TYPE Mini Review
PUBLISHED 03 February 2025
DOI 10.3389/fbinf.2025.1520382

Machine learning approaches for predicting protein-ligand binding sites from sequence data

Orhun Vural* and Leon Jololian

Department of Electrical and Computer Engineering, The University of Alabama at Birmingham, Birmingham, AL, United States

OPEN ACCESS

EDITED BY
Wen Wei,
Arizona State University, United States

REVIEWED BY
Kumar Yugandhar,
Cornell University, United States
Minh Nguyen,
Bioinformatics Institute (A∗STAR), Singapore

*CORRESPONDENCE
Orhun Vural,
✉ orhun@uab.edu

RECEIVED 31 October 2024
ACCEPTED 10 January 2025
PUBLISHED 03 February 2025

CITATION
Vural O and Jololian L (2025) Machine learning approaches for predicting protein-ligand binding sites from sequence data.
Front. Bioinform. 5:1520382.
doi: 10.3389/fbinf.2025.1520382

Proteins, composed of amino acids, are crucial for a wide range of biological functions. Proteins have various interaction sites, one of which is the protein-ligand binding site, essential for molecular interactions and biochemical reactions. These sites enable proteins to bind with other molecules, facilitating key biological functions. Accurate prediction of these binding sites is pivotal in computational drug discovery, helping to identify therapeutic targets and facilitate treatment development. Machine learning has made significant contributions to this field by improving the prediction of protein-ligand interactions. This paper reviews studies that use machine learning to predict protein-ligand binding sites from sequence data, focusing on recent advancements. The review examines various embedding methods and machine learning architectures, addressing current challenges and the ongoing debates in the field. Additionally, research gaps in the existing literature are highlighted, and potential future directions for advancing the field are discussed. This study provides a thorough overview of sequence-based approaches for predicting protein-ligand binding sites, offering insights into the current state of research and future possibilities.

KEYWORDS

protein-ligand binding sites, computational drug discovery, sequence-based methods, deep learning, binding prediction

1 Introduction

Protein-ligand binding sites are specific regions on proteins where various ligands—including small organic molecules, peptides, nucleotides, and proteins—can attach or bind (Zhao et al., 2020). Although experimental laboratory methods identify these regions with the highest accuracy, they are generally costly and time-consuming (Sadybekov and Katritch, 2023). Therefore, computational approaches to drug discovery have become increasingly important. These computational methods offer distinct advantages by reducing costs and speeding up identifying and optimizing potential drug candidates (Gupta et al., 2021). Predicting protein-ligand binding sites is a critical component of computational drug discovery, essential for pinpointing viable drug targets and advancing the development of new therapeutics (Stank et al., 2016). Recent advancements in machine learning have significantly improved this field by introducing sophisticated computational techniques to analyze the complex interactions between proteins and ligands (Xia et al., 2024). While traditional methods based on geometry, energy, or templates have been successful, deep learning has recently achieved much better results (Gagliardi et al., 2022). Deep learning models can learn complex patterns directly from raw data and generalize better across diverse datasets. Protein ligand binding sites prediction in computational models is divided into two main categories, based on input type: structure-based and sequence-based (Gamouh et al., 2023; Hosseini et al., 2024).

Structure-based methods in computational drug discovery (SBDD) utilize detailed knowledge of the spatial information of proteins and integrate chemical properties using methods such as voxel-grid techniques (Sunseri and Koes, 2020). Figure 1 presents a 3D view of the 6Y3C protein and its associated ligands (Miciaccia et al., 2021). The number of protein-ligand binding sites on a protein can vary widely, depending on the specific protein and its function. In Figure 1A, the regions highlighted in yellow, blue, and purple represent protein-ligand binding sites. Figure 1B focuses on one of the binding sites shown in Figure 1A, offering a closer view of how the ligand interacts with the binding pocket. Figure 1C highlights the specific interactions between the ligand and the surrounding amino acid residues. In recent years, deep learning techniques used to identify these regions have often approached the problem as either image segmentation or object detection within structure-based frameworks. For instance, studies like RefinePocket (Liu et al., 2023), Kalasanty (Stepniewska-Dziubinska et al., 2020), PointSite (Yan et al., 2022), and DeepPocket (Aggarwal et al., 2021) use image segmentation techniques for binding site prediction, while RecurPocket (Li et al., 2022) and FRSite (Jiang et al., 2019) employ object detection techniques. Structure-based approaches depend on high-resolution 3D protein structures from X-ray crystallography or NMR spectroscopy (Maveyraud and Mourey, 2020). These methods face challenges such as reliance on accurate structures, static views of dynamic proteins, and high time and cost demands. AlphaFold (Abramson et al., 2024) has revolutionized the determination of 3D protein structures, significantly reducing reliance on experimental methods. However, drug discovery still primarily depends on 1D amino acid sequence data for critical tasks. Advancing approaches like AlphaFold requires a deeper understanding of the 1D sequence data used as input. This topic is further explored in the Discussion and Analysis section.

