Computational and Structural Biotechnology Journal 18 (2020) 417–426

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Short review

Exploring the computational methods for protein-ligand binding site prediction

Jingtian Zhao [a], Yang Cao [b,*], Le Zhang [a,*]

[a] College of Computer Science, Sichuan University, Chengdu 610065, China
[b] Center of Growth, Metabolism and Aging, Key Laboratory of Bio-Resource and Eco-Environment of Ministry of Education, College of Life Sciences, Sichuan University, Chengdu 610065, China

ARTICLE INFO

Article history:
Received 14 October 2019
Received in revised form 23 January 2020
Accepted 11 February 2020
Available online 17 February 2020

Keywords:
Protein
Ligand binding site
Machine learning
Deep learning
Protein–ligand binding

ABSTRACT

Proteins participate in various essential processes *in vivo* via interactions with other molecules. Identifying the residues participating in these interactions not only provides biological insights for protein function studies but also has great significance for drug discoveries. Therefore, predicting protein–ligand binding sites has long been under intense research in the fields of bioinformatics and computer aided drug discovery. In this review, we first introduce the research background of predicting protein–ligand binding sites and then classify the methods into four categories, namely, 3D structure-based, template similarity-based, traditional machine learning-based and deep learning-based methods. We describe representative algorithms in each category and elaborate on machine learning and deep learning-based prediction methods in more detail. Finally, we discuss the trends and challenges of the current research such as molecular dynamics simulation based cryptic binding sites prediction, and highlight prospective directions for the near future.

1. Background and significance of protein ligand binding site research

Proteins are some of the most important elements for life. They are not only critical cellular components, but they also participate in various critical activities and processes in the life cycle of organisms, which can achieve or help achieve important biological functions. Proteins do not work independently in living organisms. They need to bind to other biomolecules or ions (such as metal ions, nucleic acids, inorganic or organic small molecules) to create specific interactions to achieve corresponding functions [1]. These molecules and ions are called ligands (Fig. 1). Particularly, intermolecular interactions between proteins and ligands, such as small compounds, occur via amino acid residues at specific positions in the protein, usually located in pocket-like regions. These specific key amino acid residues are called ligand binding sites (LBSs). LBSs have attracted much attention in the fields of molecular docking, drug-target interactions, compound design, ligand affinity prediction, and even molecular dynamics [2–6]. Thus, identification of LBSs not only helps to explore the mechanism of intermolecular interactions but also effectively explains the pathogenesis of diseases, which provides insights for drug discovery and design [7].

Compared with highly accurate but time-consuming biological experiments [8], the advantage of computational methods is that LBS predictions can be performed based on sequence and structure information without relying on annotating the biological function of protein binding residues [9]. In addition, combining multiple computational methods, or combining experimental methods with computational methods can improve both accuracy and efficiency of LBS prediction, provide valuable assistance for drug design and drug discovery researches [10–13]. The emergence of Critical Assessment of Protein Structure Prediction (CASP) [14], Continuous Automated Model EvaluatiOn (CAMEO) projects [15], Critical Assessment of Function Annotation (CAFA) [16], PDB database [17,18], and BioLip database [19] etc. have promoted the development of this field and provided some standard evaluation indicators and relatively unified concepts and definitions. According to the definition given in BioLip, if the distance between any one of the atoms in the ligand molecule and at least one of the atoms in the amino acid residue of the protein does not exceed the sum of the radii of these two atoms plus 0.5 Å, the amino acid residue is regarded as a ligand binding residue. Since the prediction of ligand binding residues is a typical dichotomy problem from an

* Corresponding authors at: No. 24 South Section 1, Yihuan Road, Chengdu, Sichuan 610065, China.
E-mail addresses: cao@scu.edu.cn (Y. Cao), zhangle06@scu.edu.cn (L. Zhang).

https://doi.org/10.1016/j.csbj.2020.02.008

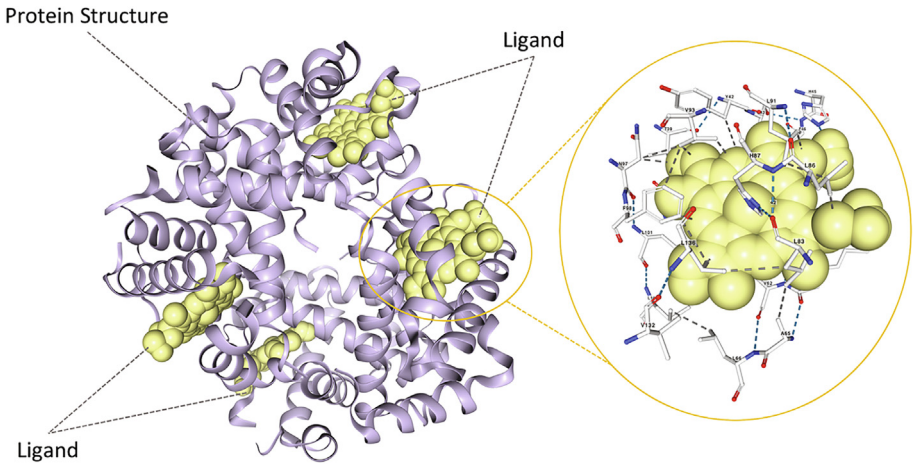

Fig. 1. 3D schematic of a protein structure and its binding ligands generated from the PDB website. The protein shown above is the crystal structure of human deoxyhaemoglobin at 1.74 Å resolution, published on PDB (Access Code: 4HHB). The amplified ligand is [HEM (PROTOPORPHYRIN IX CONTAINING FE)] 142: C with its bonds (Hydrogen, Halogen, et al).

algorithmic point of view, the evaluation index for the prediction method in this field is very similar to the index for evaluating the accuracy of the dichotomy algorithm. The common LBS prediction indicators are sensitivity (*Sen*), accuracy (*Acc*), specificity (*Spe*), precision (*Pre*), and Mattheu's correlation coefficients (*MCC*) [20], which are defined as below:

$$Sen = \frac{TP}{TP+FN} \tag{1}$$

$$Acc = \frac{TP+TN}{TP+FN+TN+FP} \tag{2}$$

$$Spe = \frac{TN}{TN+FP} \tag{3}$$

$$Pre = \frac{TP}{TP+FP} \tag{4}$$

$$MCC = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP+FP) \cdot (TP+FN) \cdot (TN+FP) \cdot (TN+FN)}} \tag{5}$$

where $TP(TruePositive)$ indicates the number of samples to which the binding site is correctly predicted, $TN(TrueNegative)$ indicates the number of samples in which the false binding site is correctly predicted, $FP(FalsePositive)$ indicates the number of samples in which the binding site was incorrectly predicted, and $FN(FalseNegative)$ indicates the number of samples in which the false binding site was incorrectly predicted [21–25].

In the last twenty years, under the promotion of CASP and other research goals, researchers have made great progress in the field of LBS predictions. A series of different prediction methods based on sequence information, structural templates, and three-dimensional structures have been developed. These methods employ various computational methods, including geometry or energy feature searching, sequence or structure similarity comparison, as well as machine learning related algorithms [26–31]. Recently, deep learning-based methods have stood out from machine learning methods and have drawn much attention in computational biology [32–34]. Some state-of-the-art LBS prediction methods that employ machine learning and deep learning algorithms show significant advances over traditional methods [35,36]. In this paper, we systematically introduce the background, principles, algorithms and performance of popular LBS prediction methods by clustering prediction methods into four groups according to their working principles. Particularly, this paper highlights the most recent progress in deep learning-based methods.

