

Application and Evaluation of Algorithms for Protein Binding Site Prediction

Zerui Song^{1*}, Renmao Zhao¹, Jing Zhang¹

¹ School of Information and Intelligent Engineering, Sanya University, Sanya, China

Email Address

1187183544@qq.com (Zerui Song)

*Correspondence: 1187183544@qq.com

Received: 25 November 2025; **Accepted:** 23 December 2025; **Published:** 15 January 2026

Abstract:

Protein binding site prediction is a central computational task for elucidating protein functional mechanisms and advancing drug discovery. This article systematically analyzes the algorithmic evolution in this field, from traditional feature engineering to end-to-end deep learning, focusing on the mechanistic characteristics and applicable boundaries of advanced methods such as graph neural networks, three-dimensional convolutional networks, and sequence-structure fusion models. The study delves into the practical strategies and prominent challenges of algorithms in specific scenarios such as small molecule drug targets, protein-protein interaction interfaces, and allosteric sites, including key issues such as allosteric site identification, binding affinity prediction, and conformational dynamics processing. By establishing a rigorous evaluation framework, this article reveals the current performance bottlenecks of algorithms in terms of data bias, insufficient dynamic representation, and lack of physical principle embedding. It also points out that future development directions will focus on equivariant neural networks, multi-conformation integrated prediction paradigms, and ligand-specific adaptive models, providing theoretical guidance and technical pathways for achieving more accurate and interpretable binding site prediction.

Keywords:

Protein Binding Site Prediction, Deep Learning, Graph Neural Networks, Ligand Specificity, Algorithm Evaluation, Structural Bioinformatics

1. Introduction

Proteins, as the primary executors of life activities, largely depend on specific interactions with other molecules (ligands) for their functional realization. Accurate identification of binding sites on the protein surface is crucial for elucidating their molecular mechanisms, promoting rational drug design, and optimizing protein engineering. Although experimental techniques such as X-ray crystallography and cryo-electron microscopy can provide high-resolution information on binding interfaces, their low throughput, high cost, and technical bottlenecks for specific protein-ligand systems make it difficult to meet the needs of large-scale, rapid discovery [1,2].

In this context, the development of efficient computational prediction methods has become a core issue in the field of structural bioinformatics [3]. Protein binding sites typically refer to regions on the protein surface that are spatially and physicochemically complementary to specific ligands, characterized by the coupling of local geometric cavities with specific electrostatic distributions, hydrophobic patches, and evolutionary conservation patterns. It is worth noting that the sites corresponding to different ligand types (e.g., small molecules, peptides, nucleic acids) exhibit significant differences in shape, size, and physicochemical properties, which poses inherent challenges to the construction of general-purpose prediction algorithms.

In recent years, with breakthroughs in deep learning technology and the accumulation of protein structure data, the field of binding site prediction has evolved from an early machine learning paradigm relying on hand-crafted features to an end-to-end deep learning paradigm capable of automatically learning complex patterns from sequences or three-dimensional structures [4]. However, current research still faces a series of severe challenges: the subjectivity of positive and negative sample definitions leads to inherent noise in the datasets; the model's ability to capture dynamic conformational changes of proteins is limited; most methods lack physical interpretability of the prediction results; and the generalization performance of the model across different protein families and ligand types still needs systematic evaluation [5].

This article aims to systematically review the mainstream algorithms for protein binding site prediction, focusing on analyzing their core mechanisms, evaluating their application effectiveness in specific biomedical scenarios, and critically discussing the current performance bottlenecks of the algorithms. Through in-depth technical review and evaluation of this field, we aim to provide a theoretical basis for computational biology researchers to select appropriate prediction tools and to provide valuable insights for future algorithm innovation directions.

2. Core Algorithmic Mechanisms of Protein Binding Site Prediction

The evolution of protein binding site prediction algorithms is essentially a shift from shallow feature engineering relying on expert knowledge to an end-to-end deep representation learning paradigm. Understanding the intrinsic mechanisms of different algorithmic paradigms is the cornerstone for evaluating their application potential and limitations.

2.1. Paradigm Based on Traditional Features and Machine Learning

Before the rise of deep learning, the core of research in this field focused on how to use domain knowledge to construct feature vectors that can effectively distinguish between binding and non-binding sites. This process, feature engineering, directly determines the upper limit of the model's performance.

