactions using connected springs.

Generally, energetic methods have higher accuracy than geometric methods and high interpretability. However, they are computationally expensive and may miss some interactions not covered by the existing probe types.

Machine learning-based

Most machine learning-based methods can be described in the following way. Firstly, they calculate feature vectors for amino acid residues in the input protein; then, the method feeds the feature

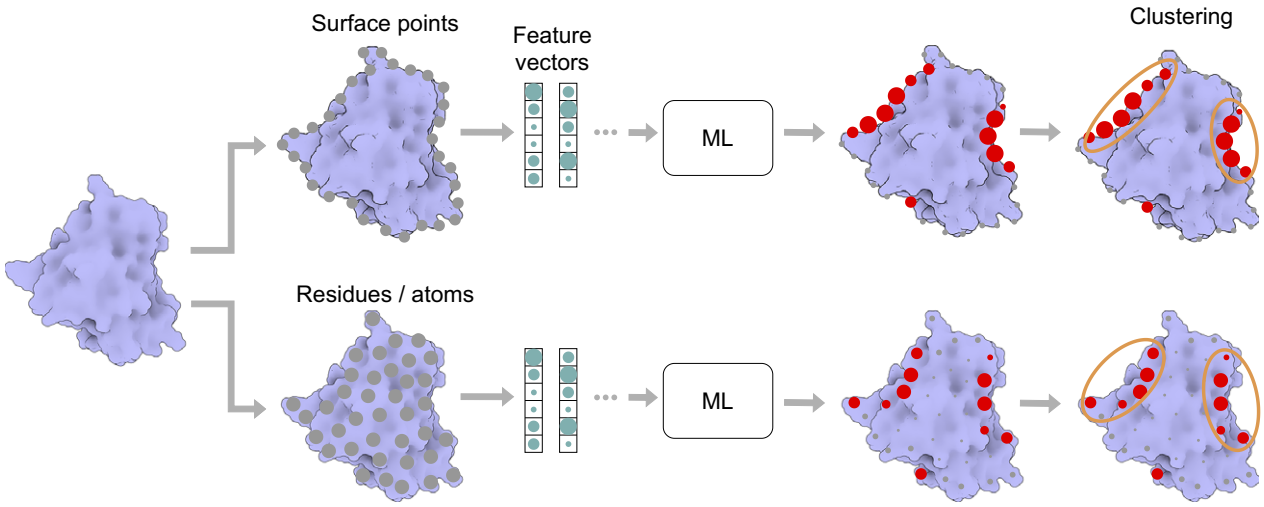

Figure 5. Schematic presentation of the machine learning-based methods. On the top, the target structure is represented as a surface, and feature vectors are calculated for the surface points. On the bottom, feature vectors are calculated for the target's residues or atoms. Then, an ML classifier predicts the binding scores for the points, residues, or atoms, based on the input feature vectors. Finally, the output predictions are filtered by a score threshold and clustered.

https://doi.org/10.1017/S003358352500006X

vectors into an ML classifier, which outputs the probability of a residue being in the binding site. Then, the method spatially clusters high-scoring residues to get the binding site composition (Figure 5). Feature vectors can contain sequential (conservation), physicochemical (electrostatics, hydrogen bonds, solvation energy, hydrophobicity, atom types), geometrical, or structural (solvent accessibility, secondary structure, local geometry) descriptors. In Gutteridge *et al.* (2003), the method identifies catalytic residues in enzymes. It calculates multiple descriptors for each residue: conservation score, relative solvent accessibility, secondary structure, and closeness to a cleft identified by Surfnet (Laskowski, 1995), and residue depth. This feature vector is used as input for a single-layer NN. Petrova and Wu (2006) calculates sequential and structural properties of residues (conservation, flexibility, solvent accessibility, position on the protein surface, hydrogen bonds, secondary structure) and classifies them using an SVM. Tong *et al.* (2009) calculate electrostatic features, geometric properties, and sequence-based conservation for each residue and classify them using a maximum-likelihood algorithm. Qiu and Wang (2011) calculates eight structural properties (solvent accessible surface area, solvation energy, hydrophobicity, depth index, protrusion index, preference, theoretical b-factor) for residues, and uses a Random Forest classifier. ISMBLab-LIG (Jian *et al.*, 2016) first calculates 3D probability density maps that describe interacting atom types around the protein surface using pre-calculated distributions from a database. Then, for each surface atom, the method collects a feature vector of surface local geometry combined with properties retrieved from the described density maps. The method uses a NN model as a classifier. GRaSP (Santana *et al.*, 2020) retrieves a set of physicochemical properties (solvent relative accessibility, atom types, interaction level) from an atomic graph, and then uses an extremely randomized tree to classify residues as binding/non-binding.

Another approach is to classify points or patches on the protein surface instead of residues. Bradford and Westhead (2005) classifies surface points, where for each point a feature vector consists of seven properties: shape index, curvedness, conservation, electrostatic potential, hydrophobicity, residue interface propensity, and solvent accessible surface area. P2Rank (Krivák and Hoksza, 2018) generates a protein surface, and projects features calculated for protein atoms to surface points. Afterward, it predicts the ligandability of each point using an RF classifier and clusters points with high scores. Similarly, SiteFerret (Gagliardi and Rocchia, 2023) first calculates a set of features for surface points, incorporating information about cavities, and then classifies them using the Isolation Forest method. It is also possible to directly classify cavities detected by a geometric method. SCREEN (Nayal and Honig, 2006) first identifies all possible cavities on the protein surface. After that, it calculates a large set of cavity descriptors of different types, such as cavity size, electrostatics, hydrogen bonding, hydrophobicity, polarity, amino acid composition, rigidity, secondary structure, and cavity shape. An RF model classifies cavities as drug-binding/non-drug-binding and selects a smaller fraction of relevant descriptors. FEATURE (Bagley and Altman, 1995, 1996; Wei and Altman, 1998, 2003) represents a set of tools for the prediction of binding sites of different types, such as calcium binding (Wei and Altman, 1998), ATP binding (Wei and Altman, 2003), serine protease active sites (Bagley and Altman, 1996), and others (Liang *et al.*, 2003). It represents microenvironments around a protein as concentric shells with centers placed on a grid and calculates physicochemical properties within these shells. These properties are compared with a set of features for known binding sites and non-binding sites. Then, the Bayesian classifier (Friedman *et al.*, 1997) is used to distinguish binding sites from non-binding ones. There are also consensus-based methods that retrieve predictions from multiple other methods, including geometric, energy-based, or template-based approaches and combine them into final predictions using an ML-based re-scoring. For example, MetaPocket2.0 (Zhang *et al.*, 2011) aggregates results from eight different methods: LIGSITE^{csc} (Huang and Schroeder, 2006), PASS (Brady and Stouten, 2000), Q-SiteFinder (Laurie and Jackson, 2005), Surfnet (Laskowski, 1995), Fpocket (Le Guilloux *et al.*, 2009), GHECOM (Kawabata, 2010), ConCavity (Capra *et al.*, 2009), and POCASA (Yu *et al.*, 2010). Another example is COACH (Yang *et al.*, 2013), which combines prediction results from TM-SITE (Yang *et al.*, 2013), S-SITE (Yang *et al.*, 2013), COFACTOR (Roy *et al.*, 2012), FINDSITE (Brylinski and Skolnick, 2011), and ConCavity (Capra *et al.*, 2009).

ML-based methods strongly rely on the dataset construction and calculated feature vectors, and can produce false positive predictions – that is, identification of 'undruggable' regions (Broomhead and Soliman, 2017). Moreover, even extensive feature engineering does not guarantee capturing all the information relevant to the binding site prediction.

Deep learning-based

The accumulation of large amounts of structural data and advancements in deep learning methods in other fields have led to the development of top-performing methods in structural bioinformatics problems, such as protein structure prediction (Jumper *et al.*, 2021; Baek *et al.*, 2021) DL-based approaches may not require hand-crafted feature engineering and may be capable of capturing relevant structural context by construction. To begin with, these methods typically operate on protein structures represented as a point cloud, a graph, or a 3D density grid. Although a graph is typically a 2D representation of a structure, the 3D information can be encoded as the feature vectors of graph nodes or edges. Therefore, most of the DL-based methods can be classified based on the structure representation, and further split into two classes: (i) where a DL model performs segmentation using the entire graph or grid; and (ii) where a DL model samples sub-graphs or sub-grids and classifies whether their centers correspond to a binding site or not. Figure 6 demonstrates a schematic representation of this idea, and we provide more details about methods in each class below.

We start with methods that sample small 3D voxel grids around a protein structure and classify whether the grid center corresponds to a binding site. DeepSite (Jiménez *et al.*, 2017) was one of the first methods developed for predicting ligand binding sites. It represents a protein structure as a 3D voxel grid with 1Å voxels in size, where each voxel contains eight channels for atoms of different types: hydrophobic, aromatic, hydrogen bond acceptor, hydrogen bond donor, positive ionizable, negative ionizable, metal, and excluded volume. Each channel of a voxel stores the occupancy value of nearby atoms of the respective type, where occupancy is calculated as $n(r) = 1 - exp\left(-(r_{vdw}/r)^{12}\right)$. Then, from the generated 3D voxel grid of a protein, subgrid cubes of size $16 \times 16 \times 16$ voxels are sampled through a sliding window. These cubes are provided as input into a 3D CNN, which outputs a probability score for the cube center being closer than 4Å to the geometric center of a binding site. In Jiang *et al.* (2019a), the authors utilized a similar approach but parameterized the input voxel grid differently. They calculate four pseudo-energy channels instead of using an occupancy-based grid: (i) a shape channel retrieved as output from the LIGSITE (Hendlich *et al.*, 1997) method; (ii) a van der Waals potential energy channel of

https://doi.org/10.1017/S003358352500006X

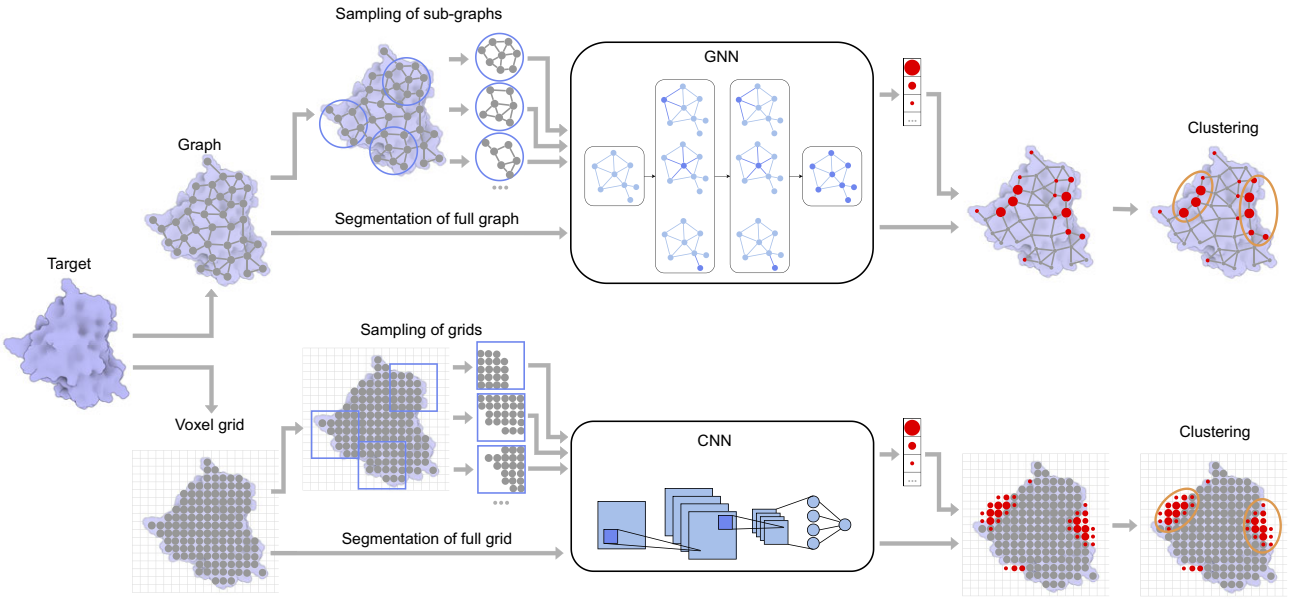

Figure 6. Schematic presentation of the DL-based methods. Most of the methods utilize graph-based or voxel grid representations of the target macromolecular structure. Then, they sample either sub-graphs or sub-grids around the structure and classify their centers as belonging to the binding site or not. Alternatively, they use segmentation models to operate with the full graph or grid.

an -CH3 probe; (iii) a hydrogen bond potential channel using an -OH probe; and (iv) an electric potential energy channel. DeepPocket (Aggarwal *et al.*, 2021) re-scores pockets predicted by Fpocket (Le Guilloux *et al.*, 2009) using a similar 3D CNN model. For this, it uses libmolgrid (Sunseri and Koes, 2020) to obtain a cubic 3D voxel grid of size 23.5Å with a voxel size of 0.5Å around a pocket and passes it into a 3D CNN model that classifies the pocket as ligandable or non-ligandable. After that, the method also generates a segmented representation of ligandable pockets by passing a larger (32Å) cubic grid into another U-Net-like (Ronneberger *et al.*, 2015) 3D CNN. DeepSurf (Mylonas *et al.*, 2021), CAT-Site (Petrovski *et al.*, 2022), and SAPocket (Wang *et al.*, 2023c) instead of using a sliding window, sample points on the protein surface, calculate voxelized representations for cubic grids centered on these points and pass these grids as input into a 3D CNN classification model. FRSite (Jiang *et al.*, 2019b) utilizes the faster R-CNN approach (Ren *et al.*, 2016): it first passes a voxelized representation of a protein into a 3D Region Proposal Network, and further feeds proposals into a 3D CNN for classification. BiteNet (Kozlovskii and Popov 2020) utilizes the YOLO approach for real-time object detection in videos (Redmon *et al.*, 2016). It first obtains a voxelized representation of an input protein and applies a 3D CNN model to it. The model splits the input grid into cells, where each cell contains predicted values for the probabilities of a binding site center being within the cell and the center of a binding site with respect to the cell. Kalasanty (Stepniewska-Dziubinska *et al.*, 2020) was one of the first methods to perform 3D segmentation of binding sites in a single pass. The method represents a protein as a 3D grid of a constant size of 70Å along each direction with a 2Å voxel size and feeds it into a 3D U-Net (Ronneberger *et al.*, 2015). There are multiple follow-up methods of the Kalasanty approach with adjusted 3D CNN model (Kandel *et al.*, 2021; Li *et al.*, 2022; Li *et al.*, 2023b; Nazem *et al.*, 2021; Liu *et al.*, 2023). PUResNet (Kandel *et al.*, 2021) modified the encoder in the U-Net model to ResNet. RefinePocket (Liu *et al.*, 2023) used an attention-enhanced encoder and a mask-guided decoder inside the U-Net. RecurPocket (Li *et al.*, 2022) and GLPocket (Li *et al.*, 2023b) used a recurrent LMSER (Least Mean Square Error Reconstruction) network with gated recurrent refinement. DUNet (Wang *et al.*, 2022b) added a DenseNet (Huang *et al.*, 2017) encoder into the U-Net model. InDeep (Mallet *et al.*, 2022) used a 3D U-Net for the prediction and segmentation of small molecule binding sites occurring on protein–protein interfaces. PointSite (Yan *et al.*, 2022) represents a point cloud-based segmentation approach: it constructs a point cloud from all protein atoms, converts it into a 3D sparse grid, and applies a segmentation model with a U-Net architecture based on submanifold sparse convolutions.

A different approach is to consider the 3D structure of a protein as a graph. Some methods sample points on the protein surface or around the protein on a grid, and analyze the graph constructed from these points, that is predicting whether a point corresponds to a binding site or not. MaSIF (Gainza *et al.*, 2020) pre-calculates physicochemical properties for the whole protein surface. Then, from this surface, the method samples patches represented as a graph with surface points as nodes and with surface point feature vectors as node feature vectors, containing geometric (shape index, distance-dependent curvature) and chemical (hydropathy, continuum electrostatics, free electrons/protons) descriptors with additional values for geodesic coordinates. Then, a GNN with geodesic convolutions is applied to the patch graph with multiple orientations to obtain an embedding vector for the input patch. The authors used this approach for different tasks: prediction of protein–ligand binding sites, prediction of protein–protein binding sites, and fast scanning of protein surfaces for identification of protein–protein binder partners. In dMaSIF (Sverrisson *et al.*, 2021), the authors further improved this approach and made the method fully differentiable without the need for memory- and computation-demanding pre-calculation of surface descriptors. SiteRadar (Evteev *et al.*, 2023) selects points on a 3D grid outside of the protein, then generates a graph with protein atoms around the selected points as the nodes, and uses a GNN to analyze the graphs, predicting whether the point center

https://doi.org/10.1017/S003358352500006X

belongs to a binding site or not. PocketAnchor (Li *et al.*, 2023c) samples a set of 'anchor' points around the protein, representing potentially ligandable positions. For each 'anchor' point, it further generates a graph with protein atoms and surface points within 6Å from the 'anchor' center. Atom and surface point nodes contain a different number of geometric and chemical features. The graphs are processed with MPNNs, outputting binding site scores for each 'anchor' point. Of note, GraphSite (Shi *et al.*, 2022) uses a graph representation of local protein regions and utilizes GNNs to classify ligand binding sites into 14 classes. A higher-level approach is to provide the whole protein graph as input into a graph neural network model for segmentation. For example, GraphBind (Xia *et al.*, 2021) uses residue centroids as node centers and structural and sequential features of residues as node features and passes the graph into a Hierarchical Graph Neural Network, which predicts a score for each residue indicating whether it is on the binding interface with nucleic acid. Similarly, GraphPLBR (Wang *et al.*, 2023d) operates on residues as nodes. FABind (Pei *et al.*, 2023) uses two separate GNNs, one for working with the protein residue graph and another operating on the ligand atomic graph. Embeddings from the two models are combined to make a ligand-specific prediction of binding residues. GrASP (Smith *et al.*, 2023) builds a graph using all protein atoms within 5Å from the surface and uses a GNN with graph attention to classify atoms as binding or non-binding. Similarly, GU-Net (Nazem *et al.*, 2023) uses all protein atoms for the graph and predicts atom scores using a U-Net-like Graph Convolutional Network (Gao and Ji, 2019). EquiPocket (Zhang *et al.*, 2023) builds graphs using both protein atoms and surface points and applies GNNs with E(3)-equivariant convolutions (Satorras *et al.*, 2021) to identify binding atoms. LigBind (Xia *et al.*, 2023) demonstrated another approach: it first pre-trains GNNs for a ligand-general binding residue predictor and a feature extractor for ligand-residue pair embeddings, and then fine-tunes ligand-specific binding residue predictors for more than 1000 ligand types from the BioLip (Yang *et al.*, 2012) database.

Recently, models for protein structure prediction have made significant advances. A breakthrough occurred in the 14th Critical Assessment of Protein Structure Prediction (CASP14) challenge (Kryshtafovych *et al.*, 2021) when AlphaFold2 (Jumper *et al.*, 2021) achieved almost experimental accuracy in the prediction of full-atom protein structures. AlphaFold2 takes an MSA and structural templates as inputs and utilizes a complicated deep-learning model with the newly introduced Evolutionary Transformer. Some approaches for binding site prediction use these methods to generate protein structure models from sequences alone, and retrieve structural features along with sequential ones for each residue (Littmann *et al.*, 2021; Ho *et al.*, 2021; Seo *et al.*, 2024). Furthermore, some models extract a structural graph from a generated protein model, where residues correspond to the graph nodes and the their features correspond to the node features. Then, the composed graph is used as the input for a graph neural network (GNN) (Yuan *et al.*, 2022a; Zhang and Xie, 2023). Previously, it was shown that AlphaFold2 can successfully predict protein–peptide structures (Chang and Perez, 2023; Tsaban *et al.*, 2022), but it was not clear whether AlphaFold2 could be used for the analysis of interactions between proteins and small molecules, as the latter were absent in the training objective for modeling. However, in some cases, AlphaFold2 predicts rotamers as if they were interacting with small molecules, suggesting that it can be used to train a binding site detection model. Moreover, as AlphaFold2 can predict protein–peptide complexes, it can be reasoned that this model can also be useful for the identification of interactions with small molecules, as they or their fragments can resemble amino-acid side chains (Polizzi and DeGrado, 2020). AF2BIND (Gazizov *et al.*, 2023) is constructed as follows: as input to the AlphaFold2 model, it provides a sequence of a target protein, the protein backbone structure as a template, and 20 'bait' amino acids as individual chains, appending them to the sequence with large offsets. The method further uses AlphaFold2 output pairwise representations between target residues and each of the twenty 'bait' amino acids as input into a logistic regression model, predicting whether a target residue is ligand binding or not. The authors demonstrated a correlation between the chemical properties of the small molecule ligands and the 20 'bait' amino acids.

