Acta Pharmaceutica Sinica B 2026;16(1):62−92

Chinese Pharmaceutical Association
Institute of Materia Medica, Chinese Academy of Medical Sciences

Acta Pharmaceutica Sinica B

www.elsevier.com/locate/apsb
www.sciencedirect.com

REVIEW

Computational approaches to druggable site identification: Current status and future perspective

Anqi Lin[a,b,†], Zhirou Zhang[a,†], Aimin Jiang[c,†], Kexin Li[d], Ying Shi[d], Hong Yang[d], Jian Zhang[d], Rongrong Liu[b], Yaxuan Wang[e], Antonino Glaviano[f], Quan Cheng[g,h], Bufu Tang[i,*], Zhengang Qiu[b,*], Peng Luo[a,*]

[a]*Donghai County People's Hospital (Affiliated Kangda College of Nanjing Medical University); Department of Oncology, Zhujiang Hospital, Southern Medical University, Lianyungang 222000, China*
[b]*Department of Oncology, the First Affiliated Hospital of Gannan Medical University, Ganzhou 341000, China*
[c]*Department of Urology, Changhai Hospital, Naval Medical University (Second Military Medical University), Shanghai 200433, China*
[d]*Department of Oncology, Zhujiang Hospital, Southern Medical University, Guangzhou 510282, China*
[e]*Department of Urology, the First Affiliated Hospital of Harbin Medical University, Harbin 150001, China*
[f]*Department of Biological, Chemical and Pharmaceutical Sciences and Technologies, University of Palermo, Palermo 90123, Italy*
[g]*Department of Neurosurgery, Xiangya Hospital, Central South University, Changsha 410008, China*
[h]*National Clinical Research Center for Geriatric Disorders, Xiangya Hospital, Central South University, Changsha 410008, China*
[i]*Department of Interventional Radiology, Zhongshan Hospital, Fudan University, Shanghai 200032, China*

Received 14 June 2025; received in revised form 1 August 2025; accepted 5 August 2025

KEY WORDS

Ligand-binding site;
Machine learning;
Molecular dynamics simulation;

Abstract With the rapid advancements in computer technology and bioinformatics, the prediction of protein−ligand-binding sites has become a central component of modern drug discovery and development. Traditional experimental methods are often constrained by long experimental cycles and high costs; therefore, the development of accurate and efficient computational methods is of paramount significance for conserving time and cost. This review comprehensively summarizes the methodological

*Corresponding authors.
E-mail addresses: tangbufu@zju.edu.cn (Bufu Tang), qiuzhengang@gmu.edu.cn (Zhengang Qiu), luopeng@smu.edu.cn (Peng Luo).
[†]These authors made equal contributions to this work.

Peer review under the responsibility of Chinese Pharmaceutical Association and Institute of Materia Medica, Chinese Academy of Medical Sciences.

https://doi.org/10.1016/j.apsb.2025.10.032

Allosteric site;
Cryptic site;
Drug discovery;
Druggability assessment;
GPCRs

advancements and current applications in the field of screening for druggable protein target sites, systematically comparing the fundamental principles, advantages, and disadvantages of four main categories of methods: structure- and sequence-based methods, machine learning-based methods, binding site feature analysis methods, and druggability assessment methods. Subsequently, by integrating classic case studies, this paper elaborately discusses the technical support and theoretical guidance afforded by the screening of protein druggable target sites for drug discovery and drug repositioning. Finally, this paper thoroughly explores the current challenges inherent in the field of protein−ligand binding site prediction, with a particular focus on future technological trends, systematically elucidating the developmental prospects and potential applications of these predictive methods.

1. Introduction

The identification and screening of protein targets are critical stages in modern drug discovery and development, possessing particular significance for enhancing the efficiency of drug research and development. A drug target refers to the site of interaction between a drug molecule and a specific biomolecule within the body, such as a key protein, which plays a crucial role in the pathogenesis and progression of diseases. The identification of drug targets is of paramount importance for elucidating disease mechanisms, revealing drug modes of action, and evaluating therapeutic efficacy[1−4]. Protein ligand-binding sites (LBSs) are typically composed of amino acid residues within approximately 5 Å of the protein−ligand atoms and generally feature a substantial binding interface and suitable hydrophobic properties[5−8]. However, this definitional method lacks an in-depth assessment of biological behaviors such as mutations, and more precise methods for defining these sites are urgently needed[9]. In recent years, the problem of drug resistance has become increasingly prominent, and the discovery of novel protein targets offers an important avenue for accelerating new drug development and mitigating drug resistance[10]. Concurrently, advancements in computational techniques such as high-throughput screening, molecular docking, and molecular dynamics simulations have enabled the large-scale screening of druggable targets[11−17].

In recent years, with the rapid advancements in computer technology and bioinformatics[18−23], significant progress has been made in methods for predicting protein binding sites and assessing druggability. Traditional experimental methods, such as X-ray crystallography, can yield high-precision structural information, but their inherent limitations include long experimental cycles, high costs, the requirement for native ligands, and difficulties in capturing dynamic conformations—issues that are particularly prominent when studying complex targets like cryptic sites[24−28]. To address these challenges, computational methods, particularly those integrating artificial intelligence and molecular dynamics simulation, have emerged as powerful tools for identifying these elusive binding sites[29]. Compared to experimental methods, computational approaches do not require reliance on known binding residues and, when combined with high-throughput screening techniques, can significantly conserve time and costs[30−32]. Notably, the deep integration of structural biology, bioinformatics, and artificial intelligence has provided powerful research tools for studying protein interaction sites. Unlike traditional methods that primarily rely on protein and ligand similarity, novel algorithms capable of integrating multi-source information have substantially improved target selectivity and affinity[33]. The introduction of artificial intelligence technology has significantly accelerated the prediction process for protein binding sites, providing new technical support for enhancing the efficiency of drug target identification[6,34,35].

Although existing protein binding site prediction methods have achieved significant progress, numerous technical bottlenecks persist concerning prediction precision, computational efficiency, and method universality. Current prediction methods are primarily based on the analysis of static protein structures and sequence information, exhibiting limitations in predicting non-deep-pocket binding sites and accounting for the dynamic changes of protein conformations, thereby failing to comprehensively reflect the biological characteristics of protein−ligand interactions[27,34,36]. "Undruggable" targets on intrinsically disordered proteins (IDPs), along with special binding sites such as covalent modification sites and allosteric regulatory sites, possess significant value for drug development; however, the predictive performance of traditional computational methods for these sites remains suboptimal[34,37−39]. In the face of rapidly growing proteomics data, current prediction methods exhibit limited predictive capabilities for proteins with unknown structures, and their prediction accuracy significantly decreases when sequence homology is low[25,27,34,40]. Therefore, developing novel prediction methods that can effectively integrate multi-dimensional data and employ innovative computational strategies to enhance the accuracy and computational efficiency of binding site identification will be a key research focus within this field moving forward.

This review systematically summarizes the methodological advancements and current application status within the field of screening for druggable sites on protein targets. The paper first emphasizes structure-based screening strategies, including methods such as geometric conformational analysis, energetic calculations, and molecular dynamics simulations. Subsequently, it systematically discusses sequence-based analysis methods, including evolutionary conservation assessment and sequence feature recognition. Furthermore, it deeply explores the application of artificial intelligence methods in this domain, focusing on the implementation strategies of machine learning, particularly deep learning techniques. On this basis, the paper systematically summarizes the feature characterization of binding sites and quantitative druggability assessment methods, and, in conjunction with typical case studies, elaborates on the practical applications of these methods in novel drug discovery and drug repositioning. The development and application of these methods provide a critical theoretical foundation and technical support for innovative drug research and

development, significantly enhancing research and development (R&D) efficiency and optimizing development costs. However, current research still faces technical bottlenecks, including limited prediction accuracy, difficulties in accurately capturing the dynamic features of protein conformations, and insufficient integration of multi-source information. Future research directions should focus on developing high-precision prediction algorithms, constructing multi-dimensional information integration platforms, and establishing experimental validation systems to enhance the accuracy and practical value of druggable target screening, thereby providing more reliable theoretical guidance for innovative drug discovery.

2. Binding site prediction methods

The prediction of binding sites constitutes a crucial foundation for screening druggable targets. In the early stages of research, binding site prediction primarily relied on manually defined rules and the comparative analysis of known sites[41]. With the rapid advancements in computer technology and bioinformatics, automated pocket identification techniques have become predominant, currently encompassing three main categories of prediction methods: structure-based, sequence-based, and machine learning-based approaches (Table 1; Fig. 1).

2.1. Structure-based methods

2.1.1. Geometric and energetic methods

2.1.1.1. Cavity search-based methods. Cavity search-based methods identify potential protein pockets by analyzing the three-dimensional geometry of the protein surface to locate large cavities on the protein structure. These methods do not require energy calculations or consideration of ligand binding, offering the advantage of enabling rapid identification of potential binding sites on the protein surface[6,42,43].

Representative tools from early geometric methods include POCKE[44] and LigSite[45]; these methods employ a geometric grid to cover the target protein and measure pocket size based on corresponding "protein−solvent−protein" events, and are characterized by fast computation and simple operation. SURFNET[46] screens for candidate pockets by generating molecular surfaces from three-dimensional (3D) coordinates and placing probe spheres in the interstices. PASS[47], in contrast, analyzes buried regions within the protein structure, enabling rapid and efficient analysis of medium-sized proteins. Fpocket[48] extracts binding pockets in proteins through Voronoi tessellation and shape analysis, ranks these pockets using partial least squares fitting, and selects predicted binding sites by combining physicochemical descriptors, druggability scores, and visual inspection. Due to its ease of use, capability for standalone analysis, and fast computation speed, Fpocket has become one of the more popular tools for analyzing protein binding pockets in recent years. However, while Fpocket achieves high recall in predicting potential protein binding pockets, it still outputs many false positives. PocketPicker[49], an automated geometry-based technique, selects non-protein points using burial level and subsequently clusters them into pockets.

Cavity search-based methods offer advantages such as high efficiency, speed, and ease of operation, but they also possess certain limitations. Firstly, these methods primarily analyze the static structure of proteins, neglecting protein flexibility and thus failing to represent dynamic changes that occur during the binding process[34]. Secondly, current methods provide a limited

Table 1 Summary of protein druggable target site screening tools.

Method	Type	Year	Algorithmic approach	Advantage	Limitation	Application
GRID	Energy-based evaluation methods	1985	Computing interactions between chemically selective probes and pockets, including electrostatic, hydrogen bonding, van der Waals, and entropic interactions.	Supporting multiple probes.	High computational cost.	Binding hotspot identification.
POCKET	Cavity-based search approaches	1992	Using modified marching cubes algorithms and regular grid points with test sphere placement to identify protein cavities and their surrounding amino acids.	Rapid, efficient, and interactive.	Prone to missing deep or small cavities.	Conventional pocket screening.
VOIDOO	Cavity-based search approaches	1994	Locating surface cavities through analysis of the molecular surface of target proteins.	Precisely delineating cavity boundaries, capable of distinguishing structural features of connectivity with the exterior.	Slow computation speed, dependent on high-resolution structures.	Conventional pocket screening.
SURFNET	Cavity-based search approaches	1995	Placing probe spheres in gaps between protein surface atoms and using probe spheres to define pockets.	Considering the conservation degrees of residues in pockets.	Requires manual adjustment of probe parameters.	Pocket visualization and interface analysis.

LIGSITE	Cavity-based search approaches	1997	Exploring protein−solvent −protein events using geometric cubic grids based on atomic coordinates.	Rapid and efficient, with reduced dependency on pocket shape and orientation.	Prone to missing complex sites.	Statistical analysis and classification of large-scale pockets.
CAST	Cavity-based search approaches	1998	Geometric computational methods based on alpha shapes and discrete flow theory.	Capable of identifying auxiliary pockets around main active sites.	Complex output results.	Conventional pocket screening.
PASS	Cavity-based search approaches	2000	Analyzing buried regions in protein structures through iterative filling of probe spheres based on spherical probes.	Detecting hidden sites.	High computational resource consumption.	Interactive molecular modeling, protein database analysis, and active virtual screening work.
PSIPRED	Sequence pattern recognition	2000	Analyzing position-specific scoring matrices (PSSMs) generated by PSI-BLAST using dual feedforward neural networks.	High-precision secondary structure prediction.	Does not directly detect binding sites, and requires a combination with other tools.	Protein structure functional annotation.
CASTp	Cavity-based search approaches	2003	Calculating pocket or cavity areas and volumes through solvent-accessible surface models (Richards surface) and molecular surface models (Connolly surface) based on alpha shapes and pocket algorithms developed in computational geometry.	Online tool.	Protein molecular weight significantly affects computation speed.	Protein site functional annotation.
ConSurf	Evolutionary conservation analysis	2003	Identifying functional regions in proteins by mapping phylogenetic information onto surfaces.	Visualization-friendly.	Dependent on multiple sequence comparison quality.	High-throughput characterization of protein functional regions.
Conseq	Evolutionary conservation analysis	2004	Evaluating conservation by combining evolutionary rates of amino acids with relative solvent accessibility.	Applicable to functional annotation of proteins lacking crystal structures.	Lacking structural feature analysis, prone to false negatives.	Large-scale sequence analysis.
Q-SiteFinder	Energy-based evaluation methods	2005	Calculating van der Waals interaction energies between probes and proteins using methyl probes.	Capable of detecting sites with smaller volumes.	Only applicable to hydrophobic probes, prone to missing sites with other features.	High-precision identification of binding sites.
MultiSeq	Multimodal fusion methods	2006	Achieving cross-fusion in the field of bioinformatics analysis through integrated cross-references serving as informal versions of ontology-driven knowledge extraction and discovery.	Performing multiple comparisons of sequence and structural information of proteins and nucleic acids, supporting multi-format inputs.	Relatively low efficiency in processing large datasets.	Binding site conservation analysis.
CAVER	Cavity-based search approaches	2006	Constructing grids and stripping to convex hulls, combined with cost functions to evaluate free space around nodes.	Identifying hidden sites and channel pathways.	Complex parameters.	Protein channel dynamics analysis.

(*continued on next page*)

Table 1 (*continued*)

Method	Type	Year	Algorithmic approach	Advantage	Limitation	Application
LIGSITEcsc	Evolutionary conservation analysis	2006	Using Connolly surfaces and defining surface-solvent-surface events, and ranking pockets according to the conservation degree of involved surface residues.	Combining pocket conservation and physicochemical properties, an extension of LIGSITE.	Not applicable to dynamic structures.	Conventional pocket screening.
PocketPicker	Cavity-based search approaches	2007	Converting binding site grid structures into descriptors of relevant vectors.	Adapting to various pocket shapes.	Not suitable for large systems.	Site structural similarity analysis.
PocketDepth	Cavity-based search approaches	2008	Clustering protein surface concave regions using depth features.	Incorporating pocket depth features.	Dependent on reference point selection.	Conventional pocket screening.
HOLLOW	Cavity-based search approaches	2008	Capturing cavity and channel profiles in protein structures by filling virtual atoms.	Channel visualization.	Long computation time.	Protein channel and internal surface characterization.
Fpocket	Cavity-based search approaches	2009	Analysis through Voronoi tesselation and alpha sphere clustering.	Open-source, efficient, and rapid.	Outputs false-positive pockets.	Conventional pocket screening.
Roll	Cavity-based search approaches	2009	Detecting protein pockets and cavities using rolling spheres.	Combining grid systems with probe spheres.	Dependent on protein 3D structures.	Conventional pocket screening.
SplitPocket	Cavity-based search approaches	2009	Combining weighted Delaunay triangulation and discrete flow algorithms.	Combining pocket ensembles, hydrophobicity, and evolutionary conservation features.	High computational complexity.	Fine pocket profiling.
SITEHOUND	Energy-based evaluation methods	2009	Identifying potential ligand binding sites by calculating non-bonded interaction energies between chemical probes and proteins.	Supporting multiple probes, open-source.	Not applicable to flexible structures.	Reverse virtual screening.
SiteMap	Energy-based evaluation methods	2009	Comprehensively evaluating site druggability by combining multiple binding metrics, including SiteScore, Dscore, and Ssize.	Integrating multiple accessibility and druggability metrics for evaluation.	Requires commercial licensing.	Drug design.
ConCavity	Evolutionary conservation analysis	2009	Creating grids, extracting pockets, and mapping residues based on protein structure and evolutionary characteristics.	Combining evolutionary sequence conservation analysis with three-dimensional structure prediction.	Dependent on sequence alignment quality, high computational complexity.	Conventional pocket screening.
RF-score	RF	2010	Based on non-parametric machine learning methods, directly learning affinity prediction patterns from existing protein–ligand complex data.	Based on non-parametric machine learning, fully data-driven.	Dependent on manual feature engineering.	Virtual screening.

McVol	Cavity-based search approaches	2010	Combining Monte Carlo random sampling methods with graph theory algorithms.	No complex parameter dependencies of traditional geometric or energy methods.	Large computational load.	Water binding sites.
MSPocket	Cavity-based search approaches	2010	Identifying potential binding regions using surface vertex normal vector directions based on molecular surface geometric features.	No grid representation required, unrestricted by protein directionality.	Complex output results.	Detecting pockets on protein solvent-excluded surfaces.
NsitePred	SVM	2011	Combining contrastive strategies with machine learning, and integrating information extracted from sequences, evolutionary profiles, structure descriptors predicted from multiple sequences, and sequence alignments.	Fusion of multi-source information.	Prediction results dependent on input sequence quality.	Nucleic acid-protein binding site prediction.
MetaPocket	Multimodal fusion methods	2011	Combining LIGSITE cs, PASS, Q-SiteFinder, and SURFNET methods.	Comprehensive complementarity of multiple algorithms.	Dependent on submethods quality; long computation time.	Applicable to difficult-to-drug target domains.
DEPTH	Cavity-based search approaches	2011	Combining residue burial degrees with solvent-accessible surface areas (SASA).	Adaptable to binding cavities of different sizes and shapes, web-based tool.	Not suitable for large systems.	For all structural analyses based on residue depth and SASA.
FTSite	Energy-based evaluation methods	2011	Screening binding sites by placing 16 different types of small-molecule probes in dense three-dimensional grids around proteins and using empirical free energy functions.	Supporting multiple probes.	Limited in detecting complex sites.	Large-scale binding site screening.
Gaussian network model	Energy-based evaluation methods	2011	Predicting structures and thermodynamic characteristics of bound states using non-bound structures based on GNM models.	Considering protein dynamics.	Does not directly detect binding pockets, and requires a combination with other tools.	Protein dynamic functional site analysis.
VOLSITE	Cavity-based search approaches	2012	Negative images of binding cavities are based on comparing encoded shapes and pharmacophore properties at regularly spaced grid points.	Relatively insensitive to changes in atomic coordinates and grid frame definitions.	Dependent on cavities with known positions of binding ligands.	Conventional pocket screening.
dPredGB	Energy-based evaluation methods	2012	Combining rapid geometric detection with evaluation of desolvation properties of putative binding pockets.	Overcoming limitations of traditional methods that rely solely on geometric features or interaction information.	Not applicable to unbound protein structures.	Conventional pocket screening.
SiteComp	Energy-based evaluation methods	2012	Describing spatial variations in interaction energies between proteins and specific probes based on molecular interaction fields (MIFs).	Visualization, interactive, online tool.	High computational cost.	Comparing differences between similar binding sites.