2. 3D structure-based LBS prediction methods

Most small ligand binding occurs in hollows or cavities on protein surfaces because high affinity can only be gained by sufficiently large interfaces [37]. This feature has been observed in spatial structures from many detailed studies of protein–ligand complexes in PDB [38]. Therefore, attempting to locate LBSs by searching for special geometry or energy features in protein structures has long been one of the most popular methods in this area. This method generally has two different implementations. One is to perform spatial geometric measurements on the protein structure to find hollows or cavities on the surface of the protein. The second is to place some probes on the surface of the protein and then to find the cavities by estimating the energy potentials between the probe and the cavities. Table 1 lists some published 3D structure-based LBS prediction methods.

The basic idea of LBS prediction methods based on spatial geometry measurements is to locate large or even the largest hollow or cavity on the protein structure by calculating and simulating some certain geometric measures from the protein structure information. Researchers have come up with many different and creative ways to accomplish this over the past few decades.

The pioneer works of spatial geometry measurement were published in the 1990s [27,40]. The idea of these methods is to place a sphere at the gap between any two protein atoms according to the three-dimensional coordinate information in the protein database to detect ligand binding residuals. In SURFNET [40], for example, the size of the sphere is adjusted to be tangent to the surface of the atom. If the sphere collides with other neighboring atoms, the sphere volume is reduced to ensure that no conflicts occur. The above process is repeated until all pairs of protein atoms have been considered. Finally, a set of spheres filled with gaps between protein atoms is found, thereby allowing the localization of large hollows or cavities in the protein molecule, which are regarded as possible ligand binding residues.

Later, in 1997, Manfred Hendlich et al. published LIGSITE [26], which sets up some regular 3D meshes to cover the target protein. Starting from each grid point, they scan a total of 7 directions, including the x, y, and z axes and the 4 grid diagonals, and then score the grid points. If both ends of the scan line in a certain direction are included in the protein area, the point may be in the pocket, and the grid point is added by one point. After all grid points have been scanned in all directions, the candidate ligand binding residues are determined based on the final score of each

Table 1
Published 3D structure-based LBS prediction methods.

Method	Type	Feature	Year
A computational procedure (with no specific name) [39]	Probe Energy-based	Contour surfaces at appropriate energy levels are calculated for each probe and displayed with the protein structure	1985
POCKET [27]	Spatial Geometry Measurement	Place spheres between atoms and surfaces of pockets are modeled using marching cubes algorithm	1992
SURFNET [40]	Spatial Geometry Measurement	Place spheres at the gap between any two protein atoms	1995
LIGSITE [26]	Spatial Geometry Measurement	Set up some regular 3D meshes to cover the target protein	1997
CAST [41]	Spatial Geometry Measurement	Calculate by using alpha shape and discrete flow theory	1998
CASTp [42,43]	Spatial Geometry Measurement	Use alpha shape and the pocket algorithm [44] developed in computational geometry	2003
QSiteFinder [45]	Probe Energy-based	Use the interaction energy between the protein and a simple van der Waals probe	2005
LIGSITE^{CSC} [46]	Spatial Geometry Measurement	An extension and implementation of the LIGSITE algorithm by using the Connolly surface	2006
VISCANA [47]	Probe Energy-based	A total energy of the molecule is evaluated by summation of fragment energies and interfragment interaction energies	2006
Fpocket [48]	Spatial Geometry Measurement	Voronoi tessellation and alpha spheres are used to detect pockets	2009
SITEHOUND [28,49]	Probe Energy-based	The carbon probe and phosphate probe used to detect interaction force between the probe and the protein	2009
MSPocket [50]	Spatial Geometry Measurement	Identify surface pocket regions according to the normal vector directions at the vertices on the surface	2010
FTSite [51]	Probe Energy-based	Use 16 different probes on these grids to detect free energy	2011
SiteComp [52]	Probe Energy-based	Discovery of subsites with different interaction properties and for fast calculations of residue contribution to binding sites	2012
LISE [53]	Spatial Geometry Measurement	Compute a score by counting geometric motifs extracted from substructures of interaction networks connecting protein and ligand atoms	2013
Patch-Surfer2. 0 [54]	Spatial Geometry Measurement	Represent and compare pockets at the level of small local surface patches that characterize physicochemical properties of the local regions	2014
CurPocket [55]	Spatial Geometry Measurement	Compute the curvature distribution of protein surface and identify the clusters of concave regions	2019

grid point. The main advantage of the LIGSITE method is its running speed, as its typical search time is between 5 and 20 s for proteins with medium sizes, so it is suitable for detecting LBSs for a large number of proteins.

The principle of the probe energy-based LBS prediction method is to first place a specific probe molecule on the protein to be tested and to measure the interaction energy signals between the probe molecule and the surrounding residues, and then to find pockets in the protein structure from the distribution of energy signal intensities. The probe energy-based prediction method usually employs different probe parameters or multiple probes at the same time to achieve better performance.

SITEHOUND is a classical probe energy-based LBS prediction method [49,28]. The method uses a box with a grid that covers the entire target protein. A carbon probe and a phosphate probe are released to the grid points and the interaction forces between the molecules of each grid point probe and the protein are calculated. The grid points with higher interaction energies are extracted and further clustered. After mapping the grid points on to the residues, the potential LBSs are determined according to the clustered residues. A dataset that contains 77 experimentally determined protein structures with known protein–ligand complexes was used to test SITEHOUND, and the result showed that in 95% of the cases, the correct binding site was located in the top three clusters.

In 2011, Chi-Ho Ngan et al. released another probe energy-based LBS prediction method, FTSite [51].The basic idea for this method is to place a dense grid around the protein, spread 16 different small molecule probes on this grids, and use the objective free energy functions to determine the appropriate position. The probes are clustered and ranked according to the average free energy value. The overlapping sites clustered by different probes are ranked by the interactions between the probe and the protein. Amino acid residues that interact with the top cluster are regarded as possible ligand binding residues. FTSite empoyed LIGSITE^{CSC} set [46] and QSiteFinder set [45] to benchmark the method which achieved the accuracy rates of 94% and 97%, respectively.

3D structure-based LBS prediction methods have been widely used for years. However, these methods strongly depend on the state of the given protein 3D structure, which means that LBSs may not be discovered if the binding pocket does not exist in the apo state but is induced by protein–ligand interaction in the holo state. In many scenarios which lack the protein structures in holo states, those methods may not be valid.

3. Template similarity-based LBS prediction methods

Protein 3D structures provide geometry and energy clues for LBSs that allow us to make predictions using a single structure of a protein. If considering that proteins are not an independent molecule, but are evolved from others, structural or functional information can be transferred between homologous or structurally similar proteins. Hence, an LBS can be predicted using the known proteins as templates to obtain similar characteristics in the query protein. Template similarity-based LBS prediction methods mainly include two types: structure template-based methods and sequence template-based methods. Table 2 lists some template similarity-based LBS prediction methods that have been published in the last twenty years.

The basic idea of the structure template-based LBS prediction method is to search for the most similar proteins in databases that have been labeled with LBSs using a structure alignment algorithm and then to transfer the known LBS from the most similar proteins onto the query protein. This method takes advantage of the increasingly accumulated protein structure databases. It could be highly reliable if proteins are of significant structural similarity.

Table 2
Published template similarity-based LBS prediction methods.