Finally, there are methods combining multiple representations of the protein. PocketMiner (Meller *et al.*, 2023) uses a geometric vector perceptron GNN (GVP-GNN) (Jing *et al.*, 2020) and a 3D CNN for the prediction of putative cryptic pockets. For this, the authors generated 40-ns simulations for 37 proteins and trained the models to predict the positions in each structure where a pocket would open during a short simulation. The authors showed that both GVP-GNN and 3D CNN work equally well.

Benchmarks

Most of the newest ML- and DL-based methods rely on scPDB (Desaphy *et al.*, 2015), PDBbind (Liu *et al.*, 2015), or BioLip (Yang *et al.*, 2012) databases for training and validation. scPDB (Desaphy *et al.*, 2015) is a large database containing ∼16,000 complexes, where each entry is annotated with calculated properties for the ligand, cavity, and interactions. PDBbind (Liu *et al.*, 2015) is a curated database of protein–ligand complexes (∼23,000), with experimentally determined binding affinity. BioLip (Yang *et al.*, 2012; Zhang *et al.*, 2024) contains ∼460,000 structures of proteins or nucleic acids, with a total of approximately ∼890,000 ligands. BioLip includes a wide range of classes of macromolecules and ligands, allowing researchers to construct various training and validation sets. In addition, it incorporates a comprehensive procedure to select relevant ligands and includes cross-references with many other databases (PDBbind (Liu *et al.*, 2015), BindingDB (Gilson *et al.*, 2016), SIFTS (Dana *et al.*, 2019), UniProt (Consortium, 2019), and DrugBank (Knox *et al.*, 2024), etc.). Note that the provided numbers are for 2024; these are likely to increase in future versions of the databases. Finally, other approaches rely on training datasets compiled from PDB, followed by structure refinement, clustering, and filtering of redundant structures.

There are two benchmark sets, COACH420 (Krivák and Hoksza, 2018) and HOLO4K (Schmidtke *et al.*, 2010), which are widely used for the comparison of binding site detection methods. COACH420 (Krivák and Hoksza, 2018) is a dataset of 420 single-chain proteins containing natural compounds and drug-like ligands. It was first created for the evaluation of the P2Rank method (Krivák and Hoksza, 2018) as a subset of a test set from (Roy *et al.*, 2012; Yang *et al.*, 2013), without proteins from the training set of P2Rank. HOLO4K (Schmidtke *et al.*, 2010), in turn, is a large set of protein–ligand complexes. It was initially composed for the validation of the PocketFinder (An *et al.*, 2005) method and later used for a comprehensive large-scale comparison of binding site prediction methods (Schmidtke *et al.*, 2010). Interestingly, originally, it comprised apo complexes; but after the work of (Krivák and Hoksza, 2018), a subset of holo complexes is mainly used. It is

https://doi.org/10.1017/S003358352500006X

important to note that, although the COACH420 and HOLO4K benchmarks are used by many methods, most of them perform additional filtering (e.g., removing irrelevant ligands or addressing data leakage between the training and test sets), resulting in slightly different subsets of COACH420 and HOLO4K. Therefore, a direct comparison of methods based on these benchmarks may not be as straightforward as it may seem. Nonetheless, one may see the performance metrics of different methods in Supplementary Tables S2–S10.

The other benchmarks include: CHEN11 (Chen *et al.*, 2011), B48/U48 (Huang and Schroeder, 2006), B210 (Huang and Schroeder, 2006), DT198 (Zhang *et al.*, 2011), ASTEX (Hartshorn *et al.*, 2007), and CASP (Lopez *et al.*, 2009; Schmidt *et al.*, 2011; Gallo Cassarino *et al.*, 2014). CHEN11 (Chen *et al.*, 2011) is a non-redundant dataset of 251 proteins, where each structure is the most representative structure of a family, with a ligand superimposed from the closest homolog in cases where a ligand is absent in the original structure. B48/U48 (Huang and Schroeder, 2006) is a small dataset of pairs of apo and holo-structures of the same protein. The Astex Diverse set (Hartshorn *et al.*, 2007) is a small benchmark for docking methods, which was used as the binding site detection benchmark (Le Guilloux *et al.*, 2009; Yan *et al.*, 2022). Some of the older classical methods used CASP8 (Lopez *et al.*, 2009), CASP9 (Schmidt *et al.*, 2011), and CASP10 (Gallo Cassarino *et al.*, 2014) benchmarks to evaluate their performance for the prediction of ligand-binding residues. However, these benchmarks are much smaller compared to the other ones described above. Finally, we would like to note that, while for older methods, the most used metrics correspond to binary classification metrics derived from the residue scores, newer methods include metrics based on the distances between the predicted binding site center and the true binding site, as well as the overlap of the predicted and true binding site cavity in the case of binding site segmentation. Supplementary Section Metrics provides more details on commonly used performance metrics for binding site prediction methods.

Protein–peptide binding sites

Protein–protein interactions (PPIs) regulate numerous essential biological pathways, making them a key class of pharmacological targets (Ruffner *et al.*, 2007). There is an increasing need to develop inhibitors of intracellular PPIs to modulate critical biological processes. However, PPIs have long been considered difficult to target (Tsomaia, 2015). On the one hand, large biologics, which are effective in targeting extracellular PPIs, cannot penetrate cell membranes to reach intracellular PPIs. On the other hand, traditional small molecule scaffolds can cross membranes but are often unsuitable for the large, shallow surfaces typical for PPI interfaces (Tsomaia, 2015). PPI interfaces exhibit distinct characteristics, such as larger contact areas ($\sim 1500-3000\text{Å}^2$ for PPI compared to $\sim 300-1000\text{Å}^2$ for protein–small molecule interactions (Smith and Gestwicki, 2012)) and the absence of deep binding pockets usually found in small molecule interactions ($\sim 270\text{Å}^3$ in volume (Buchwald, 2010)). Notably, PPI interfaces often contain smaller binding pockets ($\sim 100\text{Å}^3$ (Fuller *et al.*, 2009)) that play a crucial role in binding affinity (Clackson and Wells, 1995). Peptides and peptide-based molecules occupy a unique position between small molecules (with a molecular weight $< 0.5kDa$) and biologics ($> 150kDa$). They offer a promising therapeutic approach for targeting intracellular PPIs, as they can potentially combine the benefits of biologics, such as low toxicity, high specificity, and strong affinity, with the membrane permeability of small molecules (Tsomaia, 2015). The successful design of therapeutic peptides requires detailed knowledge of the binding sites on their protein targets. Identifying new protein–peptide binding sites could broaden the range of druggable targets, opening up new opportunities for drug discovery. Many methods for protein–peptide binding site prediction utilize approaches similar to the ones described in the previous section, but we still cover these methods here to highlight some specific characteristics. See Table 2 for an extensive list of methods for prediction of protein-peptide binding sites.

Machine learning-based

Multiple sequence-based methods calculate features for each residue (e.g., PSSM, predicted ASA, SS, physicochemical properties, and intrinsic disorder) in an input protein sequence and pass these features as input into a classical ML model (Taherzadeh *et al.*, 2016; Zhao *et al.*, 2018; Iqbal and Hoque, 2018; Shafiee *et al.*, 2022). More advanced approaches tend to rely on additional information. SPRINT-Str (Taherzadeh *et al.*, 2018) and MultiVORFFIP (Segura *et al.*, 2012) calculate structural and physicochemical descriptors for each residue in a target protein and use RF for the binary classification of residues as binding/non-binding. PINUP (Liang *et al.*, 2006) calculates structural and physicochemical descriptors for interface residues, then selects surface patches by choosing a central surface residue and 19 residues nearest to it, and then classifies the patch based on a set of features for the 20 patch residues. P2Rank-Pept (Krivák *et al.*, 2018) calculates geometrical and physicochemical descriptors for protein surface points and classifies these points using RF. PepSite (Trabuco *et al.*, 2012) uses spatial PSSMs for the identification of peptide-binding hot spots on the protein surface. For this, the method estimates the densities of protein atoms around each amino acid type in the peptide and encodes them into a 3D grid. Then, PepSite screens the target protein with these S-PSSM grids and identifies appropriate hot spots. PepBind (Zhao *et al.*, 2018) is a consensus method combining predictions from SVMpep, S-SITE, and TM-SITE.

Deep learning-based

The sequence-based approaches, such as VisualP (Wardah *et al.*, 2020) encode a window around a residue into a 2D image and apply a CNN. MTDSite (Sun *et al.*, 2021) uses a BiLSTM to predict binding residues for DNA, RNA, carbohydrates, and peptides. PepBCL (Wang *et al.*, 2022a) and PepNN-Seq (Abdin *et al.*, 2022) retrieve protein sequence embeddings from the language model ProtTrans (Elnaggar *et al.*, 2021). Similarly to the machine learning-based methods, recent deep learning-based approaches tend to incorporate different types of information into the model. PepCNN (Chandra *et al.*, 2023) represents residues using sequential and structural descriptors, along with embeddings from the ProtT5 (Elnaggar *et al.*, 2021) model, and passes them into a 1D CNN model. PepNN-Struct (Abdin *et al.*, 2022) uses a GNN with attention to extract embeddings from a graph of protein residues and uses multi-head attention to encode a peptide sequence for predicting of binding residues. The authors also demonstrated that pre-training on protein–protein complexes significantly increases the model accuracy in predicting peptide-binding residues. GraphPPepIS (Li *et al.*, 2023a) represents both protein and

https://doi.org/10.1017/S003358352500006X

Table 1. List of methods for prediction of protein–small molecule binding sites

Method name	Year	Representation type	Algorithm type	Reference
Goodford *et al.*	1985	Structure	Energetic	(Goodford *et al.*, 1985)
CavitySearch	1990	Structure	Geometric	(Ho and Marshall, 1990)
POCKET	1992	Structure	Geometric	(Levitt and Banaszak, 1992)
Delaney *et al.*	1992	Structure	Geometric	(Delaney, 1992)
Del Carpio *et al.*	1993	Structure	Geometric	(Del Carpio *et al.*, 1993)
VOIDOO	1994	Structure	Geometric	(Kleywegt and Jones, 1994)
SurfNet	1995	Structure	Geometric	(Laskowski *et al.* 1995)
Masuya *et al.*	1995	Structure	Geometric	(Masuya and Doi, 1995)
FEATURE	1995	Structure	ML-based	(Bagley and Altman, 1995, 1996; Wei and Altman, 1998, 2003; Liang *et al.*, 2003)
APROPOS	1996	Structure	Geometric	(Peters *et al.*, 1996)
LIGSITE	1997	Structure	Geometric	(Hendlich *et al.*, 1997)
Ruppert *et al.*	1997	Structure	Energetic	(Ruppert *et al.*, 1997)
de Rinaldis *et al.*	1998	Structure	Template-based	(de Rinaldis *et al.*, 1998)
Rosen *et al.*	1998	Structure	Template-based	(Rosen *et al.*, 1998)
PASS	2000	Structure	Geometric	(Brady and Stouten, 2000)
ConSurf	2001	Structure	Geometric	(Armon *et al.*, 2001)
SuperStar	2001	Structure	Energetic	(Verdonk *et al.*, 2001)
Schmitt *et al.*	2002	Structure	Template-based	(Schmitt *et al.*, 2002)
Rate4Site	2002	Structure	Geometric	(Pupko *et al.*, 2002)
Jess	2003	Structure	Template-based	(Barker and Thornton, 2003)
Spriggs *et al.*	2003	Structure	Template-based	(Spriggs *et al.*, 2003)
PINTS	2003	Structure	Template-based	(Stark *et al.*, 2003)
CAST	2003	Structure	Geometric	(Liang *et al.*, 1998; Binkowski *et al.*, 2003)
Gutteridge *et al.*	2003	Structure	ML-based	(Gutteridge *et al.*, 2003)
SiteEngine	2004	Structure	Template-based	(Shulman-Peleg *et al.*, 2004)
Panchenko *et al.*	2004	Structure	Geometric	(Panchenko *et al.*, 2004)
DrugSite	2004	Structure	Energetic	(An *et al.*, 2004)
OMD	2004	Structure	Energetic	(Bhinge *et al.*, 2004)
eF-Site	2005	Structure	Template-based	(Kinoshita and Nakamura, 2005)
Q-SiteFinder	2005	Structure	Energetic	(Laurie and Jackson, 2005)
PocketFinder	2005	Structure	Energetic	(An *et al.*, 2005)
Bradford *et al.*	2005	Structure	ML-based	(Bradford and Westhead, 2005)
SitesBase	2006	Structure	Template-based	(Gold and Jackson, 2006)
LIGSITEcsc	2006	Structure	Geometric	(Huang and Schroeder, 2006)
Coleman *et al.*	2006	Structure	Geometric	(Coleman and Sharp, 2006)
SURFNET-ConSurf	2006	Structure	Geometric	(Glaser *et al.*, 2006)
Petrova *et al.*	2006	Structure	ML-based	(Petrova and Wu, 2006)
SCREEN	2006	Structure	ML-based	(Nayal and Honig, 2006)
firestar	2007	Sequence	Template-based	(López *et al.*, 2007)
PocketPicker	2007	Structure	Geometric	(Weisel *et al.*, 2007)
PHECOM	2007	Structure	Geometric	(Kawabata and Go, 2007)
Bock *et al.*	2007	Structure	Geometric	(Bock *et al.*, 2007)

(*Continued*)

https://doi.org/10.1017/S003358352500006X

Table 1. (*Continued*)

Method name	Year	Representation type	Algorithm type	Reference
Xie *et al.*	2007	Structure	Geometric	(Xie and Bourne, 2007)
FINDSITE	2008	Structure	Template-based	(Brylinski and Skolnick, 2008)
Kim *et al.*	2008	Structure	Geometric	(Kim *et al.*, 2008)
PatchFinder	2008	Structure	Geometric	(Nimrod *et al.*, 2008)
LIBRUS	2009	Sequence	Template-based, ML-based	(Kauffman and Karypis, 2009)
SiteMap	2009	Structure	Geometric	(Halgren, 2009)
Fpocket	2009	Structure	Geometric	(Le Guilloux *et al.*, 2009)
ConCavity	2009	Structure	Geometric	(Capra *et al.*, 2009)
SiteHound	2009	Structure	Energetic	(Ghersi and Sanchez, 2009; Hernandez et al., 2009)
Tong *et al.*	2009	Structure	ML-based	(Tong *et al.*, 2009)
3DLigandSite	2010	Structure	Template-based	(Wass *et al.*, 2010; McGreig *et al.*, 2022)
ProBiS	2010	Structure	Template-based	(Konc and Janežič, 2010)
GHECOM	2010	Structure	Geometric	(Kawabata, 2010)
POCASA	2010	Structure	Geometric	(Yu *et al.*, 2010)
DoGSite	2010	Structure	Geometric	(Volkamer *et al.*, 2010)
MSPocket	2011	Structure	Geometric	(Zhu and Pisabarro, 2011)
Qiu *et al.*	2011	Structure	ML-based	(Qiu and Wang, 2011)
MetaPocket2.0	2011	Structure	ML-based	(Zhang *et al.*, 2011)
COFACTOR	2012	Structure	Template-based	(Roy *et al.*, 2012)
LISE	2012	Structure	Geometric	(Xie and Hwang, 2012)
FTSite	2012	Structure	Energetic	(Brenke *et al.*, 2009; Ngan *et al.*, 2012)
S-SITE	2013	Sequence	Template-based	(Yang *et al.*, 2013)
TM-SITE	2013	Structure	Template-based	(Yang *et al.*, 2013)
COACH	2013	Structure	ML-based	(Yang *et al.*, 2013)
G-LoSA	2013	Structure	Template-based	(Lee and Im, 2013)
LigandRFs	2014	Sequence	ML-based	(Chen *et al.*, 2014)
LigandDSES	2015	Sequence	ML-based	(Chen *et al.*, 2015)
OSML	2015	Sequence	ML-based	(Yu *et al.*, 2015)
LIBRA	2015	Structure	Template-based	(Viet Hung *et al.*, 2015)
SILCS	2015	Structure	Energetic	(Faller *et al.*, 2015)
bSiteFinder	2016	Structure	Template-based	(Gao *et al.*, 2016)
AutoSite	2016	Structure	Energetic	(Ravindranath and Sanner, 2016)
Tsujikawa *et al.*	2016	Structure	Energetic	(Tsujikawa *et al.*, 2016)
ISMBLab-LIG	2016	Structure	ML-based	(Jian *et al.*, 2016)
DeepSite	2017	Structure	DL-based	(Jiménez *et al.*, 2017)
P2Rank	2018	Structure	ML-based	(Krivák and Hoksza, 2018)
MPLs-Pred	2019	Sequence	ML-based	(Lu *et al.*, 2019)
DeepCSeqSite	2019	Sequence	DL-based	(Cui *et al.*, 2019)
Jiang *et al.*	2019	Structure	DL-based	(Jiang *et al.*, 2019a)
FRSite	2019	Structure	DL-based	(Jiang *et al.*, 2019b)
CurPocket	2020	Structure	Geometric	(Liu *et al.*, 2020a)
PlayMolecule CrypticScout	2020	Structure	Energetic	(Martinez-Rosell *et al.*, 2020)

(*Continued*)

https://doi.org/10.1017/S003358352500006X

Table 1. (*Continued*)

Method name	Year	Representation type	Algorithm type	Reference
GRaSP	2020	Structure	ML-based	(Santana *et al.*, 2020)
BiteNet	2020	Structure	DL-based	(Kozlovskii and Popov, 2020)
Kalasanty	2020	Structure	DL-based	(Stepniewska-Dziubinska et al., 2020)
CAVIAR	2021	Structure	Geometric	(Marchand *et al.*, 2021)
DeepPocket	2021	Structure	DL-based	(Aggarwal *et al.*, 2021)
DeepSurf	2021	Structure	DL-based	(Mylonas *et al.*, 2021)
PUResNet	2021	Structure	DL-based	(Kandel *et al.*, 2021)
GraphBind	2021	Structure	DL-based	(Xia *et al.*, 2021)
bindEmbed21	2021	Structure	DL-based	(Littmann *et al.*, 2021)
HoTS	2022	Sequence	DL-based	(Lee and Nam, 2022)
MDPA	2022	Structure	Energetic	(Gu *et al.*, 2022)
CAT-Site	2022	Structure	DL-based	(Petrovski *et al.*, 2022)
RecurPocket	2022	Structure	DL-based	(Li *et al.*, 2022)
DUNet	2022	Structure	DL-based	(Wang *et al.*, 2022a)
InDeep	2022	Structure	DL-based	(Mallet *et al.*, 2022)
PointSite	2022	Structure	DL-based	(Yan *et al.*, 2022)
GraphSite	2022	Structure	DL-based	(Yuan *et al.*, 2022a)
MsPBRsP	2023	Sequence	DL-based	(Li *et al.*, 2023a)
SiteFerret	2023	Structure	ML-based	(Gagliardi and Rocchia, 2023)
SAPocket	2023	Structure	DL-based	(Wang *et al.*, 2023a)
RefinePocket	2023	Structure	DL-based	(Liu *et al.*, 2023)
GLPocket	2023	Structure	DL-based	(Li *et al.*, 2023a)
SiteRadar	2023	Structure	DL-based	(Evteev *et al.*, 2023)
PocketAnchor	2023	Structure	DL-based	(Li *et al.*, 2023a)
GraphPLBR	2023	Structure	DL-based	(Wang *et al.*, 2023a)
FABind	2023	Structure	DL-based	(Pei *et al.*, 2023)
GrASP	2023	Structure	DL-based	(Smith *et al.*, 2023)
GU-Net	2023	Structure	DL-based	(Nazem *et al.*, 2023)
EquiPocket	2023	Structure	DL-based	(Zhang *et al.*, 2023)
LigBind	2023	Structure	DL-based	(Xia *et al.*, 2023)
LaMPSite	2023	Structure	DL-based	(Zhang and Xie, 2023)
AF2BIND	2023	Structure	DL-based	(Gazizov *et al.*, 2023)
PocketMiner	2023	Structure	DL-based	(Meller *et al.*, 2023)
Pseq2sites	2024	Structure	DL-based	(Seo *et al.*, 2024)

peptide structures as graphs and passes them into a GCN, extracting binding residues on both the protein and peptide sides. GAPS (Zhu *et al.*, 2023) encodes a protein into a point cloud of atoms and uses a geometric attention-based network to classify atoms as binding or non-binding. BiteNet $_{Pp}$ (Kozlovskii and Popov, 2021a) represents peptide binding sites as a set of hotspots and utilizes an approach similar to BiteNet (Kozlovskii and Popov, 2020): it encodes an input protein into a 3D voxel grid and feeds it into a 3D CNN, which splits the grid into cells containing probabilities of a peptide binding site hotspot being in the cell and hotspot center coordinates. DeepProSite (Fang *et al.*, 2023) builds a model using ESMFold (Rives *et al.*, 2021), retrieves embeddings using the ProtTrans (Elnaggar *et al.*, 2021) model, and feeds the graph into a Graph Transformer network (Ingraham *et al.*, 2019) afterward to predict protein–protein and protein–peptide binding sites.