(*continued on next page*)

Table 1 (*continued*)

Method	Type	Year	Algorithmic approach	Advantage	Limitation	Application
Cofactor	Evolutionary conservation analysis	2012	Combining global-to-local sequence and structure alignment algorithms.	High prediction accuracy.	Dependent on known template libraries, prone to missing novel sites.	Protein−ligand interaction prediction applicable to genome-wide structure and function annotation.
Target	SVM	2013	Predicting pockets using classifier ensembles and spatial clustering designs.	Independent of structural templates, applicable to proteins lacking homologous structures.	Not applicable to flexible structures.	Virtual screening.
TargetS	SVM	2013	Predicting binding residues along sequences through ligand-specific strategies, and further identifying binding sites of predicted binding residues through recursive spatial clustering algorithms.	Applicable to proteins lacking three-dimensional structural information or available homologous templates.	Complex parameter adjustments.	Targeting protein−ligand binding sites from primary sequences.
eFindSite	SVM	2013	Template-based prediction through fusion of meta-threading, clustering techniques, machine learning algorithms, and confidence estimation systems.	Using only weak homologous templates.	Dependent on meta-threading result quality.	Homologous protein site mining.
LISE	Cavity-based search approaches	2013	Calculating scores by counting geometric motifs extracted from interaction network substructures connecting protein and ligand atoms.	Considering multiple features, rapid and efficient.	Success rates for certain specific protein types (*e.g.*, hormone carriers or hormone-binding proteins) are expected to be significantly reduced.	Large-scale pocket screening.
COACH	Ensemble learning methods	2013	Integrating five independent prediction methods: TM-SITE, S-SITE, COFACTOR, FINDSITE, and ConCavity, combined with SVM for model training.	Comprehensive prediction through complementarity between various algorithms.	Dependent on known template libraries.	Conventional pocket screening.
S-SITE	Sequence pattern recognition	2013	Detecting protein templates and LBS through comparisons between binding site-specific sequence profiles.	No extensive parameters required.	Lacking structural feature analysis.	Conventional pocket screening.
TM-SITE	Sequence pattern recognition	2013	Comparing structural similarities of "ligand binding residue fragments (SSFL)" between target proteins and known template proteins.	Using SSFL residues for structural alignment, a composite scoring function.	Lower prediction accuracy.	Conventional pocket screening.
LigandRFs	RF	2014	Entirely sequence-based, utilizing random forest ensemble learning frameworks to identify protein−ligand binding residues from co-evolutionary contexts.	Identifying ligand-binding residues solely from sequence information.	Lower prediction performance for non-metallic ligands.	Applicable to target proteins lacking structural information.

KVFinder	Cavity-based search approaches	2014	Spatial segmentation of protein pockets based on adjustable grid densities and probe parameters.	Customizable parameters, open-source, and visualization-friendly.	Large probe "probe out" settings affect success rates.	Supporting PyMOL plugins.
Patch-Surfer2. 0	Cavity-based search approaches	2014	Representing and comparing pockets at the level of small local surface patches characterizing physicochemical properties of local regions.	Capable of identifying binding pockets for identical ligands.	Slow computation.	Virtual screening, similarity searching.
DeepBind	CNN	2015	Utilizing CNNs for automatic recognition of functional patterns at unknown positions in protein sequences.	Integration with high-throughput technologies.	High model complexity, lengthy training time.	Sequence specificity analysis of DNA and RNA binding proteins.
LigandDSES	RF	2015	Dynamically selecting and integrating relevant random forest sub-models for prediction based on similarities between query proteins and training subsets.	Dynamic integration of protein sequence information.	Limited prediction performance for non-metallic ligands.	Applicable to target proteins lacking structural information.
PRank	RF	2015	Extracting local physicochemical features of pocket points and training random forest classifiers.	Reducing false positive results of traditional methods.	Biased toward binding sites in the training dataset.	Post-processing tools for outputs of pocket prediction tools.
OSML	SVM	2015	Constructing prediction models dependent on input queries rather than global training sets based on dynamic learning frameworks.	Avoiding interference from redundant data in traditional global models affects performance.	Requires large amounts of training data.	Applicable to predictions of various binding types.
DeepSite	CNN	2017	Employing voxelized three-dimensional descriptors to extract physical and chemical features of protein atoms.	Open-source, allowing GPU acceleration.	High computational resource requirements.	Designed 3D descriptors applicable to broader biological structures.
P2Rank	RF	2018	Extracting protein surface points with feature vectors of physicochemical and geometric properties and performing random forest classification.	Independently open-source, useable as components for large data processing.	Results dependent on input structure quality.	Predicting novel allosteric sites.
COACH-D	SVM	2018	Combining multiple template-based and template-free methods.	Enhanced COACH server, improving prediction accuracy.	High computational resource consumption.	Conventional pocket screening.
DeepCSeqSite	CNN	2019	Automatically extracting multi-level features from low to high layers based on extensive sequence information.	No three-dimensional structural data required.	High computational cost for training.	Predicting metal ion and acid radical ion binding residues.
DeepDrug3D	CNN	2019	Learning patterns of specific molecular interactions between ligands and their protein targets.	Open-source with high classification accuracy.	Results dependent on input structure quality.	Orphan receptor ligand prediction, protein distant relationship analysis, and multi-target drug development.

(*continued on next page*)

Table 1 (*continued*)

Method	Type	Year	Algorithmic approach	Advantage	Limitation	Application
DeepPocket	CNN	2021	Combining Fpocket with CNN for rescoring and subregion segmentation of protein pockets.	Stable applicability across all provided protein structures, employing multi-step methods for site identification.	Dependent on initial detection by Fpocket.	Applicable to template methods that are unable to provide the required binding sites.
DeepSurf	CNN	2021	Integrating surface-based local three-dimensional grid representations with 3D-CNN (three-dimensional convolutional neural networks) architectures.	High positioning and overlap precision, independently open-source.	Dependent on high-quality protein structures.	Conventional pocket screening.
GraphBind	GNN	2021	Constructing structural graphs of target residues and their spatial neighborhoods, and utilizing hierarchical graph neural networks (HGNN) to learn local structures and biophysicochemical features.	Trained in an end-to-end manner, constructing structural context graphs.	Results dependent on structural quality and geometric knowledge.	Identifying protein-nucleic acid binding residues.
CAVIAR	Cavity-based search approaches	2021	Automatically identifying cavities in proteins using spatial search algorithms and further segmenting them into biologically meaningful subcavities.	Open-source, applicable to machine learning applications.	Limited handling of protein flexibility.	Structure-based virtual screening.
dSPRINT	Ensemble learning methods	2021	Ensemble of five classifiers: Gradient boosting, random forest, logistic regression, support vector machine, and neural networks.	Comprehensive prediction through complementarity between various algorithms.	High computational resource consumption.	Predicting interaction sites of proteins with DNA, RNA, ions, peptides, and small molecules.
GraphSite	GNN	2022	Integrating deep learning with graph representations of local protein structures.	Independent of ligand information, maintaining high performance in new datasets.	High computational resource requirements.	Binding site classification.
HoTS	Transformer	2022	Identifying binding regions through pre-trained transformer networks.	Independent of three-dimensional structural information.	Insufficient existing datasets, limiting model training.	Identifying drug-target interactions.
BindWeb	Ensemble learning methods	2022	Integrating GraphBind network based on GNN with DELIA model combining CNN and biLSTM, and clustering through mean shift algorithm.	Multi-model integration, comprehensively complementary.	Slow computation speed.	Protein functional site annotation.
DREAMM	Ensemble learning methods	2022	Combining machine learning, conformational sampling, and scoring algorithms.	No structural dynamic changes.	High computational resource consumption.	Predicting binding sites of peripheral membrane proteins.

DeepBindGCN	CNN	2023	Combining molecular vector representations with graph convolutional neural networks.	Rapid and efficient, independent of docking conformations, concisely preserving spatial information and physicochemical features.	High computational cost.	Large-scale virtual screening.
SiteRadar	GNN	2023	Utilizing GNNs for atomic-level modeling of protein three-dimensional structures.	No data preprocessing required.	Long computation time.	Applicable to complex sites including proteins with low sequence similarity, shallow sites, and covalent binding.
LigBind	GNN	2023	Pre-training ligand-residue pairs at the graph level through a relation-aware graph neural network framework, combined with ligand similarity classifiers to extract cross-ligand commonalities.	Superior performance on large-scale ligand-specific benchmark datasets.	Lower performance and sensitivity in structural prediction datasets.	Applicable to residue prediction tasks for ligands with scarce samples.
PocketMiner	GNN	2023	Combining graph neural networks with molecular dynamics simulations.	Rapid and efficient.	Requires molecular dynamics trajectory input.	Hidden pocket prediction.
Motif2Mol	Transformer	2023	Integrating Transformer architectures with biochemical language models and predicting novel active compounds based on sequence motifs of ligand binding sites.	Independent of three-dimensional structural information, capable of directly predicting active compounds from sequences.	Limited training sets, restricted applicability.	*De novo* drug design.
MaSIF	Transformer	2023	Based on Transformer architectures, converting protein binding interfaces into multiple two-dimensional "interface images" and performing multi-attention processing and contrastive learning.	Combining protein interface images with traditional energy terms.	Dependent on high-quality three-dimensional structures.	Protein−protein interaction interface recognition.
DoGSite3	Cavity-based search approaches	2023	Detecting concave regions on protein structure surfaces using Gaussian difference (DoG) filtering algorithms derived from image processing.	Considering multidimensional parameters, an online tool.	Lacking analysis of physicochemical factors.	Detecting atypical pockets.
The integration of auto-fusion with MPRL	Multimodal fusion methods	2024	MPRL extracts primary sequence (ESM-2), residue graphs (VGAE), and atomic point cloud (PAE) data, with auto-fusion performing multimodal integration.	Unsupervised, multimodal, symmetry-preserving protein representation learning.	Complex model training, requires large-scale data.	Dynamic pocket screening.
RHML	CNN	2024	By combining unsupervised clustering with a CNN-based	No predefined classification labels required, residue-level	Depends on large amounts of high-quality molecular	It can effectively identify potential allosteric sites,

(*continued on next page*)

Table 1 (*continued*)

Method	Type	Year	Algorithmic approach	Advantage	Limitation	Application
			classification model, accurate and interpretable classification of MD trajectory residues at the residue level is achieved in the absence of predefined classification labels.	interpretability.	dynamics data and high computational resource consumption.	providing a foundation for subsequent allosteric modulator discovery and mechanism elucidation.

Abbreviations: CNN Convolutional Neural Network, GNN Graph Neural Networks, RF Random Forest, SVM Support Vector Machine, MIFs molecular interaction fields, SSFL subsequence from the first binding residue to the last binding residue, LBS ligand binding site, PSSMs position-specific scoring matrices, GNM Gaussian Network Model, DoG Difference of Gaussian, SASA solvent-accessible surface areas, HGNN hierarchical graph neural networks, 3D-CNN three-dimensional convolutional neural network, RHML residue-intuitive hybrid machine learning.

understanding of the chemical properties of ligands and binding sites, and the geometric shape of a protein is not always correlated with ligand binding behavior, which can lead to prediction errors[34,50]. Furthermore, these methods often rely on complex binding calculations, resulting in high computational costs that limit their application in large-scale studies.

2.1.1.2. Energy-based assessment methods. Energy probes are specific small molecules or functional groups (*e.g.*, water molecules, methyl, amino, carboxyl, and hydroxyl groups) used to evaluate the interaction energy between a protein and potential ligands. The core principle of energy-based assessment methods is to map the ligand-binding surface of a protein by calculating the potential energy distribution at different probe positions[51,52]. Q-SiteFinder[53] identifies potential binding sites by assessing the interaction energy between the protein and van der Waals probes, subsequently performing cluster analysis and prioritizing sites based on their accumulated energy values and spatial proximity. FTSite[54] employs a multiple molecular probe strategy by constructing a computational grid on the protein surface, systematically placing probes onto this grid, and then forming consensus clusters through cluster analysis and score ranking, preferentially screening regions that interact with multiple probes to accurately predict potential LBSs. Energy-based assessment methods not only comprehensively consider intermolecular interactions but also exhibit strong tolerance to missing atoms in the input structure, enabling efficient large-scale binding site screening while maintaining prediction accuracy[6]. However, while these methods demonstrate excellent performance in predicting single binding sites, they still face limitations when dealing with complex systems involving multiple interaction sites. Concurrently, these methods may not fully simulate all biological factors influencing molecular interactions[55], which somewhat restricts their scope of application.

2.1.2. Molecular dynamics simulation methods

2.1.2.1. Mixed-solvent molecular dynamics (MixMD). Mixed-solvent molecular dynamics (MixMD)[56,57] is a probe-mapping technique that incorporates considerations of protein dynamic properties. Its core principle involves performing molecular dynamics simulations within a mixed solvent system composed of water and organic molecules, subsequently identifying potential drug binding sites through an analysis of the preferred binding locations of these organic solvent molecules. The advantage of mixed-solvent simulation techniques lies in their capacity to accurately delineate binding hotspots on the protein surface; these techniques have been successfully applied to the identification and characterization of active sites, allosteric sites, protein−protein interaction (PPI) interfaces, and cryptic pockets[50]. Studies have demonstrated that MixMD can precisely identify the allosteric site of GSK3β within a specific surface region, a finding that holds significant guiding importance for the development of allosteric inhibitors targeting GSK3β[50].

As previously discussed, the shape of a protein pocket is not the sole indicator of ligand binding behavior; in contrast to the geometry-based Fpocket method, MixMD can capture precise binding sites on the protein surface founded upon molecular dynamics, thereby offering improved precision[50]. When compared to the FTMap technique, which similarly maps hotspots, MixMD can more accurately capture the dynamic changes of allosteric proteins during the binding process[50,58]. However, for more complex systems that necessitate helix or domain movements, MixMD, even when employing accelerated dynamics, may not

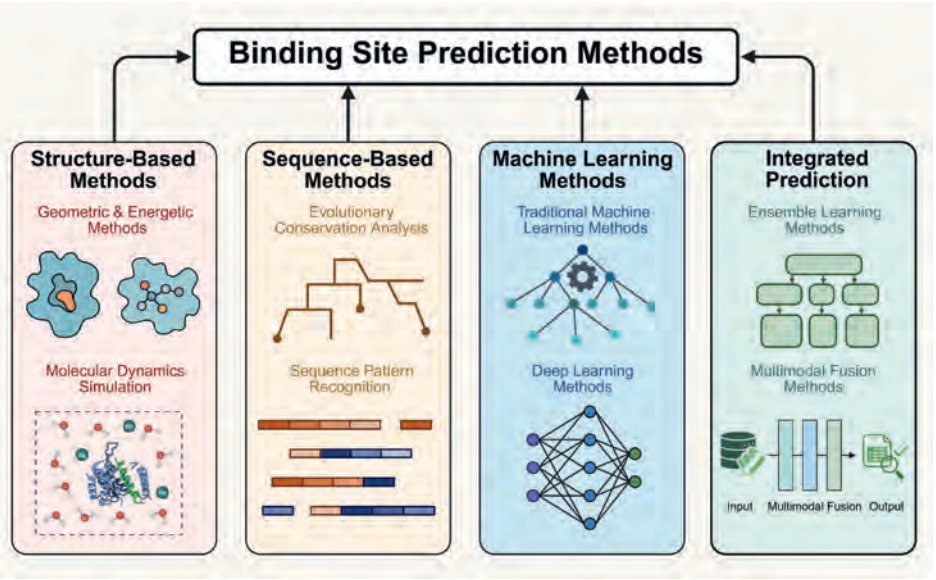

Figure 1 Overview of protein druggability target screening tools. This figure depicts the four categories of protein druggability target screening tools discussed in this section, which include structure-based methods, sequence-based methods, machine learning approaches, and integrated prediction methods that combine multiple approaches.

comprehensively capture all binding sites, consequently leading to longer computation times and higher computational costs[59].

2.1.2.2. Site-identification by ligand competitive saturation (SILCS). Site-identification by ligand competitive saturation (SILCS)[60] is a computational method that combines Grand Canonical Monte Carlo (GCMC) and Molecular Dynamics (MD) simulations to identify potential LBSs. This method explores the chemical space of a protein by simulating the interactions between the target protein and small-molecule probes representing common functional groups. Subsequently, it constructs functional group free energy maps (FragMaps) by normalizing and applying Boltzmann weighting to the probability distributions of these functional groups. FragMaps facilitate the simultaneous consideration of crucial factors such as protein conformational flexibility, functional group specificity, and protein desolvation, thereby offering multifaceted guidance and significant application value for drug discovery[61−63]. A notable advantage of the SILCS method is its capacity to employ multiple solvent probe molecules concurrently, thereby enabling a comprehensive probing and characterization of potential binding sites on the protein surface[60]. However, the selection of solvent probes necessitates careful consideration of the binding site's spatial characteristics. For example, probe molecules that are excessively large may be unable to access narrow channels within the protein, potentially leading to the omission of some critical binding sites[62].

2.1.2.3. Markov state models (MSMs). Markov state models (MSMs)[64−66] are a class of methods that partition configurations obtained from molecular dynamics simulations into multiple non-overlapping microstates based on geometric criteria. These microstates are subsequently used to construct a metastable state space wherein transitions between states follow a Markov process, ultimately facilitating a microscopic-to-macroscopic model transformation through coarse-graining. MSMs can reveal the transition mechanisms between different protein conformations, demonstrating considerable utility in fields such as allosteric site identification and protein interaction analysis.

A significant advantage of MSMs is their ability to extract long-timescale kinetic features from short-timescale molecular dynamics simulations, construct stochastic kinetic models, and enable the automatic identification of metastable states, thereby overcoming the timescale limitations of traditional MD[67,68]. Limited sampling can lead to poorly sampled microstates; the Nyström method, based on algebraic principles, overcomes this issue by mitigating interference from limited sampling in the identification of metastable states[64]. Despite their numerous advantages, the limitations of MSMs cannot be ignored: MSM sampling capability is constrained by the simulation timescale, and events that occur beyond the accessible timescale of the MSM cannot be captured. Furthermore, the accuracy of MSMs depends on the precision of the force field; although existing force fields perform well in many systems, they may exhibit limitations when addressing more complex systems. Lastly, while MSMs can establish quantitative relationships with experimental data, the direct correlation of simulation results with experimental observations remains a pressing issue requiring further methodological improvements[66].

2.1.2.4. Enhanced sampling methods. Although traditional molecular dynamics simulations can facilitate the study of protein structural changes and conformational fluctuations through conformational sampling, their sampling efficiency in exploring optimal conformations has not yet reached an ideal level. This is primarily because energy barriers within low-energy regions hinder transitions between adjacent low-energy conformations of the system[69]. In recent years, researchers have developed various enhanced sampling strategies to improve the efficiency of exploring conformational space. Among these, Gaussian accelerated Molecular Dynamics (GaMD) achieves efficient sampling of biomolecular conformations through unconstrained enhanced sampling. This method primarily smooths the original potential energy surface by

introducing a harmonic boost potential, thereby lowering the energy barriers of the system[70]. Multicanonical Molecular Dynamics (McMD), by applying a bias to accelerate dynamics, enables the study of phenomena such as intermolecular binding processes at atomic resolution, capabilities currently unattainable by experimental methods or traditional MD simulations[71,72].