Method	Type	Feature	Year
ConSurf [56]	Sequence Template-based	Phylogenetic relationships among the sequences and the similarity between the amino acids are taken into account	2003
A Sequence template-based approach with no specific name [57]	Sequence Template-based	An information-theoretic approach for estimating sequence conservation based on Jensen–Shannon divergence	2007
FINDSITE [58]	Structure Template-based	PROSPECTOR 3 threading algorithm and TMalign tool are used	2008
A two-stage template-based LBS prediction method [59]	Structure Template-based	Construct protein's 3D model and use structural clustering of ligand-containing templates on the predicted 3D model	2009
3DLigandSite [29]	Structure Template-based	MAMMOTH is used	2010
FunFOLD [60]	Structure Template-based	Use an automatic approach for cluster identification and residue selection	2011
COFACTOR [61]	Structure and Sequence Template-based	Use global-to-local sequence and structural comparison algorithm	2012
webPDBinder [62]	Structure Template-based	Search a protein structure against a library of known binding sites and a collection of control nonbinding pockets.	2013
S-SITE [31]	Sequence Template-based	Needleman–Wunsch algorithms are used	2013
TM-SITE [31]	Structure and Sequence Template-based	Mix Structure Template-based and Sequence Template-based method	2013

In 2008, a popular template-based ligand binding site prediction method, FINDSITE, was published [58]. For a given target protein sequence, FINDSITE uses the PROSPECTOR 3 threading algorithm [63,64] to identify a structural template that binds to the ligand from the PDB database and overlays the template with the target protein using TMalign [65]. Then, the LBSs that bound to the structural template are clustered and ranked as predictions. FINDSITE achieved a 67.3% success rate with 75.5% ranking accuracy on protein models that have a less than 35% sequence identity to the closest template structure. Although the prediction accuracy is comparable to some 3D structure-based LBS prediction methods, it can make some very unique LBS discoveries.

Later, in 2010, Mark N. Wass et al. developed the 3DLigandSite prediction method [29]. 3DLigandSite first used MAMMOTH [66] to score the similarity between a target protein and structural templates, and the 25 template proteins with the highest similarity to the target protein structure and their corresponding ligand information were selected as templates. Similar to FINDSITE, these templates are overlaid with the target protein, and these overlaid ligands are clustered using the Single linkage clustering algorithm. The cluster with the most template ligands was chosen as the basis for the prediction of the LBS. The performance of 3DLigandSite has been tested on CASP8 [67] targets with a set of 617 proteins from the FINDSITE test set and achieved an *MCC* of 0.64, a coverage of 71%, and an accuracy of 60%.

Up to now (December 21, 2019), 158787 protein structures have been published in the PDB [38]. However, for a large number of proteins, it is still impossible to detect their LBS using the above methods. Meanwhile, with the continuous development of sequencing technology, a huge number of protein sequences are published every year. Therefore, sequence template-based LBS prediction methods have received extensive attention. The basic idea of sequence template-based LBS prediction methods is similar to the structure template-based LBS prediction methods, that is, the alignment tool is used to align the sequence of the protein to be tested with the sequence of the known protein, and then, the template is selected according to the similarity. Finally, the ligand-binding residues of the protein to be tested are presumed by referring the known ligand-binding residues on the aligned regions.

In 2013, Yang Zhang's team published a ligand binding site prediction method called S-SITE [31], which employs the Needleman–Wunsch algorithm [68] to align the query protein to each of the proteins in the BioLip [19] database and screens similar sequences from the query protein according to the alignment result. The residues of the query protein are aligned with the template protein residues which were annotated as binding residues. Consensus voting is used to score the alignment results of the templates. Residues that received more than 25% of the votes were considered an LBS. S-SITE achieved both an *MCC* and *Pre* of 0.45 on the test datasets.

Hybrid methods have been proposed to further improve LBS predictions. A representative algorithm, TM-SITE [31], mixes the structure template-based and the sequence information-based prediction methods. The TMalign algorithm is first used to align the protein to be tested with the known template proteins. The evolutionary information of the sequence and the spatial distance information of the structure are combined to form a comprehensive scoring function to score the similarity of each template protein, and the qualified template proteins are screened from the BioLip database according to the scoring results. Finally, the ligand-binding residues of the protein being tested are predicted based on these eligible templates. TM-SITE achieved an *MCC* of 0.51 and *Pre* of 0.59 on the test datasets.

4. Traditional machine learning-based LBS prediction methods

The continuous development of computer technology has promoted the application of artificial intelligence-related theories and algorithms to other fields. In the study of protein LBS predictions, 3D structure-based and template similarity-based prediction methods have shown complementary advantages to LBS predictions. How to integrate that information and further improve the prediction accuracy is one of the urgent questions of this area. Many researchers try to use machine learning algorithms not only for carrying out LBS predictions but also for the binding affinity research, which has caused significant breakthroughs. Table 3 lists some traditional machine learning-based LBS prediction methods and a few related binding affinity research methods published in recent years. However, to focus the topic, we only detail a few representative LBS prediction methods listed above. Binding affinity related methods are elaborated on in the discussion.

As mentioned earlier, predicting protein ligand binding sites is a typical dichotomous problem from a mathematical point of view, and there is a state of sample imbalance. Among the many classic machine learning algorithms that can implement the dichotomy, the naive Bayesian algorithm needs to calculate the prior probability and does not apply to data with a correlation between samples. Although the logistic regression is simple to implement, its accuracy is poor because it tends to under-fit characteristics. Besides, although the KNN algorithm is fast and has low training costs,

Table 3
Traditional machine learning-based LBS prediction and binding affinity research methods.

Method	Machine Learning Algorithm	Year
Knowledge-based QSAR approach [69]	Kernel-Partial Least Squares (K-PLS) [70]	2004
Multi-RELIEF [71]	RELIEF algorithm [72]	2007
SFCscore [73]	multiple linear regression partial least squares analysis	2008
ATPint [74]	Support Vector Machine	2009
ConCavity [75]	K-Means algorithm	2009
MetaPocket [76]	hierarchical clustering algorithm [77]	2009
RF-Score [4]	The Random Forest algorithm	2010
MetaDBSite [78]	Support Vector Machine	2011
NsitePred [79]	Support Vector Machine	2011
NNSCORE [80,81]	Artificial Neural Network (shallow neural network [82])	2011
L1pred [30]	L1-Logreg Regression classifier	2012
TargetS [83]	Support Vector Machine	2013
eFindSite [84]	Support Vector Machine	2013
VitaPred [85]	Support Vector Machine	2013
COACH [31]	Support Vector Machine	2013
LigandRFs [86]	The Random Forest algorithm	2014
OSML [87]	Support Vector Machine	2015
LigandDSES [88]	The Random Forest algorithm	2015
PRANK [89]	The Random Forest algorithm	2015
A method for protein-ligand binding affinity prediction [90]	Gradient Boosting Regressor [91]	2018
SAnDReS [92]	Regression Analysis	2016
P2Rank [93]	The Random Forest algorithm	2018
COACH-D [94]	Support Vector Machine	2018
Taba [95]	Regression Analysis	2019

the classification effect is poor under the sample imbalance situation. Therefore, a support vector machine (SVM) stands out from many traditional machine learning algorithms by virtue of its high classification accuracy, strong generalization ability, and excellent classification ability for high-dimensional small sample data. It has become the most popular machine learning method in the field of LBS predictions. As demonstrated in Fig. 2, SVM is a supervised learning algorithm that classifies data by solving hyperplanes that can binarily classify data in space. In the past few years, SVM-based

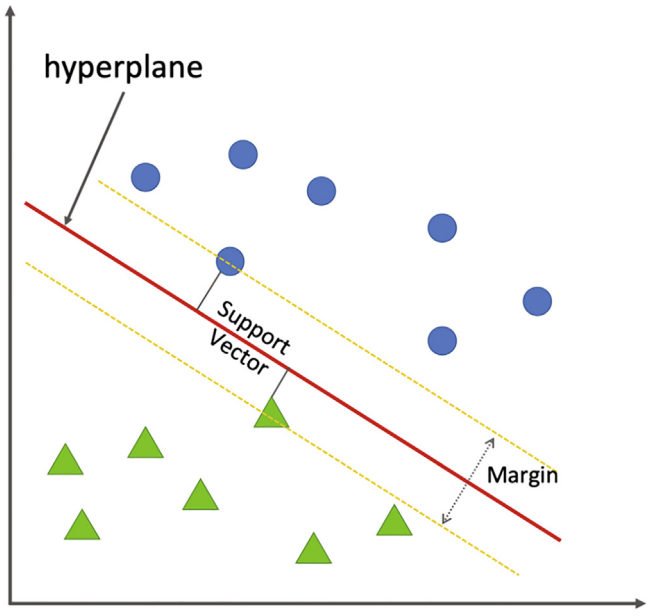

Fig. 2. A simple schematic of SVM A hyperplane divides the points into two categories.

prediction methods have been published. Three representative methods are introduced below.