Sequence-based methods utilize one-dimensional (1D) amino acid sequence data as input. The 1D sequence is a direct representation of the protein's genetic blueprint and is experimentally measurable with high reliability (Alfaro et al., 2021). Sequence-based methods are less computationally intensive, do not require high-resolution structural data, and can be applied to a wider variety of proteins, including those for which structural information is unavailable. There are many more known protein sequences than experimentally determined structures (Chelur and Priyakumar, 2022). The general process of sequence-based binding site identification begins with a given protein sequence as input, leading to the final prediction and evaluation. The first step is feature extraction, which is challenging due to the complexity and diversity of proteins. This involves converting linear sequence data into numerical vectors that accurately represent the protein's functional and structural characteristics. Effective feature extraction is critical because the quality of the numerical representation directly impacts the performance of subsequent machine learning models. These techniques include binary representation, which encodes the presence or absence of specific amino acids; physicochemical representation, which considers the chemical and physical properties of amino acids; evolution-based representation, which leverages evolutionary information from multiple sequence alignments; and structure or machine learning-based representations, which use structural data or advanced algorithms to infer relevant features (Jing et al., 2019). Once the protein sequence is converted into a numerical format, it is ready for training with machine learning models using datasets with known binding sites. These datasets provide the necessary ground truth for model training and validation. The datasets most frequently employed in the literature are sc-PDB (Desaphy et al., 2015), COACH420 (Krivák and Hoksza, 2018), HOLO4k (Krivák and Hoksza, 2018), PDBBind (Liu et al., 2015), CSAR NRC-HiQ (Dunbar et al., 2013), UniProt (UniProt Consortium, 2015), Pfam (Finn et al., 2014), BioLip (Zhang et al., 2024), and PiSite (Higurashi et al., 2009). Each dataset has its unique characteristics and specific applications, contributing to the robustness and generalizability of the trained models. For instance, COACH420, derived from the COACH (Yang et al., 2013b) test set, is a widely recognized benchmark dataset that includes 420 protein-ligand complexes. Each complex consists of a single-chain protein intricately bound to a small molecule ligand. HOLO4K: A larger and more challenging dataset with 4,009 protein-ligand complexes. It includes multi-chain structures, offering a wider range of protein binding scenarios.

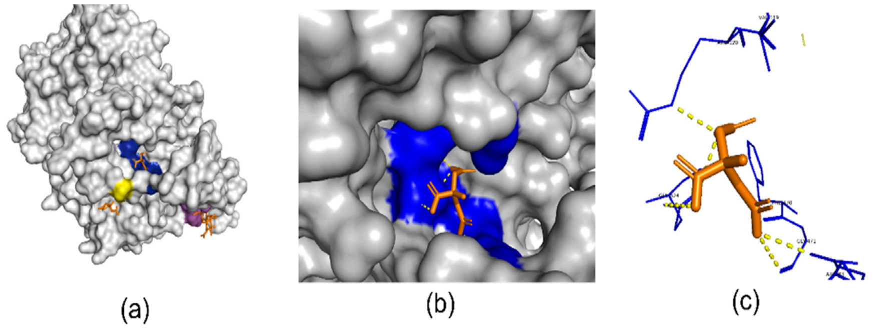

FIGURE 1
(A) Three binding site regions of 6Y3C protein in blue, yellow, and purple. **(B)** Close-up of a binding site with its ligand. **(C)** Ligand (orange) binding to a site. Generated with PyMol (Schrodinger, 2015).

TABLE 1 Sequence-based machine learning models for predicting protein-ligand binding sites.

Model	Feature extraction methods	Machine learning model[a]	Dataset	Evaluation metric	Accuracy[b]	Year
SCRIBER (Zhang and Kurgan, 2019)	ASAquick, HHblits, ANCHOR, PSIPRED, AAindex	Logistic Regression	BioLip, UniProt, Pfam	MCC	0.230	2019
DeepCSeqSite (Cui et al., 2019)	PSSpred, Anglor, Jensen-Shannon divergence (JSD), Relative entropy	Deep Convolutional Neural Network	BioLip	MCC	0.496	2019
DELIA (Xia et al., 2020)	PSI-BLAST, HHblits, SCRATCH-1D, S-SITE	ResNet + BiLSTM	BioLip, ATPBind	MCC	0.469	2020
HoTs (Lee and Nam, 2022)	1D-CNN, hierarchical recurrent neural network	CNN + Transformers	scPDB, PDBbind, COACH420, HOLO4k	Top-n success rate (%)	66.3 ± 0.9	2022
Birds (Chelur and Priyakumar, 2022)	DeepMSA, PSIPRED, SOLVPRED	ResNet	scPDB	MCC	0.568	2022
T5 GAT Ensemble (Gamouh et al., 2023)	ProtT5	Graph Neural Network + Attention	BioLip, RCSB	MCC	0.592	2023
LaMPSite (Zhang and Xie, 2023)	ESM-2, RDKit	Pooling + Clustering	scPDB, COACH420	Top-n success rate	66.02	2023
Pseq2Sites (Seo et al., 2024)	ProtTrans	CNN + Attention	COACH420, HOLO4k, CSAR	Top-n success rate	96.8	2024
Seq-InSite (Hosseini et al., 2024)	ProtT5, MSA	MLP + LSTM	PiSite	MCC	0.462	2024