Sequence features focus on evolutionary information, usually extracted through position-specific scoring matrices (PSSM) or co-evolution analysis. Binding sites are often under evolutionary pressure, and the amino acid types exhibit higher conservation in species evolution. Structural features are derived from resolved three-dimensional structures and are crucial in this paradigm. Solvent accessible surface area (SASA) is used to quantify the degree of exposure of amino acid residues, with binding sites often located in regions with higher SASA values. Geometric descriptors, such as cavity depth, curvature, and surface normal vectors, are used to characterize

the unique topological shapes of binding pockets. In addition, electrostatic potential calculations based on the Poisson-Boltzmann equation can reveal the tendency of protein surfaces to interact with charged ligands; while hydrogen bond donor/acceptor distribution and hydrophobic field analysis further depict the chemical complementarity of the interactions.

Based on the above high-dimensional feature vectors, classical machine learning algorithms such as Random Forest (RF) and Support Vector Machine (SVM) are widely used [6]. Random Forest, by constructing multiple decision trees and integrating their results, can effectively handle high-dimensional features and has some resistance to overfitting. The feature importance ranking it provides is instructive for understanding the contribution of different features. Support Vector Machine is committed to finding an optimal hyperplane in the feature space to achieve maximum margin classification. When dealing with linearly inseparable problems, the introduction of kernel functions (such as radial basis functions) enables it to capture complex nonlinear relationships. However, the fundamental limitation of this paradigm is that its performance is severely dependent on the quality and completeness of feature engineering, and hand-designed features are difficult to comprehensively capture the intricate associations between protein structure and function.

2.2. End-to-End Prediction Paradigm Based on Deep Learning

Deep learning technology, through its multi-layer neural network structure, achieves the automatic learning of hierarchical feature representations from raw data, thus eliminating the reliance on manually designed features [7].

Graph Neural Network (GNN) models naturally represent the three-dimensional structure of proteins as graph structures, where nodes correspond to amino acid residues, and edges are defined based on spatial distance (e.g., the distance between C α atoms is less than a certain threshold) or chemical interactions. GNN variants such as Message Passing Neural Networks (MPNNs) update the representation of each node by iteratively aggregating information from neighboring nodes, thereby capturing complex long-range interactions and topological environments within the protein structure. This method can directly and naturally encode three-dimensional spatial relationships and is invariant to rotations and translations of the protein structure, making it one of the most promising architectures for processing structural information currently.

Three-dimensional Convolutional Neural Network (3D-CNN) models adopt a different strategy, embedding the protein into a three-dimensional voxel grid, where each voxel channel can encode properties such as atomic density, atom type, or electrostatic potential. By performing convolutional operations on the three-dimensional grid, 3D-CNNs automatically learn local and global spatial patterns, enabling effective identification of the unique cavity shapes of binding pockets. However, this method has a huge computational overhead, and its performance is sensitive to the voxelization resolution. More importantly, standard 3D-CNNs do not possess rotational invariance, and arbitrary rotations of the protein structure may lead to different prediction results, which is usually partially alleviated by data augmentation.

Sequence-Structure Fusion Models represent the current state-of-the-art research direction, aiming to integrate information from different modalities of proteins. These

models typically employ a dual-stream architecture: one-dimensional sequence information is processed through powerful sequence models such as Transformers or Long Short-Term Memory networks (LSTMs) to capture long-range evolutionary constraints and sequence context; three-dimensional structural information is processed through GNNs or 3D-CNNs. Finally, the learned sequence representations and structural representations are fused at specific levels (e.g., through attention mechanisms or simple concatenation) to form the final feature representation for each residue, which is used for classification. This fusion strategy fully leverages the biological central dogma of “sequence determines structure, and structure determines function,” and often achieves performance superior to single-modality models in practice.

2.3. Comparison of Algorithm Mechanisms and Suitability Analysis

From a mechanistic perspective, the fundamental difference between traditional machine learning and deep learning paradigms lies in the construction of the feature space. The former relies on explicit, human-interpretable biophysical rules and expert knowledge, resulting in relatively high model transparency but a ceiling on feature expression ability. The latter relies on implicit, data-driven feature learning, possessing a stronger expressive capacity and the ability to discover complex patterns that are difficult for humans to design, but its “black box” characteristic leads to poor interpretability.