In 2011, Jingna Si et al. developed the MetaDBSite server [78], relying on sequence information to predict protein-DNA binding residues. MetaDBSite uses SVM to integrate the results of the six predictive tools: DISIS [96], DNABindR [97], BindN [98], BindN-rf [99], DP-Bind [100] and DBS-PRED [101]. The final output is superior to any single prediction method. The prediction results returned by DISIS, DNABindR, BindN, and BindN-rf are the main input parameters of SVM, while DP-Bind and DBS-PRED provide smaller score effects as auxiliary parameters. MetaDBSite achieved *ACC*, *Spe*, *Sen* of 0.77 and *MCC* of 0.32 on a test set, which is better than any of the single methods it combined.

In 2011, Ke Chen et al. published the NsitePred algorithm [79], which predicted the five most common nucleotide residues in the PDB database. The main steps of the NsitePred algorithm are to first extract the secondary structure, relative solvent accessibility and dihedral angles, determine the PSSM profile and other information from a given protein sequence to be tested, and use sliding window technology to process the information to generate an eigenvector describing the residue. These eigenvectors are used as inputs to the SVM to obtain a classification model. The model is used to predict the protein, and the SVM-based prediction results are combined with the BLAST [102] results as the final output. In the benchmarks, NsitePred showed better performance over ATPint [74] and GTPbinder [103].

In 2013, Yang Zhang's team published the SVM-based prediction method COACH [31]. It combines the structure template-based and sequence information-based prediction methods S-SITE and TM_SITE with the prediction results of the three methods of the new COFACTOR [104], FINDSITE [58], and ConCavity [75] as eigenvectors to the SVM for training and to form a classification model, and finally uses this classification model to output the prediction results. The benchmark results show that COACH outperforms other classical prediction algorithms ($MCC = 0.54$ and $Pre = 0.59$), making it the most popular protein LBS prediction method over the past few years.

5. Deep learning-based LBS prediction methods

In 2006, deep learning led the third wave of artificial intelligence [105], which far surpassed traditional machine learning in text classification, speech recognition, semantic modeling, image recognition, image segmentation and computer vision [106–109]. In some areas, it has even surpassed the human brain [110] and has become the most popular research branch in the field of machine learning. Therefore, an increasing number of researchers have seen the possibility of using deep learning techniques to solve complex problems in the fields of bioinformatics and medical research, such as small-compound-drug discovery, activity prediction, chemical structure design, bioimaging, and medical imaging-based diagnosis [35,90,111–114].

Deep learning is a complex machine learning technique that simulates the learning mechanism of the human brain by building and simulating the neural networks in the human brain and uses this mechanism to interpret data. Deep learning is mainly implemented in three ways: convolutional neural networks (CNNs), deep belief networks (DBNs) and self-encoding neural networks. Among them, CNN is the most popular approach used in fields other than computer science since it is relatively simple to use and generalize. CNN is a kind of feedforward neural network. Similar to traditional artificial neural networks (ANNs) [115], CNN is also composed of multiple neurons and each of them does a part of the calculation base on a part of the input and give a part of the output, as below:

$$f(\sum w_i x_i + b) \quad (6)$$

where x is the input, w is a set of weights, and b is the bias. $f(x)$ is the activation function, which makes the neural network approximate the nonlinear function so that the network can be used in a nonlinear model. As described in Fig. 3, CNNs are mainly composed of three layers: the convolution layer, the pooling layer and the fully connected layer. The convolutional layer is used to extract different local features of the input; it consists of several convolutional units, and the parameters of each convolutional unit are optimized by backpropagation [116]. The pooling layer cuts the high dimensional local features obtained by convolutional layers into several regions and calculates the maximum value or the average value of them so that new low dimensional features can be generated. Finally, the fully connected layer combines all the local features into global features and calculates the score for each final class.

DBN is a highly scalable deep neural network, it consists of multiple layers of Restricted Boltzmann Machine (RBM) [117], which is used to learn a probability distribution of the inputs. The DBN training process can be divided into two main steps: First, unsupervised training is performed for each layer of RBM independently. Then, a supervised classifier is set after the last layer of RBM to receive the output features of RBMs and generate classification results. The structure of DBNS is shown in Fig. 4.

In the past two years, some protein LBS prediction methods using deep learning techniques have been reported. Developing new deep learning-based prediction method has become a new hotspot in LBS prediction. Table 4 lists some deep learning-based LBS prediction methods and related studies. Some representative LBS prediction methods or LBS highly related methods are introduced below.

In 2017, J Jiménez et al. developed the DEEPSite algorithm [36] for predicting binding sites for protein ligands. The basic idea of the algorithm is to treat the protein structure as a three-dimensional image and discretize it into a mesh with certain size voxels. A series of atomic attributes, such as hydrophobicity and hydrogen bond acceptors or donors, are used as features to calculate the occupancy of each attribute on each voxel. Finally, subgrids of a certain size are sampled, and the features of the subgrid are used as inputs to the convolutional neural network. The probability of the site being labeled a binding site is output. DEEPSite was compared with Fpocket and Concavity on the same test dataset, and the result indicated that DEEPSits outperforms other methods.

In 2019, Yifeng Cui et al. developed the DeepCSeqSite algorithm [121], which used the seven characteristics of the position-specific score matrix, relative solvent accessibility, secondary structure, the dihedral angle, conservation scores, residue type and position embeddings to construct the eigenspace. Each residue in the amino

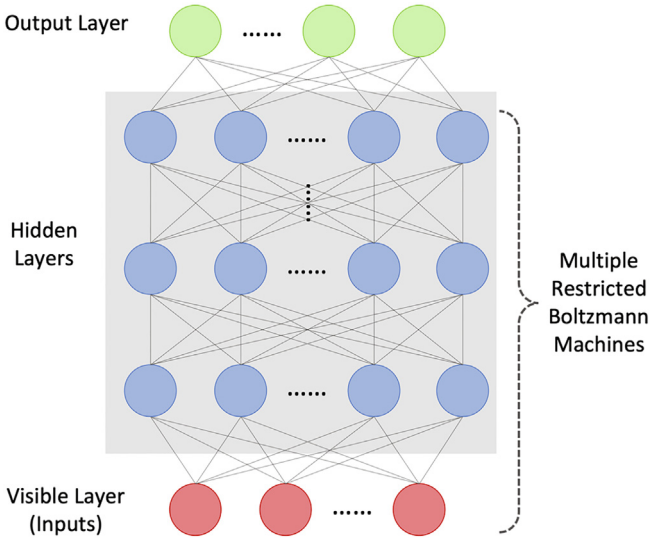

Fig. 4. A simple demonstration of deep belief network DBNs are constructed by combining multiple RBMs. Training of DBNs is performed layer by layer. The hidden layer is first inferred from the data vector, and this hidden layer is used as the input data vector of the next layer.

Table 4
Deep learning-based LBS prediction and binding affinity research methods.