[a] The Machine Learning Model column catalogs foundational models that constitute the core framework of the research presented, although the architecture of these studies may incorporate additional models.

[b] The reported results are sourced from their own publications. Please note that direct comparisons between these values may not be valid due to differences in methodologies, preprocessing steps, and testing datasets. If separate results were provided for each ligand type, their average was calculated.

In this paper, we focus on sequence-based protein-ligand binding site prediction studies that employ machine learning techniques. As seen in Table 1, we have summarized these studies by focusing on their feature extraction techniques, and machine learning models. The Analysis and Discussion section provides a detailed evaluation of the machine learning models listed in Table 1, highlighting the strengths, limitations, and research gaps of sequence-based approaches. Additionally, potential future directions are outlined in the Future Directions section.

2 Sequence-based computational methods

Proteins are composed of a set of amino acids, each represented by a unique symbol (e.g., "A" for Alanine, "G" for Glycine). Similar to human language, which consists of sequences of words that convey meaning, protein sequences are structured in specific patterns that hold significant biological information. To analyze these sequences, feature engineering techniques are employed to derive meaningful attributes from the data. Machine learning models are then trained on these features to predict protein-ligand interactions or other relevant biological properties.

2.1 Feature engineering

Sequence-based methods leverage sequence data to capture biochemical and biophysical properties without direct 3D structural information. Multiple review papers provide a detailed overview of embedding approaches for protein sequence-based structures (Jing et al., 2019; Ibtehaz and Kihara, 2023; Villegas-Morcillo et al., 2022; Zhang and Liu, 2019; Hoksza and Gamouh, 2022; Tran et al., 2023). Embedding methods have been categorized in various ways across different studies. Jing et al. (2019) classified these methods into five distinct categories based on their information sources and methodologies: binary encoding, physicochemical properties encoding, evolution-based encoding, structure-based encoding, and machine-learning encoding. We categorize embedding methods into two groups: traditional embedding methods and machine learning-based embedding methods.

Transformer-based models (Vaswani et al., 2017) have gained popularity for applying linguistic analogies to protein sequences. For example, ProtTrans (Elnaggar et al., 2021), ESM-1b (Rives et al., 2021), and ESM-MSA (Rao et al., 2021) are transformer-based protein language models used for feature extraction. ProtTrans includes models like ProtBert and ProtT5, leveraging the transformer architecture to process large-scale protein datasets and produce sequence embeddings. ProtBert has 420 million parameters and was trained on 2 billion protein sequences. ESM-1b employs a transformer-based architecture to generate embeddings for protein sequences and has been trained on 250 million protein sequences. ESM-MSA is another protein language model that uses multiple sequence alignments (MSAs) from UniRef50 (Suzek et al., 2007) as input, interleaving row and column attention. It is trained on 26 million MSAs. Other popular advanced embedding methods for protein sequences are ProtVec (Asgari and Mofrad, 2015), SeqVec (Heinzinger et al., 2019), and UniRep (Alley et al., 2019). ProtVec uses the skip-gram-based Word2Vec model (Mikolov et al., 2013) to treat amino acid k-mers like words. It is trained on a corpus of 546,790 sequences obtained from Swiss-Prot (Boutet et al., 2007). SeqVec uses the Embeddings from Language Models (ELMo) (Sarzynska-Wawer et al., 2021) approach, which generates context-aware embeddings by considering the surrounding amino acids in a sequence. UniRep, based on a multiplicative Long Short-Term Memory (mLSTM) model (Krause et al., 2016), captures essential biochemical properties by predicting the next amino acid in a sequence and is trained on approximately 24 million protein sequences from UniRef50.

In addition to these protein language models, various other methods can be employed to create feature maps from protein sequences. These techniques include 1D-CNN, calculating relative solvent accessibility (RSA), position-specific score matrix (PSSM), secondary structure (SS), token embeddings, segment embeddings, one-hot encoding, conservation scores (CS), amino acid composition (AAC), physiochemical properties, and more (Laine et al., 2021; Guo et al., 2021; Raj and Chandra, 2024). Many specialized tools and software have been developed to calculate these features, enabling the generation of comprehensive feature maps from protein sequences.