Regarding input requirements, traditional methods and structure-based deep models typically depend on known or highly accurate predicted three-dimensional structures. In contrast, advanced sequence-structure fusion models, against the backdrop of the widespread adoption of protein structure prediction models like AlphaFold2, demonstrate significant advantages: they can accept sequences as input and utilize predicted structural information for site prediction, greatly expanding their application scope to proteins whose structures have not yet been experimentally resolved. In terms of computational efficiency, models based on traditional features generally have faster training and prediction speeds, while 3D-CNNs and large fusion models require significant GPU computing resources. Therefore, algorithm selection requires balancing multiple factors such as prediction accuracy, interpretability, input data availability, and computational cost.

3. Practical Applications and Challenges of Algorithms in Specific Scenarios

The value of protein-binding site prediction algorithms is ultimately reflected in their ability to solve specific biomedical problems. However, no “universal” algorithm can be equally applied to all scenarios. The vast differences in ligand types and their interaction patterns require prediction strategies to be specifically tailored and to directly address the unique challenges that arise.

3.1. Prediction Targeting Small Molecule Drug Targets

The prediction of binding sites for small molecule drug targets is one of the most mature application areas in computational biology. Such sites typically appear as moderately deep hydrophobic cavities with well-defined boundaries on the protein surface. In practice, the success of algorithms (such as P2Rank, which is based on geometric cavity detection and machine learning ranking, or DeepSite, which is based

on 3D-CNN) largely depends on capturing the detailed features of local hydrophobicity and shape. These predictions directly serve structure-based virtual screening, enabling the rapid identification of lead compounds from massive compound libraries that are complementary to the predicted pockets, or providing new target clues for drug repurposing.

However, the challenges in this scenario are increasingly prominent. The primary problem is the identification of allosteric sites. Allosteric sites are located far from the active center of the protein, and their binding regulates protein function through long-range conformational changes. These sites are evolutionarily less conserved, have more irregular geometric features, and lack a large number of known binding templates, resulting in a significant decrease in the prediction sensitivity of mainstream algorithms. Another deeper challenge is that most current algorithms can only predict the “possibility” of binding, but it is difficult to accurately distinguish the strength of binding affinity. A successfully predicted site may bind a ligand with a weak micromolar affinity, which is of limited value in drug design, but algorithms currently struggle to robustly quantify binding strength from static structures.

3.2. Prediction of Protein-Protein Interaction (PPI) Interfaces

Protein-protein interaction (PPI) interfaces regulate the core processes of cell life activities and are important targets for developing new therapeutic strategies (such as PPI inhibitors). PPI interfaces are typically broader and flatter than small molecule binding sites, lacking obvious deep cavities, and their binding affinity depends on complementary electrostatic and hydrophobic patches formed by multiple residues, rather than local deep embedding. Therefore, successful prediction strategies (such as MaSIF) focus on segmenting and encoding the large-scale physicochemical properties of the protein surface, and using GNNs to integrate interaction networks spanning multiple residues.

The core challenge in this field stems from the dynamics and complexity of PPIs themselves. Many PPIs with important biological functions are transient weak interactions with low binding constants, which may not be stably captured in crystal structures, leading to systematic bias in the training data. A more difficult problem is the conformational change induced by binding. The conformation of a protein in the unbound state (apo) and the bound state (holo) may differ significantly, i.e., “conformational selection” and “induced fit”. The vast majority of current algorithms use a single, static apo structure as input, which undoubtedly ignores the dynamic nature of the binding process, leading to the failure to predict certain flexible interfaces.

3.3. Prediction of Allosteric Sites and Peptide Ligand Sites

Allosteric site prediction is essentially a process of discovering “unknown” regions. Due to the lack of clear sequence or structural consensus in allosteric sites, directly applying models designed for active sites often yields unsatisfactory results. Cutting-edge research is attempting to introduce additional physical or evolutionary information to enhance the signal. For example, residue interaction networks or dynamic coupling parameters obtained through molecular dynamics simulations can identify regions that have a critical impact on the overall conformational fluctuations of proteins; these regions are often associated with allosteric regulation. In addition, analyzing long-range co-evolutionary signals may also reveal residue pairs that are

functionally coupled but spatially separated, providing clues for inferring allosteric communication pathways.