Enhanced sampling methods exhibit numerous significant advantages. Firstly, these methods can accurately predict the native binding conformations of compounds and elucidate molecular binding mechanisms by analyzing binding and dissociation processes, thereby providing micro-level insights for protein binding site prediction. Furthermore, by analyzing binding mechanisms, dynamic docking can guide the modification of functional groups, optimize binding characteristics, and enhance compound affinity and specificity. More importantly, enhanced sampling methods have a broad scope of application and are suitable not only for academic research but also for drug optimization in industrial settings, aiding in the study of high-affinity, high-specificity drugs. Notably, although enhanced sampling algorithms can effectively overcome the multiple-minima problem and improve sampling efficiency, their complex implementation, high computational cost, and sensitivity to parameter settings limit their widespread application in macromolecular systems[69]. Recent studies have demonstrated that combining enhanced sampling with machine learning can overcome some of these limitations, thereby enabling the discovery of previously elusive cryptic sites. Within the extensive trajectories generated by enhanced sampling, Chen et al.[29] effectively integrated machine learning to achieve automatic mining of key conformational features, overcoming the traditional limitations of relying on prior knowledge to preset coordinates and avoiding the misidentification or omission of important conformations caused by human bias. Particularly for capturing open conformations that do not necessarily correspond to the lowest energy states or MSM macrostates, developing unbiased universal identification methods has become crucial.

2.1.2.5. Reversed allostery-driven cryptic pocket discovery. Traditional allosteric site prediction methods have primarily focused on communication signals propagating from allosteric sites to orthosteric sites. However, recent advances in allosteric theory have revealed that allosteric coupling is bidirectional, with orthosteric perturbations capable of modulating allosteric sites through reversed allosteric communication. This paradigm shift has opened new avenues for discovering cryptic allosteric sites that are not visible in static protein structures but emerge only under specific dynamic conditions. The reversed allosteric communication theory, pioneered by Ni et al.[73] proposes that signals can propagate from orthosteric sites to allosteric sites, facilitating ab initio identification of allosteric sites without prior knowledge of their locations. This approach offers a more physiologically relevant method compared to techniques that rely on artificial perturbations, as it uses natural orthosteric ligand binding as the perturbation source. The computational framework integrates large-scale accelerated molecular dynamics simulations with Markov State Models to capture conformational dynamics and identify metastable states triggered by orthosteric ligand binding. The AlloReverse web server, developed by Zha et al.[74] represents a practical implementation of this theory, providing multiscale analysis of multiple allosteric regulations and revealing hierarchical relationships among different allosteric pathways. This method has demonstrated success in identifying novel allosteric sites on proteins such as CDC42 and SIRT3, with experimental validation confirming the functional relevance of the predicted sites.

2.2. Sequence-based methods

2.2.1. Evolutionary conservation analysis

During the evolutionary process of proteins, functionally critical amino acid residues often exhibit high conservation, as natural selection pressure primarily acts on domains possessing important biological functions[75]. Based on this evolutionary theory, researchers analyze sequence variation patterns in protein families or genes across different species to thoroughly investigate their evolutionary history and conserved structural-functional features, thereby accurately identifying domains with important biological functions in proteins[76,77]. ConSurf[78], a classic LBS prediction tool based on evolutionary conservation analysis, operates on the core principle of identifying functional regions by mapping phylogenetic information onto the protein structure's surface. The latest version of ConSurf integrates an optimized algorithmic framework that not only provides more precise conservation scoring methods but also introduces a confidence index assessment system, significantly enhancing the accuracy of its prediction results. Cofactor[75] is a next-generation prediction tool that innovatively integrates structural and sequence alignment algorithms, employing structural modeling and a combined global-local similarity search strategy to accurately identify interaction sites among protein molecules or between proteins and other molecules. Through large-scale benchmark tests and blind test experiments, Cofactor has been demonstrated to be significantly superior to existing state-of-the-art prediction methods in terms of prediction accuracy.

2.2.2. Sequence pattern recognition

With the rapid development of high-throughput sequencing technologies, vast amounts of novel protein sequence data are determined and publicly released each year. Consequently, sequence template-based LBS prediction methods have garnered widespread attention. The core principle of these methods is to use sequence alignment algorithms to perform homology analysis between the target protein sequence and known protein sequences, followed by selecting the optimal template based on sequence similarity. Subsequently, potential LBSs on the target protein are predicted by analyzing information from experimentally verified ligand-binding residues in the template[51]. PSIPRED[79], a widely validated prediction tool, can perform high-precision secondary structure prediction based on input protein sequences. The synergistic application of two complementary algorithms, TM-SITE and S-SITE, significantly enhances the prediction accuracy of protein binding sites. S-SITE[80] is based on binding-specificity sequence profile-profile alignments, while TM-SITE primarily performs structural comparison of subsets of continuously distributed residues in the target protein pocket; their combined use can significantly improve both prediction reliability and coverage.

Sequence-based prediction methods rely solely on one-dimensional amino acid sequence information for analysis, requiring no three-dimensional structural data support, and thus offer the advantage of high computational efficiency. However, LBSs exhibit high conservation at the spatial or conformational level, whereas sequence-level conservation is relatively lower; this lack of structural information limits the prediction accuracy of sequence-based methods. With the development of machine

learning, vast amounts of sequence information can be processed and its constituent patterns extracted, enabling sequence-based methods to overcome some of their previous limitations[27,55,80].

In traditional computational methods, structure-based approaches often require detailed molecular structure information, while sequence-based methods rely on the comparative analysis of amino acid sequences; both types of methods can be inefficient and cumbersome when dealing with complex biological systems. Through learning from large volumes of sequence data and model optimization, machine learning can analyze complex protein binding site information, extract key features, identify underlying patterns, and significantly improve prediction efficiency[81,82], thereby proving useful for efficiently processing massive datasets, reducing computational costs, and accelerating large-scale screening processes[83].

2.3. *Machine learning methods*

Given the rapid advancements in artificial intelligence technology, machine learning has demonstrated immense potential in the field of drug discovery[84−86]. Early research primarily focused on the construction of association models and descriptors for target proteins and ligands, utilizing classification algorithms, such as support vector machines, for binding site prediction. More recently, researchers have developed diverse deep neural network models to generate string representations of novel compounds, thereby fostering the synergistic integration of structural drug design and deep learning technologies[87].

2.3.1. *Traditional machine learning methods*

Traditional machine learning methods for LBS primarily employ two main strategies: one being structure-based direct prediction methods that entail extracting and computationally analyzing the spatial and chemical features of the target protein for feature recognition, and the other being rule-based prediction methods that predict and rank candidate sites[81]. Machine learning methods offer several advantages. They can effectively handle complex data that is challenging to interpret using solely physical or geometrical approaches. Additionally, they efficiently process vast amounts of structural, environmental, and feature information associated with protein binding sites. Moreover, their computational cost and time efficiency are considerably superior to those of traditional prediction methods, rendering them particularly suitable for large-scale site screening[83]. However, machine learning methods also present certain limitations. Their predictive performance is highly dependent on the accuracy of feature engineering, which can potentially lead to false positives, such as the incorrect identification of non-druggable regions. Concurrently, an over-reliance on data correlation at the expense of establishing causality can compromise prediction accuracy, underscoring the importance of enhancing model generalization capabilities to effectively address practical challenges in virtual screening[55,88].

Among traditional machine learning methods, Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting Decision Tree (GBDT) are three of the most representative algorithms. This section will provide a systematic analysis of the fundamental principles, application scenarios, and respective advantages and disadvantages of these methods.

2.3.1.1. Support vector machine (SVM). Support vector machine (SVM)[89] is a machine learning algorithm for binary classification problems; its core mechanism involves transforming input vectors into a high-dimensional feature space through nonlinear mapping and constructing a linear decision boundary within this space. This decision boundary possesses unique mathematical properties that ensure the algorithm's high generalization ability, thereby enabling accurate differentiation of sample data belonging to different classes. Owing to its excellent classification performance and high-dimensional data processing capabilities, SVM has demonstrated significant advantages in the field of protein binding site prediction[89]. Furthermore, SVM can integrate multidimensional information—such as geometric features, interaction potentials, offsets, conservation scores, and properties of the surrounding pocket—to provide more comprehensive and accurate binding site predictions that are significantly superior to traditional prediction methods[90].

Machine learning prediction tools based on support vector machines primarily include Target[91], eFindSite[92], and COACH[93], among others; these tools combine protein structure modeling with the identification of homologous proteins with bound ligands in the PDB database to infer the binding sites of target proteins[94]. Target exhibits high accuracy in protein−ligand binding site prediction due to its powerful feature extraction capabilities and classification techniques, but it is computationally time-consuming. eFindSite can effectively handle low-quality protein structures (such as homology models and weak homology templates) and demonstrates excellent performance in large-scale proteome analysis, particularly in scenarios where only the target protein sequence information is available. COACH, by integrating multiple prediction methods, not only fully utilizes structural and sequence information but also adapts to structural data of varying quality, thereby demonstrating excellent predictive performance in complex systems[94]. The recently developed SVM prediction tool, COACH-D[95]—an enhanced version of the COACH server—predicts LBSs by combining multiple template-based and template-free methods, and constructs protein−ligand complexes using consensus prediction and molecular docking algorithms. Benchmark test results indicate that COACH-D significantly reduces spatial clashes between ligands and protein structures (by 85%) through molecular docking and outperforms other LBS prediction methods in blind tests, thereby providing a more accurate prediction tool for drug discovery research.

2.3.1.2. Random forest (RF). The random forest (RF) method[96] effectively addresses the limitations of traditional decision tree methods by constructing multiple decision trees through random sampling in different subspaces of the feature space. The RF algorithm can efficiently handle hundreds to thousands of molecular descriptors and enhances the model's generalization ability and robustness through ensemble learning, rendering it particularly suitable for handling large-scale imbalanced datasets and demonstrating superior sensitivity and overall performance in bioactive compound classification tasks[97,98]. P2Rank[99] is a protein binding site prediction tool based on the random forest algorithm, which achieves prediction by classifying protein surface atoms as belonging to or constituting binding sites. This method comprehensively analyzes the local environmental features of solvent-accessible surface (SAS) points—including physicochemical properties, geometric configurations, and statistical information—and assigns scores to these points by combining the feature vectors of neighboring atoms. Compared to existing methods that require input or integration with other bioinformatics tools or databases, P2Rank does not require complex

preprocessing steps and supports automated prediction *via* a command-line interface. Furthermore, P2Rank can directly process multi-chain structures and identify potential binding sites formed by multi-chain residues, exhibiting outstanding performance in terms of computational efficiency and prediction accuracy, which makes it particularly suitable for the rapid and precise prediction of novel allosteric sites[99−102].

2.3.1.3. Gradient boosting decision tree (GBDT). The gradient boosting decision tree (GBDT)[103], an ensemble supervised machine learning algorithm extensively utilized for classification and regression problems, iteratively trains decision trees through the integration of multiple weak learners and optimizes the loss function *via* the gradient descent method, thereby continuously enhancing the model's predictive accuracy. As an advanced iteration of the gradient boosting algorithm, the Extreme Gradient Boosting algorithm (XGBoost)[104] significantly enhances model training efficiency through the introduction of parallel computing and distributed learning mechanisms. Concurrently, this algorithm optimizes out-of-core computation strategies, which enables the processing of large-scale data under memory-constrained conditions, thereby facilitating significant breakthroughs in system scalability and overall performance. Moreover, the XGBoost algorithm supports custom objective functions and evaluation metrics, a feature that further enhances the model's predictive accuracy[105,106]. The SimBoost algorithm acquires features by constructing target-target similarity, drug−drug similarity, and drug−target interaction networks, and then employs these features as input to a gradient boosting regression tree for training[107].

2.3.2. Deep learning methods

As an advanced machine learning technique, deep learning utilizes complex learning mechanisms, achieved by constructing artificial neural networks inspired by the human brain's structure, and has demonstrated breakthrough progress in multiple fields, including image classification and natural language processing[51,108−112]. Deep learning models can automatically learn from fundamental protein features (including amino acid sequences, three-dimensional structures, and surface characteristics) to efficiently identify binding sites. Through the integration of neural network models with spatial information, deep learning techniques can effectively identify and predict potential binding pockets.

2.3.2.1. Convolutional neural networks (CNNs). The convolutional neural network (CNNs)[108], a representative deep learning model, is widely used in image and video processing; however, specific processing strategies are often required when addressing non-Euclidean data such as protein structures and three-dimensional point cloud data. The basic workflow for applying CNNs to protein analysis includes data preprocessing and integration of binding site information, construction of the convolutional neural network model, and selection of classification models based on protein binding site feature metrics[113]. Through multi-scale spatial feature extraction, convolutional neural networks can effectively capture local and global feature patterns in data, thereby exhibiting strong feature adaptability. Furthermore, through the optimization of weight matrices during the learning process, CNNs can minimize prediction errors, demonstrating excellent prediction accuracy in protein binding site prediction tasks[114,115]. DeepSurf[116] innovatively integrates surface-based local three-dimensional grid representations with a three-dimensional convolutional neural networks (3D-CNNs) architecture, achieving prediction accuracies of 67% and 44.5% on the Coach420 and Holo4K standard datasets, respectively, and significantly outperforming other existing deep learning tools. DeepBindGCN employs an innovative graph representation method, wherein protein pocket residues and ligand atoms are constructed as graph nodes, and it maintains spatial structural information and physicochemical features through adjacency relationships. Distinct from traditional methods, DeepBindGCN discards conformational search strategies, thereby substantially enhancing prediction efficiency; a single prediction, for instance, requires only 0.0004 to 0.0012 s during virtual screening processes[117]. Beyond direct druggability prediction, CNNs have also been developed for complex conformational analysis and cryptic site identification. For example, Chen et al.[29] developed a residue-intuitive hybrid machine learning (RHML) framework that incorporates an interpretable CNN multiclassifier. This RHML model addresses the challenge of the lack of predefined category labels in molecular dynamics (MD) trajectories by combining unsupervised clustering with CNNs. Importantly, it employs a Local Interpretable Model-Agnostic Explanation (LIME) framework to provide residue-level interpretability, thereby identifying key residues that define specific conformational states, including those associated with cryptic allosteric site opening. This interpretability feature is crucial for translating complex deep learning outputs into actionable biological insights, thereby advancing drug discovery efforts.

2.3.2.2. Graph neural networks (GNNs). Graph neural networks (GNNs)[118,119] are a class of deep learning architectures specifically designed for processing non-Euclidean, graph-structured data. GNNs have demonstrated significant advantages in handling non-Euclidean data and have been successfully applied across various research fields, including text classification, traffic flow prediction, and physical system simulation. In the biomedical field, GNNs have been widely applied to tasks such as quantum property prediction, molecular fingerprint generation, protein interface identification, and drug−target interaction prediction. This approach not only simplifies the molecular modeling process but also exhibits rotational and translational invariance, while concurrently enhancing model training efficiency and effectively handling complex node feature information. However, GNNs continue to face challenges in three-dimensional voxel segmentation tasks; their computational efficiency is lower than that of 3D-CNNs, and they may incur greater computational overhead[81]. GraphSite[120] combines deep learning with graph representations of local protein structures, achieving precise classification of LBSs *via* an improved graph neural network architecture. This method effectively extracts structural features, physicochemical properties, and evolutionary information of binding pockets using neural weighted message passing layers, significantly reducing the risk of overfitting and improving classification accuracy. This approach can achieve prediction without ligand information, exhibits strong generalization capability, effectively identifies non-binding sites, significantly reduces the false positive rate, and shows considerable promise for practical application.

2.3.2.3. Transformer. The transformer model, a deep learning architecture founded upon a self-attention mechanism, was initially applied in the field of natural language processing. The core principle of this model involves dynamically assigning attention weights to different sequence positions, thereby effectively capturing the relative importance of various segments of the sequence[121,122]. In the domain of protein binding site prediction,

the Transformer model analogizes protein sequences to language text. It learns sequence context information through pre-training to construct efficient biological feature representations, which are subsequently applied to tasks such as structure prediction, functional analysis, and stability assessment, thereby significantly advancing the development of protein representation learning. The introduction of Transformers has significantly enhanced the capacity of machine learning models to capture long-range dependencies between sequence elements, facilitating a deeper understanding of sequence features[123]. The Motif2Mol model[87] integrates the Transformer architecture with a biochemical language model and is capable of predicting novel active compounds based on the sequence motifs of LBSs. This model demonstrated excellent performance in large-scale validation involving 228 human kinase sequence motifs and protein kinase inhibitor (PKI) data, not only achieving high prediction accuracy rapidly but also consistently maintaining a low validation loss. In 2024, Ryuichiro Ishitani and colleagues[81] developed an innovative method by fusing a rule-based approach with a graph transformer neural network to construct a LBS prediction model. Compared to traditional 3D-CNNs, this model can incorporate multiple atomic and residue features as node attributes, enhancing its ability to process complex input data and demonstrating significant advantages in predicting complex protein binding sites.

2.4. Integrated prediction by combining multiple methods

2.4.1. Ensemble learning methods

In the early stages of machine learning development, researchers primarily relied on single models to address various prediction tasks. However, single models, such as support vector machines, often exhibit limitations, including overfitting, underfitting, and low computational efficiency when processing complex, high-dimensional data. Ensemble learning models, by leveraging the strengths of multiple learning algorithms, can predict protein binding sites more comprehensively and accurately. COACH[93] integrates five independent prediction methods—TM-SITE, S-SITE, COFACTOR, FINDSITE, and ConCavity—and employs a Support Vector Machine (SVM) for model training, leveraging the complementarity of these algorithms to achieve genome-wide annotation from protein sequence to structure and function. In a test set of 500 non-redundant proteins, COACH's predictions demonstrated a 12.5% improvement in the Matthews Correlation Coefficient (MCC) value compared to single algorithms, and it exhibited excellent performance in the CAMEO experiment with an AUC (Area under the ROC curve) score of 0.87, highlighting its superior accuracy and coverage relative to single algorithms. dSPRINT[124], an ensemble machine learning method, predicts not only protein−LBSs but also interaction sites with DNA, RNA, ions, and peptides, significantly expanding its prediction scope; concurrently, this method can identify binding sites and predict the binding properties of structural domains. In the field of protein−membrane interaction research, peripheral membrane proteins (PMPs) are important therapeutic targets for various diseases; however, due to the complexity of the protein−membrane interface, the lack of structural information, and the absence of computational workflows, PMPs are often categorized as difficult-to-drug proteins. The DREAMM[125] server employs an ensemble machine learning algorithm specifically designed for membrane interface identification and binding pocket prediction in PMPs. Additionally, the XGBoost model[126] integrates the advantages of tree-based and linear models. In research on the development of CYP450 inhibitors, investigators utilized XGBoost and other learning models to analyze molecular descriptors and fingerprint features. The XGBoost model subsequently demonstrated optimal classification performance on the test set (average prediction precisionof 90.4%), surpassing a deep learning model developed using a multitask deep autoencoder neural network (88.5%)[127].