Method	Main Goal	Network Type	Year
A deep learning framework for modeling structural features of RNA-binding protein targets [118]	Binding references modeling of RNA-binding proteins	DBN	2015
DeepBind [119]	Sequence specificities prediction of DNA- and RNA-binding proteins	CNN	2015
DeepDTA [3]	Drug-target interaction identification	CNN	2018
K_{DEEP} [120]	Protein-ligand binding affinity prediction	CNN	2018
DEEPSite [36]	LBS Prediction	CNN	2017
DeepCSeqSite [121]	LBS Prediction	CNN	2019
DeepConv-DTI [122]	Drug-target interaction identification	CNN	2019
DeepDrug3D [35]	Binding pockets characterization and classification	CNN	2019
Onionnet [123]	Protein-ligand binding affinity prediction	CNN	2019

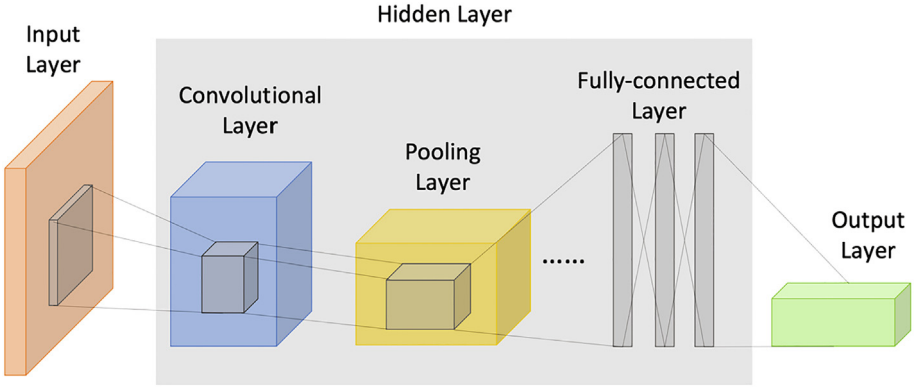

Fig. 3. A simple model of a convolutional neural network Hidden Layers are used to generate the classification result (multiple convolutional layers and pooling layers can be set in a CNN).

acid sequence is embedded in the eigenspace such that the amino acid sequence is converted to a feature map, and then the map is used as an input to the convolutional neural network. The output of the network is the predicted result of protein ligand binding residues. Instead of using any template, including the three-dimensional structure, DeepCSeqSite directly predicts the binding sites of protein ligands. Its performance on test datasets is significantly better than COACH, the most accurate SVM-based prediction method mentioned above.

Recently, Ingoo Lee et al. reported the DeepConv-DTI prediction model [122] to identify interactions between drugs and targets. The idea of the model is to input the entire protein sequence into a convolutional neural network, convolve the various amino acid subsequences of the protein to capture how the protein matches the local residue pattern participating in the DTI, and use that as the input to the higher layer network to build the model and extract features. The new features will connect the model to the drug signature and predict the likelihood of DTI through a higher fully connected layer in the network. By further optimizing the model, it achieves better predictions of performance. Through the model, new features will be linked to drug characteristics and predict the likelihood of DTI through a higher fully connected layer in the network. Finally, the model is further optimized to achieve better predictive performance. As a result, the local features detected by DeepConv-DTI show better performance than other protein descriptors, such as CTD and SW scores according to the authors.

In 2019, Limeng Pu et al. presented DeepDrug3D [35], a new deep learning-based binding pockets characterization and classification algorithm, which can classify nucleotide- and heme-binding sites by learning the patterns of specific molecular interactions between ligands and their protein targets. First, the ligand–protein complexes are converted into 3D pocket grids, and the physicochemical properties of binding pockets are considered and characterized. These 3D pocket grids are then voxelized into a 3D image with 14 channels. These voxels are used as inputs for a designed convolutional neural network to get the classification result. DeepDrug3D was tested on the PDB dataset of nucleotide- and heme-binding sites and achieved an accuracy of 95%, which is much better than volume- and shape-based approaches.

6. Discussion

From the long history of LBS prediction methods, we have seen that the research focus of LBS predictions has shifted from analyzing simple 3D structure features and sequence/structure similarities to the integration of multiple features. Machine learning algorithms [21,22,24,124–130] have played a critical role in this process. Particularly, the application of deep learning algorithms has begun to show great value in LBS predictions. Furthermore, information about binding affinity and crystal structures can be used as inputs to machine learning or deep learning algorithms to help complete the LBS prediction, which makes LBS predictions more closely integrated with areas such as affinity prediction and molecular docking [23,131].

With the continuous publication of more excellent machine learning and deep learning-based LBS prediction methods, other biological studies using these methods, such as protein structure and function prediction, protein–protein interaction site prediction, and drug design, have also made new breakthroughs [132–137]. For instance, in 2015, COACH was used in drug design studies targeting MARK4 regulatory enzymes related to cancer, type 2 diabetes and many other diseases [138]. In 2019, DeepDTA was used to research protein kinases to help develop a predictive model which can estimate kinase-ligand pKi values [139].

New solutions often bring new challenges and problems while solving problems. Although deep learning-based LBS prediction methods have been used and applied in the past 2 years, there are still some problems and deficiencies to this type of solution. A key problem is that deep learning algorithms often require extremely high training costs (expensive computing resources, huge training sets, etc.) compared with traditional machine learning algorithms [140,141].

Studies have also been inconclusive about whether deep learning approaches are always superior to traditional machine learning algorithms in all cases. In fact, traditional machine learning algorithms and even some 3D structure-based binding affinity prediction methods are constantly being optimized. For instance, some methods can predict binding affinity based on the known crystal structure of a specific ligand or a protein can accurately identify the key LBS [131,142–144]. Additionally, the performance of deep learning algorithms is similar to traditional machine learning algorithms in some cases with low dimensional or small amounts of data. Thus, how to take advantage of deep learning to obtain the best solution for LBS predictions in the near future is still an open question.

In addition, researchers also think that the series of LBS prediction methods mentioned in the article cannot completely solve the problem of LBS detection since there exist some cryptic sites that are not evident in the unbound protein but form upon ligand binding [145]. Conformational change is critical to reveal these cryptic sites. Thus, detecting cryptic binding sites has received lots of attention in the past few years, and molecular dynamics simulations have become one of the most popular methods for conformational sampling in this field [2,5,146–148]. For instance, Bowman and Geissler built Markov state models from molecular dynamics (MD) simulations that can identify prospective cryptic sites [149], and a series of studies have been carried out by Gorfe's team to find hidden binding sites in Ras proteins using probe-based molecular dynamics simulations [150–153]. We believe that in the future, the advanced machine learning or deep learning approaches together with protein conformational sampling technique is also likely to become a new development direction in the field of LBS prediction.

CRediT authorship contribution statement

Jingtian Zhao: Investigation, Writing - original draft. **Yang Cao:** Conceptualization, Resources, Writing - review & editing. **Le Zhang:** Writing - review & editing, Supervision, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by National Natural Science Foundation of China [number 61372138 and 81973243], and the National Science and Technology Major Project [2018ZX10201002].

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2020.02.008.