Template- and energy-based methods

There are a few template-based and energy-based approaches. For example, SPOT-peptide (Litfin *et al.*, 2019) and InterPep

https://doi.org/10.1017/S003358352500006X

Table 2. List of methods for prediction of protein–peptide binding sites

Method name	Year	Representation type	Algorithm type	Reference
PINUP	2006	Structure	ML-based	(Liang *et al.*, 2006)
Multi-VORFFIP	2012	Structure	ML-based	(Segura *et al.*, 2012)
PepSite	2012	Structure	ML-based	(Trabuco *et al.*, 2012)
PeptiMap	2013	Structure	Energetic	(Lavi *et al.*, 2013)
Verschueren *et al.*	2013	Structure	Energetic	(Verschueren *et al.*, 2013)
ACCLUSTER	2014	Structure	Energetic	(Yan and Zou, 2014)
SPRINT	2016	Sequence	ML-based	(Taherzadeh *et al.*, 2016)
SVMpep	2018	Sequence	ML-based	(Zhao *et al.*, 2018)
PBRpredict	2018	Sequence	ML-based	(Iqbal and Hoque, 2018)
SPRINT-Str	2018	Structure	ML-based	(Taherzadeh *et al.*, 2018)
P2Rank-Pept	2018	Structure	ML-based	(Krivák *et al.*, 2018)
PepBind	2018	Structure	ML-based	(Zhao *et al.*, 2018)
SPOT-peptide	2019	Structure	Template-based	(Litfin *et al.*, 2019)
InterPep	2019	Structure	Template-based	(Johansson-Åkhe *et al.*, 2019)
Visual	2020	Sequence	ML-based	(Wardah *et al.*, 2020)
MTDSite	2020	Sequence	ML-based	(Sun *et al.*, 2021)
$BiteNet_{Pp}$	2020	Structure	DL-based	(Kozlovskii and Popov, 2021a)
SPPPred	2022	Sequence	ML-based	(Shafiee *et al.*, 2022)
PepBCL	2022	Sequence	DL-based	(Wang *et al.*, 2022a)
PepNN-Seq	2022	Sequence	DL-based	(Abdin *et al.*, 2022)
PepNN-Struct	2022	Structure	DL-based	(Abdin *et al.*, 2022)
PepCNN	2023	Structure	DL-based	(Chandra *et al.*, 2023)
GraphPPepIS	2023	Structure	DL-based	(Li *et al.*, 2023a)
GAPS	2023	Structure	DL-based	(Zhu *et al.*, 2023)
DeepProSite	2023	Structure	DL-based	(Fang *et al.*, 2023)

(Johansson-Åkhe *et al.*, 2019) screen a query protein against a database of known protein–peptide complexes. Energy-based methods sample small molecule probes around a protein and cluster low-energy conformations to get final predictions. PeptiMap (Lavi *et al.*, 2013) adapts the FTmap (Brenke *et al.*, 2009) method for protein–small molecule binding site prediction with additional post-processing for filtering out irrelevant sites. ACCLUSTER (Yan and Zou, 2014) scans a protein surface with 20 amino acid probes. In Verschueren *et al.* (2013), the method uses polypeptide fragments from the BriX (Vanhee *et al.*, 2011) database mapped around the target protein and generates ensembles of energetically favorable protein–peptide complexes.

Benchmarks

For protein–peptide binding sites, the most widely used benchmark is TS125, which is a test set from SPRINT-Seq (Taherzadeh *et al.*, 2016), constructed as a non-redundant subset of 1,279 protein–peptide complexes from the BioLip database (Yang *et al.*, 2012). Other benchmarks include TS092, TS251, and TS639. TS092 is a test benchmark from PepNN (Abdin *et al.*, 2022), designed as a subset of protein–peptide complexes from the PDB, submitted after a specific date and having a sequence identity lower than 30% with all protein targets in the training set. The TS251 benchmark from InterPep (Johansson-Åkhe *et al.*, 2019) was constructed such that the TM-score (Zhang and Skolnick, 2005) of the protein structures is lower than 0.5 with all the structures in the template database. Finally, TS639 from PepBind (Zhao *et al.*, 2018) is a different subset of T1279, used for training and validation of SPRINT-Seq (Taherzadeh *et al.*, 2016), described above. Table 3 lists performance metrics (AUC and MCC, see also Supplementary Section Metrics) for the protein–peptide binding site prediction methods. As one can see, the top methods are ML- or DL-based, with $BiteNet_{Pp}$ (Kozlovskii and Popov, 2021a) being the top-performing-one.

Nucleic acid–small molecule binding sites

RNA molecules are emerging as a significant class of pharmacological targets (Warner *et al.*, 2018). Efforts in RNA-targeting drug discovery span various approaches, such as designing stabilizers for DNA G-quadruplexes (Ortiz de Luzuriaga *et al.*, 2021), developing antibiotics that target riboswitches (Panchal and Brenk, 2021), using antisense RNA (McClorey and Wood, 2015), and creating RNA-targeting antivirals. RNA targets that expand the druggable genome, including those associated with 'undruggable' proteins or non-coding microRNAs, hold particular promise (Matsui and Corey, 2017). However, the development of RNA-targeted drugs faces significant challenges, such as limited chemical diversity and the dynamic nature of RNA structures (Falese *et al.*, 2021). To advance RNA-targeting drug discovery, efficient tools for detecting structure-specific RNA-small molecule binding sites are needed.

There are many approaches targeting binding sites on proteins; however, there is a limited number of methods for nucleic acids. Table 4 provides a list of methods for prediction of nucleic acid-small molecule binding sites.

Knowledge-based

Firstly, there are several knowledge-based methods. Rsite (Zeng *et al.*, 2015) and Rsite2 (Zeng and Cui, 2016) calculate distances between nucleotides based on tertiary and secondary structures, respectively, and determine nucleotides that are the most distant from others as the binding nucleotides. Similarly, RBind (Wang,

https://doi.org/10.1017/S003358352500006X

Table 3. Performance of protein–peptide binding site detection methods on test benchmarks retrieved from Kozlovskii and Popov (2021a), Abdin *et al.* (2022), and Fang *et al.* (2023)

Dataset	Method	AUC	MCC
TS125	$BiteNet_{Pp}$ (Kozlovskii and Popov, 2021a)	0.91	0.49
	DeepProSite (Fang *et al.*, 2023)	0.88	0.45
	PepNN-Struct (Abdin *et al.*, 2022)	0.89	0.39
	PepBCL (Wang *et al.*, 2022a)	0.81	0.39
	PepBind (Zhao *et al.*, 2018)	0.79	0.37
	P2Rank-Pept (Krivák *et al.*, 2018)	0.85	0.35
	MTDSite (Sun *et al.*, 2021)	0.76	0.30
	SPRINT-Str (Taherzadeh *et al.*, 2018)	0.78	0.29
	PeptiMap (Lavi *et al.*, 2013)	0.63	0.27
	PepNN-Seq (Abdin *et al.*, 2022)	0.79	0.26
	Multi-VORFFIP (Segura *et al.*, 2012)	0.78	0.21
	SPRINT-Seq (Taherzadeh *et al.*, 2016)	0.68	0.20
	PepSite (Trabuco *et al.*, 2012)	0.61	0.20
	Visual (Wardah *et al.*, 2020)	0.73	0.17
TS092	PepNN-Struct (Abdin *et al.*, 2022)	0.86	0.41
	PepNN-Seq (Abdin *et al.*, 2022)	0.78	0.27
	PBRPredict (Iqbal and Hoque, 2018)	0.59	0.08
TS251	PepNN-Struct (Abdin *et al.*, 2022)	0.83	0.37
	PepNN-Seq (Abdin *et al.*, 2022)	0.77	0.28
	Interpep (Johansson-Åkhe *et al.*, 2019)	0.79	
TS639	DeepProSite (Fang *et al.*, 2023)	0.86	0.40
	PepNN-Struct (Abdin *et al.*, 2022)	0.87	0.35
	PepBind (Zhao *et al.*, 2018)	0.77	0.35
	PepBCL (Wang *et al.*, 2022a)	0.80	0.31
	PepNN-Seq (Abdin *et al.*, 2022)	0.80	0.25

Jian, *et al.*, 2018b) calculates the degree and closeness of nodes in a nucleotide network and determines binding nucleotides as those with values exceeding a specified threshold. RNetsite (Liu *et al.*, 2024) represents an RNA molecule as a graph and calculates local (degree, neighborhood connectivity) and global (betweenness centrality, closeness, and eccentricity) properties for each node of the graph. Then, each node is classified as binding or non-binding based on the property statistics computed from a reference set of RNA molecules.

Energetic

To the best of our knowledge, only two methods use an energy-based approach. SILCS-RNA (Kognole *et al.*, 2022) runs simulations of a target macromolecule in a mixed solvent with eight different probes. From these simulations, the method calculates a 3D grid with energy maps, which can be used for binding site identification, docking, and binding affinity evaluation tasks. SHAMAN (Panei *et al.*, 2024) is also a probe-based approach, but adds a metadynamics enhanced-sampling technique to explore wider conformational changes of the input RNA molecule.

Machine learning-based

Machine learning-based methods for binding site detection in nucleic acids have emerged very recently. RNAsite (Su *et al.*, 2021) calculates sequential features (e.g., conservation from MSA) and structural features (e.g., topological properties, solvent accessibility, and Laplacian norm) for each nucleotide and passes them into an RF classifier to distinguish between binding and non-binding nucleotides. Similarly, DrugPred_RNA (Rekand and Brenk, 2021) calculates a set of simple structural descriptors such as size, shape, and polarity for a pocket and uses an XGBoost model (Chen and Guestrin, 2016) to classify it as druggable or non-druggable. As descriptors are constructed in a macromolecule type-agnostic way, the model is first pre-trained on a protein dataset and then fine-tuned for binding sites in RNAs.

Deep learning-based

RLBind (Wang, Zhou, *et al.*, 2023b) calculates local and global sequential features (e.g., nucleotide types and evolutionary conservation) and structural features (e.g., network topological properties, biochemical properties, and ASAs), retrieves a window of 11 nucleotides for each position, and feeds it into a 1D CNN that classifies the position as binding/non-binding. RNet (Möller *et al.*, 2022) utilizes an approach similar to DeepSite for predicting binding sites in proteins (Jiménez *et al.*, 2017). It represents a macromolecule structure as an $80 \times 80 \times 80 Å^3$ 3D voxel grid with eight channels representing different atom types: carbon, nitrogen, oxygen, phosphorus, sulfur, fluorine, bromine, and iodine. The method passes this grid as input into a 3D CNN model, predicting ligandability scores for voxels of size $4 \times 4 \times 4 Å^3$. Binding sites are retrieved by clustering predicted ligandable voxels. The authors pre-trained the model on protein binding sites and fine-tuned it to RNAs. MultiModRLBP (Wang *et al.*, 2024) uses a relational GCN to obtain features from a nucleotide structure graph and a pre-trained language model (RNABert (Kalicki and Haritaoglu, 2022)) to get embeddings from an RNA sequence. The model concatenates these structural and sequential features and feeds the resulting vector into a small neural network of fully connected layers to obtain a prediction for each nucleotide. $BiteNet_N$ (Kozlovskii and Popov, 2021b) predicts binding site centers on both RNA and DNA macromolecules. To train the model, the authors composed the largest dataset of $\sim$2000 nucleic acid–small molecule structures. First, the method converts an input nucleic acid macromolecule structure into a voxel-based representation. Then, a 3D CNN model takes this grid as input and produces a set of binding site centers and coordinates, along with a binding score for each nucleotide.

Benchmarks

One of the most widely used benchmarks for RNA-ligand binding site detection methods is the TE18 test set from RNAsite (Su *et al.*, 2021). Another benchmark is RB19 from RBind (Wang, Jian, *et al.*, 2018b). Note that, typically, methods use only a subset of these test sets to avoid sharing similar complexes with the training sets. Most recently, the authors of SHAMAN (Panei *et al.*, 2024) created a test set based on seven RNA complexes: riboswitches (FMN, THF, TPP, and dG) and viral RNAs (HIV-1 TAR, HCV-IRES-IIa, and IAV). They also introduced different strategies to evaluate the methods' performance based on the holo or apo structures of these complexes. Supplementary Tables S11–S16 list the performance of RNA-small molecule binding site detection methods.

https://doi.org/10.1017/S003358352500006X

Protein–ion binding site prediction

Ions are crucial for various physiological processes, such as enzymatic function, signal transduction, and muscle contraction, through their interactions with proteins (Al-Fartusie and Mohssan, 2017). Ions can bind to protein-active sites (Andreini *et al.*, 2008), stabilize or trigger conformational changes in protein structures (Dudev and Lim, 2014; Jernigan *et al.*, 1994), regulate the activity of DNA/RNA polymerases (De Baaij *et al.*, 2015), or affect the concentration-dependent aggregation rate of proteins (Poulson *et al.*, 2020). In addition, ions can act as allosteric modulators. For instance, sodium ions modulate G protein-coupled receptors (Katritch *et al.*, 2014), while in calcium-sensing receptors (CaSR), Ca^{2+}, and Mg^{2+} serve as activators, Cl^- acts as a positive allosteric modulator, and SO_4^{2-}/PO_4^{3-} act as negative modulators (Liu *et al.*, 2020a). Chloride ions (Cl^-) also modulate mGluRs (metabotropic glutamate receptors) (Tora *et al.*, 2015), and calcium ions (Ca^{2+}) influence nAChRs (nicotinic acetylcholine receptors) (Changeux, 2018). Therefore, understanding protein–ion interactions, particularly ion binding sites, is critical to deciphering protein function. Ion binding sites differ from protein–ligand and protein–peptide binding sites in several ways. First, the size of ion binding sites is generally smaller, as small molecules or peptides typically interact with more residues on the protein surface. Furthermore, ion-binding sites are often more adaptable than those of ligands (Chakrabarti, 1993). Another distinction is that many ions require specific coordination geometries with protein atoms. For example, Zn^{2+} binding sites are typically formed by residues such as Cys, His, Asp, or Glu, and are coordinated by four or five atoms, adopting a distorted-tetrahedral or trigonal-bipyramidal geometry (Auld, 2001). Various computational approaches have been proposed to identify ion binding sites, as summarized in Table 5.

Sequence-based

Similarly to sequence-based methods that predict protein–small molecule binding sites, almost all of the sequence-based methods for the identification of binding sites for ions utilize a machine learning-based approach (Chen *et al.*, 2013; Shu *et al.*, 2008; Lippi *et al.*, 2008; Passerini *et al.*, 2011; Ferrè and Clote, 2006; Passerini *et al.*, 2007; Haberal and Oğul, 2017, 2019; Qiao and Xie, 2019; Yu *et al.*, 2013, 2015; Li *et al.*, 2019a; Li *et al.*, 2019b; Yan *et al.*, 2019; Jiang *et al.*, 2016; Ding *et al.*, 2017; Srivastava and Kumar, 2018; Zhao *et al.*, 2019; Essien *et al.*, 2019; Sun *et al.*, 2022). First, they move a sliding window along the input sequence and calculate sequential features for each position. Features can include: evolutionary information such as position-specific scoring matrix (PSSM) or conservation score, predicted secondary structure, and predicted solvent accessibility of residues. Then, these features are fed into an SVM, RF, AdaBoost, or a simple NN. ZincExplorer (Chen *et al.*, 2013) combines a machine learning approach with a templates-based search of known binders to identify Zn-binding sites. IBayes_Zinc (Li *et al.*, 2019a) uses previously described sequence descriptors and predictions from other methods (ZincExplorer (Chen *et al.*, 2013), ZincFinder (Passerini *et al.*, 2007), and ZincPred (Shu *et al.*, 2008)) as input into a Bayesian algorithm to predict Zn sites. MetalPredator (Valasatava *et al.*, 2016) searches through a database of Pfam domains for Fe-S clustering binding and metal binding fragments from MetalPDB (Andreini *et al.*, 2012). ZINCCLUSTER (Ajitha *et al.*, 2018) first creates a database of all monopeptides, dipeptides, and tripeptides and assigns a Z-score for each of them to be Zn-binding based on a dataset. Then, it screens an input sequence with pentapeptides and retrieves a Z-score from the database for two central dipeptides and three tripeptides. The method considers this fragment to be Zn-binding if the average Z-score of dipeptides and tripeptides is higher than zero. With advancements in deep learning, transformer-based models have been developed for ion binding site prediction. IonPred (Essien *et al.*, 2023) employs a transformer architecture to predict ion binding sites directly from protein sequences. M-Ionic (Shenoy *et al.*, 2024) leverages residue embeddings generated by the pre-trained protein language model ESM-2 (Lin *et al.*, 2023) to identify binding sites for various ions. Similarly, LMetalSite (Yuan *et al.*, 2022b) utilizes residue embeddings from ProtTrans (Elnaggar *et al.*, 2021) for the prediction of binding sites specific to Zn^{2+}, Ca^{2+}, Mg^{2+}, and Mn^{2+}.

Template-based

Many methods aim to find fragments of an input structure that are present in the template database of known ion-binding sites. MIB (Lin *et al.*, 2016) and (Lu *et al.*, 2012) use a fragment transformation method to search for parts of an input protein that are present in a database of binding residue templates for multiple ion types. For this, they split residues in the input structure and the template into residue triplets, measured triplet pair similarity and performed clustering of triplets similar to binding ones to get the final predictions. FindSite-metal (Brylinski and Skolnick, 2011) utilizes TM-align (Zhang and Skolnick, 2005; Pandit and Skolnick, 2008) to align template fragments onto the input structure, clusters the obtained alignments, and outputs residue binding scores as the fraction of templates including corresponding positions. TEMSP (Zhao *et al.*, 2011) creates a database of Zn-binding templates from all pairs of residues interacting with this ion. Then, for an input protein, it screens all residue pairs and detects the ones present in the template library. After that, matched pairs are combined into 'pairs-of-pairs', which are further filtered using predefined geometrical thresholds to get the final predictions. In Garg and Pal (2021), the authors used a geometric hashing technique to match query structures with templates of binding sites for different ion types. In Schymkowitz *et al.* (2005b), the method creates a database of canonical positions of water molecules or ions with respect to

Table 4. List of methods for prediction of nucleic acid–small molecule binding sites

Method name	Year	Algorithm type	Reference
Rsite	2015	Knowledge-based	(Zeng *et al.*, 2015)
Rsite2	2016	Knowledge-based	(Zeng and Cui, 2016)
RBind	2018	Knowledge-based	(Wang, Jian, *et al.*, 2018b)
RNAsite	2021	ML-based	(Su *et al.*, 2021)
$BiteNet_N$	2021	DL-based	(Kozlovskii and Popov, 2021b)
DrugPred_RNA	2021	ML-based	(Rekand and Brenk, 2021)
SILCS-RNA	2022	Energetic	(Kognole *et al.*, 2022)
RNet	2022	DL-based	(Möller *et al.*, 2022)
SHAMAN	2023	DL-based	(Panei *et al.*, 2024)
RLBind	2023	DL-based	(Wang, Zhou, *et al.*, 2023b)
MultiModRLBP	2023	DL-based	(Wang *et al.*, 2024)

https://doi.org/10.1017/S003358352500006X

protein atom triads. Then, it screens the surface of the protein with these triads, clusters favorable points, and performs optimization of positions using the empirical force field. GASS-Metal (Paiva *et al.*, 2022) uses a genetic algorithm for the effective search of structural patterns similar to ion binding sites from a curated database of templates.

Machine learning-based

Apart from the sequence-based machine learning approaches, most of the structure-based machine learning methods (Sodhi *et al.*, 2004; Bordner, 2008; Zheng *et al.*, 2012; Ireland and Martin, 2021; Song and Jiang, 2023) calculate sequential and structural features for each residue and feed them into an SVM, RF, or neural network classifier. For example, FEATURE (Ebert and Altman, 2008) constructs concentric radial shells for the atomic environments, calculates physicochemical features inside each of them, and uses Bayesian learning to differentiate whether an environment corresponds to a Zn-binding site or not. IonCom (Hu *et al.*, 2016) combines predictions from the sequence-based approach IonSeq and other tools for binding site prediction: COFACTOR (Roy *et al.*, 2012), TM-SITE, S-SITE, and COACH (Yang *et al.*, 2013), and trains a classifier on top of them. PinMyMetal (Zheng *et al.*, 2024) uses geometrical features, including residue properties, interatomic distances, bond angles, and atomic types as input into the ensemble ML model predicting Zn^{2+} binding sites.