2.4.2. Multimodal fusion methods

Because protein binding site prediction involves complex physicochemical properties, sequence information, and structural details, among other multidimensional data, traditional methods struggle to effectively integrate these diverse types of information, resulting in significantly limited prediction accuracy. Multimodal fusion methods refer to the integration of multisource information, including protein sequence and structure, to fully leverage the complementarity and diversity of various data types. This approach facilitates a comprehensive analysis of protein characteristics and binding behaviors. An exemplary multimodal fusion method is an LSTM-based classifier[128]. This method innovatively uses volumetric representation to integrate the three-dimensional structure of proteins with the biological attributes of amino acids (*e.g.*, hydrophobicity, isoelectric point, and charge). It also extracts sequence features using autocovariance and joint-triad algorithms and finally constructs a multimodal input feature set for joint modeling and prediction. Fusing protein sequence and structural information can help reveal evolutionary changes in sequence, structure, and function. MultiSeq[129], a unified bioinformatics analysis environment, achieves cross-fusion of these domains by employing integrated cross-referencing, which functions as an informal version of ontology-driven knowledge extraction and discovery. Specifically, MultiSeq supports automatic metadata download and data redundancy removal, which aids in accelerating the evolutionary analysis of sequence and structural data. Furthermore, it can be integrated with various common bioinformatics tools (*e.g.*, HMMER) and possesses multifunctional input/output capabilities, thereby reducing task execution time and cost. Auto-Fusion is an adaptive fusion technique that effectively merges multimodal features by learning to compress information from different modalities while preserving their contextual information[130]. MPRL is an innovative framework that learns unified unsupervised representations of proteins by combining primary and tertiary structural data, thereby overcoming the limitations of traditional pre-training methods. This tool utilizes Auto-Fusion to synthesize joint representations, thereby enhancing the capacity for multimodal information extraction and creating robust and comprehensive protein representations[131]. In a series of task executions, MPRL achieved an accuracy of up to 83.9% in enzyme identification and performed exceptionally well in analyzing protein fold classification, demonstrating strong adaptability and competitiveness.

3. Binding site feature analysis

A systematic analysis of the structural and functional features of predicted binding sites (Fig. 2) not only facilitates a comprehensive assessment of their druggability as pharmaceutical targets but also reveals their potential biological functions, thereby offering a theoretical basis for drug molecule design and optimization and consequently accelerating the new drug development process.

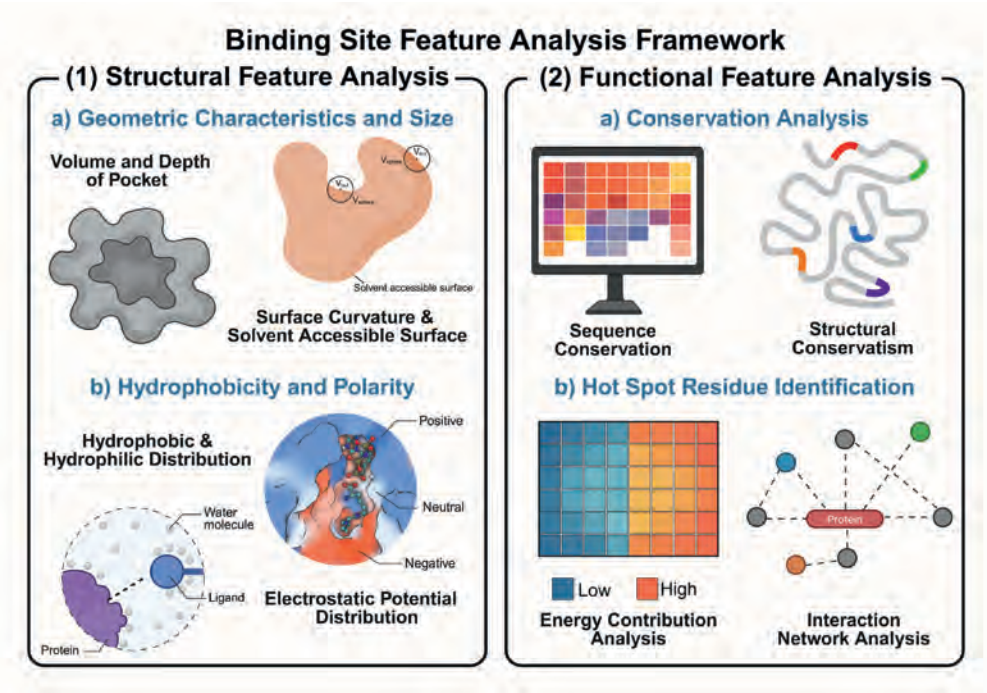

Figure 2 Binding site feature analysis. This figure depicts the parameters considered during binding site feature analysis. These encompass structural features (including geometry and size, hydrophobicity, and polarity) and functional features (such as sequence and structural conservation, and hotspot residue identification).

3.1. Structural feature analysis

Shape descriptors that quantify parameters of protein structures, including features such as volume, depth, surface curvature, and accessibility, are widely applied in the study and comparison of protein−ligand interactions, PPI, and their binding sites. The biological function of a protein is primarily determined by the three-dimensional structure of its specific binding sites. These functional binding sites are predominantly distributed on the protein's molecular surface, and their specificity is jointly determined by the spatial arrangement of dozens of non-hydrogen atoms and their physicochemical properties[132]. Most small-molecule ligands typically bind to large cavities or concavities on the protein surface, given that a sufficiently large interaction interface is necessary to generate strong binding affinity[133]. Therefore, identifying LBSs based on the geometric features and spatial dimensions of protein binding pockets remains a prevalent strategy in this research field.

3.1.1. Geometric shape and size

3.1.1.1. Pocket volume and depth. As early as the 1990s, protein binding site prediction methods based on geometric shape and spatial dimensions had achieved significant progress; these methods are primarily categorized into grid-based and grid-free approaches. The fundamental principle of grid-based methods (*e.g.*, POCKET[44] and LigSite[45]) is to cover the protein structure with a three-dimensional grid and identify possible binding pockets by searching for protein−solvent−protein interfaces at the grid points. Grid-free methods mainly include those based on Voronoi diagrams and probe spheres; for instance, SURFNET[46] identifies potential binding pockets by analyzing three-dimensional (3D) spatial coordinates and placing probe spheres in the interstices of the molecular surface. Fpocket[48] identifies and extracts binding pockets in proteins using a Voronoi surface tessellation algorithm combined with geometric shape analysis. With advancements in research, combining machine learning with molecular dynamics simulations to detect and analyze pocket volume and depth has gradually become a prominent research focus, and this approach has demonstrated high computational efficiency and prediction accuracy. MDpocket[134], building upon the binding pockets identified by Fpocket, assesses the conformational stability of potential binding pockets by analyzing trajectory data from the final 5 ns of GaMD simulations[70]. P2Rank[99,102] employs a random forest algorithm to discretize the protein surface into a set of points. These points are subsequently annotated with the physicochemical properties and geometric features of surrounding atoms and residues, thereby improving the accuracy of binding site prediction.

3.1.1.2. Surface curvature and accessibility. In the early 21st century, researchers proposed an innovative method for measuring protein surface curvature. This method involves fitting a sphere to a protein surface patch using the least squares fitting (LSF) method, with the radius serving as the curvature metric. Furthermore, the problem can be simplified to a plane fitting problem through geometric inversion transformation[135]. The solvent accessible surface area (ASA) of a protein is an important parameter that characterizes the degree of exposure of amino acids on the protein surface. ASA is not only significant for maintaining protein stability but was also commonly used in early studies to predict accessibility based on amino acid sequences, in a manner similar to secondary structure prediction methods[136]. ASAView[137] is an algorithmic tool that represents the solvent accessibility of amino acid residues using two-dimensional helical wheel diagrams and bar charts. This tool provides a visual display of the protein and its chains, which

facilitates rapid assessment of the exposed surface area and the overall topological distribution of residues within the protein. With enhanced computational capabilities and improved statistical analysis methods, researchers have incorporated ASA to improve the prediction accuracy of binding sites. For example, InterProSurf predicts potential interacting residues based on solvent accessible surface area, interface propensity scales, and a clustering algorithm. It also utilizes five quantitative descriptors to analyze the physicochemical differences between interface and surface regions, thereby effectively identifying potential functional sites[138]. As research has progressed, many tools have gradually adopted methods that integrate multiple physicochemical properties to predict protein binding sites. The CASTp server[139,140] utilizes a 1.4 Å probe radius to calculate protein active sites, evaluate solvent accessible surface area, and measure pocket volume, area, and degree of opening. Additionally, it further validates the credibility of prediction results in conjunction with the SCFBio and Depth servers. PocketPicker[49] identifies potential binding pockets in proteins by calculating the "buriedness" of binding sites and analyzing their volume and accessibility based on grid points. This method can not only predict the location of binding sites but also provide detailed shape descriptions, including pocket depth and spatial distribution.

3.1.2. Hydrophobicity and polarity

3.1.2.1. Hydrophobic/hydrophilic distribution. Protein−ligand interaction sites typically exhibit significant hydrophobic characteristics[8]. Therefore, analyzing the distribution of hydrophobic and hydrophilic regions on the protein surface can provide an important basis for identifying potential binding sites. In early studies, HINT[141] quantitatively assessed protein−ligand interactions by integrating hydrophobic atomic constants (a) with the solvent-accessible surface area (S). This method, based on the CLOGP theory proposed by Hansch and Leo, quantifies atomic hydrophobicity using hydrophobic atomic constants (a) and utilizes the solvent-accessible surface area (S) to correct for the interaction effects of molecules in exposed or buried states. The HINT system ultimately generates an interaction score (where positive values indicate favorable interactions and negative values indicate unfavorable interactions) to characterize the hydrophobic/hydrophilic properties of binding sites. Recently, Joel Graef's research team[142] developed DoGSite3, a grid-based automated pocket detection tool that employs a Difference of Gaussians (DoG) filtering algorithm for binding pocket prediction; this method comprehensively considers multidimensional parameters such as pocket volume, hydrophobicity, enclosure, and depth. This method achieved 99% coverage of binding sites in the SCPDB database and ranked highest in accuracy assessments among various pocket prediction methods.

3.1.2.2. Electrostatic potential distribution. The electrostatic properties of a protein are primarily determined by the spatial distribution and relative proportions of its polar and charged residues. Electrostatic interactions play a crucial role in the formation of protein−protein complexes, not only by enhancing binding specificity but also in the process of optimizing binding affinity; high-affinity antibodies often exhibit stronger electrostatic characteristics[143]. Implicit solvent methods[144,145] are important computational approximation techniques that consider the average effect of the solvent environment on the system rather than explicitly simulating solvent molecules; these methods have achieved significant results in research areas such as PPI, protein−ligand binding thermodynamics, conformational scoring, and peptide and protein folding. Among these, the Poisson-Boltzmann (PB) electrostatics method[145] is widely used in biological research; this method can estimate the electrostatic solvation energy of solutes and characterize the electrostatic potential distribution on macromolecular surfaces by solving the PB equation[146]. The APBS software package is one of the primary tools for evaluating electrostatic interactions in biomolecular systems within an implicit solvent environment. To enhance its utility, researchers developed the iAPBS interface library, which enables seamless integration of APBS with mainstream molecular dynamics simulation programs (such as Amber, CHARMM, and NAMD), thereby incorporating the calculation of electrostatic and solvation energies and forces into biomolecular simulations[147]. With the increasing demand for electrostatic calculations, the GUI-based PBEQ-Solver tool was developed. This tool not only solves the Poisson-Boltzmann equation and visualizes the electrostatic potential distribution of proteins, but also calculates electrostatic potentials, solvation free energies, protein interaction energies, and pK_a values of titratable residues, while supporting calculations in both aqueous and membrane environments[148].

3.2. Functional feature analysis

3.2.1. Conservation analysis

3.2.1.1. Sequence conservation. Unlike traditional evolutionary sequence conservation analysis, sequence conservation analysis based on functional features primarily integrates sequence conservation information with experimentally validated functional data. For instance, FireStar[149] predicts protein functional residues by extracting structural information from remote homologs and combining it with the local sequence conservation of small-molecule ligand-binding residues from the FireDB database, along with catalytic residue annotation information from the Catalytic Site Atlas (CSA).

3.2.1.2. Structural conservation. Compared to sequence conservation, protein binding sites often exhibit higher conservation in their spatial structures; consequently, analysis methods based on structural conservation can more accurately predict potential binding sites[27,150]. The conserved features of binding pockets aid in the discovery of new binding sites; however, high sequence and structural conservation can also lead to cross-recognition issues among targets. For example, the protein kinase family generally possesses a highly conserved ATP-binding pocket. This structural feature can cause kinase inhibitors to exhibit broad target overlap upon binding, leading to non-specific binding and adverse reactions[151].

3.2.2. Hotspot residue identification

3.2.2.1. Energy contribution analysis. Hotspot residues are key amino acid residues that significantly contribute to the binding free energy during PPIs or protein−ligand binding processes[152]. The Molecular Mechanics Poisson-Boltzmann Surface Area (MM-PBSA) method can accurately predict and identify hotspot residues by considering conformational flexibility during the protein binding process. The FTMap[142,153] algorithm utilizes 16 small-molecule chemical probes of varying sizes, shapes, and polarities to identify and locate binding hotspots on the protein surface

through energy clustering analysis, thereby determining optimal binding sites.

3.2.2.2. Interaction network analysis. Residue interaction networks (RINs) represent protein structures as network models where amino acid residues serve as nodes and interactions between residues serve as edges. Results indicate that RINs play a crucial role in bioinformatics applications[154]. These networks can function as a protein interaction detection method, often incorporating a graph attention framework, and are capable of constructing condition-specific PPI networks and capturing key interaction features, thereby enabling precise identification of protein functions.

4. Druggability assessment methods

To achieve stable binding between proteins and ligands, it is necessary not only to examine the geometric features and key residue distribution of protein binding pockets but also to assess their physicochemical compatibility with potential ligands[152]. Therefore, protein druggability assessment is crucial for screening high-quality targets and primarily includes methods based on physicochemical properties and machine learning (Fig. 3).

4.1. Physicochemical property-based assessment

4.1.1. Pocket volume, hydrophobicity, and other parameters

4.1.1.1. SiteMap. SiteMap[155], a specialized computational tool for protein structure analysis, can effectively identify protein binding sites and assess their druggability by systematically analyzing protein structures to evaluate the spatial size, structural characteristics, and solvent exposure of LBSs[156]. The fundamental principle of this tool is to establish a multidimensional scoring system by thoroughly analyzing the geometric features and physicochemical properties of protein structures, thereby comprehensively evaluating the druggability characteristics and ligand-binding potential of these sites. The SiteMap system

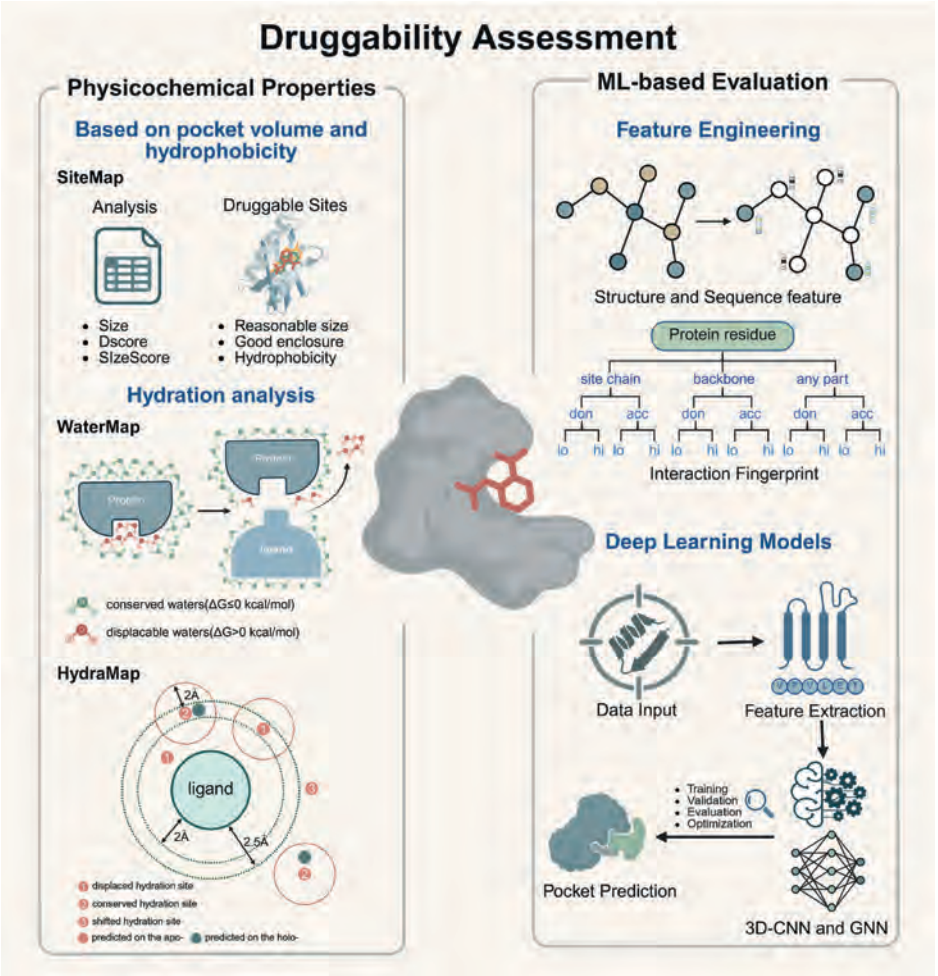

Figure 3 Druggability assessment methods. This figure illustrates the core mechanisms of druggability assessment methods. SiteMap scores potential druggable targets based on parameters such as pocket volume and hydrophobicity, using three core metrics: SiteScore, Dscore, and Size. WaterMap determines the likelihood of water molecules being displaced by ligands by analyzing the thermodynamic properties of water molecules on the protein surface. HydraMap performs hydration analysis by comparing the intrinsic potential energies of apo- and holo-proteins. Machine learning-based methods include feature engineering (which extracts structural and sequential features and interaction fingerprints) as well as deep learning models based on 3D-CNNs (three-dimensional convolutional neural networks) and GNNs (Graph Neural Networks).

integrates multiple scoring function modules, primarily including three core indicators: SiteScore, Dscore, and Size. Among these, SiteScore is specifically used to evaluate ligand-binding potential; Dscore focuses on quantitatively characterizing the druggability of the binding site and assessing its capacity to accommodate drug molecules; and the Size indicator is primarily used for accurately measuring the spatial dimensions of the binding site, providing an important reference for assessing the suitability of specific ligand molecules[156]. In large-scale validation studies, SiteMap has demonstrated excellent predictive performance, accurately identifying over 96% of known binding sites, with an accuracy rate exceeding 98% for predicting tightly bound sites[157]. In the field of binding site analysis, this tool can accurately distinguish between ligand-binding and non-binding sites, while also providing quantitative and visual analysis data for drug molecule design. It effectively guides ligand structure optimization to enhance drug molecule efficacy and improve their physicochemical properties. Although SiteMap has certain limitations in identifying the allosteric sites of some kinases[50], its efficient and precise computational performance has made it an indispensable analytical platform in drug discovery and lead compound optimization processes, playing a significant role in virtual screening and binding site assessment [157].

4.1.2. Hydration analysis

4.1.2.1. WaterMap. WaterMap[158,159], a computational tool based on molecular dynamics, can thoroughly analyze the thermodynamic properties of water molecules on a protein surface using statistical mechanical principles. This method offers significant advantages in characterizing solvent effects, accurately assessing the spatial positions and energy parameters (including entropy, enthalpy, and free energy) of water molecules, and predicting their impact on ligand-binding affinity. WaterMap plays a crucial role in structure-guided drug design and virtual screening studies, effectively revealing the mechanisms of water molecule-mediated intermolecular interactions in protein binding pockets. The primary advantage of this method is its ability to comprehensively consider the dynamic behavior of water molecules, thereby aiding in the design and optimization of drug molecules with higher activity. However, the accuracy of this method in predicting the spatial positions and energy parameters of water molecules possesses certain limitations; therefore, it is recommended that it be used in conjunction with computational methods such as MM-GB/SA(Molecular Mechanics Generalized Born Surface Area) or FEP(Free Energy Perturbation) to achieve higher prediction accuracy.