Prediction of peptide ligand sites faces difficulties of another dimension. As ligands, peptides have binding modes that fall between small molecules and proteins: they may form local helical or extended conformations, binding in an elongated, shallow groove. The geometric characteristics of these sites are not as obvious as small molecule pockets, but not as flat as PPI interfaces. The challenge for models lies in the need to simultaneously handle the structure of the protein receptor and have some tolerance for the flexibility of potential peptide binding conformations. Some methods attempt to address this challenge by simulating peptide conformation ensembles or learning short linear motif features in known protein-peptide complexes, but the prediction accuracy is still far lower than that of small molecule targets, highlighting the fundamental difficulty of dealing with ligand flexibility and binding mode diversity.

4. Systematic Evaluation and Performance Bottleneck Analysis

Objective and systematic evaluation of protein binding site prediction algorithms is a crucial step in advancing the field. Only through in-depth analysis of the performance boundaries and failure modes of each algorithm on a unified benchmark can reliable guidance be provided for their optimization and application.

4.1. Construction of Evaluation Datasets and Benchmarks

Reliable evaluation begins with high-quality, unbiased datasets. Currently, commonly used benchmark datasets in the field, such as Benchmark_DS and sc-PDB derived from PDB, screen protein-ligand complexes through a series of criteria (such as structural resolution and ligand properties), providing a basis for performance comparison. However, the inherent limitations of these datasets themselves profoundly affect the evaluation results. The most central problem is data bias: the structures in the PDB database are not randomly sampled, but are overly inclined towards proteins that are easy to crystallize and structurally stable, and are rich in certain popular target families (such as kinases and GPCRs). This leads to a sharp decline in the generalization ability of models trained on such data when facing structurally unique or sparse protein families. Another fundamental challenge is the construction of negative samples. Binding site prediction is essentially a highly imbalanced classification problem, where positive samples (binding residues) are far fewer than negative samples (non-binding residues). Simply defining non-binding residues as surface residues other than binding residues may mix in unannotated potential binding sites or regions with similar physicochemical properties to real sites. This "noise" can overestimate the true performance of the model. Therefore, constructing more rigorous and biologically meaningful negative sample sets, for example, by defining non-binding regions through steric hindrance or experimental validation, is an important direction for improving the rigor of the evaluation.

4.2. Evaluation Metric System and Interpretation

Selecting appropriate evaluation metrics is essential for correctly interpreting algorithm performance. Metrics such as precision, recall, and F1-score provide insights from different perspectives. Precision measures the proportion of truly positive samples among those predicted as positive, focusing on the reliability of the prediction results; recall measures the proportion of all true positive samples that are

successfully predicted, focusing on the model's ability to discover. The F1-score, as the harmonic mean of precision and recall, is a commonly used metric for comprehensive evaluation.

In the context of highly imbalanced data, the area under the receiver operating characteristic curve (AUC-ROC), although widely used, may give an overly optimistic evaluation due to the majority of negative samples. In contrast, the area under the precision-recall curve (AUC-PR) is more sensitive to the proportion of positive samples and can provide a more discriminating performance measure in imbalanced classification problems. In addition, the Matthews correlation coefficient (MCC), as a balanced indicator that comprehensively considers true positives, false positives, true negatives, and false negatives, its evaluation results are usually more reliable than the F1-score when the class distribution is extremely imbalanced.

4.3. Horizontal Comparison of Mainstream Algorithm Performance

A horizontal comparison of mainstream algorithms under a unified dataset and evaluation framework can reveal their relative advantages and disadvantages. For example, in benchmarks like sc-PDB, advanced methods based on graph neural networks (e.g., P2Rank, MaSIF) and sequence-structure fusion models (e.g., ScanNet) typically exhibit significantly higher AUC-PR and F1-scores than traditional machine learning methods relying on handcrafted features and purely geometric methods. This confirms the powerful ability of deep learning in automatically capturing complex binding patterns.