Deep learning-based

GraphBind (Xia *et al.*, 2021) is a GNN-based model that predicts binding sites for Ca^{2+}, Mn^{2+}, and Mg^{2+}. DeepProSite (Fang *et al.*, 2023), as mentioned in Section Deep learning-based, uses ESMFold (Rives *et al.*, 2021), ProtTrans (Elnaggar *et al.*, 2021), and a GNN for the prediction of different types of binding sites, including those for Ca^{2+}, Mn^{2+}, and Mg^{2+}. In Gamouh *et al.* (2023), the authors used embeddings from ProtTrans (Elnaggar *et al.*, 2021) as features for the graph nodes and used GNN to predict binding sites for nucleotides and ions: Ca^{2+}, Mg^{2+}, Mn^{2+}, Fe^{3+}, and Zn^{2+}. DELIA (Xia *et al.*, 2020) first constructs a feature vector as the combination of outputs from two other models: (i) BiLSTM which takes as the input the sequence-based features; and (ii) 2D CNN ResNet model, which takes as the input the distance matrix. The constructed feature vectors are used as the input to the next fully connected layer, which outputs the probability of each residue being binding or non-binding. Metal3D (Dürr *et al.*, 2023) employs a 3D CNN to predict the probability density of Zn^{2+} binding across the protein structure. MoM (Laveglia *et al.*, 2023) utilizes a GNN to classify local protein environments composed of Cys, His, Asp, and Glu residues, determining whether these environments are likely to bind Zn^{2+}. BindWeb (Xia *et al.*, 2022) is a consensus method combining predictions from GraphBind (Xia *et al.*, 2021) and DELIA (Xia *et al.*, 2020) models.

Other

There are several geometric methods that search positions with surrounding atoms whose geometry resembles the ion coordination shell. GRE4Zn (Liu *et al.*, 2014) utilizes the fact that most known Zn-binding sites comprise sets of four or three residues with distinctly specific geometries. GaudiMM Metals (Sciortino *et al.*, 2019) retrieves information about acceptable coordination shell geometries for a set of ions and implements them as an additional objective for optimization with ion presence in the GaudiMM platform (Rodríguez-Guerra Pedregal *et al.*, 2017). BioMetAll (Sánchez-Aparicio *et al.*, 2020) constructs a grid of metal probes around a protein and checks each grid position to see if the amino acid environment matches geometric constraints determined from statistics in a dataset of protein structures. The method obtains final predictions from the clustering of relevant points. Also, there are methods that utilize an energetic approach. For example, BION (Shashikala *et al.*, 2021) calculates electrostatic potential maps with a gaussian-smooth dielectric function term to predict the positions of non-specifically surface-bound ions.

It is important to note that many of the ion binding site identification methods consider only Cys, His, Glu, and Asp residues (Chen *et al.*, 2013; Shu *et al.*, 2008; Passerini *et al.*, 2007), as these four amino acids are involved in the coordination shell of a bound ion in many cases. Moreover, MetalDetector (Lippi *et al.*, 2008; Passerini *et al.*, 2011) and DeepMBS (Haberal and Oğul, 2017, 2019) operate only with His and Cys, and DiANNA (Ferrè and Clote, 2006) work solely with Cys residues. On the other hand, many methods have been developed to predict binding regions for specific ions. For example, multiple approaches aim to predict Zn-binding regions (Chen *et al.*, 2013; Shu *et al.*, 2008; Passerini *et al.*, 2007; Haberal and Oğul, 2017, 2019; Li *et al.*, 2019a; Li *et al.*, 2019b; Yan *et al.*, 2019; Ajitha *et al.*, 2018).

It is worth noting that for some ions (e.g., Ca^{2+}, Mg^{2+}, Na^{+}, and K^{+}), the performance metrics are much lower compared to others, as can be seen in Supplementary Table S19. As pointed out in Lu *et al.* (2012) and Qiao and Xie (2019), this can be caused by the higher variability of these binding sites in terms of amino acid composition and structure. Indeed, in Qiao and Xie (2019), the authors calculate the frequency difference index, defined as the average difference in the ratio of binding and non-binding residues of each type among the 20 amino acid types, and observed that the index values are much lower for Ca^{2+}, Mg^{2+}, Na^{+}, and K^{+} compared to other ions.

Benchmarks

We observed that different methods use various benchmarks for evaluation; here, we list benchmarks that were used by several methods. The Passerini dataset (Passerini *et al.*, 2006) is a dataset containing 2,727 sequences with 687 protein chains bound to a metal atom. There are four methods that used it as a training or validation set for the prediction of Zn-binding sites (ZincFinder (Passerini *et al.*, 2007), ZincPred (Shu *et al.*, 2008), ZincExplorer (Chen *et al.*, 2013), DeepMBS (Haberal and Oğul, 2017)). However, note that these methods calculated different metrics on different sets of residues (e.g., Cys and His or Cys, His, Glu, and Asp). The Zhao dataset (Zhao *et al.*, 2011) is a dataset used for training and validation of a template-based method TEMSP, consisting of ∼600 protein targets with bound Zn ions. Although many methods use this dataset as an independent test set, some methods retrieved only a subset from it. SSWPNN (Li *et al.*, 2019b) provides the most complete comparison of methods on this dataset (see Supplementary Table S21). Furthermore, for the validation of SSWPNN, the authors also collected a second independent test set from PDB consisting of 213 protein chains with 1,017 Zn-binding sites, and compared SSWPNN with five other approaches for the prediction of Zn binding sites on the Zhao and SSWPNN datasets (see Supplementary Tables S21 and S22). ZincBindDB (Ireland and Martin, 2019) is the largest database of Zn binding sites (about 35,000 binding sites from about 16,000 structures), that was used for training and validation of the

https://doi.org/10.1017/S003358352500006X

Table 5. List of methods for prediction of protein–ion binding sites

Method name	Year	Representation	Algorithm type	Ion types	Reference
MetSite	2004	Structure	ML-based	Ca^{2+}, Cu^{2+}, Fe^{3+}, Mg^{2+}, Mn^{2+}, Zn^{2+}	(Sodhi *et al.*, 2004)
Schymkowitz *et al.*	2005	Structure	Template-based	Mg^{2+}, Ca^{2+}, Zn^{2+}, Mn^{2+}, Cu^{2+}	(Schymkowitz *et al.*, 2005a)
DiANNA	2006	Sequence	ML-based	Fe^{3+}, Zn^{2+}, Cd^{2+}	(Ferrè and Clote, 2006)
ZincFinder	2007	Sequence	ML-based	Zn^{2+}	(Passerini *et al.*, 2007)
ZincPred	2008	Sequence	ML-based	Zn^{2+}	(Shu *et al.*, 2008)
MetalDetector	2008	Sequence	ML-based	Any metal ion	(Lippi *et al.*, 2008; Passerini *et al.*, 2011)
SitePredict	2008	Structure	ML-based	Ca^{2+}, Cu^{2+}, Fe^{3+}, Mg^{2+}, Mn^{2+}, Zn^{2+}	(Bordner, 2008)
FEATURE	2008	Structure	ML-based	Zn^{2+}	(Ebert and Altman, 2008)
FindSite-metal	2011	Structure	Template-based	Ca^{2+}, Co^{2+}, Cu^{2+}, Fe^{3+}, Mg^{2+}, Mn^{2+}, Ni^{2+}, Zn^{2+}	(Brylinski and Skolnick, 2011)
TEMSP	2011	Structure	Template-based	Zn^{2+}	(Zhao *et al.*, 2011)
Lu *et al.*	2012	Structure	Template-based	Ca^{2+}, Cu^{2+}, Fe^{3+}, Mg^{2+}, Mn^{2+}, Zn^{2+}	(Lu *et al.*, 2012)
ZincIdentifier	2012	Structure	ML-based	Zn^{2+}	(Zheng *et al.*, 2012)
ZincExplorer	2013	Sequence	ML-based	Zn^{2+}	(Chen *et al.*, 2013)
TargetS	2013	Sequence	ML-based	Ca^{2+}, Zn^{2+}, Mg^{2+}, Mn^{2+}, Fe^{3+}	(Yu *et al.*, 2013)
GRE4Zn	2014	Structure	Geometric	Zn^{2+}	(Liu *et al.*, 2014)
MetalS^3	2014	Structure	Template-based	Any metal ion	(Valasatava *et al.*, 2014)
OSML	2015	Sequence	ML-based	Ca^{2+}, Mg^{2+}, Mn^{2+}, Fe^{3+}, Zn^{2+}	(Yu *et al.*, 2015)
MetalPredator	2016	Sequence	ML-based	Fe-S clusters	(Valasatava *et al.*, 2016)
IonSeq	2016	Sequence	ML-based	Zn^{2+}, Cu^{2+}, Fe^{2+}, Fe^{3+}, Ca^{2+}, Mg^{2+}, Mn^{2+}, Na^{+}, K^{+}, CO_3^{2-}, NO_2^{-}, SO_4^{2-}, PO_4^{3-}	(Hu *et al.*, 2016)
IonCom	2016	Structure	ML-based	Zn^{2+}, Cu^{2+}, Fe^{2+}, Fe^{3+}, Ca^{2+}, Mg^{2+}, Mn^{2+}, Na^{+}, K^{+}, CO_3^{2-}, NO_2^{-}, SO_4^{2-}, PO_4^{3-}	(Hu *et al.*, 2016)
MIB	2016	Structure	Template-based	Ca^{2+}, Cu^{2+}, Fe^{3+}, Mg^{2+}, Mn^{2+}, Zn^{2+}, Cd^{2+}, Fe^{2+}, Ni^{2+}, Hg^{2+}, Co^{2+}, Cu^{+}	(Lin *et al.*, 2016)
DeepMBS	2017	Sequence	ML-based	Zn^{2+}	(Haberal and Oğul, 2017, 2019)
EC-RUS	2017	Sequence	ML-based	Ca^{2+}, Mg^{2+}, Mn^{2+}, Fe^{3+}, Zn^{2+}	(Ding *et al.*, 2017)
ZINCCLUSTER	2018	Sequence	ML-based	Zn^{2+}	(Ajitha *et al.*, 2018)
ZincBinder	2018	Sequence	ML-based	Zn^{2+}	(Srivastava and Kumar, 2018)
MIonSite	2019	Sequence	ML-based	Zn^{2+}, Ca^{2+}, Mg^{2+}, Mn^{2+}, Fe^{3+}, Cu^{2+}, Fe^{2+}, Co^{2+}, Na^{+}, K^{+}, Cd^{2+}, Ni^{2+}	(Qiao and Xie, 2019)
IBayes_Zinc	2019	Sequence	ML-based	Zn^{2+}	(Li *et al.*, 2019a)
SSWPNN	2019	Sequence	ML-based	Zn^{2+}	(Li *et al.*, 2019b)
ZnMachine	2019	Sequence	ML-based	Zn^{2+}	(Yan *et al.*, 2019)
SXGBsite	2019	Sequence	ML-based	Ca^{2+}, Mg^{2+}, Mn^{2+}, Fe^{3+}, Zn^{2+}	(Zhao *et al.*, 2019)
GaudiMM Metals	2019	Structure	Geometric	Mg^{2+}, Ca^{2+}, Mn^{2+}, Fe^{3+}, Co^{2+}, Ni^{2+}, Cu^{2+}, Zn^{2+}	(Sciortino *et al.*, 2019)
ZinCaps	2019	Sequence	DL-based	Zn^{2+}	(Essien *et al.*, 2019)
DELIA	2020	Structure	DL-based	Ca^{2+}, Mn^{2+}, Mg^{2+}	(Xia *et al.*, 2020)
BioMetAll	2020	Structure	Geometric	Any metal ion	(Sánchez-Aparicio *et al.*, 2020)
Garg *et al.*	2021	Structure	Template-based	Zn^{2+}, Cu^{2+}, Fe^{3+}, Ca^{2+}, Mg^{2+}	(Garg and Pal, 2021)
BION	2021	Structure	Energetic	Ca^{2+}, Zn^{2+}, Cl^{-}, Mg^{2+}	(Shashikala *et al.*, 2021)
ZincBindPredict	2021	Structure	ML-based	Zn^{2+}	(Ireland and Martin, 2021)
GraphBind	2021	Structure	DL-based	Ca^{2+}, Mn^{2+}, Mg^{2+}	(Xia *et al.*, 2021)
BindWeb	2022	Structure	DL-based	Ca^{2+}, Mn^{2+}, Mg^{2+}	(Xia *et al.*, 2022)
LMetalSite	2022	Sequence	DL-based	Zn^{2+}, Ca^{2+}, Mn^{2+}, Mg^{2+}	(Yuan *et al.*, 2022a)

(*Continued*)

https://doi.org/10.1017/S003358352500006X

Table 5. (*Continued*)

Method name	Year	Representation	Algorithm type	Ion types	Reference
GASS-Metal	2022	Structure	Template-based	Zn^{2+}, Fe^{3+}, Mg^{2+}, Ca^{2+}, Mn^{2+}, K^{+}, Na^{+}, Cu^{2+}, Ni^{2+}, Co^{2+}, Gd^{3+}, Hg^{2+}, Sb^{3+}, Cd^{2+}	(Paiva *et al.*, 2022)
Sun *et al.*	2022	Sequence	DL-based	Ca^{2+}, Mg^{2+}	(Sun *et al.*, 2022)
DeepProSite	2023	Structure	DL-based	Ca^{2+}, Mn^{2+}, Mg^{2+}	(Fang *et al.*, 2023)
Gamouh *et al.*	2023	Structure	DL-based	Ca^{2+}, Mn^{2+}, Mg^{2+}, Fe^{3+}, Zn^{2+}	(Gamouh *et al.*, 2023)
M-Ionic	2023	Sequence	DL-based	Ca^{2+}, Mn^{2+}, Mg^{2+}, Zn^{2+}, Cu^{2+}, PO_4^{3-}, SO_4^{2-}, Fe^{2+}, Fe^{3+}, Co^{2+}	(Shenoy *et al.*, 2024)
IonPred	2023	Sequence	DL-based	Zn^{2+}, Cu^{2+}, Fe^{2+}, Fe^{3+}, Ca^{2+}, Mg^{2+}, Mn^{2+}, Na^{+}, K^{+}, CO_3^{2-}, NO_2^{-}, SO_4^{2-}, PO_4^{3-}	(Essien *et al.*, 2023)
Song *et al.*	2023	Structure	ML-based	Ca^{2+}, Cu^{2+}, K^{+}, Mg^{2+}, Na^{+}, Zn^{2+}	(Song and Jiang, 2023)
Metal3D	2023	Structure	DL-based	Zn^{2+}	(Dürr *et al.*, 2023)
MoM	2023	Structure	DL-based	Zn^{2+}	(Laveglia *et al.*, 2023)
PinMyMetal	2024	Structure	DL-based	Zn^{2+}	(Zheng *et al.*, 2024)

ZincBindPredict method (Ireland and Martin, 2021) (however, it was not used by the other methods). Note that a newer version of the database contains more samples ($\sim$ 40,000 binding sites from $\sim$ 16,000 structures); so one may expect improved performance for newer methods. The BION dataset (Petukh *et al.*, 2013) contains binding sites for Ca^{2+}, Zn^{2+}, Cl^{-}, and Mg^{2+} ions from 446 protein structures. In Shashikala *et al.* (2021), the authors used this dataset to compare the performances of BION (Petukh *et al.*, 2013) and BION-2 (Shashikala *et al.*, 2021) methods with forcefield-based tools from VMD (Humphrey *et al.*, 1996) and Fold-X (Schymkowitz *et al.*, 2005a). In Hu *et al.* (2016), the authors created a large dataset of 2,100 protein structures in complex with 3,075 ions (Zn^{2+}, Cu^{2+}, Fe^{2+}, Fe^{3+}, Ca^{2+}, Mg^{2+}, Mn^{2+}, Na^{+}, K^{+}, CO_3^{2-}, NO_2^{-}, SO_4^{2-}, and PO_4^{3-}) retrieved from the BioLip database (Yang *et al.*, 2012). The authors used it for 5-fold cross-validation of IonSeq and IonCom methods, and there are available scores for several methods on this benchmark (see Supplementary Table S18). In MIonSite (Qiao and Xie, 2019), the authors created a large dataset of 7,676 sequences for training and 274 sequences for an independent test set. These sets include ions of multiple types: Zn^{2+}, Ca^{2+}, Mg^{2+}, Mn^{2+}, Fe^{3+}, Cu^{2+}, Fe^{2+}, Co^{2+}, Na^{+}, K^{+}, Cd^{2+}, and Ni^{2+}. MIonSite was compared with other methods on their test set (see Supplementary Table S19). The authors also created a small dataset (BTD) of 10 proteins with metal ion-binding sites and 10 proteins without metal ion-binding sites for additional comparison with other methods (see Supplementary Table S20). TargetS (Yu *et al.*, 2013) used the BioLip database (Yang *et al.*, 2012) to assemble training and validation sets with metal ion binding sites (Ca^{2+}, Zn^{2+}, Mg^{2+}, Mn^{2+}, and Fe^{3+}) and nucleotides with 3,779 and 642 ion-bound protein sequences, respectively. The authors used an independent test set to compare TargetS with ligand-specific predictors and an alignment-based predictor (see Supplementary Table S23). Garg and Pal (Garg and Pal, 2021) assembled datasets for five metal ions (Cu^{2+}, Fe^{3+}, Ca^{2+}, Mg^{2+}, and Zn^{2+}) and split them into training and testing sets with 1,079 and 268 structures in total, respectively. The authors compared their method with IonCom (Hu *et al.*, 2016) and MIB (Lin *et al.*, 2016) (see Supplementary Table S24). In Yan *et al.* (2019), the authors prepared the zn1436 dataset of proteins with bound zinc ions for a comparison of the ZnMachine method with ZincExplorer (Chen *et al.*, 2013) (see Supplementary Table S25). In Yuan *et al.* (2022b), the authors compiled a test set from the BioLip database (Yang, 2012), consisting of 211, 183, 235, and 57 protein chains bound to Zn^{2+}, Ca^{2+}, Mg^{2+}, and Mn^{2+} ions, respectively (see Supplementary Table S27). Similarly, the authors of M-Ionic (Shenoy *et al.*, 2024) constructed an independent test set from BioLip, but reported performance metrics only for LMetalSite (Yuan *et al.*, 2022b) and M-Ionic (Shenoy *et al.*, 2024) on this dataset (see Supplementary Table S28). Among the benchmarks used by a single method, one may notice the BioMetAll test set (Sánchez-Aparicio *et al.*, 2020), which consists of 53 crystallographic structures containing the two-histidine one-carboxylate motif (FTM). This is an interesting benchmark since its structures vary in size, and this motif may bind multiple types of metal ions (Cd^{2+}, Co^{2+}, Cu^{2+}, Fe^{3+}, Hg^{2+}, Mg^{2+}, Mn^{2+}, Ni^{2+}, Ru^{3+}, and Zn^{2+}). Note that almost all methods operate with residue-based scores as their performance metric, and only GaudiMM Metals (Sciortino *et al.*, 2019) and BION (Shashikala *et al.*, 2021) used distance-based scores, which are likely more representative of the ion binding site prediction problem.

Despite the large variety of methods and benchmarks (see Table 5), one can see from Supplementary Tables S18 and S19 that MIonSite (Qiao and Xie, 2019) and IonCom (Hu *et al.*, 2016) demonstrate better performance for different ions. Interestingly, MIonSite is a sequence-based method, and IonCom is a structure-based method; therefore, it would be interesting to see if a combined approach shows even better results. As for the Zn-specific predictors, Supplementary Table S22 shows that SSWPNN (Li *et al.*, 2019b) outperforms other methods.

Challenges

The composition of high-quality and diverse labeled datasets and benchmarks is one of the biggest challenges in the binding site detection problem. As one can see from the Benchmarks subsections, typically, there are no unified training-validation sets and test benchmarks to perform a rigorous comparison of the developed methods. Moreover, the existing training and test splits often contain data leakage (let alone structural artifacts), resulting in likely overly optimistic performance metrics, that should not be compared between the different methods. Unfortunately, despite the exponential growth of available experimental structures in PDB, in some cases, the test benchmarks are too small to make

https://doi.org/10.1017/S003358352500006X

statistically solid conclusions. For example, the commonly used RNA-small molecule binding site benchmark, TE18, contains just 18 structures (see Section Nucleic acid–small molecule binding sites Benchmarks). Related to this, another challenge to consider, especially with respect to the ML and DL methods, is overfitting. Overfitting is a common problem in machine learning, where models perform well on training data but fail to generalize to the unseen cases. Deficient training sets or incorrect training-validation-test splitting can also result in models showing artificially high score values (Kapoor and Narayanan, 2023). Thus, to overcome these challenges, not only high-quality datasets are required but also unified training-validation-test splits. This will also help to make the comparison of different methods more rigorous. However, other hidden biases may remain. For example, using pre-trained language models (LMs) may lead to data leakage, as the LM model itself might have seen data similar to the training set.