4.1.2.2. HydraMap. HydraMap[160,161] is an analysis tool based on the weighted potential of mean force (W-PMF) derived from protein−ligand complex crystal structures. It generates density maps by evaluating the positional probabilities of water probes in the binding pocket and converts these maps into explicit hydration sites using cluster analysis. In the validation of its ability to reproduce known hydration sites, this method achieved a success rate of approximately 70%. Although HydraMap offers significant advantages in computational efficiency and prediction accuracy, its limitations include failing to evaluate the statistical potential of ligand−water interactions and neglecting the rearrangement effects of water molecules during ligand binding. These factors may affect the accuracy of the prediction results. To address these issues, researchers have introduced molecular dynamics (MD) simulation techniques to evaluate the statistical potential of ligand−water interactions and compare the intrinsic potentials of apo- and holo-proteins, thereby enhancing prediction accuracy.

4.2. Machine learning-based assessment

4.2.1. Feature engineering

4.2.1.1. Structural and sequence feature extraction. Given the rapid development of machine learning technology, data-driven methods that leverage protein sequence and structural features can effectively identify binding sites and their potential ligands[162], thereby providing an important basis for target druggability assessment. The novel computational tool DeepProSite[27] integrates protein structural and sequence information, utilizes ESMFold and pre-trained language models (such as Transformer) to construct protein feature representations, and subsequently employs graph node classification algorithms and a Graph Transformer to deeply analyze the relationship network among amino acid residues, thereby achieving precise prediction of protein binding sites. Researchers have developed a deep learning model based on convolutional neural networks (CNNs) that innovatively combines one-dimensional SMILES representations with the sequence information of protein binding pockets rich in structural motifs. This model demonstrates significantly superior computational efficiency and accuracy in protein−ligand interaction prediction compared to traditional machine learning methods[163].

4.2.1.2. Interaction fingerprint features. Interaction fingerprints (IFPs) are a digital, one-dimensional method for representing the interactions between amino acid residues in the binding site of a protein−ligand complex and the ligand molecule. This method transforms the three-dimensional structural information of protein−ligand complexes into a series of binary bits (0 or 1), where each bit specifically characterizes the presence or absence of a particular type of interaction[164]. Molecular Surface Interaction Fingerprinting (MaSIF)[165] and its differentiable variant dMaSIF[166] achieve efficient identification of complementary interacting surfaces by systematically analyzing the geometric conformations and chemical properties of protein surfaces, significantly reducing computational costs. Moreover, they can also be used for pocket functional classification. However, the use of both methods requires computationally expensive pre-processing in their original implementations[167]. MDFit[168] generates compiled simulation fingerprints (SimFPs), assisting users in resolving key protein−ligand interactions and ranking ligands based on compatibility. Recent machine learning models, by integrating the interaction fingerprints generated from molecular docking simulations, provide new informational dimensions for protein−ligand interaction studies. Among these approaches, utilizing different feature types for similarity analysis of training and test sets for various subtypes can improve the accuracy of CYP450 inhibition prediction[169]. Compared to traditional static protein−ligand interaction fingerprints, MDFit, by incorporating molecular dynamics simulations, can more comprehensively assess the dynamic behavior of ligands. This capability provides more effective features for machine learning models, thereby improving drug efficacy prediction accuracy and deepening the understanding of action mechanisms. Furthermore,

MDFit's lower dependence on chemical similarity allows for broader sampling in larger ligand libraries, further enhancing prediction accuracy[168].

4.2.2. *Deep learning models*

Traditional machine learning methods, such as support vector machines (SVMs), primarily rely on manually engineered feature inputs, and their feature transformation processes focus on partitioning the sample space. In contrast, deep convolutional neural networks with stacked hierarchical architecture (DCS-SI) can autonomously extract low-dimensional features from data and progressively integrate them into high-dimensional feature representations, significantly enhancing the model's feature representational capabilities. Furthermore, deep learning methods can be trained without explicit data sampling processes and exhibit characteristics markedly superior to those of traditional machine learning methods in handling imbalanced data distributions, automatic feature extraction, and capturing long-range dependencies[170].

4.2.2.1. 3D convolutional neural networks. 3D-CNNs can effectively capture the spatial features of local protein regions by performing convolutional operations on input three-dimensional structural descriptors[171]. DeepCSeqSite, by integrating residue prediction methods, achieves precise identification of residues at potential LBSs on the protein surface. Validated on 151 non-redundant protein samples and 3 extended test sets, DeepCSeqSite demonstrated significant performance advantages, with its MCC and precision metrics improving by at least 0.05% and 15%, respectively, compared to existing methods[55,172]. The AK-Score algorithm employs convolutional layers to extract protein−ligand interaction features and, combined with Grad-CAM technology, enables visual analysis of the contribution of these features to the prediction results[124]. DeepSite[171] uses voxelized three-dimensional descriptors to extract the physicochemical features of protein atoms and outputs the probability of binding site locations *via* a 3D convolutional neural network. In a benchmark test set composed of 7622 proteins from the SCPDB database, DeepSite's DVO index (0.648) was significantly superior to those of Fpocket and Concavity, indicating that DeepSite can provide more accurate binding site predictions than the other two methods. DeepSite learns entirely from data, independent of traditional rule-based or physics-based methods. Furthermore, this method can be accelerated by GPUs, significantly enhancing the model's generalization ability and prediction efficiency.

4.2.2.2. Graph neural networks. Graph neural networks (GNNs) take protein atomic point clouds as direct input and require no complex hyperparameter configuration, which greatly simplifies the modeling process. Due to their inherent rotational and translational invariance, no additional data augmentation is needed, thereby enhancing model training efficiency. This method can flexibly handle node features, effectively supporting the use of complex, multidimensional node features, particularly the high-dimensional outputs of protein language models. However, GNNs are more challenging to apply and less efficient than 3D-CNNs in 3D voxel segmentation around binding sites. Handling node diversity and maintaining rotational invariance can lead to increased computational overhead[81]. GraphSite[120], an innovative deep learning-based method, utilizes GNNs to achieve a graphical representation of local protein structures for the classification of LBSs. This method employs neural weighted message-passing layers to effectively capture the structural, physicochemical, and evolutionary features of binding pockets, thereby reducing the risk of model overfitting and improving classification accuracy. In calculations performed on a negative dataset of surface pockets not bound to any ligands, GraphSite achieved a median probability of 0.93, further validating the accuracy of GraphSite's classification. GraphBind [173], on the other hand, employs a hierarchical graph neural network to embed contextual information from protein structures into graph features for identifying nucleic acid-binding residues. It also separately trains predictors for DNA- and RNA-binding residues, thereby constructing datasets of DNA- and RNA-binding proteins. In the DNA-binding test set, t-SNE projection results showed that original graph feature vectors could not distinguish between binding and non-binding residues, whereas the latent graph feature vectors learned by GraphBind clustered most binding residues together, fully demonstrating GraphBind's superiority in identifying nucleic acid-binding residues.

5. Application cases

Protein binding site prediction and target druggability assessment methods play crucial roles in multiple key stages of modern drug discovery and development, and hold particular significance in novel drug target discovery and drug repositioning research. These computational methods provide a reliable theoretical basis for subsequent target validation and drug molecule design by systematically analyzing potential binding sites within protein three-dimensional structures and their druggability features.

5.1. *Novel drug target discovery*

5.1.1. *Classical target binding sites*

5.1.1.1. Kinase inhibitor design. Over the past two decades, the design of kinase inhibitors[151,174,175] has made significant breakthroughs, primarily benefiting from the flourishing development of genomics, structural biology, and high-throughput screening technologies. These advancements have enabled kinase inhibitors to demonstrate significant pharmacological value in the treatment of various diseases. However, despite numerous research achievements, critical issues such as insufficient selectivity, the development of resistance, and side effects continue to constrain their clinical application. Benefiting from improvements in human genome annotation and the expansion of kinase profiling technologies, the research and development of kinase inhibitors have made further breakthroughs, especially in the design of highly selective inhibitors. However, drug development targeting different kinase-driven mutations still requires further in-depth optimization. Future drug research and development is trending towards personalization and precision, wherein the combination strategy of multi-target kinase inhibitors is expected to become an important breakthrough for overcoming drug resistance. Furthermore, with the continuous development of structural chemistry and computational chemistry, the molecular design of kinase inhibitors will become more precise. This advancement will not only help improve their therapeutic effects but also significantly reduce the occurrence of adverse reactions.

5.1.1.2. G protein-coupled receptor targeting. G protein-coupled receptors (GPCRs)[176,177] are one of the most important

classes of current drug targets, accounting for approximately 35% of US Food and Drug Administration (FDA)-approved drugs. GPCRs represent the largest superfamily of membrane proteins and can be classified into five distinct subfamilies: the rhodopsin-like family (Class A), the secretin/adhesion family (Class B), the glutamate-like family (Class C), the smoothened/frizzled family (Class F), and the taste 2 family (Class T). GPCRs consist of seven transmembrane helical domains (7TM) that are connected by three extracellular loops (ECL1-3) and three intracellular loops (ICL1-3), playing a crucial role in signal transduction from extracellular to intracellular environments, thereby regulating various physiological processes[178]. GPCRs possess structurally diverse binding pockets capable of forming highly specific interactions with various ligands, a characteristic that renders them ideal targets for drug design. To facilitate the research and development of GPCR-targeted drugs, the GPCRdb project has constructed a systematic, template-based database that covers extensive GPCR structural information. Currently, numerous computational tools and algorithms have been developed for predicting GPCR binding sites. Examples of these tools include PocketMatch[179], eF-seek[180], and Patch-Surfer [181]; these methods assess site similarity by analyzing the geometric features and electrostatic properties of binding pockets[182]. In addition to traditional prediction methods, ligand interaction fingerprint technology, which leverages global analysis and activation state specificity, is emerging as a novel research tool for defining GPCR docking sites[183]. However, the low subtype specificity and significant adverse effects of GPCR ligands limit their therapeutic potential, and current drug development strategies primarily focus on addressing these two critical limitations. With the advancement of computational and experimental technologies, particularly artificial intelligence (AI)-driven approaches, a deeper understanding of GPCR structure, function, and signal transduction has been achieved, further promoting the development of new drugs, which primarily include selective ligands, biased ligands, and allosteric ligands, multi-target drugs, and antibody drugs[178].

The trace amine-associated receptor (TAAR) family belongs to class A G-protein-coupled receptors (GPCRs) and exerts extensive biological functions in neurological and metabolic systems. In a 2023 study, the research team utilized cryo-electron microscopy to resolve high-resolution ligand−TAAR1 complex structures, and through systematic comparison of nine complex structures, they not only identified a highly conserved core binding site on TAAR1 but also revealed two expandable subpockets distributed around this site that can accommodate structurally diverse small molecule ligands. More importantly, through mutational validation, conformational comparison, and molecular dynamics simulations, the study further confirmed the roles of these pockets in functional selectivity during signal transduction, establishing the structural foundation for drug polypharmacology. This work clearly demonstrates that incorporating high-resolution structural information cannot only clarify the location and nature of binding sites but also reveal hidden subpockets and their regulatory roles, thereby providing a structural basis for rational drug design[184].

5.1.2. Novel allosteric regulatory sites

Protein allosteric sites, serving as critical functional regulatory regions, play an essential role in diverse fields including drug discovery, gene transcription regulation, disease diagnosis, and enzyme catalysis, thereby offering substantial opportunities for the development of innovative therapeutic strategies[185]. Methods for identifying allosteric sites have made substantial advances in recent years and can be classified into two principal categories: experimental techniques and computational methods. Experimental techniques encompass X-ray crystallography, nuclear magnetic resonance spectroscopy, hydrogen-deuterium exchange mass spectrometry (HDXMS), and the integration of X-ray crystallography with high-throughput screening. Computational methods involve molecular dynamics simulations, residue topology analysis (RTA) algorithms, elastic network models, and normal mode analysis (NMA); simultaneously, machine learning approaches based on allosteric site features have also gained widespread application[185,186].

5.1.2.1. Protein kinase allosteric inhibitors. Protein kinase domains are important targets in drug discovery; however, traditional orthosteric protein kinase binding sites are plagued by issues such as high sequence conservation and susceptibility to drug resistance, which severely restrict the development of corresponding drugs. Allosteric sites offer advantages such as high selectivity and reduced off-target side effects; in-depth research into these sites holds promise for overcoming the aforementioned limitations[187−190]. However, the identification of allosteric sites still faces numerous challenges: first, compared to orthosteric sites, allosteric sites lack clear structural feature definitions[187]; second, it is difficult to accurately assess their functional impact using only static protein structures, a problem particularly prominent for allosteric regulation induced by structural changes[189]. Colabind[34], a molecular dynamics simulation method that utilizes a cloud computing platform, employs molecular probes for coarse-grained simulations and can effectively identify binding sites on protein structures, particularly known drug allosteric sites and PPI sites. A recently developed method allows for the generation of flexible allosteric protein conformations with multiple binding sites; this method has been experimentally validated on adenylate kinase (ADK), thus providing an innovative research approach for targeting allosteric proteins[191]. Furthermore, the introduction of artificial intelligence technology has provided robust technical support for the identification and development of allosteric sites[34].

5.1.2.2. GPCR allosteric modulators. G protein-coupled receptors (GPCRs), a class of transmembrane proteins renowned for their high conformational flexibility, exhibit significant dynamic characteristics within their LBSs. The traditional view, which considers only a single binding pocket, is insufficient for comprehensively elucidating their functional features; therefore, accurately modeling GPCR binding sites is of paramount importance for drug research and development[192]. GPCR allosteric modulators typically bind to dynamic cryptic pockets on proteins, and their selectivity primarily depends on the differences in the conformational 'open' states of these cryptic pockets[193]. The high selectivity of allosteric modulators makes it possible not only to significantly reduce or avoid the risk of drug overdose but also to achieve tissue-specific targeting, thereby enhancing therapeutic efficacy[192,194]. Although a few GPCR allosteric modulators have been approved for clinical use, their systematic development and broad application in drug discovery remain limited due to technical challenges. In recent studies, Chen et al.[29] employed an integrated approach combining RHML and MD to reveal a previously unreported cryptic allosteric site on the β2-adrenergic receptor (β2AR) within the G protein-coupled receptor (GPCR) family, thereby identifying a negative allosteric modulator (ZINC5042). Through rigorous experimental validation, ZINC5042 demonstrated effects comparable to those observed with G protein bias and exhibited unique pathway

specificity, thereby demonstrating potential as a next-generation β-blocker and novel pharmacological tool compound. This indicates that the computational workflow combining RHML and MD possesses significant potential for discovering allosteric sites and drugs and elucidating their regulatory mechanisms in other target proteins.

5.1.3. *Covalent inhibitor targeting strategy*

Covalent inhibitors achieve prolonged and highly selective inhibition through selective covalent modification of nucleophilic residues (Cys/Lys/Tyr, etc.) within target proteins, currently demonstrating therapeutic value across numerous human diseases, particularly in anticancer drug development[195−197]. Although traditional Cys-targeting strategies have achieved success (such as EGFR/BTK inhibitors), their therapeutic efficacy remains compromised due to limitations including residue scarcity and drug-resistant mutations[198,199]. Non-cysteine covalent inhibitors targeting alternative nucleophilic residues (such as lysine, tyrosine, and serine) are progressively revealing their pharmaceutical value and may further address drug resistance mutations and toxicity challenges. Presently, emerging non-cysteine targeting technologies (such as sulfonyl fluoride warheads and reversible boronic acid designs) are significantly expanding the target space of covalent drugs[197]. Protein binding site screening tools have the potential to accelerate the rational design of next-generation covalent inhibitors by precisely identifying modifiable residues and optimizing warhead binding patterns, thereby providing breakthrough pathways for "undruggable" targets. CavityPlus is a web server based on three-dimensional protein structures that identifies binding sites on protein surfaces through the CAVITY algorithm and ranks them by combining ligandability and druggability scores; the CovCys module can automatically detect druggable cysteine residues, which is particularly useful for identifying novel binding sites for the design of covalent allosteric ligands[200]. CoDEL technology was initially applied to the discovery of covalent inhibitors targeting cysteine (Cys); subsequently, researchers discovered tyrosine (Tyr)-targeting covalent inhibitors through the combined ABPP-CoDEL strategy, thus significantly expanding the scope of covalent drug development. Recently, researchers have innovatively integrated lysine (Lys) reactivity proteomics data to guide CoDEL screening and successfully discovered lysine-targeting covalent inhibitors with novel structures and diverse reaction mechanisms. This advancement extends the applicability of covalent targeting strategies to important non-cysteine residues, thereby opening new avenues for the design of non-Cys covalent inhibitors[201−204].

5.2. *Drug repositioning*

5.2.1. *Drug repurposing based on binding site similarity*

Drug repurposing, also known as drug repositioning, refers to the research strategy of discovering new therapeutic applications for existing drugs[205,206]. Binding site similarity is a key determinant in drug repurposing, founded on the principle that ligands can interact with multiple protein binding pockets that share common structural features, exhibiting "cross-reactivity" or "cross-sensitivity"[207]. Identifying binding pockets with similar amino acid arrangement patterns allows for the prediction of potential off-target effects of specific ligands, thereby guiding the development of drug molecules with fewer adverse reactions. Concurrently, the characterization and analysis of amino acid arrangement features within binding pockets provide an important basis for structure-based drug design, contributing to the development of highly selective drug molecules targeting specific protein families or individual proteins[205]. SiteMine, a binding site search tool that utilizes databases and structural alignment, exhibits high sensitivity and accuracy; however, its performance is limited by specific application domains and database completeness[208]. ProCare is an innovative computational method that employs point cloud registration algorithms to assess the similarity between protein binding pockets. This method, through the calculation and utilization of local descriptors, possesses unique advantages in identifying local similarities between binding cavities of different sizes[209]. A recently developed, highly scalable binding site alignment algorithm enables rapid and accurate alignment at the residue level, providing an effective tool for proteome-scale identification of similar binding sites[210].

5.2.2. *Virtual screening and repositioning strategies*

In large-scale drug virtual screening, the core task is to efficiently and accurately identify high-affinity binders from vast small molecule libraries; this binding affinity is primarily influenced by the protein's binding pocket, the ligand's spatial conformation, and the residue and atom types involved[117]. Predictive analysis of high-ranking binding sites, identified during the exploration of binding pockets, provides an important basis for subsequent virtual screening[211]. DataDTA, a deep learning-based affinity prediction tool, integrates protein pocket descriptors, compound SMILES strings, and algebraic graph features, and utilizes a dual interactive aggregation neural network for multi-scale interaction feature learning, thereby achieving accurate predictions of protein−ligand binding affinity[8].

Specialized databases such as Cavbase and CavitySpace can analyze the geometric shapes and chemical similarity features of protein binding sites through clustering methods. ProBis-Dock, on the other hand, performs 3D spatial labeling of binding sites and uses these labels as modeling input, subsequently constructing a unified protein−ligand feature space using template-matched ligands. These methods not only aid in the identification of protein binding sites but also allow for the querying of sites with small molecules, thereby providing a theoretical basis for the implementation of drug repositioning strategies[207].

6. Challenges and prospects

Although significant progress has been made in methods for screening protein druggable sites, numerous challenges remain, necessitating innovative breakthroughs in methodology and technology to further enhance prediction accuracy, efficiency, and applicability (Fig. 4).