References

[1] Chen K, Mizianty MJ, Kurgan L. ATPsite: sequence-based prediction of ATP-binding residues. Proteome science. BioMed Central; 2011.
[2] Durrant JD, McCammon JA. Molecular dynamics simulations and drug discovery. BMC Biol 2011;9:71.
[3] Öztürk H, Özgür A, Ozkirimli E. DeepDTA: deep drug–target binding affinity prediction. Bioinformatics 2018;34:i821–9.
[4] Ballester PJ, Mitchell JB. A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. Bioinformatics 2010;26:1169–75.
[5] Seco J, Luque FJ, Barril X. Binding site detection and druggability index from first principles. J Med Chem 2009;52:2363–71.
[6] Heo L, Shin W-H, Lee MS, Seok C. GalaxySite: ligand-binding-site prediction by using molecular docking. Nucleic Acids Res 2014;42:W210–4.
[7] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–10.
[8] Vajda S, Guarnieri F. Characterization of protein-ligand interaction sites using experimental and computational methods. Curr Opin Drug Discov Devel 2006;9:354.
[9] Marrone TJ, Briggs A, James M, McCammon JA. Structure-based drug design: computational advances. Annual Rev Pharmacol Toxicol 1997;37:71–90.
[10] Kubinyi H. Combinatorial and computational approaches in structure-based drug design. Curr Opin Drug Discov Devel 1998;1:16–27.
[11] Zhang Z, Li Y, Lin B, Schroeder M, Huang B. Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction. Bioinformatics 2011;27:2083–8.
[12] Tong AHY et al. A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. Science 2002;295:321–4.
[13] Henrich S, Salo-Ahen OM, Huang B, Rippmann FF, Cruciani G, Wade RC. Computational approaches to identifying and characterizing protein binding sites for ligand design. J Mol Recognit 2010;23:209–19.
[14] Moult J, Pedersen JT, Judson R, Fidelis K. A large-scale experiment to assess protein structure prediction methods. Proteins 1995;23:ii–v.
[15] Haas J et al. (2013) The Protein Model Portal—a comprehensive resource for protein structure and model information. Database 2013.
[16] Radivojac P et al. A large-scale evaluation of computational protein function prediction. Nat Methods 2013;10:221.
[17] Bernstein FC et al. The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol 1977;112:535–42.
[18] Berman HM, Bourne PE, Westbrook J, Zardecki C. The protein data bank. In: Protein structure. CRC Press; 2003. p. 394–410.
[19] Yang J, Roy A, Zhang Y. BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions. Nucleic Acids Res 2012;41:D1096–103.
[20] Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta (BBA)-Protein Struct 1975;405:442–51.
[21] Zhang L et al. Computed tomography angiography-based analysis of high-risk intracerebral haemorrhage patients by employing a mathematical model. BMC Bioinf 2019;20:193.
[22] Zhang L, Dai Z, Yu J, Xiao M. CpG-Island-based annotation and analysis of human house-keeping genes. Briefings Bioinform 2019.
[23] Li J, Fu A, Zhang L. An overview of scoring functions used for protein-ligand interactions in molecular docking. Interdiscip Sci 2019;11:320–8.
[24] Zhang L et al. Building up a robust risk mathematical platform to predict colorectal cancer. Complexity 2017;2017:14.
[25] Xia Y et al. Exploring the key genes and signaling transduction pathways related to the survival time of glioblastoma multiforme patients by a novel survival analysis model. BMC Genomics 2017;18:950.
[26] Hendlich M, Rippmann F, Barnickel G. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. J Mol Graph Model 1997;15:359–63.
[27] Levitt DG, Banaszak LJ. POCKET: a computer graphies method for identifying and displaying protein cavities and their surrounding amino acids. J Mol Graph 1992;10:229–34.
[28] Hernandez M, Ghersi D, Sanchez R. SITEHOUND-web: a server for ligand binding site identification in protein structures. Nucleic Acids Res 2009;37:W413–6.
[29] Wass MN, Kelley LA, Sternberg MJ. 3DLigandSite: predicting ligand-binding sites using similar structures. Nucleic Acids Res 2010;38:W469–73.
[30] Dou Y, Wang J, Yang J, Zhang C. L1pred: a sequence-based prediction tool for catalytic residues in enzymes with the L1-logreg classifier. PLoS ONE 2012;7:e35666.
[31] Yang J, Roy A, Zhang Y. Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. Bioinformatics 2013;29:2588–95.
[32] Min S, Lee B, Yoon S. Deep learning in bioinformatics. Briefings Bioinf 2017;18:851–69.
[33] Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics 2014;31:761–3.
[34] Almagro Armenteros JJ, Sønderby CK, Sønderby SK, Nielsen H, Winther O. DeepLoc: prediction of protein subcellular localization using deep learning. Bioinformatics 2017;33:3387–95.
[35] Pu L, Govindaraj RG, Lemoine JM, Wu H-C, Brylinski M. DeepDrug3D: classification of ligand-binding pockets in proteins with a convolutional neural network. PLoS Comput Biol 2019;15:e1006718.
[36] Jiménez J, Doerr S, Martínez-Rosell G, Rose AS, De Fabritiis G. DeepSite: protein-binding site predictor using 3D-convolutional neural networks. Bioinformatics 2017;33:3036–42.
[37] Sotriffer C, Klebe G. Identification and mapping of small-molecule binding sites in proteins: computational tools for structure-based drug design. Il Farmaco 2002;57:243–51.
[38] Rose PW et al. The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. Nucleic Acids Res 2014;43:D345–56.
[39] Goodford PJ. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. J Med Chem 1985;28:849–57.
[40] Laskowski RA. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. J Mol Graph 1995;13:323–30.
[41] Liang J, Woodward C, Edelsbrunner H. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. Protein Sci 1998;7:1884–97.
[42] Dundas J, Ouyang Z, Tseng J, Binkowski A, Turpaz Y, Liang J. CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. Nucleic Acids Res 2006;34:W116–8.
[43] Binkowski TA, Naghibzadeh S, Liang J. CASTp: computed atlas of surface topography of proteins. Nucleic Acids Res 2003;31:3352–5.
[44] Edelsbrunner H, Facello M, Liang J. On the definition and the construction of pockets in macromolecules. Discrete Appl Math 1998;88:83–102.
[45] Laurie AT, Jackson RM. Q-SiteFinder: an energy-based method for the prediction of protein–ligand binding sites. Bioinformatics 2005;21:1908–16.
[46] Huang B, Schroeder M. LIGSITE csc: predicting ligand binding sites using the Connolly surface and degree of conservation. BMC Struct Biol 2006;6:19.
[47] Amari S et al. VISCANA: visualized cluster analysis of protein– ligand interaction based on the ab initio fragment molecular orbital method for virtual ligand screening. J Chem Inf Model 2006;46:221–30.
[48] Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: an open source platform for ligand pocket detection. BMC Bioinf 2009;10:168.
[49] Ghersi D, Sanchez R. Improving accuracy and efficiency of blind protein-ligand docking by focusing on predicted binding sites. Proteins 2009;74:417–24.
[50] Zhu H, Pisabarro MT. MSPocket: an orientation-independent algorithm for the detection of ligand binding pockets. Bioinformatics 2010;27:351–8.
[51] Ngan C-H, Hall DR, Zerbe B, Grove LE, Kozakov D, Vajda S. FTSite: high accuracy detection of ligand binding sites on unbound protein structures. Bioinformatics 2011;28:286–7.
[52] Lin Y, Yoo S, Sanchez R. SiteComp: a server for ligand binding site analysis in protein structures. Bioinformatics 2012;28:1172–3.
[53] Xie Z-R, Liu C-K, Hsiao F-C, Yao A, Hwang M-J. LISE: a server using ligand-interacting and site-enriched protein triangles for prediction of ligand-binding sites. Nucleic Acids Res 2013;41:W292–6.
[54] Zhu X, Xiong Y, Kihara D. Large-scale binding ligand prediction by improved patch-based method Patch-Surfer2. 0. Bioinformatics 2014;31:707–13.
[55] Liu Y, Grimm M, Dai W-T, Hou M-C, Xiao Z-X, Cao Y. CB-Dock: a web server for cavity detection-guided protein–ligand blind docking. Acta Pharmacol Sin 2019:1–7.
[56] Glaser F et al. ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. Bioinformatics 2003;19:163–4.
[57] Capra JA, Singh M. Predicting functionally important residues from sequence conservation. Bioinformatics 2007;23:1875–82.
[58] Brylinski M, Skolnick J. A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. Proc Natl Acad Sci 2008;105:129–34.
[59] Oh M, Joo K, Lee J. Protein-binding site prediction based on three-dimensional protein modeling. Proteins 2009;77:152–6.
[60] Roche DB, Tetchner SJ, McGuffin LJ. FunFOLD: an improved automated method for the prediction of ligand binding residues using 3D models of proteins. BMC Bioinf 2011;12:160.
[61] Roy A, Zhang Y. Recognizing protein-ligand binding sites by global structural alignment and local geometry refinement. Structure 2012;20:987–97.
[62] Bianchi V, Mangone I, Ferre F, Helmer-Citterich M, Ausiello G. webPDBinder: a server for the identification of ligand binding sites on protein structures. Nucleic Acids Res 2013;41:W308–13.
[63] Skolnick J, Kihara D, Zhang Y. Development and large scale benchmark testing of the PROSPECTOR_3 threading algorithm. Proteins 2004;56:502–18.
[64] Skolnick J, Kihara D. Defrosting the frozen approximation: PROSPECTOR—a new approach to threading. Proteins 2001;42:319–31.
[65] Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33:2302–9.
[66] Ortiz AR, Strauss CE, Olmea O. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. Protein Sci 2002;11:2606–21.
[67] Lopez G, Ezkurdia I, Tress ML. Assessment of ligand binding residue predictions in CASP8. Proteins 2009;77:138–46.