One rather technical but important thing that also prevents a rigorous comparison of binding site detection methods is the use of different performance metrics for their evaluation. First of all, many papers present accuracy, precision, recall, or ROC AUC metrics, which may be misleading. Indeed, precision or recall metrics are high when the model outputs either the lowest or highest number of positive predictions, respectively, and accuracy or ROC AUC tends to be 1 when there is a high imbalance in binary labels. MCC is a much more suitable metric for binding site detection; however, it still depends on the choice of the threshold for determining binary labels. On the other hand, AP or AURPC, which is the area under the precision-recall curve, is much more convenient for the classification of binding residues, as it takes into account the ranking of predictions and does not depend on class imbalance. Also, instead of calculating precision on the top-N predictions for each structure (see DCC and DCA metrics in Supplementary Section Metrics), one can use AP for assessing model performance using a distance-based criterion as well (Kozlovskii and Popov, 2020), following this idea from object detection in computer vision (Everingham *et al.*, 2010). More reliable performance metrics can be roughly divided into three categories: distance-, volume-, and residue-based. Distance-based metrics define a prediction as successful if the distance from the predicted binding site center to the true center or any atom of the binding site or ligand is lower than a threshold value, which is usually set to 4Å. Volume-based metrics calculate the overlap between the predicted and true binding site cavities. Finally, residue-based metrics rely on binary classification metrics calculated from residue scores. There is no one-size-fits-all solution, however. For example, residue-based scores may be misleading for protein–ion binding sites. Indeed, the number of interacting residues is small; thus, the impact of a single residue on the metric is high. Furthermore, the definition of interacting residues varies too, resulting in different metric values for the same predictions but with a slightly different set of true labels. Similarly, a residue-based metric may be misleading for nucleic acid–small molecule binding site prediction, though with the opposite reasoning. In this case, the number of nucleotides in the structure is typically small; thus, the binding site covers a larger portion of nucleotide residues. As a result, residue-based metrics may become insensitive to very different predictions. In contrast, distance- and volume-based metrics have been shown to be informative enough for protein–small molecule binding site predictions. Therefore, distance-based metrics would be more robust for protein–ion and nucleic acid–small molecule binding site detection problems. However, as one can see from Sections Nucleic acid–small molecule binding sites Benchmarks and Protein–ion binding site prediction Benchmarks, most of the methods rely on residue-based metrics. On the other hand, residue-based metrics can be more suitable for protein–peptide binding site detection methods compared to distance- and volume-based metrics, because of the large size of protein–peptide binding site interfaces.

Another challenge is the interpretability of ML and DL-based binding site detection models. While these methods could achieve superior accuracy compared to classical approaches, their predictions often lack clear mechanistic explanations, making it difficult to extract meaningful insights about the underlying molecular interactions (Murdoch *et al.*, 2019; Vecchietti *et al.*, 2024). This issue can be particularly relevant when understanding why a model identifies a particular region as a potential binding site is essential for hypothesis-driven drug design. Unlike physics-based or first-principle methods, which rely on well-understood physical and chemical principles, DL models operate as complex, non-linear transformations of input data, obscuring the contributions of individual molecular features. This ambiguity also hampers debugging models, detecting biases in datasets, and ensuring reliable generalization across diverse molecular structures. Recent advances in explainable AI (XAI) (Jiménez-Luna *et al.*, 2020; Bhatt *et al.*, 2024), such as feature attribution techniques (e.g., SHAP (Lundberg and Lee, 2017), LRP (Montavon *et al.*, 2019), Grad-CAM (Selvaraju *et al.*, 2017)) and attention mechanisms in transformer-based models (Wiegreffe and Pinter, 2019), have been proposed to increase interpretability, but their application to binding site prediction remains limited. Incorporating interpretable ML techniques into binding site detection could improve trust in DL-based predictions and enhance their practical usability in drug discovery pipelines.

In drug discovery, one of the challenges is to assess the 'druggability' or 'ligandability' of the detected binding sites. Currently, there are no strict criteria for the 'druggable' binding sites. Similarly to the characterization of drug-like molecules using the rule of 5, Ghose filter, or other heuristics, one can compose such knowledge-based criteria for the binding sites based on their properties. While an exhaustive review of binding site characterization methods is beyond the scope of this article, it is noteworthy that several computational tools have been developed to analyze specific properties of binding sites. These tools assess attributes such as volume, surface area, and flexibility, and often identify sub-pockets within larger pockets (Durrant *et al.*, 2014; Guerra *et al.*, 2021), typically, employing approaches similar to those discussed in Section Protein–small molecule binding sites Geometric. Moreover, certain geometric binding site prediction methods inherently provide volume estimations of binding sites (Kawabata and Go, 2007; Capra *et al.*, 2009; Le Guilloux *et al.*, 2009; Zhu and Pisabarro, 2011), and some methods analyze pockets throughout molecular dynamics (MD) trajectories, offering dynamic insights into binding site properties (Craig *et al.*, 2011; Schmidtke *et al.*, 2011; Paramo *et al.*, 2014; Laurent *et al.*, 2015; Wagner *et al.*, 2017; Chen *et al.*, 2019; Lv and Cao, 2024). Other tools, like MOLE (Pravda *et al.*, 2018), CAVER (Stourac *et al.*, 2019), and others (Yaffe *et al.*, 2008; Lee and Helms, 2012) aim at the characterization of protein tunnels, channels, and pores.

Last but not least challenge is the prospective validation of the developed methods. Given the aforementioned challenges that can lead to over-optimistic performance on the retrospective benchmarks, the real-world application is of crucial importance. However, such case studies are quite rare (Popov *et al.*, 2024; Naz *et al.*, 2015) and absent for most of the developed methods. In this regard, community-driven challenges, such as CASP (https://predictioncenter.org) and CACHE (https://cache-challenge.org), may comprise targets with

https://doi.org/10.1017/S003358352500006X

previously unpublished binding sites and, thus, provide an opportunity to demonstrate the predictive power of the developed methods.

Trends and future directions

It is no wonder that machine learning-based approaches are gradually displacing the first-principle methods, and while older research focuses more on searching for the most powerful features, newer research is more focused on exploring various neural network architectures. Moreover, with the advances in large language models, it has become common to utilize embeddings produced by, for example, protein language models, as the feature vectors for the downstream task of the binding site detection. While this idea seems promising, extensive exploration of this research direction is computationally expensive, requiring significant hardware resources and time, which can limit accessibility for some research groups.

When applying or testing binding site detection methods, an important question to address is the flexibility of the target. Naturally, one expects that a method should detect a binding site in the holo conformation of the target. But in practice, one needs to discover novel binding sites given the unbound conformation. At what point in the imaginary trajectory between the unbound and bound conformations should a method detect the binding site? The answer to this question likely depends on the considered set of ligands for a particular target, such that the method should detect a binding site in the structure, if the corresponding conformation is within a certain vicinity of the bound conformation for at least one of the ligands. The vicinity can be simply defined as all conformations within a given RMSD threshold relative to the bound conformations. Constructing such a benchmark of conformational ensembles would be a valuable step forward for the development of robust binding site detection approaches. Currently, one typically addresses the flexibility issue by generating multiple conformations of the target using molecular dynamics or another method and applying binding site detection to them (Kozlovskii and Popov, 2020; Martinez-Rosell *et al*., 2020; Meller *et al*., 2023; Panei *et al*., 2024). The development of spatiotemporal methods to analyze target binding sites and their dynamics is a valuable direction for future research.

There are other types of binding sites besides those covered in this review. For example, specific models have been developed for protein-nucleotide binding sites (Chauhan *et al*., 2009; Chen *et al*., 2012; Kusuma *et al*., 2019), carbohydrate binding sites (Canner *et al*., 2023), vitamin binding sites (Panwar *et al*., 2013), catalytic sites (Dou *et al*., 2012), as well as water positions (Zaucha *et al*., 2020; Park and Seok, 2022). As for RNA targets, there are methods to predict RNA-ion binding sites, including MetalionRNA (Philips *et al*., 2012), MgNet (Zhou and Chen, 2022), Metal3DRNA (Zhao *et al*., 2023), as well as machine learning methods to predict RNA-protein binding sites using only sequence data (Choi and Han, 2013; Panwar and Raghava, 2015; Tuvshinjargal *et al*., 2016; Choi *et al*., 2017; Zhan *et al*., 2018; Pan *et al*., 2020; Tahir *et al*., 2021; Zhao *et al*., 2022), sequential and secondary structure (Uhl *et al*., 2019), or sequential and tertiary data (Luo *et al*., 2017). Some of the sequence-based methods rely on large databases with experimental data on RNA-protein binding, such as RNAcompete (Ray *et al*., 2009), CLIP-seq, and RIP-seq (Ray *et al*., 2013). Notably, there are approaches to predict DNA binding sites (e.g., DeepBind (Alipanahi *et al*., 2015) and DeepSTF (Ding *et al*., 2023)), trained on datasets from protein binding microarrays (PBMs) (Mukherjee *et al*., 2004), ChIP-seq (Kharchenko *et al*., 2008), or HT-SELEX (Jolma *et al*., 2010). We would like to separately note that protein-covalent ligand binding sites constitute a special case of protein–small molecule binding sites. Covalent ligands may be useful as a therapeutic modality in various diseases; hence, the prediction of this type of binding site is relevant in covalent drug discovery (Boike *et al*., 2022). A ligand can form a covalent bond with target residues (commonly Cys, Ser, and Lys, but there are other cases as well) upon binding, which imposes strict constraints on the binding site detection problem. There are several databases of covalent agents, including CovalentInDB (Du *et al*., 2021) and CovPDB (Gao *et al*., 2022), and there are methods for predicting the ability of cysteines to form a covalent bond with ligands (Zhang *et al*., 2016, 2017; Zhao *et al*., 2017; Du *et al*., 2022; Gao and Günther, 2023). Other examples include methods to predict macromolecular binding sites, such as protein-nucleic acids (Hendrix *et al*., 2021; Wei *et al*., 2022; Yuan *et al*., 2022a; Liu and Tian, 2023; Roche *et al*., 2023; Song *et al*., 2023; Zhu and Yu, 2023) or protein–protein (Fout *et al*., 2017; Gainza *et al*., 2020; Dai and Bailey-Kellogg, 2021; Renaud *et al*., 2021; Sverrisson *et al*., 2021; Tubiana *et al*., 2022; Gao *et al*., 2023; Jha *et al*., 2023; Krapp *et al*., 2023). Most of these methods are based on approaches similar to the ones described here. Given the large variety of binding site types on one hand and the advances in multi-modal and multi-task machine learning approaches on the other hand, we expect that the next-generation binding site prediction methods will operate across different types of macromolecular structures as well as their binding counterparts. Although there is currently no strong evidence that this will improve model accuracy, one may expect that a single model trained on comprehensive datasets could have stronger generalization ability and robustness. We observed that some methods implicitly explore this idea already; for example, RNet (Liu *et al*., 2024) and BiteNet$_{Pp}$ (Kozlovskii and Popov, 2021a) started with protein–small molecule binding site detection models and fine-tuned them for RNA-small molecule and protein–peptide binding site models, respectively.

Finally, the discovery of novel binding sites may come from an orthogonal direction. For example, global molecular docking approaches search for the optimal configuration of two binding partners without prior knowledge of the corresponding binding site. Molecular docking methods have been developed for different types of macromolecules and ligands, including those described here in the context of the binding site detection problem. Needless to say, the molecular docking field and virtual ligand screening, in general, are also affected by the machine-learning era (Fadahunsi *et al*., 2024). Moreover, breakthroughs in protein structure prediction by DeepMind (https://deepmind.google/technologies/alphafold/) have also opened new opportunities to solve binding site detection problems. Particularly, one promising direction is the development of co-folding approaches that aim to predict the 3D structure of the molecular complex starting from a 1D (sequence) or 2D (graph) representation of its subunits. The most recent examples of such approaches include AlphaFold3 (Abramson *et al*., 2024), RoseTTaFold All-Atom (Krishna *et al*., 2024), or NeuralPLexer (Qiao *et al*., 2024). Although their predictive power has yet to be assessed on independent test benchmarks to date, one may expect the rise of end-to-end approaches to solving structure prediction, binding site detection, and molecular docking problems simultaneously in the future.

https://doi.org/10.1017/S003358352500006X

List of abbreviations

ML	machine learning
DL	deep learning
RF	random forest
SVM	support vector machine
MSA	multiple sequence alignment
NN	neural network
CNN	convolutional neural network
GNN	graph neural network
GCN	graph convolutional network
MPNN	message passing neural network
GRU	gated recurrent unit
LSTM	long short-term memory
RMSD	root-mean-square deviation
MD	molecular dynamics
SS	secondary structure
ASA	accessible surface area
SASA	solvent accessible surface area
RSASA	relative solvent accessible surface area
PSSM	position specific scoring matrix
AF2	AlphaFold2
ATP	adenosine triphosphate
RNA	ribonucleic acid
DNA	deoxyribonucleic acid
PDB	Protein Data Bank
NMR	nuclear magnetic resonance
pLM	protein language model

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/S003358352500006X.

Acknowledgements. This study was done within the PROXIDRUGS consortium. PROXIDRUGS is as part of the initiative 'Clusters4Future' funded by the Federal Ministry of Education and Research BMBF (03ZU2109JE; 03ZU2109ID).

Competing interests. The authors declare no competing interests.

References

Abdin O, ***et al.*** (2022) PepNN: a deep attention model for the identification of peptide binding sites. *Communications Biology* **5**, 503.

Abramson J, ***et al.*** (2024) Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, **630**(8016), 493–500.

Adamczak R, **Porollo A and Meller J** (2004) Accurate prediction of solvent accessibility using neural networks–based regression. *Proteins: Structure, Function, and Bioinformatics* **56**, 753–767.

Aggarwal R, ***et al.*** (2021) DeepPocket: ligand binding site detection and segmentation using 3D convolutional neural networks. *Journal of Chemical Information and Modeling* **62**, 5069–5079.

Ahmad S, **Gromiha MM and Sarai A** (2003) Real value prediction of solvent accessibility from amino acid sequence. *Proteins: Structure, Function, and Bioinformatics* **50**, 629–635.

Ajitha M, ***et al.*** (2018) Development of metal-active site and Zinccluster tool to predict active site pockets. *Proteins: Structure, Function, and Bioinformatics* **86**, 322–331.

Alberts B (2017) *Molecular Biology of the Cell*. New York: Garland Science.

Al-Fartusie FS and Mohssan SN (2017) Essential trace elements and their vital roles in the human body. *Indian Journal of Advanced Chemical Sciences* **5**, 127–136.

Alipanahi B, ***et al.*** (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology* **33**, 831–838.

An J, **Totrov M and Abagyan R** (2004) Comprehensive identification of 'druggable' protein ligand binding sites. *Genome Informatics* **15**, 31–41.

An J, **Totrov M and Abagyan R** (2005) Pocketome via comprehensive identification and classification of ligand binding envelopes. *Molecular & Cellular Proteomics* **4**, 752–761.

Andreini C, ***et al.*** (2008) Metal ions in biological catalysis: from enzyme databases to general principles. *JBIC Journal of Biological Inorganic Chemistry* **13**, 1205–1218.

Andreini C, ***et al.*** (2012) MetalPDB: a database of metal sites in biological macromolecular structures. *Nucleic Acids Research* **41**, D312–D319.

Armon A, **Graur D and Ben-Tal N** (2001) ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *Journal of Molecular Biology* **307**, 447–463.

Auld DS (2001) Zinc coordination sphere in biochemical zinc sites. In Maret, W. *Zinc Biochemistry, Physiology, and Homeostasis*, Dordrecht: Springer pp. 85–127.

Baek M, ***et al.*** (2021) Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876.

Bagley SC and Altman RB (1995) Characterizing the microenvironment surrounding protein sites. *Protein Science* **4**, 622–635.

Bagley SC and Altman RB (1996) Conserved features in the active site of nonhomologous serine proteases. *Folding and Design* **1**, 371–379.

Barker JA and Thornton JM (2003) An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. *Bioinformatics* **19**, 1644–1649.

Békés M, **Langley DR and Crews CM** (2022) PROTAC targeted protein degraders: the past is prologue. *Nature Reviews Drug Discovery* **21**, 181–200.

Benet LZ, ***et al.*** (2016) BDDCS, the Rule of 5 and drugability. *Advanced Drug Delivery Reviews* **101**, 89–98.

Bhatt R, **Koes DR and Durrant JD** (2024) CENsible: interpretable insights into small-molecule binding with context explanation networks. *Journal of Chemical Information and Modeling* **64**, 4651–4660.

Bhinge A, ***et al.*** (2004) Accurate detection of protein: ligand binding sites using molecular dynamics simulations. *Structure* **12**, 1989–1999.

Binkowski TA, **Naghibzadeh S and Liang J** (2003) CASTp: computed atlas of surface topography of proteins. *Nucleic Acids Research* **31**, 3352–3355.

Bock ME, **Garutti C and Guerra C** (2007) Effective labeling of molecular surface points for cavity detection and location of putative binding sites. In Markstein, Peter and Xu, Ying *Computational Systems Bioinformatics: (Volume 6)*, pp. 263–274. Singapore: World Scientific.

Boike L, **Henning NJ and Nomura DK** (2022) Advances in covalent drug discovery. *Nature Reviews Drug Discovery* **21**, 881–898.

Bondugula R and Xu D (2007) MUPRED: a tool for bridging the gap between template-based methods and sequence profile-based methods for protein secondary structure prediction. *Proteins: Structure, Function, and Bioinformatics* **66**, 664–670.

Bordner AJ (2008) Predicting small ligand binding sites in proteins using backbone structure. *Bioinformatics* **24**, 2865–2871.

Bradford JR and Westhead DR (2005) Improved prediction of protein–protein binding sites using a support vector machines approach. *Bioinformatics* **21**, 1487–1494.

Brady GP and Stouten PFW (2000) Fast prediction and visualization of protein binding pockets with PASS. *Journal of Computer-Aided Molecular Design* **14**, 383–401.

Brandes N, ***et al.*** (2022) ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics* **38**, 2102–2110.

Brenke R, ***et al.*** (2009) Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques. *Bioinformatics* **25**, 621–627.

Bron C and Kerbosch J (1973) Algorithm 457: finding all cliques of an undirected graph. *Communications of the ACM* **16**, 575–577.

Brooks BR, ***et al.*** (1983) CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry* **4**, 187–217.

Broomhead NK and Soliman ME (2017) Can we rely on computational predictions to correctly identify ligand binding sites on novel protein drug targets? Assessment of binding site prediction methods and a protocol for validation of predicted binding sites. *Cell Biochemistry and Biophysics* **75**, 15–23.

Brylinski M and Skolnick J (2008) A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proceedings of the National Academy of Sciences* **105**, 129–134.

Brylinski M and Skolnick J (2011) FINDSITE-Metal: integrating evolutionary information and machine learning for structure-based metal-binding site prediction at the proteome level. *Proteins: Structure, Function, and Bioinformatics* **79**, 735–751.

Buchwald P (2010) Small-molecule protein–protein interaction inhibitors: therapeutic potential in light of molecular size, chemical space, and ligand binding efficiency considerations. *IUBMB Life* **62**, 724–731.

Canner SW, **Shanker S and Gray JJ** (2023) Structure-based neural network protein–carbohydrate interaction predictions at the residue level. *Frontiers in Bioinformatics* **3**, 1186531.

Capra JA, ***et al.*** (2009) Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Computational Biology* **5**, e1000585.

Chakrabarti P (1993) Anion binding sites in protein structures. *Journal of Molecular Biology* **234**, 463–482.

Chalmers MJ, ***et al.*** (2006) Probing protein ligand interactions by automated hydrogen/deuterium exchange mass spectrometry. *Analytical Chemistry* **78**, 1005–1014.

Chandra A, ***et al.*** (2023) PepCNN deep learning tool for predicting peptide binding residues in proteins using sequence, structural, and language model features. *Scientific Reports* **13**, 20882.

Chang L and Perez A (2023) Ranking peptide binders by affinity with AlphaFold. *Angewandte Chemie* **135**, e202213362.

Changeux JP (2013) The concept of allosteric modulation: an overview. *Drug Discovery Today: Technologies* **10**, e223–e228.

Changeux JP (2018) The nicotinic acetylcholine receptor: a typical 'allosteric machine'. *Philosophical Transactions of the Royal Society, B: Biological Sciences* **373**, 20170174.

Changeux JP and Christopoulos A (2016) Allosteric modulation as a unifying mechanism for receptor function and regulation. *Cell* **166**, 1084–1102.

Chauhan JS, **Mishra NK and Raghava GPS** (2009) Identification of ATP binding residues of a protein from its primary sequence. *BMC Bioinformatics* **10**, 1–9.

Chen K, ***et al.*** (2011) A critical comparative assessment of predictions of protein-binding sites for biologically relevant organic compounds. *Structure* **19**, 613–621.

Chen K, **Mizianty MJ and Kurgan L** (2012) Prediction and analysis of nucleotide-binding residues using sequence and sequence-derived structural descriptors. *Bioinformatics* **28**, 331–341.

Chen P, ***et al.*** (2015) A sequence-based dynamic ensemble learning system for protein ligand-binding site prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **13**, 901–912.