6.1. *Improving prediction accuracy through model ensemble and algorithmic enhancements*

Enhancing prediction accuracy is a paramount focus of current research, demanding multidimensional optimization and refinement of existing methodological frameworks. The integration of multi-source information can offer a more comprehensive research perspective, effectively enhancing the reliability and accuracy of research outcomes. Ensemble learning or deep learning techniques can facilitate the effective integration of diverse information, significantly boosting prediction accuracy and generalization capabilities[5,36,93,129]. Furthermore, integrating ligand-binding data and

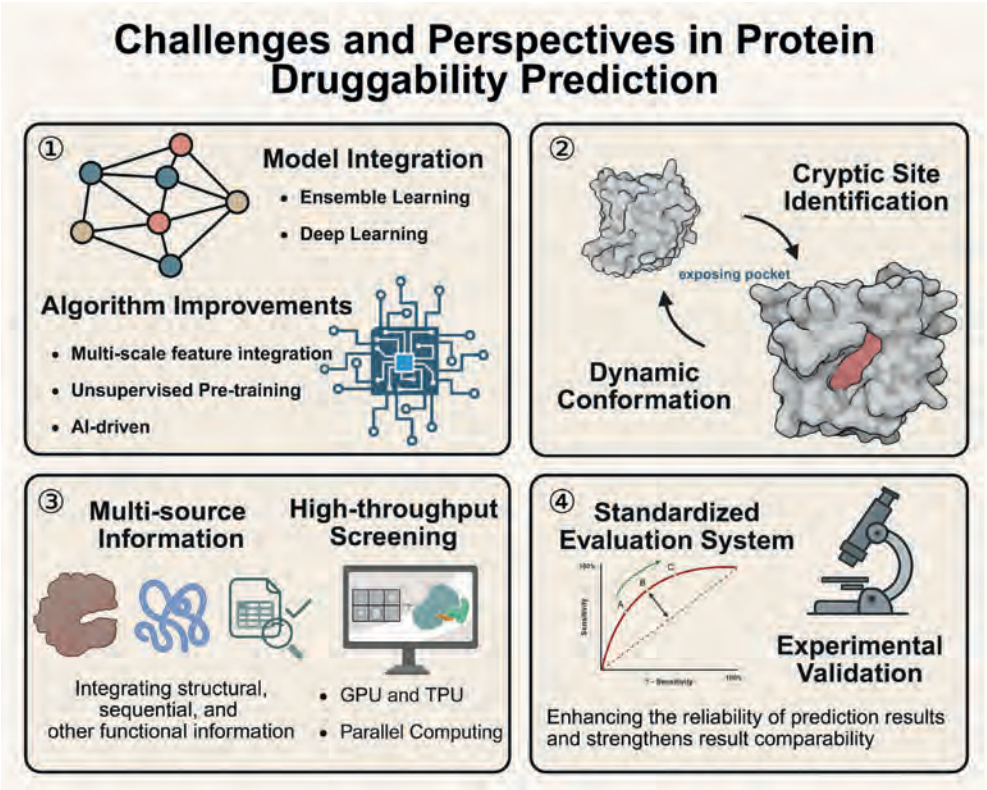

Figure 4 Challenges and future directions in the field of druggable target screening. This figure outlines the current challenges encountered in the research field of druggable target screening methods and four key future directions. It emphasizes model ensembles and algorithmic improvements, the simulation of protein dynamic conformations and mining of cryptic sites, multisource information integration and high-throughput screening, and the construction of efficient computational-experimental validation pipelines and standardized evaluation systems.

PPI information through methods such as machine learning and molecular docking can effectively predict protein binding sites, identify functional characteristic sites, and unveil potential regulatory mechanisms[212–215]. By amalgamating experimental data with computational predictions, such as on the Colabind platform, the plasticity of binding sites can be investigated in depth[26,34,174]. The current limitations of methods such as machine learning and deep learning necessitate more efficient algorithms to overcome them; for instance, unsupervised pre-training and multi-scale feature integration can significantly enhance the generality of protein representations, as well as prediction speed and efficiency[107,131,216]. Moreover, studies have designed specialized models targeting influenza A, GPCRs, kinases, and other specific targets. These models, by combining molecular dynamics simulations, AI(artificial intelligence), deep learning, and structural dynamics techniques, have greatly accelerated the identification of relevant protein binding sites and drug development[169,176,187,192,217–220].

6.2. Simulation of protein dynamic conformations and mining of cryptic sites

The conformational dynamics of proteins significantly influence binding site identification and druggability assessment, thus necessitating the development of more precise methods to capture and predict this dynamism. Constructing efficient molecular dynamics simulation methods by integrating innovative algorithms such as MSM, QSKR, FISST, and aLMMD can markedly accelerate protein conformational sampling, elucidate allosteric regulatory mechanisms, and substantially enhance drug screening efficiency[10,68,221,222]. Among these, Markov State Models (MSMs) can effectively assess transition states in protein conformational changes, and enhanced sampling models can accelerate rare event sampling; both approaches are widely applied in predicting allosteric sites[71,223–225]. Cryptic pockets are transient binding sites formed during protein conformational changes that are difficult to capture using experimental methods such as X-ray crystallography, requiring atomistic simulations to identify these sites[28,226]. Advanced algorithms developed in recent years, for instance, by combining elastic network models, evolutionary coupling analysis, and molecular complex characterization, can improve the identification accuracy of cryptic pockets and allosteric sites[28,76,227,228]. Establishing allosteric site databases can facilitate research into and validation of cryptic pockets. Existing databases such as KLIFS, PDBbind, BioLip, MolMovDB, ASBench, and CASBench are accelerating the development of tools for identifying and predicting cryptic sites[5,218,229,230]. Notably, compared to traditional methods, the computational framework for allosteric site prediction based on the reversed allosteric communication theory cannot only achieve unbiased detection of allosteric sites under physiological conditions but also identify allosteric sites *de novo* without requiring orthosteric site information. This provides a solid theoretical foundation and clear research approach for identifying cryptic binding sites[73].

6.3. Achieving precise target localization through multi-source information integration

Effective integration of multi-source information is a critical step in enhancing prediction performance, highlighting the urgent need for

the development of innovative methods to fuse multi-type and multi-scale data. In this context, the integration of structural, sequential, and functional information can comprehensively reveal the dynamic features of binding sites and their functional relevance[5,25,231]. However, efficiently integrating multimodal information and processing massive datasets still presents significant challenges, while simultaneously the development of multimodal deep learning models offers new opportunities for robust information extraction[27,163,188]. To systematically integrate heterogeneous information, researchers have proposed various innovative feature representation methods—including optimizing active site definitions and data augmentation techniques, constructing ligand-based human protein pocket characterization methods, and "anchor" message passing strategies—to enhance the performance and generalization ability of site prediction[207,232,233]. Biomedical knowledge graph technologies that have been recently developed, such as BioKG, provide a systematic data foundation for multi-source information integration[234]. Furthermore, multi-scale modeling methods, by fusing data and models from different scales, achieve a comprehensive description of the multi-level interactions and dynamic changes in protein−ligand binding. Tools based on multi-model integration, such as CryptoSite—which combines quantum mechanics, molecular mechanics, and coarse-graining—can rapidly predict cryptic sites and handle multi-scale interactions, effectively overcoming the computational bottlenecks of traditional methods[107]. Cross-scale algorithms, by integrating various techniques such as molecular dynamics, bioinformatics, and stochastic simulations, achieve precise modeling and simulation from atomic to biological system scales, significantly improving the efficiency and specificity of drug development[107].

6.4. Driving high-throughput site screening via hardware acceleration and parallel computing technologies

Given the exponential growth of protein data, developing efficient high-throughput screening methods has become a critical research focus. Existing prediction methods face limitations such as high computational costs and insufficient accuracy in predicting unbound state structures, necessitating algorithmic optimization to reduce computational complexity. GPUs and TPUs, by enabling half-precision optimization, computational acceleration, and memory efficiency improvements, have substantially enhanced the computational performance and prediction accuracy of biomolecular system simulations and binding site predictions[172,210,235−237]. Furthermore, large-scale parallel computing technologies, which facilitate the concurrent processing of multiple tasks, provide robust computational support for the development of high-throughput screening methods. Employing data-parallel and model-parallel strategies, the supervised Molecular Dynamics (suMD) method, combined with parallel computing techniques to accelerate the conformational sampling process, not only significantly improves computational efficiency but also ensures consistency with experimental data and model accuracy[68,169]. The Fragment Molecular Orbital (FMO) method, based on quantum mechanical principles, demonstrates a high correlation with experimental data and excellent computational accuracy by considering electron correlation[238].

6.5. Construction of efficient computational-experimental validation pipelines and standardized evaluation systems

The seamless integration of computational and experimental methods is key to enhancing the reliability and practical value of prediction results[239]. The cross-validation and integrated application of computational methods (such as molecular docking and molecular dynamics simulations) contribute to improving prediction accuracy and efficiency, thereby enabling large-scale screening of drug targets[30−32,34]. Iterative optimization strategies founded upon both computation and experimentation not only integrate the advantages of experimental techniques and computational simulations but also deeply elucidate the intrinsic relationship between molecular structural changes and function[94,240]. The application of high-throughput experimental technologies provides an effective means of validation for large-scale prediction results. Notably, ESMFold[25], by introducing large-scale pre-trained protein language models, significantly accelerates the prediction process while improving prediction accuracy[25]. Constructing larger training datasets and systematically studying the relationship between dataset size and prediction performance can effectively circumvent overfitting issues and enhance model prediction capabilities, thereby providing a scientific basis for tool selection in specific application scenarios. Establishing standardized evaluation systems is crucial for unifying comparison standards across different tools and enhancing the reliability of prediction results. Constructing high-quality benchmark datasets can effectively prevent overfitting and improve prediction capabilities, thus providing important support for tool selection and standardized evaluation in specific application scenarios[81,168,208].

7. Conclusions

Screening for protein druggable targets, owing to its high efficiency and low cost, is progressively becoming a central methodology in target discovery. Currently, mainstream screening strategies primarily encompass four categories: structure- and sequence-based methods, machine learning-based methods, binding site feature analysis methods, and druggability assessment methods. While these methods enhance the efficiency and accuracy of candidate target screening, they also provide a solid foundation for drug research and development. Nevertheless, this field still confronts numerous challenges, and future research is urgently required to explore these areas further and accelerate the process of novel drug development.

Acknowledgments

This work was funded by the National Natural Science Foundation of China (82260546) and the Natural Science Foundation of Jiangxi Province (20224BAB206062, China). All the figures were created based on the tools provided by Biorender.com. We acknowledge Sogen Biotechnology, Zhengzhou, Henan Province, and the SolvingLab team for their technical support and assistance.

Author contributions

Writing-original draft, Anqi Lin, Zhirou Zhang, Aimin Jiang; Conceptualization, Bufu Tang, Zhengang Qiu, Peng Luo; Investigation, Anqi Lin, Zhirou Zhang, Aimin Jiang; Writing-review and editing, Anqi Lin, Zhirou Zhang, Aimin Jiang, Kexin Li, Ying Shi, Hong Yang, Jian Zhang, Rongrong Liu, Yaxuan Wang, Antonino Glaviano, Quan Cheng, Bufu Tang, Zhengang Qiu, Peng Luo; Visualization, Anqi Lin, Zhirou Zhang, Aimin Jiang;

Supervision, Bufu Tang, Zhengang Qiu, Peng Luo; All authors have read and agreed to the published version of the manuscript.

Conflicts of interest

The authors declare no conflicts of interest.

References

1. Eder J, Sedrani R, Wiesmann C. The discovery of first-in-class drugs: origins and evolution. *Nat Rev Drug Discov* 2014;**13**:577−87.
2. Shen Q, Cheng F, Song H, Lu W, Zhao J, An X, et al. Proteome-scale investigation of protein allosteric regulation perturbed by somatic mutations in 7000 cancer genomes. *Am J Hum Genet* 2017;**100**: 5−20.
3. Dastan D, Soleymanekhtiari S, Ebadi A. Peptidic compound as DNA binding agent: *in silico* fragment-based design, machine learning, molecular modeling, synthesis, and DNA binding evaluation. *Protein Pept Lett* 2024;**31**:332−44.
4. Cao Q, Ding Y, Xu Y, Li M, Zheng R, Cao Z, et al. Small-molecule anti-COVID-19 drugs and a focus on China's homegrown mind-eudesivir (vv116). *Front Med* 2023;**17**:1068−79.
5. Carpenter KA, Altman RB. Databases of ligand-binding pockets and protein−ligand interactions. *Comput Struct Biotechnol J* 2024;**23**: 1320−38.
6. Gu L, Li B, Ming D. A multilayer dynamic perturbation analysis method for predicting ligand−protein interactions. *BMC Bioinf* 2022; **23**:456.
7. Yan X, Lu Y, Li Z, Wei Q, Gao X, Wang S, et al. PointSite: a point cloud segmentation tool for identification of protein ligand binding atoms. *J Chem Inf Model* 2022;**62**:2835−45.
8. Zhu Y, Zhao L, Wen N, Wang J, Wang C. DataDTA: a multi-feature and dual-interaction aggregation framework for drug−target binding affinity prediction. *Bioinformatics* 2023;**39**:btad560.
9. Nandigrami P, Fiser A. Assessing the functional impact of protein binding site definition. *bioRxiv* 2023. 2023.01.26.525812.
10. Xu C, Zhang X, Zhao L, Verkhivker GM, Bai F. Accurate characterization of binding kinetics and allosteric mechanisms for the HSP90 chaperone inhibitors using AI-augmented integrative biophysical studies. *JACS Au* 2024;**4**:1632−45.
11. Macarron R, Banks MN, Bojanic D, Burns DJ, Cirovic DA, Garyantes T, et al. Impact of high-throughput screening in biomedical research. *Nat Rev Drug Discov* 2011;**10**:188−95.
12. Kitchen DB, Decornez H, Furr JR, Bajorath J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat Rev Drug Discov* 2004;**3**:935−49.
13. Chen J, Lin A, Jiang A, Qi C, Liu Z, Cheng Q, et al. Computational frameworks transform antagonism to synergy in optimizing combination therapies. *npj Digit Med* 2025;**8**:44.
14. Chen J, Lin A, Luo P. Advancing pharmaceutical research: a comprehensive review of cutting-edge tools and technologies. *Curr Pharmaceut Anal* 2024;**21**:1−19.
15. Segal E, Nissenbaum J, Peretz M, Golan Lev T, Cashman R, Philip H, et al. Genome-wide screen for anticancer drug resistance in haploid human embryonic stem cells. *Cell Prolif* 2023;**56**:e13475.
16. Nogales C, Mamdouh ZM, List M, Kiel C, Casas AI, Schmidt HHHW. Network pharmacology: curing causal mechanisms instead of treating symptoms. *Trends Pharmacol Sci* 2022;**43**:136−50.
17. Jamal GA, Jahangirian E, Tarrahimofrad H. Expression, purification, and evaluation of antibody responses and antibody−immunogen complex simulation of a designed multi-epitope vaccine against SARS-CoV-2. *Protein Pept Lett* 2024;**31**:619−38.
18. Zhang J, Li H, Tao W, Zhou J. GseaVis: an R package for enhanced visualization of gene set enrichment analysis in biomedicine. *Media Res* 2025;**1**:131−5.
19. Fang Y, Kong Y, Rong G, Luo Q, Liao W, Zeng D. Systematic investigation of tumor microenvironment and antitumor immunity with iobr. *Media Res* 2025;**1**:136−40.
20. Xu H, Chen C, Lu ZX, Nie Z. Rcssci*:* an R package for visualization of restricted cubic spline. *Media Res* 2025;**1**:201−6.
21. Xie F, Niu Y, Lian L, Wang Y, Zhuang A, Yan G, et al. Multi-omics joint analysis revealed the metabolic profile of retroperitoneal liposarcoma. *Front Med* 2024;**18**:375−93.
22. Liu Y, Zhang S, Liu K, Hu X, Gu X. Advances in drug discovery based on network pharmacology and omics technology. *Curr Pharmaceut Anal* 2024;**21**:33−43.
23. Wang Z, Zhao Y, Zhang L. Emerging trends and hot topics in the application of multi-omics in drug discovery: a bibliometric and visualized study. *Curr Pharmaceut Anal* 2024;**21**:20−32.
24. Liu C, Wen X. Mass spectrometry in covalent drug discovery. *Curr Mol Pharmacol* 2024;**17**:e18761429319065.
25. Yuan Q, Tian C, Yang Y. Genome-scale annotation of protein binding sites *via* language model and geometric deep learning. *eLife* 2024;**13**: RP93695.
26. Verkhivker G, Alshahrani M, Gupta G. Exploring conformational landscapes and cryptic binding pockets in distinct functional states of the SARS-CoV-2 Omicron BA. 1 and BA. 2 trimers: mutation-induced modulation of protein dynamics and network-guided prediction of variant-specific allosteric binding sites. *Viruses* 2023;**15**. 2009.
27. Fang Y, Jiang Y, Wei L, Ma Q, Ren Z, Yuan Q, et al. DeepProSite: structure-aware protein binding site prediction using ESMFold and a pretrained language model. *Bioinformatics* 2023;**39**: btad718.
28. Otazu K, Olivos-Ramirez GE, Fernández-Silva PD, Vilca-Quispe J, Vega-Chozo K, Jimenez Avalos GM, et al. The malaria box molecules: a source for targeting the RBD and NTD cryptic pocket of the spike glycoprotein in SARS-CoV-2. *J Mol Model* 2024; **30**:217.
29. Chen X, Wang K, Chen J, Wu C, Mao J, Song Y, et al. Integrative residue-intuitive machine learning and MD approach to unveil allosteric site and mechanism for β2AR. *Nat Commun* 2024;**15**:8130.
30. Carnero A. High throughput screening in drug discovery. *Clin Transl Oncol* 2006;**8**:482−90.
31. Hou T, Xu X. Recent development and application of virtual screening in drug discovery: an overview. *Curr Pharm Des* 2004;**10**: 1011−33.
32. 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.
33. Sankar S, Vasudevan S, Chandra N. CRD: a *de novo* design algorithm for the prediction of cognate protein receptors for small molecule ligands. *Structure* 2024;**32**:362-75.e4.
34. Andreev G, Kovalenko M, Bozdaganyan ME, Orekhov PS. Colabind: a cloud-based approach for prediction of binding sites using coarse-grained simulations with molecular probes. *J Phys Chem B* 2024; **128**:3211−9.
35. Fu L, Jia G, Liu Z, Pang X, Cui Y. The applications and advances of artificial intelligence in drug regulation: a global perspective. *Acta Pharm Sin B* 2025;**15**:1−14.
36. Singh K, Malik YS. ANN based prediction of LBSs outside deep cavities to facilitate drug designing. *Curr Res Struct Biol* 2024;**7**: 100144.
37. Kakarala KK, Jamil K. Identification of novel allosteric binding sites and multi-targeted allosteric inhibitors of receptor and non-receptor tyrosine kinases using a computational approach. *J Biomol Struct Dyn* 2022;**40**:6889−909.
38. Chen T, Dumas M, Watson R, Vincoff S, Peng C, Zhao L, et al. PepMLM: target sequence-conditioned generation of therapeutic peptide binders *via* span masked language modeling. *ArXiv* 2024. arXiv:2310.03842v3.