[68] Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970;48:443–53.
[69] Deng W, Breneman C, Embrechts MJ. Predicting protein– ligand binding affinities using novel geometrical descriptors and machine-learning methods. J Chem Inf Comput Sci 2004;44:699–703.
[70] Rosipal R, Trejo LJ. Kernel partial least squares regression in reproducing kernel hilbert space. J Mach Learn Res 2001;2:97–123.
[71] Ye K, Anton Feenstra K, Heringa J, IJzerman AP, Marchiori E. Multi-RELIEF: a method to recognize specificity determining residues from multiple sequence alignments using a Machine-Learning approach for feature weighting. Bioinformatics 2007;24:18–25.
[72] Kononenko I. Estimating attributes: analysis and extensions of RELIEF, in European conference on machine learning. Springer; 1994.
[73] Sotriffer CA, Sanschagrin P, Matter H, Klebe G. SFCscore: scoring functions for affinity prediction of protein–ligand complexes. Proteins 2008;73:395–419.
[74] Chauhan JS, Mishra NK, Raghava GP. Identification of ATP binding residues of a protein from its primary sequence. BMC Bioinf 2009;10:434.
[75] Capra JA, Laskowski RA, Thornton JM, Singh M, Funkhouser TA. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. PLoS Comput Biol 2009;5:e1000585.
[76] Huang B. MetaPocket: a meta approach to improve protein ligand binding site prediction. OMICS 2009;13:325–30.
[77] Bandyopadhyay S, Coyle EJ. An energy efficient hierarchical clustering algorithm for wireless sensor networks. IEEE INFOCOM 2003. Twenty-second annual joint conference of the IEEE computer and communications societies (IEEE Cat. No. 03CH37428) Vol. 3. IEEE; 2003.
[78] Si J, Zhang Z, Lin B, Schroeder M, Huang B. MetaDBSite: a meta approach to improve protein DNA-binding sites prediction. BMC Syst Biol 2011;5:S7.
[79] Chen K, Mizianty MJ, Kurgan L. Prediction and analysis of nucleotide-binding residues using sequence and sequence-derived structural descriptors. Bioinformatics 2011;28:331–41.
[80] Durrant JD, McCammon JA. NNScore 2.0: a neural-network receptor–ligand scoring function. J Chem Inf Model 2011;51:2897–903.
[81] Durrant JD, McCammon JA. NNScore: a neural-network-based scoring function for the characterization of protein–ligand complexes. J Chem Inf Model 2010;50:1865–71.
[82] Siu K-Y, Bruck J. Neural computation of arithmetic functions. Proc IEEE 1990;78:1669–75.
[83] Yu D-J, Hu J, Yang J, Shen H-B, Tang J, Yang J-Y. Designing template-free predictor for targeting protein-ligand binding sites with classifier ensemble and spatial clustering. IEEE/ACM Trans Comput Biol Bioinf 2013;10:994–1008.
[84] Brylinski M, Feinstein WP. eFindSite: improved prediction of ligand binding sites in protein models using meta-threading, machine learning and auxiliary ligands. J Comput Aided Mol Des 2013;27:551–67.
[85] Panwar B, Gupta S, Raghava GP. Prediction of vitamin interacting residues in a vitamin binding protein using evolutionary information. BMC Bioinf 2013;14:44.
[86] Chen P, Huang JZ, Gao X. LigandRFs: random forest ensemble to identify ligand-binding residues from sequence information alone. BMC bioinformatics. BioMed Central; 2014.
[87] Yu D-J, Hu J, Li Q-M, Tang Z-M, Yang J-Y, Shen H-B. Constructing query-driven dynamic machine learning model with application to protein-ligand binding sites prediction. IEEE Trans Nanobiosci 2015;14:45–58.
[88] Chen P et al. A sequence-based dynamic ensemble learning system for protein ligand-binding site prediction. IEEE/ACM Trans Comput Biol Bioinf 2015;13:901–12.
[89] Krivák R, Hoksza D. Improving protein-ligand binding site prediction accuracy by classification of inner pocket points using local features. J Cheminf 2015;7:12.
[90] Cang Z, Wei GW. Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction. Int J Numer Meth Biomed Eng 2018;34:e2914.
[91] Pedregosa F et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825–30.
[92] Morrone Xavier M et al. SAnDReS a computational tool for statistical analysis of docking results and development of scoring functions. Comb Chem High Throughput Screening 2016;19:801–12.
[93] Krivák R, Hoksza D. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. J Cheminf 2018;10:39.
[94] Wu Q, Peng Z, Zhang Y, Yang J. COACH-D: improved protein–ligand binding sites prediction with refined ligand-binding poses through molecular docking. Nucleic Acids Res 2018;46:W438–42.
[95] da Silva AD, Bitencourt-Ferreira G, de Azevedo WF. Taba: A tool to analyze the binding affinity. J Comput Chem 2019.
[96] Ofran Y, Mysore V, Rost B. Prediction of DNA-binding residues from sequence. Bioinformatics 2007;23:i347–53.
[97] Yan C, Terribilini M, Wu F, Jernigan RL, Dobbs D, Honavar V. Predicting DNA-binding sites of proteins from amino acid sequence. BMC Bioinf 2006;7:262.
[98] Wang L, Brown SJ. BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. Nucleic Acids Res 2006;34:W243–8.
[99] Wang L, Yang MQ, Yang JY. Prediction of DNA-binding residues from protein sequence information using random forests. BMC Genomics 2009;10:S1.
[100] Hwang S, Gou Z, Kuznetsov IB. DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. Bioinformatics 2007;23:634–6.
[101] Ahmad S, Gromiha MM, Sarai A. Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. Bioinformatics 2004;20:477–86.
[102] Kent WJ. BLAT—the BLAST-like alignment tool. Genome Res 2002;12:656–64.
[103] Chauhan JS, Mishra NK, Raghava GP. Prediction of GTP interacting residues, dipeptides and tripeptides in a protein from its evolutionary information. BMC Bioinf 2010;11:301.
[104] Roy A, Yang J, Zhang Y. COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. Nucleic Acids Res 2012;40:W471–7.
[105] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science 2006;313:504–7.
[106] Amodei D. et al. (2016) Deep speech 2: End-to-end speech recognition in english and mandarin,in International conference on machine learning Vol.
[107] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.
[108] Papandreou G, Chen L-C, Murphy KP, Yuille AL. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. Proceedings of the IEEE international conference on computer vision, 2015.
[109] Voulodimos A, Doulamis N, Doulamis A, Protopapadakis E. (2018) Deep learning for computer vision: a brief review. Comput Intel Neurosci 2018.
[110] Wang F-Y et al. Where does AlphaGo go: From church-turing thesis to AlphaGo thesis and beyond. IEEE/CAA J Autom Sin 2016;3:113–20.
[111] Greenspan H, Van Ginneken B, Summers RM. Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique. IEEE Trans Med Imaging 2016;35:1153–9.
[112] Sun W, Zheng B, Qian W. Computer aided lung cancer diagnosis with deep learning algorithms. Medical imaging 2016: computer-aided diagnosis Vol. 9785. International Society for Optics and Photonics; 2016.
[113] Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. Drug Discovery Today 2018;23:1241–50.
[114] Cheng J-Z et al. Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans. Sci Rep 2016;6:24454.
[115] Kleene SC. "Representation of events in nerve nets and finite automata," RAND PROJECT AIR FORCE SANTA MONICA CA, 1951.
[116] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Cognitive Model 1988;5:1.
[117] Smolensky P. Chapter 6: information processing in dynamical systems: foundations of harmony theory, Parallel distributed processing: explorations in the microstructure of cognition 1.
[118] Zhang S et al. A deep learning framework for modeling structural features of RNA-binding protein targets. Nucl Acids Res 2015;44. e32-e32.
[119] Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. Nat Biotechnol 2015;33:831.
[120] Jimenez J, Skalic M, Martinezrosell G, De Fabritiis G. KDEEP: protein-ligand absolute binding affinity prediction via 3D-convolutional neural networks. J Chem Inf Model 2018;58:287–96.
[121] Cui Y, Dong Q, Hong D, Wang X. Predicting protein-ligand binding residues with deep convolutional neural networks. BMC Bioinf 2019;20:93.
[122] Lee I, Keum J, Nam H. DeepConv-DTI: prediction of drug-target interactions via deep learning with convolution on protein sequences. PLoS Comput Biol 2019;15:e1007129.
[123] Zheng L, Fan J, Mu Y. (2019) OnionNet: a multiple-layer inter-molecular contact based convolutional neural network for protein-ligand binding affinity prediction, arXiv preprint arXiv:1906.02418.
[124] Zhang L, Bai W, Yuan N, Du Z. Comprehensively benchmarking applications for detecting copy number variation. PLoS Comput Biol 2019;15:e1007069.
[125] Zhang L et al. Discovery of a ruthenium complex for the theranosis of glioma through targeting the mitochondrial DNA with bioinformatic methods. Int J Mol Sci 2019;20:4643.
[126] Zhang L et al. Revealing dynamic regulations and the related key proteins of myeloma-initiating cells by integrating experimental data into a systems biological model. Bioinformatics 2019.
[127] Zhang L et al. EZH2-, CHD4-, and IDH-linked epigenetic perturbation and its association with survival in glioma patients. J Mol Cell Biol 2017;9:477–88.
[128] Zhang L et al. Investigation of mechanism of bone regeneration in a porous biodegradable calcium phosphate (CaP) scaffold by a combination of a multi-scale agent-based model and experimental optimization/validation. Nanoscale 2016;8:14877–87.
[129] Zhang L, Xiao M, Zhou J, Yu J. Lineage-associated underrepresented permutations (LAUPs) of mammalian genomic sequences based on a Jellyfish-based LAUPs analysis application (JBLA). Bioinformatics 2018;34:3624–30.
[130] Zhang L, Zhang S. Using game theory to investigate the epigenetic control mechanisms of embryo development: Comment on: "Epigenetic game theory: How to compute the epigenetic control of maternal-to-zygotic transition" by Qian Wang et al. Phys Life Rev 2017;20:140–2.
[131] Levin NMB, Pintro VO, Bitencourt-Ferreira G, de Mattos BB, de Castro Silvério A, de Azevedo Jr WF. Development of CDK-targeted scoring functions for prediction of binding affinity. Biophys Chem 2018;235:1–8.