Chen P, **Huang JZ and Gao X** (2014) LigandRFs: random forest ensemble to identify ligand-binding residues from sequence information alone. In *BMC Bioinformatics* **15**, 1–12.

Chen S, ***et al.*** (2024) RNA-binding small molecules in drug discovery and delivery: an overview from fundamentals. *Journal of Medicinal Chemistry* **67**, 16002–16017.

Chen T and Guestrin C (2016) XGBoost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Chen Z, ***et al.*** (2019) D3Pockets: a method and web server for systematic analysis of protein pocket dynamics. *Journal of Chemical Information and Modeling* **59**, 3353–3358.

Chen Z, ***et al.*** (2013) ZincExplorer: an accurate hybrid method to improve the prediction of zinc-binding sites from protein sequences. *Molecular BioSystems* **9**, 2213–2222.

Cheng H, **Sen TZ**, **Jernigan RL and Kloczkowski A** (2007) Consensus data mining (CDM) protein secondary structure prediction server: combining GOR V and fragment database mining (FDM). *Bioinformatics* **23**, 2628–2630.

Choi D, ***et al.*** (2017) Predicting protein-binding regions in RNA using nucleotide profiles and compositions. *BMC Systems Biology* **11**, 1–12.

Choi S and Han K (2013) Predicting protein-binding RNA nucleotides using the feature-based removal of data redundancy and the interaction propensity of nucleotide triplets. *Computers in Biology and Medicine* **43**, 1687–1697.

Clackson T and Wells JA (1995) A hot spot of binding energy in a hormone-receptor interface. *Science* **267**, 383–386.

Coleman RG and Sharp KA (2006) Travel depth, a new shape descriptor for macromolecules: application to ligand binding. *Journal of Molecular Biology* **362**, 441–458.

Consortium U (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research* **47**, D506–D515.

Cortes C and Vapnik VN (1995) Support-vector networks. *Machine Learning* **20**, 273–297.

Craig IR, ***et al.*** (2011) Pocket-space maps to identify novel binding-site conformations in proteins. *Journal of Chemical Information and Modeling* **51**, 2666–2679.

Cui Y, ***et al.*** (2019) Predicting protein-ligand binding residues with deep convolutional neural networks. *BMC Bioinformatics* **20**, 1–12.

Dai B and Bailey-Kellogg C (2021) Protein interaction interface region prediction by geometric deep learning. *Bioinformatics* **37**, 2580–2588.

Dana JM, ***et al.*** (2019) SIFTS: updated structure integration with function, taxonomy and sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. *Nucleic Acids Research* **47**, D482–D489.

De Baaij JHF, **Hoenderop GJ and Bindels RJM** (2015) Magnesium in man: implications for health and disease. *Physiological Reviews* **95**(1) 1–46

Delaney JS (1992) Finding and filling protein cavities using cellular logic operations. *Journal of Molecular Graphics* **10**, 174–177.

Del Carpio CA, **Takahashi Y and Sasaki SI** (1993) A new approach to the automatic identification of candidates for ligand receptor sites in proteins: (I) search for pocket regions. *Journal of Molecular Graphics* **11**, 23–29.

Denessiouk KA, **Rantanen VV and Johnson MS** (2001) Adenine recognition: a motif present in ATP-, CoA-, NAD-, NADP-, and FAD-dependent proteins. *Proteins: Structure, Function, and Bioinformatics* **44**, 282–291.

Desaphy J, ***et al.*** (2015) sc-PDB: a 3D-database of ligandable binding sites—10 years on. *Nucleic Acids Research* **43**, D399–D404.

Devlin J, ***et al.*** (2018) *BERT: pre-training of deep bidirectional transformers for language understanding.* arXiv preprint.arXiv:1810.04805.

Devos D and Valencia A (2000) Practical limits of function prediction. *Proteins: Structure, Function, and Bioinformatics* **41**, 98–107.

Dewey JA, ***et al.*** (2023) Molecular glue discovery: current and future approaches. *Journal of Medicinal Chemistry* **66**, 9278–9296.

Ding P, ***et al.*** (2023) DeepSTF: predicting transcription factor binding sites by interpretable deep neural networks combining sequence and shape. *Briefings in Bioinformatics*, **24**(4) bbad231.

Ding Y, **Tang J and Guo F** (2017) Identification of protein–ligand binding sites by sequence information and ensemble classifier. *Journal of Chemical Information and Modeling* **57**, 3149–3161.

Di Pietro O, ***et al.*** (2017) Unveiling a novel transient druggable pocket in BACE-1 through molecular simulations: conformational analysis and binding mode of multisite inhibitors. *PLoS One* **12**, e0177683.

Dor O and Zhou Y (2007) Real-SPINE: an integrated system of neural networks for real-value prediction of protein structural properties. *Proteins: Structure, Function, and Bioinformatics* **68**, 76–81.

Dou Y, ***et al.*** (2012) L1pred: a sequence-based prediction tool for catalytic residues in enzymes with the L1-Logreg classifier. *PLoS One* **7**, e35666.

Doyle SK, ***et al.*** (2016) Advances in discovering small molecules to probe protein function in a systems context. *Current Opinion in Chemical Biology* **30**, 28–36.

Du H, ***et al.*** (2021) CovalentInDB: a comprehensive database facilitating the discovery of covalent inhibitors. *Nucleic Acids Research* **49**, D1122–D1129.

Du H, ***et al.*** (2022) Proteome-wide profiling of the covalent-druggable cysteines with a structure-based deep graph learning network. *Research* **2022**, 9873564.

Dudev T and Lim C (2014) Competition among metal ions for protein binding sites: determinants of metal ion selectivity in proteins. *Chemical Reviews* **114**, 538–556.

Dürr SL, **Levy A and Rothlisberger U** (2023) Metal3D: a general deep learning framework for accurate metal ion location prediction in proteins. *Nature Communications* **14**, 2713.

Durrant JD, ***et al.*** (2014) POVME 2.0: an enhanced tool for determining pocket shape and volume characteristics. *Journal of Chemical Theory and Computation* **10**, 5047–5056.

Ebert JC and Altman RB (2008) Robust recognition of zinc binding sites in proteins. *Protein Science* **17**, 54–65.

Eguida M and Rognan D (2022) Estimating the similarity between protein pockets. *International Journal of Molecular Sciences* **23**, 12462.

Elnaggar A, ***et al.*** (2021) ProtTrans: toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**, 7112–7127.

Essien C, ***et al.*** (2023) Prediction of protein ion–ligand binding sites with ELECTRA. *Molecules* **28**, 6793.

Essien C, **Wang D and Xu D** (2019) Capsule network for predicting zinc binding sites in metalloproteins. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2337–2341.

Everingham M, ***et al.*** (2010) The Pascal Visual Object Classes (VOC) challenge. *International Journal of Computer Vision* **88**, 303–338.

Evteev SA, **Ereshchenko AV and Ivanenkov YA** (2023) SiteRadar: utilizing graph machine learning for precise mapping of protein–ligand-binding sites. *Journal of Chemical Information and Modeling* **63**, 1124–1132.

Fadahunsi AA, ***et al.*** (2024) Revolutionizing drug discovery: an AI-powered transformation of molecular docking. *Medicinal Chemistry Research*, **33**(12) 2187–2203.

Falese JP, **Donlic A and Hargrove AE** (2021) Targeting RNA with small molecules: from fundamental principles towards the clinic. *Chemical Society Reviews* **50**, 2224–2243.

Faller CE, ***et al.*** (2015) Site identification by ligand competitive saturation (SILCS) simulations for fragment-based drug design. In Klon, Anthony E. *Fragment-Based Methods in Drug Discovery*, New York, NY: Springer New York 75–87.

Fang Y, ***et al.*** (2023) DeepProSite: structure-aware protein binding site prediction using ESMFold and pretrained language model. *Bioinformatics* **39**, btad718.

Faraggi E, ***et al.*** (2012) SPINE X: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. *Journal of Computational Chemistry* **33**, 259–267.

Ferrè F and Clote P (2006) DiANNA 1.1: an extension of the DiANNA web server for ternary cysteine classification. *Nucleic Acids Research* **34**, W182–W185.

Ferré S, ***et al.*** (2014) G protein–coupled receptor oligomerization revisited: functional and pharmacological perspectives. *Pharmacological Reviews* **66**, 413–434.

Fout A, ***et al.*** (2017) Protein interface prediction using graph convolutional networks. *Advances in Neural Information Processing Systems* **30**, 6530–6539.

Friedman N, **Geiger D and Goldszmidt M** (1997) Bayesian network classifiers. *Machine Learning* **29**, 131–163.

Fuller JC, **Burgoyne NJ and Jackson RM** (2009) Predicting druggable binding sites at the protein–protein interface. *Drug Discovery Today* **14**, 155–161.

Gagliardi L and Rocchia W (2023) SiteFerret: beyond simple pocket identification in proteins. *Journal of Chemical Theory and Computation* **19**, 5242–5259.

Gainza P, ***et al.*** (2020) Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods* **17**, 184–192.

Gallo Cassarino T, **Bordoli L and Schwede T** (2014) Assessment of ligand binding site predictions in CASP10. *Proteins: Structure, Function, and Bioinformatics* **82**, 154–163.

Gamouh H, **Hoksza D and Novotny M** (2023) *Hybrid protein-ligand binding residue prediction with protein language models: does the structure matter?* bioRxiv, *2023–08*.

Gao H and Ji S (2019) Graph U-Nets. In *International Conference on Machine Learning*, 2083–2092. PMLR.

Gao J, ***et al.*** (2016) BSiteFinder, an improved protein-binding sites prediction server based on structural alignment: more accurate and less time-consuming. *Journal of Cheminformatics* **8**, 1–10.

Gao M and Günther S (2023) HyperCys: a structure- and sequence-based predictor of hyper-reactive druggable cysteines. *International Journal of Molecular Sciences* **24**, 5960.

Gao M, ***et al.*** (2022) CovPDB: a high-resolution coverage of the covalent protein–ligand interactome. *Nucleic Acids Research* **50**, D445–D450.

Gao Z, ***et al.*** (2023) Hierarchical graph learning for protein–protein interaction. *Nature Communications* **14**, 1093.

Garg A, **Kaur H and Raghava GPS** (2005) Real value prediction of solvent accessibility in proteins using multiple sequence alignment and secondary structure. *Proteins: Structure, Function, and Bioinformatics* **61**, 318–324.

Garg A and Pal D (2021) Inferring metal binding sites in flexible regions of proteins. *Proteins: Structure, Function, and Bioinformatics* **89**(9) 1125–1133.

Gazizov A, ***et al.*** (2023) *AF2BIND: predicting ligand-binding sites using the pair representation of AlphaFold2.* bioRxiv, *2023–10*.

Gelis L, ***et al.*** (2012) Prediction of a ligand-binding niche within a human olfactory receptor by combining site-directed mutagenesis with dynamic homology modeling. *Angewandte Chemie International Edition* **51**, 1274–1278.

Ghersi D and Sanchez R (2009) EasyMIFS and SiteHound: a toolkit for the identification of ligand-binding sites in protein structures. *Bioinformatics* **25**, 3185–3186.

Gilson MK, ***et al.*** (2016) BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Research* **44**, D1045–D1053.

Glaser F, ***et al.*** (2006) A method for localizing ligand binding pockets in protein structures. *Proteins: Structure, Function, and Bioinformatics* **62**, 479–488.

Gold ND and Jackson RM (2006) SitesBase: a database for structure-based protein–ligand binding site comparisons. *Nucleic Acids Research* **34**, D231–D234.

Goodford PJ (1985) A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *Journal of Medicinal Chemistry* **28**, 849–857.

Gu L, **Li B and Ming D** (2022) A multilayer dynamic perturbation analysis method for predicting ligand–protein interactions. *BMC Bioinformatics* **23**, 456.

Guerra JVdS, ***et al.*** (2021) pyKVFinder: an efficient and integrable Python package for biomolecular cavity detection and characterization in data science. *BMC Bioinformatics* **22**, 607.

Gutteridge A, **Bartlett GJ and Thornton JM** (2003) Using a neural network and spatial clustering to predict the location of active sites in enzymes. *Journal of Molecular Biology* **330**, 719–734.

Haberal I and Oğul H (2019) Prediction of protein metal binding sites using deep neural networks. *Molecular Informatics* **38**, 1800169.

Haberal I and Oğul H (2017) DeepMBS: prediction of protein metal binding-site using deep learning networks. In *2017 Fourth International Conference on Mathematics and Computers in Sciences and Industry (MCSI)*, 21–25.

Halgren TA (2009) Identifying and characterizing binding sites and assessing druggability. *Journal of Chemical Information and Modeling* **49**, 377–389.

Hardy JA and Wells JA (2004) Searching for new allosteric sites in enzymes. *Current Opinion in Structural Biology* **14**, 706–715.

Hartshorn MJ, ***et al.*** (2007) Diverse, high-quality test set for the validation of protein–ligand docking performance. *Journal of Medicinal Chemistry* **50**, 726–741.

Heffernan R, ***et al.*** (2015) Improving prediction of secondary structure, local backbone angles and solvent accessible surface area of proteins by iterative deep learning. *Scientific Reports* **5**, 11476.

Hendlich M, **Rippmann F and Barnickel G** (1997) LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *Journal of Molecular Graphics and Modelling* **15**, 359–363.

Hendrix SG, ***et al.*** (2021) DeepDISE: DNA binding site prediction using a deep learning method. *International Journal of Molecular Sciences* **22**, 5510.

Henrich S, ***et al.*** (2010) Computational approaches to identifying and characterizing protein binding sites for ligand design. *Journal of Molecular Recognition: An Interdisciplinary Journal* **23**, 209–219.

Hernandez M, **Ghersi D and Sanchez R** (2009) SITEHOUND-Web: a server for ligand binding site identification in protein structures. *Nucleic Acids Research* **37**, W413–W416.

Ho CMW and Marshall GR (1990) Cavity search: an algorithm for the isolation and display of cavity-like binding regions. *Journal of Computer-Aided Molecular Design* **4**, 337–354.

Ho QT, ***et al.*** (2021) FAD-BERT: improved prediction of FAD binding sites using pre-training of deep bidirectional transformers. *Computers in Biology and Medicine* **131**, 104258.

Ho TK (1995) Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition* **1**, 278–282. IEEE.

https://doi.org/10.1017/S003358352500006X

Hopkins AL and Groom CR (2002) The druggable genome. *Nature Reviews Drug Discovery* **1**, 727.
Hu X, ***et al.*** (2016) Recognizing metal and acid radical ion-binding sites by integrating ab initio modeling with template-based transferals. *Bioinformatics* **32**, 3260–3269.
Huang B and Schroeder M (2006) LIGSITE CSC: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Structural Biology* **6**, 1–11.
Huang G, ***et al.*** (2017) Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4700–4708.
Huey R, ***et al.*** (2007) A semiempirical free energy force field with charge-based desolvation. *Journal of Computational Chemistry* **28**, 1145–1152.
Humphrey W, **Dalke A and Schulten K** (1996) VMD: visual molecular dynamics. *Journal of Molecular Graphics* **14**, 33–38.
Ingraham J, ***et al.*** (2019) Generative models for graph-based protein design. *Advances in Neural Information Processing Systems* **32**, 15820–15831.
Iqbal S and Hoque MT (2018) PBRpredict-suite: a suite of models to predict peptide-recognition domain residues from protein sequence. *Bioinformatics* **34**, 3289–3299.
Ireland SM and Martin ACR (2019) ZincBind—the database of zinc binding sites. *Database* **2019**, baz006.
Ireland SM and Martin ACR (2021) ZincBindPredict—prediction of zinc binding sites in proteins. *Molecules* **26**, 966.
Jernigan RL, **Raghunathan G and Bahar I** (1994) Characterization of interactions and metal ion binding sites in proteins. *Current Opinion in Structural Biology* **4**, 256–263.
Jha K, **Karmakar S and Saha S** (2023) Graph-BERT and language model-based framework for protein–protein interaction identification. *Scientific Reports* **13**, 5663.
Jian JW, ***et al.*** (2016) Predicting ligand binding sites on protein surfaces by 3-dimensional probability density distributions of interacting atoms. *PLoS One* **11**, e0160315.
Jiang M, ***et al.*** (2019a) A novel protein descriptor for the prediction of drug binding sites. *BMC Bioinformatics* **20**, 1–13.
Jiang M, ***et al.*** (2019b) FrSite: protein drug binding site prediction based on Faster R–CNN. *Journal of Molecular Graphics and Modelling* **93**, 107454.
Jiang Z, ***et al.*** (2016) Identification of Ca^{2+}-binding residues of a protein from its primary sequence. *Genetics and Molecular Research* **15**(2), gmr.15027618.
Jiménez J, ***et al.*** (2017) DeepSite: protein-binding site predictor using 3D-convolutional neural networks. *Bioinformatics* **33**, 3036–3042.
Jiménez-Luna J, **Grisoni F and Schneider G** (2020) Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence* **2**, 573–584.
Jing B, ***et al.*** (2020) *Learning from protein structure with geometric vector perceptrons.* arXiv preprint.arXiv:2009.01411.
Johansson-Åkhe I, **Mirabello C and Wallner B** (2019) Predicting protein-peptide interaction sites using distant protein complexes as structural templates. *Scientific Reports* **9**, 1–13.
Jolma A, ***et al.*** (2010) Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Research* **20**, 861–873.
Jumper J, ***et al.*** (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589.
Kalicki CH and Haritaoglu ED (2022) *RNABERT: RNA family classification and secondary structure prediction with BERT pretrained on RNA sequences.*
Kandel J, **Tayara H and Chong KT** (2021) PUResNet: prediction of protein-ligand binding sites using deep residual neural network. *Journal of Cheminformatics* **13**, 1–14.
Kapoor S and Narayanan A (2023) Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* **4**(9), 100804.
Katritch V, ***et al.*** (2014) Allosteric sodium in class A GPCR signaling. *Trends in Biochemical Sciences* **39**, 233–244.
Kauffman C and Karypis G (2009) LIBRUS: combined machine learning and homology information for sequence-based ligand-binding residue prediction. *Bioinformatics* **25**, 3099–3107.
Kawabata T (2010) Detection of multiscale pockets on protein surfaces using mathematical morphology. *Proteins: Structure, Function, and Bioinformatics* **78**, 1195–1211.
Kawabata T and Go N (2007) Detection of pockets on protein surfaces using small and large probe spheres to find putative ligand binding sites. *Proteins: Structure, Function, and Bioinformatics* **68**, 516–529.
Kawashima S, ***et al.*** (2007) AAindex: amino acid index database, progress report 2008. *Nucleic Acids Research* **36**, D202–D205.
Kharchenko PV, **Tolstorukov MY and Park PJ** (2008) Design and analysis of ChIP-Seq experiments for DNA-binding proteins. *Nature Biotechnology* **26**, 1351–1359.
Kim D, ***et al.*** (2008) Pocket extraction on proteins via the Voronoi diagram of spheres. *Journal of Molecular Graphics and Modelling* **26**, 1104–1112.
Kinoshita K and Nakamura H (2005) Identification of the ligand binding sites on the molecular surface of proteins. *Protein Science* **14**, 711–718.
Kleywegt GJ and Jones TA (1994) Detection, delineation, measurement and display of cavities in macromolecular structures. *Acta Crystallographica Section D: Biological Crystallography* **50**, 178–185.
Knox C, ***et al.*** (2024) DrugBank 6.0: the DrugBank knowledgebase for 2024. *Nucleic Acids Research* **52**, D1265–D1275.
Kognole AA, **Hazel A and MacKerell AD Jr** (2022) SILCS-RNA: toward a structure-based drug design approach for targeting RNAs with small molecules. *Journal of Chemical Theory and Computation* **18**(9), 5672–5691.
Konc J and Janežič D (2010) ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment. *Bioinformatics* **26**, 1160–1168.
Kozlovskii I and Popov P (2020) Spatiotemporal identification of druggable binding sites using deep learning. *Communications Biology* **3**, 1–12.
Kozlovskii I and Popov P (2021a) Protein–peptide binding site detection using 3D convolutional neural networks. *Journal of Chemical Information and Modeling* **61**, 3814–3823.
Kozlovskii I and Popov P (2021b) Structure-based deep learning for binding site detection in nucleic acid macromolecules. *NAR Genomics and Bioinformatics* **3**, lqab111.
Krapp LF, ***et al.*** (2023) PeSTo: parameter-free geometric deep learning for accurate prediction of protein binding interfaces. *Nature Communications* **14**, 2175.
Krishna R, ***et al.*** (2024) Generalized biomolecular modeling and design with RosettaFold all-atom. *Science* **384**, eadl2528.
Krivák R and Hoksza D (2018) P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *Journal of Cheminformatics* **10**, 1–12.
Krivák R, **Jendele L and Hoksza D** (2018) Peptide-binding site prediction from protein structure via points on the solvent accessible surface. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 645–650.
Kryshtafovych A, ***et al.*** (2021) Critical assessment of methods of protein structure prediction (CASP)—round XIV. *Proteins: Structure, Function, and Bioinformatics* **89**, 1607–1617.
Kusuma RMI, ***et al.*** (2019) Prediction of ATP-binding sites in membrane proteins using a two-dimensional convolutional neural network. *Journal of Molecular Graphics and Modelling* **92**, 86–93.
Laskowski RA (1995) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *Journal of Molecular Graphics* **13**, 323–330.
Laskowski RA, **Gerick F and Thornton JM** (2009) The structural basis of allosteric regulation in proteins. *FEBS Letters* **583**, 1692–1698.
Laurent B, ***et al.*** (2015) Epock: rapid analysis of protein pocket dynamics. *Bioinformatics* **31**, 1478–1480.
Laurie ATR and Jackson RM (2005) Q-SiteFinder: an energy-based method for the prediction of protein–ligand binding sites. *Bioinformatics* **21**, 1908–1916.
Laurie AT and Jackson RM (2006) Methods for the prediction of protein-ligand binding sites for structure-based drug design and virtual ligand screening. *Current Protein and Peptide Science* **7**, 395–406.
Laveglia V, ***et al.*** (2023) Hunting down zinc(II)-binding sites in proteins with distance matrices. *Bioinformatics* **39**, btad653.
Lavi A, ***et al.*** (2013) Detection of peptide-binding sites on protein surfaces: the first step toward the modeling and targeting of peptide-mediated interactions. *Proteins: Structure, Function, and Bioinformatics* **81**, 2096–2105.
Lawson ADG (2012) Antibody-enabled small-molecule drug discovery. *Nature Reviews Drug Discovery* **11**, 519.