39. Du H, Jiang D, Gao J, Zhang X, Jiang L, Zeng Y, et al. Proteome-wide profiling of the covalent-druggable cysteines with a structure-based deep graph learning network. *Research (Wash D C)* 2022; **2022**:9873564.
40. Khaghani F, Hemmati M, Ebrahimi M, Salmaninejad A. Emerging multi-omic approaches to the molecular diagnosis of mitochondrial disease and available strategies for treatment and prevention. *Curr Genom* 2024;**25**:358−79.
41. Eguida M, Rognan D. Estimating the similarity between protein pockets. *Int J Mol Sci* 2022;**23**:12462.
42. Yoshino R, Yasuo N, Hagiwara Y, Ishida T, Inaoka DK, Amano Y, et al. Discovery of a hidden Trypanosoma cruzi spermidine synthase binding site and inhibitors through in silico, *in vitro*, and X-ray crystallography. *ACS Omega* 2023;**8**:25850−60.
43. Song L, Gao S, Ye B, Yang M, Cheng Y, Kang D, et al. Medicinal chemistry strategies towards the development of non-covalent SARS-CoV-2 Mpro inhibitors. *Acta Pharm Sin B* 2024;**14**:87−109.
44. Levitt DG, Banaszak LJ. POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *J Mol Graph* 1992;**10**:229−34.
45. 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.
46. Laskowski RA. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J Mol Graph* 1995;**13**:307−8.
47. Brady GP, Stouten PF. Fast prediction and visualization of protein binding pockets with PASS. *J Comput Aided Mol Des* 2000;**14**: 383−401.
48. Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: an open source platform for ligand pocket detection. *BMC Bioinf* 2009;**10**: 168.
49. Weisel M, Proschak E, Schneider G. PocketPicker: analysis of ligand binding-sites with shape descriptors. *Chem Cent J* 2007;**1**:7.
50. DasGupta D, Mehrani R, Carlson HA, Sharma S. Identifying potential LBSs on glycogen synthase kinase 3 using atomistic cosolvent simulations. *ACS Appl Bio Mater* 2024;**7**:588−95.
51. Zhao J, Cao Y, Zhang L. Exploring the computational methods for protein−ligand binding site prediction. *Comput Struct Biotechnol J* 2020;**18**:417−26.
52. Goodford PJ. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J Med Chem* 1985;**28**:849−57.
53. Laurie ATR, Jackson RM. Q-sitefinder: an energy-based method for the prediction of protein−LBSs. *Bioinformatics* 2005;**21**:1908−16.
54. Ngan C-H, Hall DR, Zerbe B, Grove LE, Kozakov D, Vajda S. FTSite: high accuracy detection of LBSs on unbound protein structures. *Bioinformatics* 2012;**28**:286−7.
55. Jeevan K, Palistha S, Tayara H, Chong KT. PUResNetV2.0: a deep learning model leveraging sparse representation for improved ligand binding site prediction. *J Cheminf* 2024;**16**:66.
56. Makley LN, Johnson OT, Ghanakota P, Rauch JN, Osborn D, Wu TS, et al. Chemical validation of a druggable site on Hsp27/Hspb1 using *in silico* solvent mapping and biophysical methods. *Bioorg Med Chem* 2021;**34**:115990.
57. Ghanakota P, DasGupta D, Carlson HA. Free energies and entropies of binding sites identified by mixmd cosolvent simulations. *J Chem Inf Model* 2019;**59**:2035−45.
58. Ghanakota P, Carlson HA. Moving beyond active-site detection: Mixmd applied to allosteric systems. *J Phys Chem B* 2016;**120**: 8685−95.
59. Smith RD, Carlson HA. Identification of cryptic binding sites using MixMD with standard and accelerated molecular dynamics. *J Chem Inf Model* 2021;**61**:1287−99.
60. Guvench O, MacKerell AD. Computational fragment-based binding site identification by ligand competitive saturation. *PLoS Comput Biol* 2009;**5**:e1000435.
61. Goel H, Hazel A, Yu W, Jo S, MacKerell AD. Application of site-identification by ligand competitive saturation in computer-aided drug design. *New J Chem* 2022;**46**:919−32.
62. Yu W, Weber DJ, MacKerell AD. Integrated covalent drug design workflow using site identification by ligand competitive saturation. *J Chem Theor Comput* 2023;**19**:3007−21.
63. Mousaei M, Kudaibergenova M, MacKerell AD, Noskov S. Assessing HERG1 blockade from Bayesian machine-learning-optimized site identification by ligand competitive saturation simulations. *J Chem Inf Model* 2020;**60**:6489−501.
64. Yao Y, Cui RZ, Bowman GR, Silva DA, Sun J, Huang X. Hierarchical nyström methods for constructing Markov state models for conformational dynamics. *J Chem Phys* 2013;**138**:174106.
65. Gu S, Silva DA, Meng L, Yue A, Huang X. Quantitatively characterizing the ligand binding mechanisms of choline binding protein using Markov state model analysis. *PLoS Comput Biol* 2014;**10**: e1003767.
66. Pande VS, Beauchamp K, Bowman GR. Everything you wanted to know about Markov state models but were afraid to ask. *Methods* 2010;**52**:99−105.
67. Chodera JD, Singhal N, Pande VS, Dill KA, Swope WC. Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *J Chem Phys* 2007; **126**:155101.
68. Hardie A, Cossins BP, Lovera S, Michel J. Deconstructing allostery by computational assessment of the binding determinants of allosteric ptp1b modulators. *Commun Chem* 2023;**6**:125.
69. Nakajima N, Nakamura H, Kidera A. Multicanonical ensemble generated by molecular dynamics simulation for enhanced conformational sampling of peptides. *J Phys Chem B* 1997;**101**:817−24.
70. Hasse T, Zhang Z, Huang YMM. Molecular dynamics study reveals key disruptors of MEIG1−PACRG interaction. *Proteins* 2023;**91**: 555−66.
71. Bekker G-J, Araki M, Oshima K, Okuno Y, Kamiya N. Accurate binding configuration prediction of a G-protein-coupled receptor to its antagonist using multicanonical molecular dynamics-based dynamic docking. *J Chem Inf Model* 2021;**61**:5161−71.
72. Nakajima N, Higo J, Kidera A, Nakamura H. Flexible docking of a ligand peptide to a receptor protein by multicanonical molecular dynamics simulation. *Chem Phys Lett* 1997;**278**:297−301.
73. Ni D, Wei J, He X, Rehman AU, Li X, Qiu Y, et al. Discovery of cryptic allosteric sites using reversed allosteric communication by a combined computational and experimental strategy. *Chem Sci* 2021; **12**:464−76.
74. Zha J, Li Q, Liu X, Lin W, Wang T, Wei J, et al. AlloReverse: multiscale understanding among hierarchical allosteric regulations. *Nucleic Acids Res* 2023;**51**:W33−8.
75. Roy A, Zhang Y. Recognizing protein−LBSs by global structural alignment and local geometry refinement. *Structure* 2012;**20**:987−97.
76. Xie J, Zhang W, Zhu X, Deng M, Lai L. Coevolution-based prediction of key allosteric residues for protein function regulation. *eLife* 2023;**12**:e81850.
77. Jakubec D, Skoda P, Krivak R, Novotny M, Hoksza D. PrankWeb 3: accelerated ligand-binding site predictions for experimental and modelled protein structures. *Nucleic Acids Res* 2022;**50**:W593−7.
78. Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E, et al. ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics* 2003;**19**: 163−4.
79. McGuffin LJ, Bryson K, Jones DT. The psipred protein structure prediction server. *Bioinformatics* 2000;**16**:404−5.
80. 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.
81. Ishitani R, Takemoto M, Tomii K. Protein ligand binding site prediction using graph transformer neural network. *PLoS One* 2024;**19**: e0308425.

82. Upadhyay A, Pal D, Kumar A. Deciphering target protein cascade in Salmonella typhi biofilm using genomic data mining, and protein−protein interaction. *Curr Genom* 2023;**24**:100−9.
83. Carbery A, Buttenschoen M, Skyner R, von Delft F, Deane CM. Learnt representations of proteins can be used for accurate prediction of small molecule binding sites on experimentally determined and predicted protein structures. *J Cheminf* 2024;**16**:32.
84. Zhang N, Zhang H, Liu Z, Dai Z, Wu W, Zhou R, et al. An artificial intelligence network-guided signature for predicting outcome and immunotherapy response in lung adenocarcinoma patients based on 26 machine learning algorithms. *Cell Prolif* 2023;**56**:e13409.
85. Fu X, Yang C, Su Y, Liu C, Qiu H, Yu Y, et al. Machine learning enables comprehensive prediction of the relative protein abundance of multiple proteins on the protein corona. *Research (Wash D C)* 2024;**7**:487.
86. Wang X, Li J, Ning J. Comprehensive multi-omics integration reveals B-cell-derived ell2 as a novel diagnostic and prognostic biomarker in sepsis. *Media Res* 2025;**1**:170−4.
87. Yoshimori A, Bajorath J. Motif2Mol: prediction of new active compounds based on sequence motifs of LBSs in proteins using a biochemical language model. *Biomolecules* 2023;**13**:833.
88. Zhu H, Yang J, Huang N. Assessment of the generalization abilities of machine-learning scoring functions for structure-based virtual screening. *J Chem Inf Model* 2022;**62**:5485−502.
89. Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995;**20**: 273−97.
90. Wong GY, Leung FHF, Ling SH. Predicting protein−ligand binding site using support vector machine with protein properties. *IEEE ACM Trans Comput Biol Bioinf* 2013;**10**:1517−29.
91. Yu DJ, Hu J, Yang J, Shen H-B, Tang J, Yang J-Y. Designing template-free predictor for targeting protein−LBSs with classifier ensemble and spatial clustering. *IEEE ACM Trans Comput Biol Bioinf* 2013;**10**:994−1008.
92. Feinstein WP, Brylinski M. EFindSite: enhanced fingerprint-based virtual screening against predicted LBSs in protein models. *Mol Inform* 2014;**33**:135−50.
93. Edgar RC, Sjölander K. COACH: profile-profile alignment of protein families using hidden markov models. *Bioinformatics* 2004;**20**: 1309−18.
94. McGreig JE, Uri H, Antczak M, Sternberg MJE, Michaelis M, Wass MN. 3DLigandSite: structure-based prediction of protein−LBSs. *Nucleic Acids Res* 2022;**50**:W13−20.
95. Wu Q, Peng Z, Zhang Y, Yang J. COACH-d: improved protein−LBSs prediction with refined ligand-binding poses through molecular docking. *Nucleic Acids Res* 2018;**46**:W438−42.
96. Ho TK. Random decision forests. In: *Proceedings of 3rd international conference on document analysis and recognition*; 1995. p. 278−82.
97. Kwofie SK, Agyenkwa-Mawuli K, Broni E, Miller Iii WA, Wilson MD. Prediction of antischistosomal small molecules using machine learning in the era of big data. *Mol Divers* 2022;**26**: 1597−607.
98. Fawagreh K, Gaber MM, Elyan E. Random forests: from early developments to recent advancements. *Syst Sci Control Eng* 2014;**2**: 602−9.
99. Krivák R, Hoksza D. P2Rank: machine learning based tool for rapid and accurate prediction of LBSs from protein structure. *J Cheminf* 2018;**10**:39.
100. Kandel J, Tayara H, Chong KT. PUResNet: prediction of protein−LBSs using deep residual neural network. *J Cheminf* 2021; **13**:65.
101. Nazem F, Ghasemi F, Fassihi A, Rasti R, Dehnavi AM. A GU-net-based architecture predicting ligand−protein-binding atoms. *J Med Signals Sens* 2023;**13**:1−10.
102. Graef J, Ehrt C, Rarey M. Binding site detection remastered: enabling fast, robust, and reliable binding site detection and descriptor calculation with dogsite3. *J Chem Inf Model* 2023;**63**: 3128−37.
103. Zhang Y, Haghani A. A gradient boosting method to improve travel time prediction. *Transport Res C Emerg Technol* 2015;**58**: 308−24.
104. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*; 2016. p. 785−94.
105. Zhao Z, Xu Y, Zhao Y. SXGBsite: prediction of protein−LBSs using sequence information and extreme gradient boosting. *Genes* 2019;**10**:965.
106. Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, et al. Xgboost: extreme gradient boosting. *R package* version ;**1**:1−4.
107. Lei C, Lu Z, Wang M, Li M. StackCPA: a stacking model for compound-protein binding affinity prediction based on pocket multi-scale features. *Comput Biol Med* 2023;**164**:107131.
108. Kim P. *Convolutional neural network*. MATLAB Deep Learning; 2017. p. 121−47.
109. Jiao L, Zhao J, Wang C, Liu X, Liu F, Li L, et al. Nature-inspired intelligent computing: a comprehensive survey. *Research* 2024;**7**:442 (Wash D C).
110. Yang R, Zhang J, Zhan F, Yan C, Lu S, Zhu Z, et al. Artificial intelligence efficiently predicts gastric lesions, Helicobacter pylori infection and lymph node metastasis upon endoscopic images. *Chin J Cancer Res* 2024;**36**:489−502.
111. Zhang J, Yao H, Lai C, Sun X, Yang X, Li S, et al. A novel multi-modal prediction model based on DNA methylation biomarkers and low-dose computed tomography images for identifying early-stage lung cancer. *Chin J Cancer Res* 2023;**35**:511−25.
112. Yu P, Wang C, Zhang H, Zheng G, Jia G, Liu Zet al., Deep learning-based automatic pipeline system for predicting lateral cervical lymph node metastasis in patients with papillary thyroid carcinoma using computed tomography: a multi-center study *Chin J Cancer Res* 2024;**36**:545−61.
113. Petrovski ŽH, Hribar Lee B, Bosnić Z. CAT-site: predicting protein binding sites using a convolutional neural network. *Pharmaceutics* 2022;**15**:119.
114. Zeng H, Edwards MD, Liu G, Gifford DK. Convolutional neural network architectures for predicting DNA−protein binding. *Bioinformatics* 2016;**32**:i121−7.
115. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* 2015;**12**:931−4.
116. Mylonas SK, Axenopoulos A, Daras P. DeepSurf: a surface-based deep learning approach for the prediction of LBSs on proteins. *Bioinformatics* 2021;**37**:1681−90.
117. Zhang H, Saravanan KM, Zhang JZH. DeepBindGCN: integrating molecular vector representation with graph convolutional neural networks for protein−ligand interaction prediction. *Molecules* 2023;**28**: 4691.
118. Graph neural networks: a review of methods and applications. *AI Open* 2020;**1**:57−81.
119. Kipf TN, Welling M. *Semi-supervised classification with graph convolutional networks*. 2017.
120. Shi W, Singha M, Pu L, Srivastava G, Ramanujam J, Brylinski M. GraphSite: ligand binding site classification with deep graph learning. *Biomolecules* 2022;**12**:1053.
121. Zhang Y, Liu C, Liu M, Liu T, Lin H, Huang CB, et al. Attention is all you need: utilizing attention in AI-enabled drug discovery. *Brief Bioinform* 2024;**25**:bbad467.
122. Mou M, Pan Z, Zhou Z, Zheng L, Zhang H, Shi S, et al. A transformer-based ensemble framework for the prediction of protein−protein interaction sites. *Research (Wash D C)* 2023;**6**:240.
123. Ashish V, Noam S, Niki P, Jakob U, Llion J, Aidan N, et al. Attention is all you need. *Adv Neural Inf Process Syst* 2017;**30**:6000−10.
124. Etzion-Fuchs A, Todd DA, Singh M. DSPRINT: predicting DNA, RNA, ion, peptide and small molecule interaction sites within protein domains. *Nucleic Acids Res* 2021;**49**:e78.
125. Chatzigoulas A, Cournia Z. DREAMM: a web-based server for drugging protein−membrane interfaces as a novel workflow for targeted drug design. *Bioinformatics* 2022;**38**:5449−51.