[132] Yang J, Zhang Y. I-TASSER server: new development for protein structure and function predictions. Nucleic Acids Res 2015;43:W174–81.
[133] Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: protein structure and function prediction. Nat Methods 2015;12:7.
[134] Li G-Q, Liu Z, Shen H-B, Yu D-J. Target M6A: identifying N 6-methyladenosine sites from RNA sequences via position-specific nucleotide propensities and a support vector machine. IEEE Trans Nanobiosci 2016;15:674–82.
[135] Wei Z-S, Yang J-Y, Shen H-B, Yu D-J. A cascade random forests algorithm for predicting protein-protein interaction sites. IEEE Trans Nanobiosci 2015;14:746–60.
[136] Wei Z-S, Han K, Yang J-Y, Shen H-B, Yu D-J. Protein–protein interaction sites prediction by ensembling SVM and sample-weighted random forests. Neurocomputing 2016;193:201–12.
[137] Wass MN, Barton G, Sternberg MJ. CombFunc: predicting protein function using heterogeneous data sources. Nucleic Acids Res 2012;40:W466–70.
[138] Naz F, Shahbaaz M, Bisetty K, Islam A, Ahmad F, Hassan MI. Designing new kinase inhibitor derivatives as therapeutics against common complex diseases: structural basis of microtubule affinity-regulating kinase 4 (MARK4) inhibition. OMICS 2015;19:700–11.
[139] Govinda K, Hassan MM, Sirimulla S. KinasepKipred: a predictive model for estimating ligand-kinase inhibitor constant (pKi). BioRxiv 2019:798561.
[140] Goodfellow I, Bengio Y, Courville A. Deep learning. MIT press; 2016.
[141] LeCun Y, Bengio Y, Hinton G. Deep learning. nature 2015;521:436.
[142] de Ávila MB, Bitencourt-Ferreira G, de Azevedo WF. Structural basis for inhibition of enoyl-[acyl carrier protein] reductase (InhA) from Mycobacterium tuberculosis. Curr Med Chem 2019.
[143] Volkart PA, Bitencourt-Ferreira G, Souto AA, de Azevedo WF. Cyclin-dependent kinase 2 in cellular senescence and cancer. A structural and functional review. Curr Drug Targets 2019;20:716–26.
[144] de Ávila MB, Xavier MM, Pintro VO, de Azevedo Jr WF. Supervised machine learning techniques to predict binding affinity. A study for cyclin-dependent kinase 2. Biochem Biophys Res Commun 2017;494:305–10.
[145] Cimermancic P et al. CryptoSite: expanding the druggable proteome by characterization and prediction of cryptic binding sites. J Mol Biol 2016;428:709–19.
[146] Guterres H, Lee HS, Im W. Ligand-binding-site structure refinement using molecular dynamics with restraints derived from predicted binding site templates. J Chem Theory Comput 2019;15:6524–35.
[147] Bowman GR, Bolin ER, Hart KM, Maguire BC, Marqusee S. Discovery of multiple hidden allosteric sites by combining Markov state models and experiments. Proc Natl Acad Sci 2015;112:2734–9.
[148] Udi Y et al. Unraveling hidden regulatory sites in structurally homologous metalloproteases. J Mol Biol 2013;425:2330–46.
[149] Bowman GR, Geissler PL. Equilibrium fluctuations of a single folded protein reveal a multitude of potential cryptic allosteric sites. Proc Natl Acad Sci 2012;109:11681–6.
[150] McCarthy M, Prakash P, Gorfe AA. Computational allosteric ligand binding site identification on Ras proteins. Acta Biochim Biophy Sin 2015;48:3–10.
[151] Prakash P, Hancock JF, Gorfe AA. Binding hotspots on K-ras: Consensus ligand binding sites and other reactive regions from probe-based molecular dynamics analysis. Proteins 2015;83:898–909.
[152] Prakash P, Sayyed-Ahmad A, Gorfe AA. pMD-membrane: a method for ligand binding site identification in membrane-bound proteins. PLoS Comput Biol 2015;11:e1004469.
[153] Prakash P, Zhou Y, Liang H, Hancock JF, Gorfe AA. Oncogenic K-Ras binds to an anionic membrane in two distinct orientations: a molecular dynamics analysis. Biophys J 2016;110:1125–38.