https://doi.org/10.1017/S003358352500006X

Lee HS and Im W (2013) Ligand binding site detection by local structure alignment and its performance complementarity. *Journal of Chemical Information and Modeling* **53**, 2462–2470.
Lee I and Nam H (2022) Sequence-based prediction of protein binding regions and drug–target interactions. *Journal of Cheminformatics* **14**, 1–15.
Lee PH and Helms V (2012) Identifying continuous pores in protein structures with PROPORES by computational repositioning of gating residues. *Proteins: Structure, Function, and Bioinformatics* **80**, 421–432.
Le Guilloux V, **Schmidtke P and Tuffery P** (2009) Fpocket: an open-source platform for ligand pocket detection. *BMC Bioinformatics* **10**, 1–11.
Leis S, **Schneider S and Zacharias M** (2010) In silico prediction of binding sites on proteins. *Current Medicinal Chemistry* **17**, 1550–1562.
Levitt DG and Banaszak LJ (1992) POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *Journal of Molecular Graphics* **10**, 229–234.
Li H, **Pi D and Chen C** (2019a) An improved prediction model for zinc-binding sites in proteins based on Bayesian method. *Tehnički Vjesnik* **26**, 1422–1426.
Li H, ***et al.*** (2019b) A novel prediction method for zinc-binding sites in proteins by an ensemble of SVM and sample-weighted probabilistic neural network. *IEEE Access* **7**, 186147–186157.
Li K, ***et al.*** (2023a) Simultaneous prediction of interaction sites on the protein and peptide sides of complexes through multilayer graph convolutional networks. *Journal of Chemical Information and Modeling* **63**, 2251–2262.
Li P, ***et al.*** (2022) RecurPocket: recurrent LMSER network with gating mechanism for protein binding site detection. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 334–339. IEEE.
Li P, ***et al.*** (2023b) GLPocket: a multi-scale representation learning approach for protein binding site prediction. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 4821–4828.
Li S, ***et al.*** (2023c) PocketAnchor: learning structure-based pocket representations for protein-ligand interaction prediction. *Cell Systems* **14**, 692–705.
Li Y, ***et al.*** (2023d) *MsPBRsP: multi-scale protein binding residues prediction using language model.* bioRxiv, *2023–02*.
Liang J, **Woodward C and Edelsbrunner H** (1998) Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Science* **7**, 1884–1897.
Liang MP, ***et al.*** (2003) WebFEATURE: an interactive web tool for identifying and visualizing functional sites on macromolecular structures. *Nucleic Acids Research* **31**, 3324–3327.
Liang S, ***et al.*** (2006) Protein binding site prediction using an empirical scoring function. *Nucleic Acids Research* **34**, 3698–3707.
Liao J, ***et al.*** (2022) In silico methods for identification of potential active sites of therapeutic targets. *Molecules* **27**, 7103.
Lin HN, ***et al.*** (2005) HYPROSP II-A knowledge-based hybrid method for protein secondary structure prediction based on local prediction confidence. *Bioinformatics* **21**, 3227–3233.
Lin YF, ***et al.*** (2016) MIB: metal ion-binding site prediction and docking server. *Journal of Chemical Information and Modeling* **56**, 2287–2291.
Lin Z, ***et al.*** (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130.
Lippi M, ***et al.*** (2008) MetalDetector: a web server for predicting metal-binding sites and disulfide bridges in proteins from sequence. *Bioinformatics* **24**, 2094–2095.
Litfin T, **Yang Y and Zhou Y** (2019) Spot-Peptide: template-based prediction of peptide-binding proteins and peptide-binding sites. *Journal of Chemical Information and Modeling* **59**, 924–930.
Littmann M, ***et al.*** (2021) Protein embeddings and deep learning predict binding residues for various ligand classes. *Scientific Reports* **11**, 23916.
Liu H, ***et al.*** (2024) RNet: a network strategy to predict RNA binding preferences. *Briefings in Bioinformatics* **25**, bbad482.
Liu H, ***et al.*** (2020a) Illuminating the allosteric modulation of the calcium-sensing receptor. *Proceedings of the National Academy of Sciences* **117**, 21711–21722.
Liu X and Ciulli A (2023) Proximity-based modalities for biology and medicine. *ACS Central Science* **9**, 1269–1284.
Liu Y, ***et al.*** (2020b) CB-Dock: a web server for cavity detection-guided protein–ligand blind docking. *Acta Pharmacologica Sinica* **41**, 138–144.
Liu Y, ***et al.*** (2023) RefinePocket: an attention-enhanced and mask-guided deep learning approach for protein binding site prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **20**(5) 3314–3321.
Liu Y and Tian B (2023) *Protein-DNA binding sites prediction based on pre-trained protein language model and contrastive learning.* arXiv preprint. arXiv:2306.15912.
Liu Z, ***et al.*** (2014) Computationally characterizing and comprehensive analysis of zinc-binding sites in proteins. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* **1844**, 171–180.
Liu Z, ***et al.*** (2015) PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics* **31**, 405–412.
Lopez G, **Ezkurdia I and Tress ML** (2009) Assessment of ligand binding residue predictions in CASP8. *Proteins: Structure, Function, and Bioinformatics* **77**, 138–146.
López G, **Valencia A and Tress ML** (2007) Firestar—prediction of functionally important residues using structural templates and alignment reliability. *Nucleic Acids Research* **35**, W573–W577.
Lu C, ***et al.*** (2019) MPLs-Pred: predicting membrane protein-ligand binding sites using hybrid sequence-based features and ligand-specific models. *International Journal of Molecular Sciences* **20**, 3120.
Lu CH, ***et al.*** (2012) Prediction of metal ion–binding sites in proteins using the fragment transformation method. *PLoS One* **7**, e39252.
Lu S, ***et al.*** (2018) Discovery of hidden allosteric sites as novel targets for allosteric drug design. *Drug Discovery Today* **23**, 359–365.
Ludlow RF, ***et al.*** (2015) Detection of secondary binding sites in proteins using fragment screening. *Proceedings of the National Academy of Sciences* **112**, 15910–15915.
Lundberg SM and Lee SI (2017) A unified approach to interpreting model predictions. In Guyon I, Von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S and Garnett R (eds.), *Advances in Neural Information Processing Systems*, Vol. **30**. Red Hook, NY, USA: Curran Associates, Inc.
Luo J, ***et al.*** (2017) RPI-Bind: a structure-based method for accurate identification of RNA-protein binding sites. *Scientific Reports* **7**, 614.
Lv N and Cao Z (2024) Subpocket-based analysis approach for the protein pocket dynamics. *Journal of Chemical Theory and Computation* **20**, 4909–4920.
Lyons J, ***et al.*** (2014) Predicting backbone Cα angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network. *Journal of Computational Chemistry* **35**, 2040–2046.
Mallet V, ***et al.*** (2022) InDeep: 3D fully convolutional neural networks to assist in silico drug design on protein–protein interactions. *Bioinformatics* **38**, 1261–1268.
Marchand JR, ***et al.*** (2021) CAVIAR: a method for automatic cavity detection, description and decomposition into subcavities. *Journal of Computer-Aided Molecular Design* **35**, 737–750.
Martinez-Rosell G, ***et al.*** (2020) PlayMolecule CrypticScout: predicting protein cryptic sites using mixed-solvent molecular simulations. *Journal of Chemical Information and Modeling* **60**, 2314–2324.
Masuya M and Doi J (1995) Detection and geometric modeling of molecular surfaces and cavities using digital mathematical morphological operations. *Journal of Molecular Graphics* **13**, 331–336.
Matsui M and Corey DR (2017) Non-coding RNAs as drug targets. *Nature Reviews Drug Discovery* **16**, 167–179.
McClorey G and Wood MJ (2015) An overview of the clinical application of antisense oligonucleotides for RNA-targeting therapies. *Current Opinion in Pharmacology* **24**, 52–58.
McGreig JE, ***et al.*** (2022) 3DLigandSite: structure-based prediction of protein–ligand binding sites. *Nucleic Acids Research* **50**, W13–W20.
Meller A, ***et al.*** (2023) Predicting the locations of cryptic pockets from single protein structures using the PocketMiner graph neural network. *Biophysical Journal* **122**, 445a.
Ming D and Wall ME (2006) Interactions in native binding sites cause a large change in protein dynamics. *Journal of Molecular Biology* **358**, 213–223.
Möller L, ***et al.*** (2022) Translating from proteins to ribonucleic acids for ligand-binding site detection. *Molecular Informatics* **41**, 2200059.
Montavon G, ***et al.*** (2019) Layer-wise relevance propagation: an overview. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Cham, Switzerland: Springer International Publishing 193–209.

Moodie SL, **Mitchell JBO and Thornton JM** (1996) Protein recognition of adenylate: an example of a fuzzy recognition template. *Journal of Molecular Biology* **263**, 486–500.

Morris GM, ***et al.*** (2009) AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *Journal of Computational Chemistry* **30**, 2785–2791.

Mukherjee S, ***et al.*** (2004) Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nature Genetics* **36**, 1331–1339.

Murdoch WJ, ***et al.*** (2019) Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences* **116**, 22071–22080.

Mylonas SK, **Axenopoulos A and Daras P** (2021) DeepSurf: a surface-based deep learning approach for the prediction of ligand binding sites on proteins. *Bioinformatics* **37**, 1681–1690.

Nayal M and Honig B (2006) On the nature of cavities on protein surfaces: application to the identification of drug-binding sites. *Proteins: Structure, Function, and Bioinformatics* **63**, 892–906.

Naz F, ***et al.*** (2015) Designing new kinase inhibitor derivatives as therapeutics against common complex diseases: structural basis of microtubule affinity-regulating kinase 4 (MARK4) inhibition. *OMICS: A Journal of Integrative Biology* **19**, 700–711.

Nazem F, ***et al.*** (2021) 3D U-Net: a voxel-based method in binding site prediction of protein structure. *Journal of Bioinformatics and Computational Biology* **19**, 2150006.

Nazem F, ***et al.*** (2023) A GU-Net-based architecture predicting ligand–protein-binding atoms. *Journal of Medical Signals and Sensors* **13**, 1.

Nemethy G, ***et al.*** (1992) Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. *The Journal of Physical Chemistry* **96**, 6472–6484.

Ngan CH, ***et al.*** (2012) FTSite: high accuracy detection of ligand binding sites on unbound protein structures. *Bioinformatics* **28**, 286–287.

Nimrod G, ***et al.*** (2008) Detection of functionally important regions in 'hypothetical proteins' of known structure. *Structure* **16**, 1755–1763.

Ortiz de Luzuriaga I, **Lopez X and Gil A** (2021) Learning to model G-quadruplexes: current methods and perspectives. *Annual Review of Biophysics* **50**, 209–243.

Paiva VA, ***et al.*** (2022) GASS-Metal: identifying metal-binding sites on protein structures using genetic algorithms. *Briefings in Bioinformatics* **23**, bbac178.

Pan X, ***et al.*** (2020) RBPsuite: RNA-protein binding sites prediction suite based on deep learning. *BMC Genomics* **21**, 1–8.

Panchal V and Brenk R (2021) Riboswitches as drug targets for antibiotics. *Antibiotics* **10**, 45.

Panchenko AR, **Kondrashov F and Bryant S** (2004) Prediction of functional sites by analysis of sequence and structure conservation. *Protein Science* **13**, 884–892.

Pandit SB and Skolnick J (2008) Fr-TM-Align: a new protein structural alignment method based on fragment alignments and the TM-score. *BMC Bioinformatics* **9**, 1–11.

Panei FP, **Gkeka P and Bonomi M** (2024) Identifying small-molecules binding sites in RNA conformational ensembles with Shaman. *Nature Communications* **15**, 5725.

Panwar B, **Gupta S and Raghava GPS** (2013) Prediction of vitamin interacting residues in a vitamin binding protein using evolutionary information. *BMC Bioinformatics* **14**, 1–14.

Panwar B and Raghava GPS (2015) Identification of protein-interacting nucleotides in an RNA sequence using composition profile of tri-nucleotides. *Genomics* **105**, 197–203.

Paramo T, ***et al.*** (2014) Efficient characterization of protein cavities within molecular simulation trajectories: Trj_cavity. *Journal of Chemical Theory and Computation* **10**, 2151–2164.

Park S and Seok C (2022) GalaxyWater-CNN: prediction of water positions on the protein structure by a 3D-convolutional neural network. *Journal of Chemical Information and Modeling* **62**, 3157–3168.

Passerini A, ***et al.*** (2007) Predicting zinc binding at the proteome level. *BMC Bioinformatics* **8**, 1–13.

Passerini A, **Lippi M and Frasconi P** (2011) MetalDetector V2.0: predicting the geometry of metal binding sites from protein sequence. *Nucleic Acids Research* **39**, W288–W292.

Passerini A, ***et al.*** (2006) Identifying cysteines and histidines in transition-metal-binding sites using support vector machines and neural networks. *Proteins: Structure, Function, and Bioinformatics* **65**, 305–316.

Pei J and Grishin NV (2004) Combining evolutionary and structural information for local protein structure prediction. *Proteins: Structure, Function, and Bioinformatics* **56**, 782–794.

Pei Q, ***et al.*** (2023) *FABind: fast and accurate protein-ligand binding.* arXiv preprint.arXiv:2310.06763.

Peters KP, **Fauck J and Frömmel C** (1996) The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. *Journal of Molecular Biology* **256**, 201–213.

Petrova NV and Wu CH (2006) Prediction of catalytic residues using support vector machine with selected protein sequence and structural properties. *BMC Bioinformatics* **7**, 1–12.

Petrovski ŽH, **Hribar-Lee B and Bosnić Z** (2022) CAT-Site: predicting protein binding sites using a convolutional neural network. *Pharmaceutics* **15**, 119.

Petukh M, **Kimmet T and Alexov E** (2013) BION web server: predicting non-specifically bound surface ions. *Bioinformatics* **29**, 805–806.

Philips A, ***et al.*** (2012) MetalionRNA: computational predictor of metal-binding sites in RNA structures. *Bioinformatics* **28**, 198–205.

Polizzi NF and DeGrado WF (2020) A defined structural unit enables de novo design of small-molecule–binding proteins. *Science* **369**, 1227–1233.

Popov P, ***et al.*** (2024) Unraveling viral drug targets: a deep learning-based approach for the identification of potential binding sites. *Briefings in Bioinformatics* **25**, bbad459.

Poulson BG, ***et al.*** (2020) Aggregation of biologically important peptides and proteins: inhibition or acceleration depending on protein and metal ion concentrations. *RSC Advances* **10**, 215–227.

Pravda L, ***et al.*** (2018) MOLEonline: a web-based tool for analyzing channels, tunnels and pores (2018 update). *Nucleic Acids Research* **46**, W368–W373.

Pupko T, ***et al.*** (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics* **18**, S71–S77.

Qiao L and Xie D (2019) MIonSite: ligand-specific prediction of metal ion-binding sites via enhanced Adaboost algorithm with protein sequence information. *Analytical Biochemistry* **566**, 75–88.

Qiao Z, ***et al.*** (2024) State-specific protein–ligand complex structure prediction with a multiscale deep generative model. *Nature Machine Intelligence* **6**, 195–208.

Qiu Z and Wang X (2011) Improved prediction of protein ligand-binding sites using random forests. *Protein and Peptide Letters* **18**, 1212–1218.

Ravindranath PA and Sanner MF (2016) AutoSite: an automated approach for pseudo-ligands prediction—from ligand-binding sites identification to predicting key ligand atoms. *Bioinformatics* **32**, 3142–3149.

Ray D, ***et al.*** (2009) Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nature Biotechnology* **27**, 667–670.

Ray D, ***et al.*** (2013) A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**, 172–177.

Redmon J, ***et al.*** (2016) You only look once: unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 779–788.

Rekand IH and Brenk R (2021) DrugPred_RNA—a tool for structure-based druggability predictions for RNA binding sites. *Journal of Chemical Information and Modeling* **61**(8), 4068–4081.

Ren S, ***et al.*** (2016) Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39**, 1137–1149.

Renaud N, ***et al.*** (2021) DeepRank: a deep learning framework for data mining 3D protein-protein interfaces. *Nature Communications* **12**, 7068.

Rinaldis M de, ***et al.*** (1998) Three-dimensional profiles: a new tool to identify protein surface similarities. *Journal of Molecular Biology* **284**, 1211–1221.

Rives A, ***et al.*** (2021) Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* **118**, e2016239118.

Roche DB, **Brackenridge DA and McGuffin LJ** (2015) Proteins and their interacting partners: an introduction to protein–ligand binding site prediction methods. *International Journal of Molecular Sciences* **16**, 29829–29842.

https://doi.org/10.1017/S003358352500006X

Roche R, ***et al.*** (2023) EquiPNAS: improved protein-nucleic acid binding site prediction using protein-language-model-informed equivariant deep graph neural networks. *bioRxiv* **2023**, 2023–09.

Rodríguez-Guerra Pedregal J, ***et al.*** (2017) GaudiMM: a modular multi-objective platform for molecular modeling. *Wiley Online Library* **38**(24), 2118–2126.

Rodwell VW, **Bender D and Botham KM** (2018) *Harper's Illustrated Biochemistry*. New York, NY, USA: McGraw-Hill.

Ronneberger O, **Fischer P and Brox T** (2015) U-Net: convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III *18*, 234–241.

Rosen M, ***et al.*** (1998) Molecular shape comparisons in searches for active sites and functional similarity. *Protein Engineering* **11**, 263–277.

Roy A, **Yang J and Zhang Y** (2012) COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Research* **40**, W471–W477.

Ruffner H, **Bauer A and Bouwmeester T** (2007) Human protein–protein interaction networks and the value for drug discovery. *Drug Discovery Today* **12**, 709–716.

Ruppert J, **Welch W and Jain AN** (1997) Automatic identification and representation of protein binding sites for molecular docking. *Protein Science* **6**, 524–533.

Sánchez-Aparicio JE, ***et al.*** (2020) BioMetAll: identifying metal-binding sites in proteins from backbone preorganization. *Journal of Chemical Information and Modeling* **61**, 311–323.

Santana CA, ***et al.*** (2020) GRaSP: a graph-based residue neighborhood strategy to predict binding sites. *Bioinformatics* **36**, i726–i734.

Satorras VG, **Hoogeboom E and Welling M** (2021) E(N) equivariant graph neural networks. In *International Conference on Machine Learning*, 9323–9332. PMLR.

Schmidt T, ***et al.*** (2011) Assessment of ligand-binding residue predictions in CASP9. *Proteins: Structure, Function, and Bioinformatics* **79**, 126–136.

Schmidtke P, ***et al.*** (2011) MDpocket: open-source cavity detection and characterization on molecular dynamics trajectories. *Bioinformatics* **27**, 3276–3285.

Schmidtke P, ***et al.*** (2010) Large-scale comparison of four binding site detection algorithms. *Journal of Chemical Information and Modeling* **50**, 2191–2200.

Schmitt S, **Kuhn D and Klebe G** (2002) A new method to detect related function among proteins independent of sequence and fold homology. *Journal of Molecular Biology* **323**, 387–406.

Schymkowitz J, ***et al.*** (2005a) The FoldX web server: an online force field. *Nucleic Acids Research* **33**, W382–W388.