126. Ouzounis S, Panagiotopoulos V, Bafiti V, Zoumpoulakis P, Cavouras D, Kalatzis I, et al. A robust machine learning framework built upon molecular representations predicts CYP450 inhibition: toward precision in drug repurposing. *OMICS* 2023;**27**:305−14.
127. Wu Z, Lei T, Shen C, Wang Z, Cao D, Hou T. ADMET evaluation in drug discovery. 19. Reliable prediction of human cytochrome P450 inhibition using artificial intelligence approaches. *J Chem Inf Model* 2019;**59**:4587−601.
128. Jha K, Saha S. Amalgamation of 3D structure and sequence information for protein−protein interaction prediction. *Sci Rep* 2020;**10**: 19171.
129. Roberts E, Eargle J, Wright D, Luthey-Schulten Z. MultiSeq: unifying sequence and structure data for evolutionary analysis. *BMC Bioinf* 2006;**7**:382.
130. Sahu G, Vechtomova O. *Adaptive fusion techniques for multimodal data*. 2021.
131. Nguyen VTD, Hy TS. Multimodal pretraining for unsupervised protein representation learning. *Biol Methods Protoc* 2024;**9**:bpae043.
132. Peters KP, Fauck J, Frömmel C. The automatic search for LBSs in proteins of known three-dimensional structure using only geometric criteria. *J Mol Biol* 1996;**256**:201−13.
133. 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.
134. Schmidtke P, Bidon-Chanal A, Luque FJ, Barril X. MDpocket: open-source cavity detection and characterization on molecular dynamics trajectories. *Bioinformatics* 2011;**27**:3276−85.
135. Coleman RG, Burr MA, Souvaine DL, Cheng AC. An intuitive approach to measuring protein surface curvature. *Proteins* 2005;**61**: 1068−74.
136. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. *Proteins* 1994;**20**:216−26.
137. Ahmad S, Gromiha M, Fawareh H, Sarai A. ASAView: database and tool for solvent accessibility representation in proteins. *BMC Bioinf* 2004;**5**:51.
138. Negi SS, Braun W. Statistical analysis of physical-chemical properties and prediction of protein−protein interfaces. *J Mol Model* 2007; **13**:1157−67.
139. Mamun TI, Bourhia M, Neoaj T, Akash S, Azad MAK, Hossain MS, et al. Structure based functional identification of an uncharacterized protein from Coxiella burnetii involved in adipogenesis. *Sci Rep* 2024;**14**:16789.
140. Gorityala N, Baidya AS, Sagurthi SR. Genome mining of *Mycobacterium tuberculosis*: targeting SufD as a novel drug candidate through *in silico* characterization and inhibitor screening. *Front Microbiol* 2024;**15**:1369645.
141. Kellogg GE, Abraham DJ. Development of empirical molecular interaction models that incorporate hydrophobicity and hydropathy. the hint paradigm. *Analusis* 1999;**27**:19−22.
142. Bennett GM, Starczewski J, Dela Cerna MVC. *In silico* identification of putative druggable pockets in PRl3, a significant oncology target. *Biochem Biophys Rep* 2024;**39**:101767.
143. Sinha N, Smith Gill SJ. Electrostatics in protein binding and function. *Curr Protein Pept Sci* 2002;**3**:601−14.
144. Roux B, Simonson T. Implicit solvent models. *Biophys Chem* 1999; **78**:1−20.
145. Roux B, Bernèche S, Im W. Ion channels, permeation, and electrostatics: insight into the function of KcsA. *Biochemistry* 2000;**39**: 13295−306.
146. Honig B, Nicholls A. Classical electrostatics in biology and chemistry. *Science* 1995;**268**:1144−9.
147. Konecny R, Baker NA, McCammon JA. IAPBS: a programming interface to adaptive Poisson-Boltzmann solver (APBS). *Comput Sci Discov* 2012;**5**:015005.
148. Jo S, Vargyas M, Vasko-Szedlar J, Roux B, Im W. PBEQ-solver for online visualization of electrostatic potential of biomolecules. *Nucleic Acids Res* 2008;**36**:W270−5.
149. Lopez G, Maietta P, Rodriguez JM, Valencia A, Tress ML. Firestar—Advances in the prediction of functionally important residues. *Nucleic Acids Res* 2011;**39**:W235−41.
150. He J, Duan Q, Ran C, Fu T, Liu Y, Tan W. Recent progress of aptamer−drug conjugates in cancer therapy. *Acta Pharm Sin B* 2023; **13**:1358−70.
151. Luo Y, Liu Y, Peng J. Calibrated geometric deep learning improves kinase-drug binding predictions. *Nat Mach Intell* 2023;**5**:1390−401.
152. Ghoula M, Naceri S, Sitruk S, Flatters D, Moroy G, Camproux AC. Identifying promising druggable binding sites and their flexibility to target the receptor-binding domain of SARS-CoV-2 spike protein. *Comput Struct Biotechnol J* 2023;**21**:2339−51.
153. Ricci F, Gitto R, Pitasi G, De Luca L. *In silico* insights towards the identification of SARS-CoV-2 nsp13 helicase druggable pockets. *Biomolecules* 2022;**12**:482.
154. Wang W, Zhang Y, Liu D, Zhang H, Wang X, Zhou Y. Prediction of DNA-binding protein-drug-binding sites using residue interaction networks and sequence feature. *Front Bioeng Biotechnol* 2022;**10**: 822392.
155. Halgren TA. Identifying and characterizing binding sites and assessing druggability. *J Chem Inf Model* 2009;**49**:377−89.
156. Garisetti V, Dhanabalan AK, Dasararaju G. Discovery of potential taar1 agonist targeting neurological and psychiatric disorders: an *in silico* approach. *Int J Biol Macromol* 2024;**264**:130528.
157. Halgren T. New method for fast and accurate binding-site identification and analysis. *Chem Biol Drug Des* 2007;**69**:146−8.
158. Cappel D, Sherman W, Beuming T. Calculating water thermodynamics in the binding site of proteins—Applications of watermap to drug discovery. *Curr Top Med Chem* 2017;**17**:2586−98.
159. Kaczor AA, Zięba A, Matosiuk D. The application of watermap-guided structure-based virtual screening in novel drug discovery. *Expet Opin Drug Discov* 2024;**19**:73−83.
160. Li Y, Zhang Z, Wang R. HydraMap v.2: prediction of hydration sites and desolvation energy with refined statistical potentials. *J Chem Inf Model* 2023;**63**:4749−61.
161. Li Y, Gao Y, Holloway MK, Wang R. Prediction of the favorable hydration sites in a protein binding pocket and its application to scoring function formulation. *J Chem Inf Model* 2020;**60**: 4359−75.
162. Ribeiro AJ, Riziotis IG, Borkakoti N, Thornton JM. Enzyme function and evolution through the lens of bioinformatics. *Biochem J* 2023; **480**:1845−63.
163. D'souza S, Prema KV, Balaji S, Shah R. Deep learning-based modeling of drug−target interaction prediction incorporating binding site information of proteins. *Interdiscip Sci* 2023;**15**:306−15.
164. Wang DD, Chan MT, Yan H. Structure-based protein−ligand interaction fingerprints for binding affinity prediction. *Comput Struct Biotechnol J* 2021;**19**:6291−300.
165. Gainza P, Sverrisson F, Monti F, Rodolà E, Boscaini D, Bronstein MM, et al. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat Methods* 2020;**17**:184−92.
166. Sverrisson F, Feydy J, Correia BE, Bronstein MM. *Fast end-to-end learning on protein surfaces*. 2021. p. 15272−81.
167. Isert C, Atz K, Schneider G. Structure-based drug design with geometric deep learning. *Curr Opin Struct Biol* 2023;**79**:102548.
168. Brueckner AC, Shields B, Kirubakaran P, Suponya A, Panda M, Posy SL, et al. MDFit: automated molecular simulations workflow enables high throughput assessment of ligands−protein dynamics. *J Comput Aided Mol Des* 2024;**38**:24.
169. Naceri S, Marc D, Blot R, Flatters D, Camproux A-C. Druggable pockets at the RNA interface region of influenza A virus NS1 protein are conserved across sequence variants from distinct subtypes. *Biomolecules* 2022;**13**:64.
170. Xu Z, Wang X, Zeng S, Ren X, Yan Y, Gong Z. Applying artificial intelligence for cancer immunotherapy. *Acta Pharm Sin B* 2021;**11**: 3393−405.

171. 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.
172. Huang Y, Zhang H, Jiang S, Yue D, Lin X, Zhang J, et al. DSDP: a blind docking strategy accelerated by GPUs. *J Chem Inf Model* 2023; **63**:4355−63.
173. Xia Y, Xia CQ, Pan X, Shen HB. GraphBind: protein structural context embedded rules learned by hierarchical graph neural networks for recognizing nucleic-acid-binding residues. *Nucleic Acids Res* 2021;**49**:e51.
174. Zehra Hussain A, AlAjmi MF, Ishrat R, Hassan MI. Enriching anticancer drug pipeline with potential inhibitors of cyclin-dependent kinase-8 identified from natural products. *OMICS* 2024; **28**:478−88.
175. Basnet R, Bahadur Basnet B, Gupta R, Basnet T, Khadka S, Shan Alam M. Mammalian target of rapamycin (mTOR) signalling pathway—a potential target for cancer intervention: a short overview. *Curr Mol Pharmacol* 2024;**17**:e310323215268.
176. Kim S, Mollaei P, Antony A, Magar R, Barati Farimani A. GPCR-bert: interpreting sequential design of G protein-coupled receptors using protein language models. *J Chem Inf Model* 2024;**64**:1134−44.
177. Karelina M, Noh JJ, Dror RO. How accurately can one predict drug binding modes using AlphaFold models?. *eLife* 2023;**12**:RP89386.
178. Cheng L, Xia F, Li Z, Shen C, Yang Z, Hou H, et al. Structure, function and drug discovery of GPCR signaling. *Mol Biomed* 2023;**4**: 46.
179. Yeturu K, Chandra N. PocketMatch: a new algorithm to compare binding sites in protein structures. *BMC Bioinf* 2008;**9**:543.
180. Kinoshita K, Murakami Y, Nakamura H. EF-seek: prediction of the functional sites of proteins by searching for similar electrostatic potential and molecular surface shape. *Nucleic Acids Res* 2007;**35**: W398−402.
181. Shin W-H, Bures MG, Kihara D. PatchSurfers: two methods for local molecular property-based binding ligand prediction. *Methods* 2016; **93**:41−50.
182. Dankwah KO, Mohl JE, Begum K, Leung MY. What makes GPCRs from different families bind to the same ligand?. *Biomolecules* 2022; **12**:863.
183. Szwabowski GL, Griffing M, Mugabe EJ, O'Malley D, Baker LN, Baker DL, et al. G protein-coupled receptor−ligand pose and functional class prediction. *Int J Mol Sci* 2024;**25**:6876.
184. Xu Z, Guo L, Yu J, Shen S, Wu C, Zhang W, et al. Ligand recognition and G-protein coupling of trace amine receptor Taar1. *Nature* 2023;**624**:672−81.
185. Raza SHA, Zhong R, Yu X, Zhao G, Wei X, Lei H. Advances of predicting allosteric mechanisms through protein contact in new technologies and their application. *Mol Biotechnol* 2024;**66**: 3385−97.
186. Wu N, Yaliraki SN, Barahona M. Prediction of protein allosteric signalling pathways and functional residues through paths of optimised propensity. *J Mol Biol* 2022;**434**:167749.
187. Lee JY, Gebauer E, Seeliger MA, Bahar I. Allo-targeting of the kinase domain: insights from *in silico* studies and comparison with experiments. *Curr Opin Struct Biol* 2024;**84**:102770.
188. Rana N, Patel D, Parmar M, Mukherjee N, Jha PC, Manhas A. Targeting allosteric binding site in methylenetetrahydrofolate dehydrogenase 2 (MTHFD2) to identify natural product inhibitors *via* structure-based computational approach. *Sci Rep* 2023;**13**:18090.
189. La Sala G, Pfleger C, Käck H, Wissler L, Nevin P, Böhm K, et al. Combining structural and coevolution information to unveil allosteric sites. *Chem Sci* 2023;**14**:7057−67.
190. Wang A, Zhang Y, Lv X, Liang G. Therapeutic potential of targeting protein tyrosine phosphatases in liver diseases. *Acta Pharm Sin B* 2024;**14**:3295−311.
191. Basciu A, Athar M, Kurt H, Neville C, Malloci G, Muredda FC, et al. Predicting binding events in very flexible, allosteric, multi-domain proteins. *bioRxiv* 2024. Available from: https://doi.org/10.1101/2024.06.02.597018.
192. Liessmann F, Künze G, Meiler J. Improving the modeling of extracellular ligand binding pockets in rosettagpcr for conformational selection. *Int J Mol Sci* 2023;**24**:7788.
193. Meller A, Lotthammer JM, Smith LG, Novak B, Lee LA, Kuhn CC, et al. Drug specificity and affinity are encoded in the probability of cryptic pocket opening in myosin motor domains. *eLife* 2023;**12**: e83602.
194. Zhu C, Lan X, Wei Z, Yu J, Zhang J. Allosteric modulation of G protein-coupled receptors as a novel therapeutic strategy in neuropathic pain. *Acta Pharm Sin B* 2024;**14**:67−86.
195. Boike L, Henning NJ, Nomura DK. Advances in covalent drug discovery. *Nat Rev Drug Discov* 2022;**21**:881−98.
196. Zhang T, Hatcher JM, Teng M, Gray NS, Kostic M. Recent advances in selective and irreversible covalent ligand development and validation. *Cell Chem Biol* 2019;**26**:1486−500.
197. Zheng L, Li Y, Wu D, Xiao H, Zheng S, Wang G, et al. Development of covalent inhibitors: principle, design, and application in cancer. *MedComm Oncol* 2023;**2**:e56.
198. Khazanov NA, Carlson HA. Exploring the composition of protein−LBSs on a large scale. *PLoS Comput Biol* 2013;**9**: e1003321.
199. Woyach JA, Furman RR, Liu TM, Ozer HG, Zapatka M, Ruppert AS, et al. Resistance mechanisms for the Bruton's tyrosine kinase inhibitor ibrutinib. *N Engl J Med* 2014;**370**:2286−94.
200. Xu Y, Wang S, Hu Q, Gao S, Ma X, Zhang W, et al. CavityPlus: a web server for protein cavity detection with pharmacophore modelling, allosteric site identification and covalent ligand binding ability prediction. *Nucleic Acids Res* 2018;**46**:W374−9.
201. Li L, Su M, Lu W, Song H, Liu J, Wen X, et al. Triazine-based covalent DNA-encoded libraries for discovery of covalent inhibitors of target proteins. *ACS Med Chem Lett* 2022;**13**:1574−81.
202. Wang X, Xiong L, Zhu Y, Liu S, Zhao W, Wu X, et al. Covalent DNA-encoded library workflow drives discovery of SARS-CoV-2 nonstructural protein inhibitors. *J Am Chem Soc* 2024;**146**: 33983−96.
203. Jiang L, Liu S, Jia X, Gong Q, Wen X, Lu W, et al. ABPP-codel: activity-based proteome profiling-guided discovery of tyrosine-targeting covalent inhibitors from DNA-encoded libraries. *J Am Chem Soc* 2023;**145**:25283−92.
204. Wu X, Li S, Liang T, Yu Q, Zhang Y, Liu J, et al. Proteome-wide data guides the discovery of lysine-targeting covalent inhibitors using DNA-encoded chemical libraries. *Angew Chem Int Ed Engl* 2025;**64**: e202505581.
205. Pallante L, Cannariato M, Androutsos L, Zizzi EA, Bompotas A, Hada X, et al. VirtuousPocketome: a computational tool for screening protein−ligand complexes to identify similar binding sites. *Sci Rep* 2024;**14**:6296.
206. Ma Y, Zhou Y, Jiang D, Dai W, Li J, Deng C, et al. Integration of human organoids single-cell transcriptomic profiles and human genetics repurposes critical cell type-specific drug targets for severe COVID-19. *Cell Prolif* 2024;**57**:e13558.
207. Stevenson GA, Kirshner D, Bennion BJ, Yang Y, Zhang X, Zemla A, et al. Clustering protein binding pockets and identifying potential drug interactions: a novel ligand-based featurization method. *J Chem Inf Model* 2023;**63**:6655−66.
208. Reim T, Ehrt C, Graef J, Günther S, Meents A, Rarey M. SiteMine: large-scale binding site similarity searching in protein structure databases. *Arch Pharmazie* 2024;**357**:2300661.
209. Eguida M, Rognan D. Unexpected similarity between HIV-1 reverse transcriptase and tumor necrosis factor binding sites revealed by computer vision. *J Cheminf* 2021;**13**:90.
210. Sankar S, Chandran Sakthivel N, Chandra N. Fast local alignment of protein pockets (flapp): a system-compiled program for large-scale binding site alignment. *J Chem Inf Model* 2022;**62**:4810−9.
211. Papadopoulos MGE, Perhal AF, Medel-Lacruz B, Ladurner A, Selent J, Dirsch VM, et al. Discovery and characterization of small-molecule tgr5 ligands with agonistic activity. *Eur J Med Chem* 2024; **276**:116616.

212. Yu W, Li B, Chen L, Chen Q, Song Q, Jin X, et al. Gigantol ameliorates DSS-induced colitis *via* suppressing β2 integrin mediated adhesion and chemotaxis of macrophage. *J Ethnopharmacol* 2024; **328**:118123.
213. Utgés JS, MacGowan SA, Ives CM, Barton GJ. Classification of likely functional class for LBSs identified from fragment screening. *Commun Biol* 2024;**7**:320.
214. Vujica L, Lončar J, Mišić L, Lučić B, Radman K, Mihaljević I, et al. Environmental contaminants modulate transport activity of zebrafish (Danio rerio) multidrug and toxin extrusion protein 3 (Mate3/SLc47a2.1). *Sci Total Environ* 2023;**901**:165956.
215. Mohri M, Moghadam A, Burketova L, Ryšánek P. Genome-wide identification of the opsin protein in *Leptosphaeria maculans* and comparison with other fungi (pathogens of Brassica napus). *Front Microbiol* 2023;**14**:1193892.
216. Yuan M, Shen A, Fu K, Guan J, Ma Y, Qiao Q, et al. ProteinMAE: masked autoencoder for protein surface self-supervised learning. *Bioinformatics* 2023;**39**:btad724.
217. Chen Y, Lin YCD, Luo Y, Cai X, Qiu P, Cui S, et al. Quantitative model for genome-wide cyclic AMP receptor protein binding site identification and characteristic analysis. *Brief Bioinform* 2023;**24**: bbad138.
218. Wu N, Strömich L, Yaliraki SN. Prediction of allosteric sites and signaling: insights from benchmarking datasets. *Patterns (N Y)* 2022; **3**:100408.
219. Shen Z, Yan YH, Yang S, Zhu S, Yuan Y, Qiu Z, et al. ProfKin: a comprehensive web server for structure-based kinase profiling. *Eur J Med Chem* 2021;**225**:113772.
220. Yang H, Wang Y, Liu W, He T, Liao J, Qian Z, et al. Genome-wide pan-GPCR cell libraries accelerate drug discovery. *Acta Pharm Sin B* 2024;**14**:4296−311.
221. Singh Y, Hocky GM. Improved prediction of molecular response to pulling by combining force tempering with replica exchange methods. *ArXiv* 2023;**128**:706−15.
222. Tze Yang Ng J, Tan YS. Accelerated ligand-mapping molecular dynamics simulations for the detection of recalcitrant cryptic pockets and occluded binding sites. *J Chem Theor Comput* 2022;**18**: 1969−81.
223. Rubina Moin ST, Haider S. Identification of a cryptic pocket in methionine aminopeptidase-II using adaptive bandit molecular dynamics simulations and Markov state models. *ACS Omega* 2024;**9**: 28534−45.
224. Sarkar DK, Surpeta B, Brezovsky J. Incorporating prior knowledge in the seeds of adaptive sampling molecular dynamics simulations of ligand transport in enzymes with buried active sites. *J Chem Theor Comput* 2024;**20**:5807−19.
225. Odstrcil RE, Dutta P, Liu J. Prediction of the peptide-TIM3 binding site in inhibiting TIM3-galectin 9 binding pathways. *J Chem Theor Comput* 2023;**19**:6500−9.
226. Qiu Y, Liu Q, Tu G, Yao XJ. Discovery of the cryptic sites of SARS-CoV-2 papain-like protease and analysis of its druggability. *Int J Mol Sci* 2022;**23**:11265.
227. Kumar A, Kaynak BT, Dorman KS, Doruker P, Jernigan RL. Predicting allosteric pockets in protein biological assemblages. *Bioinformatics* 2023;**39**:btad275.
228. Feng Z, Liang T, Wang S, Chen M, Hou T, Zhao J, et al. Binding characterization of GPCRs-modulator by molecular complex characterizing system (MCCS). *ACS Chem Neurosci* 2020;**11**: 3333−45.
229. Sim J, Kwon S, Seok C. HProteome-bsite: predicted binding sites and ligands in human 3D proteome. *Nucleic Acids Res* 2023;**51**:D403−8.
230. Peng C, Zhang X, Xu Z, Chen Z, Yang Y, Cai T, et al. D3PM: a comprehensive database for protein motions ranging from residue to domain. *BMC Bioinf* 2022;**23**:70.
231. Long X, Sun K, Lai S, Liu Y, Su J, Chen W, et al. Artificial intelligence and anti-cancer drugs' response. *Acta Pharm Sin B* 2025;**15**: 3355−71.
232. Li S, Tian T, Zhang Z, Zou Z, Zhao D, Zeng J. PocketAnchor: learning structure-based pocket representations for protein−ligand interaction prediction. *Cell Syst* 2023;**14**:692−705.e6.
233. Born J, Shoshan Y, Huynh T, Cornell WD, Martin EJ, Manica M. On the choice of active site sequences for kinase-ligand affinity prediction. *J Chem Inf Model* 2022;**62**:4295−9.
234. Zhang Y, Sui X, Pan F, Yu K, Li K, Tian S, et al. BioKG: a comprehensive, large-scale biomedical knowledge graph for AI-powered, data-driven biomedical research. *bioRxiv* 2023. 2023.10.13.562216.
235. He Y, Wang S, Zeng S, Zhu J, Xu D, Han W, et al. NRIMD, a web server for analyzing protein allosteric interactions based on molecular dynamics simulation. *J Chem Inf Model* 2024;**64**: 7176−83.
236. Deng J, Cui Q. Efficient sampling of cavity hydration in proteins with nonequilibrium grand canonical Monte Carlo and polarizable force fields. *J Chem Theor Comput* 2024;**20**:1897−911.
237. Buch I, Giorgino T, De Fabritiis G. Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proc Natl Acad Sci U S A* 2011;**108**:10184−9.
238. Yuan Z, Zhang M, Chang L, Chen X, Ruan S, Shi S, et al. Discovery of a novel shp2 allosteric inhibitor using virtual screening, FMO calculation, and molecular dynamic simulation. *J Mol Model* 2024; **30**:131.
239. Cao X, Zheng W, Qiang Y, Yao N, Zuo F, Qiu S. Tumor organoid model and its pharmacological applications in tumorigenesis prevention. *CMP* 2023;**16**:435−47.
240. Fatima S, Mehrafrooz B, Boggs DG, Ali N, Singh S, Thielges MC, et al. Conformation-dependent hydrogen-bonding interactions in a switchable artificial metalloprotein. *Biochemistry* 2024;**63**: 2040−50.