Detailed comparisons can also reveal the specificity of algorithms. Certain algorithms excel in predicting small molecule binding sites but experience performance degradation when dealing with flatter protein-protein interaction interfaces. Other models may have high recall rates in identifying catalytic residues but insufficient precision, leading to more false positives. This performance difference highlights the “scenario dependence” described in Chapter 3, cautioning researchers to select algorithms based on specific tasks rather than overall rankings.

4.4. In-depth Analysis of Performance Bottlenecks and Error Sources

Going beyond performance figures to delve into the root causes of errors is a prerequisite for breaking through existing bottlenecks. These bottlenecks exist at multiple levels:

Constraints at the data level are fundamental. In addition to the aforementioned data bias and noise, the lack of sufficient training data limits the full potential of data-driven models, especially deep learning models. The diversity of protein structures and the complexity of binding patterns require sample sizes far exceeding currently available data to be fully learned.

There are two core shortcomings at the algorithm model level. The first is insufficient representation of conformational dynamics. The vast majority of models use a single static structure as input, completely ignoring the dynamic fluctuations of proteins during the execution of their functions. Binding events often involve conformational changes in local loop regions and relative movements of domains. This dynamic information is lost in static snapshots, naturally limiting the model's ability to predict flexible or allosteric sites. The second is insufficient embedding of physicochemical principles. Current models are mostly pure statistical learning. Although they can correlate patterns from data, they do not embed energy calculations,

thermodynamic laws, and other physical constraints as prior knowledge into the model. This causes prediction results to sometimes violate basic physical laws, affecting their robustness and generalization ability in real-world scenarios.

Simplification at the task definition level also constitutes an obstacle. Treating binding site prediction as a simple residue-level binary classification problem is an oversimplification of its complex biophysical reality. This paradigm ignores the strength of binding (affinity), cooperativity, and the chemical properties of ligands, leading to the limitations of a “one model predicts all” paradigm. Future developments may need to evolve toward more refined paradigms such as conditional prediction (given a specific ligand) or regression prediction (predicting binding strength).

Through this systematic evaluation and in-depth analysis of bottlenecks, we can clearly recognize that performance improvement depends not only on more sophisticated model architectures but also on higher quality data, the integration of protein dynamic characteristics, and a more profound abstraction of the essence of the task.

5. Conclusions

This study systematically reviews the technological evolution in the field of protein binding site prediction, from traditional machine learning to deep learning. It analyzes the core strategies, application effectiveness, and inherent limitations of different algorithms when dealing with specific scenarios such as small molecules, PPI interfaces, and allosteric sites. Systematic evaluation shows that although advanced algorithms perform well in benchmark tests, their performance is still limited by data quality and bias, the neglect of protein dynamics, and the inherent problems of the “universal model” paradigm.

Conflicts of Interest

The author declares that there is no conflict of interest regarding the publication of this article.

Funding

This work was supported by Training Program for Young and Middle-aged Teachers (Research-related) of Sanya University (Project No. USYJSPY24-52).

References

- [1] Zhang, S.D. Research on Protein-Ligand Binding Site Prediction Algorithm Based on Geometric Deep Learning. Shandong University, 2024.
- [2] Sun, N. Prediction of Protein Modification Sites and Biological Activity Based on Pre-trained Algorithm. Xiamen University, 2022.
- [3] Chen, Z. Algorithm Development for Protein Functional Sites Prediction Based on Sequence Information. China Agricultural University, 2014.
- [4] Xu, X.Y. Research on Protein Phosphorylation and Hydroxylation Site Prediction Algorithm Based on Hybrid Deep Learning Model. Zhejiang Sci-Tech University, 2019.

- [5] Wang, Z.Y. Prediction of Protein Phosphorylation Sites Based on Deep Learning. Jilin University, 2024.
- [6] Wei, X.; Jia, J.H. Prediction of Protein K-Caprolactamization Sites Based on Ensemble Support Vector Machine. *Journal of Jingdezhen University*, 2021, 36(03), 81-84 + 91.
- [7] Geng, J. Research on Protein Interaction and Its Site Prediction Methods Based on Deep Learning. Northwest A&F University, 2025.



© 2026 by the author(s); licensee International Technology and Science Publications (ITS), this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (<http://creativecommons.org/licenses/by/4.0/>)