Schymkowitz JWH, ***et al.*** (2005b) Prediction of water and metal binding sites and their affinities by using the Fold-X force field. *Proceedings of the National Academy of Sciences* **102**, 10147–10152.

Sciortino G, ***et al.*** (2019) Simple coordination geometry descriptors allow to accurately predict metal-binding sites in proteins. *ACS Omega* **4**, 3726–3731.

Segura J, **Jones PF and Fernandez-Fuentes N** (2012) A holistic in silico approach to predict functional sites in protein structures. *Bioinformatics* **28**, 1845–1850.

Selvaraju RR, ***et al.*** (2017) Grad-CAM: visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 618–626.

Seo S, ***et al.*** (2024) Pseq2Sites: enhancing protein sequence-based ligand binding-site prediction accuracy via the deep convolutional network and attention mechanism. *Engineering Applications of Artificial Intelligence* **127**, 107257.

Shafiee S, **Fathi A and Taherzadeh G** (2022) SPPPred: sequence-based protein-peptide binding residue prediction using genetic programming and ensemble learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **20**(3), 2029–2040.

Shashikala HB, ***et al.*** (2021) BION-2: predicting positions of non-specifically bound ions on protein surface by a Gaussian-based treatment of electrostatics. *International Journal of Molecular Sciences* **22**, 272.

Shenoy A, ***et al.*** (2024) M-Ionic: prediction of metal-ion-binding sites from sequence using residue embeddings. *Bioinformatics* **40**, btad782.

Shi W, ***et al.*** (2022) GraphSite: ligand binding site classification with deep graph learning. *Biomolecules* **12**, 1053.

Shu N, **Zhou T and Hovmöller S** (2008) Prediction of zinc-binding sites in proteins from sequence. *Bioinformatics* **24**, 775–782.

Shulman-Peleg A, **Nussinov R and Wolfson HJ** (2004) Recognition of functional sites in protein structures. *Journal of Molecular Biology* **339**, 607–633.

Simões T, ***et al.*** (2017) Geometric detection algorithms for cavities on protein surfaces in molecular graphics: a survey. *Computer Graphics Forum* **36**, 643–683.

Smith MC and Gestwicki JE (2012) Features of protein-protein interactions that translate into potent inhibitors: topology, surface area and affinity. *Expert Review of Molecular Medicine* **14**, e16.

Smith Z, ***et al.*** (2023) *Graph attention site prediction (Grasp): identifying druggable binding sites using graph neural networks with attention. bioRxiv*.

Sodhi JS, ***et al.*** (2004) Predicting metal-binding site residues in low-resolution structural models. *Journal of Molecular Biology* **342**, 307–320.

Song C and Jiang J (2023) A novel prediction method for metal-ion binding sites in protein sequence based on ensemble learning. In *Proceedings of the 2022 5th International Conference on Algorithms, Computing and Artificial Intelligence*. New York, NY, USA: Association for Computing Machinery, 1–7.

Song Y, ***et al.*** (2023) Accurately identifying nucleic-acid-binding sites through geometric graph learning on language model predicted structures. *Briefings in Bioinformatics* **24**, bbad360.

Southey MWY and Brunavs M (2023) Introduction to small molecule drug discovery and preclinical development. *Frontiers in Drug Discovery* **3**, 1314077.

Spriggs RV, **Artymiuk PJ and Willett P** (2003) Searching for patterns of amino acids in 3D protein structures. *Journal of Chemical Information and Computer Sciences* **43**, 412–421.

Srivastava A and Kumar M (2018) Prediction of zinc binding sites in proteins using sequence derived information. *Journal of Biomolecular Structure and Dynamics* **36**(16), 4413–4423.

Stark A, **Sunyaev S and Russell RB** (2003) A model for statistical significance of local similarities in structure. *Journal of Molecular Biology* **326**, 1307–1316.

Stepniewska-Dziubinska MM, **Zielenkiewicz P and Siedlecki P** (2020) Improving detection of protein-ligand binding sites with 3D segmentation. *Scientific Reports* **10**, 5035.

Stourac J, ***et al.*** (2019) Caver Web 1.0: identification of tunnels and channels in proteins and analysis of ligand transport. *Nucleic Acids Research* **47**, W414–W422.

Su H, **Peng Z and Yang J** (2021) Recognition of small molecule-RNA binding sites using RNA sequence and structure. *Bioinformatics* **37**(1), 36–42.

Sun K, ***et al.*** (2022) Predicting Ca2+ and Mg2+ ligand binding sites by deep neural network algorithm. *BMC Bioinformatics* **22**, 324.

Sun Z, ***et al.*** (2021) To improve prediction of binding residues with DNA, RNA, carbohydrate, and peptide via multi-task deep neural networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **19**, 3735–3743.

Sun Z, ***et al.*** (2020) Structure-based analysis of cryptic-site opening. *Structure* **28**, 223–235.

Sunseri J and Koes DR (2020) Libmolgrid: graphics processing unit accelerated molecular gridding for deep learning applications. *Journal of Chemical Information and Modeling* **60**, 1079–1084.

Sverrisson F, ***et al.*** (2021) Fast end-to-end learning on protein surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15272–15281.

Taherzadeh G, ***et al.*** (2016) Sequence-based prediction of protein–peptide binding sites using support vector machine. *Journal of Computational Chemistry* **37**, 1223–1229.

Taherzadeh G, ***et al.*** (2018) Structure-based prediction of protein–peptide binding regions using random forest. *Bioinformatics* **34**, 477–484.

Tahir M, ***et al.*** (2021) KDeepBind: prediction of RNA-proteins binding sites using convolution neural network and k-gram features. *Chemometrics and Intelligent Laboratory Systems* **208**, 104217.

Tan X, ***et al.*** (2024) Molecular glue-mediated targeted protein degradation: a novel strategy in small-molecule drug development. *iScience* **27**(9), 110712.

Todd AE, **Orengo CA and Thornton JM** (2001) Evolution of function in protein superfamilies, from a structural perspective. *Journal of Molecular Biology* **307**, 1113–1143.

https://doi.org/10.1017/S003358352500006X

Tong W, ***et al.*** (2009) Partial order optimum likelihood (POOL): maximum likelihood prediction of protein active site residues using 3D structure and sequence properties. *PLoS Computational Biology* **5**, e1000266.

Tong Y, **Childs-Disney JL and Disney MD** (2024) Targeting RNA with small molecules, from RNA structures to precision medicines: IUPHAR Review: 40. *British Journal of Pharmacology* **181**, 4152–4173.

Tora AS, ***et al.*** (2015) Allosteric modulation of metabotropic glutamate receptors by chloride ions. *The FASEB Journal* **29**, 4174–4188.

Trabuco LG, ***et al.*** (2012) PepSite: prediction of peptide-binding sites from protein surfaces. *Nucleic Acids Research* **40**, W423–W427.

Tsaban T, ***et al.*** (2022) Harnessing protein folding neural networks for peptide–protein docking. *Nature Communications* **13**, 176.

Tsomaia N (2015) Peptide therapeutics: targeting the undruggable space. *European Journal of Medicinal Chemistry* **94**, 459–470.

Tsujikawa H, ***et al.*** (2016) Development of a protein–ligand-binding site prediction method based on interaction energy and sequence conservation. *Journal of Structural and Functional Genomics* **17**, 39–49.

Tubiana J, **Schneidman-Duhovny D and Wolfson HJ** (2022) ScanNet: an interpretable geometric deep learning model for structure-based protein binding site prediction. *Nature Methods* **19**, 730–739.

Tuvshinjargal N, ***et al.*** (2016) PRIdictor: protein–RNA interaction predictor. *Biosystems* **139**, 17–22.

Uhl M, ***et al.*** (2019) *GraphProt2: a graph neural network-based method for predicting binding sites of RNA-binding proteins. bioRxiv*, 850024.

Ullmann JR (1976) An algorithm for subgraph isomorphism. *Journal of the ACM (JACM)* **23**, 31–42.

Utgés JS and Barton GJ (2024) Comparative evaluation of methods for the prediction of protein–ligand binding sites. *Journal of Cheminformatics* **16**, 126.

Valasatava Y, ***et al.*** (2016) MetalPredator: a web server to predict iron–sulfur cluster binding proteomes. *Bioinformatics* **32**, 2850–2852.

Valasatava Y, ***et al.*** (2014) MetalS3, a database-mining tool for the identification of structurally similar metal sites. *JBIC Journal of Biological Inorganic Chemistry* **19**, 937–945.

Van Der Spoel D, ***et al.*** (2005) GROMACS: fast, flexible, and free. *Journal of Computational Chemistry* **26**, 1701–1718.

Vanhee P, ***et al.*** (2011) BriX: a database of protein building blocks for structural analysis, modeling and design. *Nucleic Acids Research* **39**, D435–D442.

Vaswani A, ***et al.*** (2017) Attention is all you need. *Advances in Neural Information Processing Systems* **30**, 5998–6008.

Vecchietti LF, ***et al.*** (2024) *Recent advances in interpretable machine learning using structure-based protein representations.* arXiv Preprint. arXiv:2409.17726.

Verdonk ML, ***et al.*** (2001) SuperStar: improved knowledge-based interaction fields for protein binding sites. *Journal of Molecular Biology* **307**, 841–859.

Verschueren E, ***et al.*** (2013) Protein-peptide complex prediction through fragment interaction patterns. *Structure* **21**, 789–797.

Viet Hung L, ***et al.*** (2015) LIBRA: ligand binding site recognition application. *Bioinformatics* **31**, 4020–4022.

Volkamer A, ***et al.*** (2010) Analyzing the topology of active sites: on the prediction of pockets and subpockets. *Journal of Chemical Information and Modeling* **50**, 2041–2052.

Wade RC, **Clark KJ and Goodford PJ** (1993) Further development of hydrogen bond functions for use in determining energetically favorable binding sites on molecules of known structure. 1. Ligand probe groups with the ability to form two hydrogen bonds. *Journal of Medicinal Chemistry* **36**, 140–147.

Wagner JR, ***et al.*** (2016) Emerging computational methods for the rational discovery of allosteric drugs. *Chemical Reviews* **116**, 6370–6390.

Wagner JR, ***et al.*** (2017) POVME 3.0: software for mapping binding pocket flexibility. *Journal of Chemical Theory and Computation* **13**, 4584–4592.

Wang J, ***et al.*** (2018a) Druggable negative allosteric site of P2X3 receptors. *Proceedings of the National Academy of Sciences* **115**, 4939–4944.

Wang J, ***et al.*** (2024) MultiModRLBP: a deep learning approach for multimodal RNA-small molecule ligand binding sites prediction. *IEEE Journal of Biomedical and Health Informatics* **28**(8), 4995–5006.

Wang J, ***et al.*** (2023a) *Computational modeling of binding site for RNA-ligand complex by learning multi-modal features. Authorea Preprints*.

Wang K, ***et al.*** (2018b) RBind: computational network method to predict RNA binding sites. *Bioinformatics* **34**, 3131–3136.

Wang K, ***et al.*** (2023b) RLBind: a deep learning method to predict RNA–ligand binding sites. *Briefings in Bioinformatics* **24**, bbac486.

Wang R, ***et al.*** (2022a) Predicting protein–peptide binding residues via interpretable deep learning. *Bioinformatics* **38**, 3351–3360.

Wang T, **He Y and Zhu F** (2023c) SAPocket: finding pockets on protein surfaces with a focus towards position and voxel channels. *Expert Systems with Applications* **227**, 120235.

Wang W, ***et al.*** (2023d) GraphPLBR: protein-ligand binding residue prediction with deep graph convolution network. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **20**(3), 2223–2232.

Wang X, ***et al.*** (2022b) *DUnet: a deep learning guided protein-ligand binding pocket prediction.* bioRxiv, *2022–08*.

Wardah W, ***et al.*** (2020) Predicting protein-peptide binding sites with a deep convolutional neural network. *Journal of Theoretical Biology*, **496**, 110278.

Warner KD, **Hajdin CE and Weeks KM** (2018) Principles for targeting RNA with drug-like small molecules. *Nature Reviews Drug Discovery* **17**, 547–558.

Wass MN, **Kelley LA and Sternberg MJE** (2010) 3DLigandSite: predicting ligand-binding sites using similar structures. *Nucleic Acids Research* **38**, W469–W473.

Wei J, ***et al.*** (2022) Protein–RNA interaction prediction with deep learning: structure matters. *Briefings in Bioinformatics* **23**, bbab540.

Wei L and Altman RB (1998) Recognizing protein binding sites using statistical descriptions of their 3D environments. In *Pacific Symposium on Biocomputing*, pp. 497–508.

Wei L and Altman RB (2003) Recognizing complex, asymmetric functional sites in protein structures using a Bayesian scoring function. *Journal of Bioinformatics and Computational Biology* **1**, 119–138.

Weisel M, **Proschak E and Schneider G** (2007) PocketPicker: analysis of ligand binding-sites with shape descriptors. *Chemistry Central Journal* **1**, 1–17.

Wiegreffe S and Pinter Y (2019) *Attention is not not explanation.* arXiv preprint. arXiv:1908.04626.

Wilson CA, **Kreychman J and Gerstein M** (2000) Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *Journal of Molecular Biology* **297**, 233–249.

Wood MJ and Hirst JD (2005) Protein secondary structure prediction with dihedral angles. *Proteins: Structure, Function, and Bioinformatics* **59**, 476–481.

Xia CQ, **Pan X and Shen HB** (2020) Protein–ligand binding residue prediction enhancement through hybrid deep heterogeneous learning of sequence and structure data. *Bioinformatics* **36**, 3018–3027.

Xia Y, **Pan X and Shen HB** (2023) LigBind: identifying binding residues for over 1000 ligands with relation-aware graph neural networks. *Journal of Molecular Biology* **435**, 168091.

Xia Y, ***et al.*** (2022) BindWeb: a web server for ligand binding residue and pocket prediction from protein structures. *Protein Science* **31**, e4462.

Xia Y, ***et al.*** (2021) GraphBind: protein structural context embedded rules learned by hierarchical graph neural networks for recognizing nucleic-acid-binding residues. *Nucleic Acids Research* **49**, e51.

Xie L and Bourne PE (2007) A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand binding sites. *BMC Bioinformatics* **8**, 1–13.

Xie ZR and Hwang MJ (2012) Ligand-binding site prediction using ligand-interacting and binding site-enriched protein triangles. *Bioinformatics* **28**, 1579–1585.

Xue B, ***et al.*** (2008) Real-value prediction of backbone torsion angles. *Proteins: Structure, Function, and Bioinformatics* **72**, 427–433.

Yaffe E, ***et al.*** (2008) MolAxis: a server for identification of channels in macromolecules. *Nucleic Acids Research* **36**, W210–W215.

Yan C and Zou X (2014) Predicting peptide binding sites on protein surfaces by clustering chemical interactions. *Journal of Computational Chemistry* **36**, 49–61.

Yan R, ***et al.*** (2019) Prediction of zinc-binding sites using multiple sequence profiles and machine learning methods. *Molecular Omics* **15**, 205–215.

Yan X, ***et al.*** (2022) PointSite: a point cloud segmentation tool for identification of protein ligand binding atoms. *Journal of Chemical Information and Modeling* **62**, 2835–2845.

Yang J, **Roy A and Zhang Y** (2012) BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic Acids Research* **41**, D1096–D1103.

Yang J, **Roy A and Zhang Y** (2013) Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics* **29**, 2588–2595.

Yao H, ***et al.*** (2003) An accurate, sensitive, and scalable method to identify functional sites in protein structures. *Journal of Molecular Biology* **326**, 255–261.

Yaseen A and Li Y (2014) Context-based features enhance protein secondary structure prediction accuracy. *Journal of Chemical Information and Modeling* **54**, 992–1002.

Yu AM, **Choi YH and Tu MJ** (2020) RNA drugs and RNA targets for small molecules: principles, progress, and challenges. *Pharmacological Reviews* **72**, 862–898.

Yu DJ, ***et al.*** (2015) Constructing query-driven dynamic machine learning model with application to protein-ligand binding sites prediction. *IEEE Transactions on Nanobioscience* **14**, 45–58.

Yu DJ, ***et al.*** (2013) Designing template-free predictor for targeting protein-ligand binding sites with classifier ensemble and spatial clustering. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **10**, 994–1008.

Yu J, ***et al.*** (2010) Roll: a new algorithm for the detection of protein pockets and cavities with a rolling probe sphere. *Bioinformatics* **26**, 46–52.

Yuan Q, ***et al.*** (2022a) AlphaFold2-aware protein–DNA binding site prediction using graph transformer. *Briefings in Bioinformatics* **23**, bbab564.

Yuan Q, ***et al.*** (2022b) Alignment-free metal ion-binding site prediction from protein sequence through pretrained language model and multi-task learning. *Briefings in Bioinformatics* **23**, bbac444.

Yuan Z and Huang B (2004) Prediction of protein accessible surface areas by support vector regression. *Proteins: Structure, Function, and Bioinformatics* **57**, 558–564.

Zaucha J, ***et al.*** (2020) Deep learning model predicts water interaction sites on the surface of proteins using limited-resolution data. *Chemical Communications* **56**, 15454–15457.

Zeng P and Cui Q (2016) Rsite2: an efficient computational method to predict the functional sites of noncoding RNAs. *Scientific Reports* **6**, 19016.

Zeng P, ***et al.*** (2015) Rsite: a computational method to identify the functional sites of noncoding RNAs. *Scientific Reports* **5**, 9179.

Zhan ZH, ***et al.*** (2018) Accurate prediction of ncRNA-protein interactions from the integration of sequence and evolutionary information. *Frontiers in Genetics* **9**, 458.

Zhang C, ***et al.*** (2024) BioLiP2: an updated structure database for biologically relevant ligand–protein interactions. *Nucleic Acids Research* **52**, D404–D412.

Zhang S and Xie L (2023) Protein language model-powered 3D ligand binding site prediction from protein sequence. In *NeurIPS 2023 AI for Science Workshop*.

Zhang W, **Pei J and Lai L** (2017) Statistical analysis and prediction of covalent ligand targeted cysteine residues. *Journal of Chemical Information and Modeling* **57**, 1453–1460.

Zhang Y, ***et al.*** (2023) *EquiPocket: an E (3)-equivariant geometric graph neural network for ligand binding site prediction. arXiv Preprint.* arXiv:2302.12177.

Zhang Y and Skolnick J (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research* **33**, 2302–2309.

Zhang Y, ***et al.*** (2016) Identification of covalent binding sites targeting cysteines based on computational approaches. *Molecular Pharmaceutics* **13**, 3106–3118.

Zhang Z, ***et al.*** (2011) Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction. *Bioinformatics* **27**, 2083–2088.

Zhao J, **Cao Y and Zhang L** (2020) Exploring the computational methods for protein-ligand binding site prediction. *Computational and Structural Biotechnology Journal* **18**, 417–426.

Zhao W, ***et al.*** (2011) Structure-based de novo prediction of zinc-binding sites in proteins of unknown function. *Bioinformatics* **27**, 1262–1268.

Zhao X, **Zhang Y and Du X** (2022) DFpin: Deep learning–based protein-binding site prediction with feature-based non-redundancy from RNA level. *Computers in Biology and Medicine* **142**, 105216.

Zhao Y, ***et al.*** (2023) Identification of metal ion-binding sites in RNA structures using deep learning method. *Briefings in Bioinformatics* **24**(2), bbad049.

Zhao Z, ***et al.*** (2017) Determining cysteines available for covalent inhibition across the human kinome. *Journal of Medicinal Chemistry* **60**, 2879–2889.

Zhao Z, **Peng Z and Yang J** (2018) Improving sequence-based prediction of protein–peptide binding residues by introducing intrinsic disorder and a consensus method. *Journal of Chemical Information and Modeling* **58**, 1459–1468.

Zhao Z, **Xu Y and Zhao Y** (2019) Sxgbsite: Prediction of protein–ligand binding sites using sequence information and extreme gradient boosting. *Genes* **10**, 965.

Zheng C, ***et al.*** (2012) An integrative computational framework based on a two-step random forest algorithm improves prediction of zinc-binding sites in proteins. *PLoS One* **7**, e49716.

Zheng H, ***et al.*** (2024) *PinMyMetal: A hybrid learning system to accurately model metal binding sites in macromolecules.* Research Square, *rs.3.rs–3908734*.

Zhou Y and Chen SJ (2022) Graph deep learning locates magnesium ions in RNA. *QRB Discovery* **3**, e20.

Zhu C, ***et al.*** (2023) *GAPS: Geometric attention-based networks for peptide binding sites identification by the transfer learning approach.* bioRxiv, *2023–12*.

Zhu H and Pisabarro MT (2011) MSPocket: An orientation-independent algorithm for the detection of ligand binding pockets. *Bioinformatics* **27**(3) 351–358.

Zhu Y and Yu DJ (2023) *ULDNA: Integrating unsupervised multi-source language models with LSTM-attention network for protein-DNA binding site prediction.* bioRxiv, *2023–05*.