2.2 Methodological approaches

Table 1 lists studies that focus on sequence-based protein binding site prediction. In this section, we provide an overview of each model included in Table 1, highlighting the feature extraction techniques employed, the specific machine learning algorithms applied.

SCRIBER (Zhang and Kurgan, 2019) converts input protein sequences into profiles representing structural, evolutionary, and physicochemical properties. These profiles include relative solvent accessibility (RSA) values predicted by ASAquick (Faraggi et al., 2014), which calculates solvent accessibility scores using only sequence-based features without relying on 3D protein structures and predicts the ASA for each residue based on encoded sequence features. Other features include evolutionary conservation values from HHblits (Remmert et al., 2012), relative amino acid propensity (RAAP) scores, protein-binding disorder from ANCHOR (Dosztányi et al., 2009), secondary structure from PSIPRED (Buchan et al., 2013), a sequence-based tool. Physicochemical properties (charge, hydrophobicity, and polarity) from the AAindex resource (Kawashima et al., 2007). SCRIBER employs a logistic regression model (Cramer, 2002) to predict protein-binding residues. SCRIBER processes a protein in approximately 45 s, significantly faster than PSI-BLAST, which takes 194 s, and PSI-BLAST combined with SANN (Joo et al., 2012), which requires 246 s.

DeepCSeqSite (Cui et al., 2019) leverages a Deep Convolutional Neural Network along with position-specific score matrix (PSSM), relative solvent accessibility (RSA), and secondary structure (SS) anticipated through PSSpred (Yan et al., 2013). RSA, a numeric value (often between 0 and 1), indicates how much of a residue's surface is solvent-exposed versus buried. PSSpred uses neural networks to predict secondary structure elements, such as alpha-helices, beta-sheets, and coils, directly from sequence data. These elements, combined with positional embeddings, are used to build a detailed feature map from the protein sequence. To further enhance prediction accuracy, additional features such as conservation scores—calculated via Jensen-Shannon divergence (JSD) and relative entropy—residue type and dihedral angles, with predictions made by ANGLOR (Wu and Zhang, 2008), are incorporated.

DELIA (Xia et al., 2020) predicts protein–ligand binding residues using a hybrid model of convolutional neural networks (CNNs) (LeCun and Bengio, 1995) and bidirectional long short-term memory networks (BiLSTMs) (Schuster and Paliwal, 1997). It processes both 1D sequence feature vectors and 2D distance matrices to analyze amino acid sequences alongside protein spatial structures. DELIA utilizes sequence-based insights by integrating PSSMs from PSI-BLAST for evolutionary insights, fast and accurate evolutionary data from HHblits, secondary structure, and solvent accessibility predictions from SCRATCH-1D (Cheng et al., 2005), as well as binding propensities from S-SITE (Yang et al., 2013a). The SCRATCH software generates predictions for secondary structure and solvent accessibility using the amino acid sequence provided.

HoTS (Lee and Nam, 2022), employs a hierarchical recurrent neural network and 1D-CNN for protein sequence embedding to predict binding regions and drug–target interactions. HoTS leverages both CNN and transformer-based models, utilizing CNN layers to identify sequential motifs and transformers to model interdependencies. It also employs fully connected layers for accurately predicting binding regions.

Birds (Chelur and Priyakumar, 2022), utilizes a ResNet (He et al., 2016) architecture to predict a protein's binding site based on the protein's sequence information. This study employs a variety of techniques to extract information from protein sequences and construct a feature map, including token, positional, and segment embeddings, as well as multiple sequence alignments (MSAs) from DeepMSA (Zhang et al., 2020). From these MSAs, the position-specific score matrix (PSSM), Secondary Structure (SS), and Information Content (IC) were derived. Additionally, the Relative Solvent Accessibility (RSA) of each amino acid was determined by SOLVPRED from MetaPSICOV 2.0 (Jones et al., 2015).

T5 GAT Ensemble (Gamouh et al., 2023) predicts protein ligand binding sites with a hybrid approach combining sequence and structure data. This approach incorporates protein language models (pLMs) for sequence analysis and Graph Neural Networks (GNNs) (Scarselli et al., 2008) for structural insights, utilizing ProtT5-XL-UniRef50 (Elnaggar et al., 2021) to generate amino acid sequence embeddings. These embeddings serve as node features in the protein graph. The construction of the protein graph leverages the Python Deep Graph Library (DGL) (Wang et al., 2019), facilitating a sophisticated approach to modeling protein structures. In this graph, nodes are designated for individual residues, and edges define the spatial proximity between these residues. To determine the most suitable architecture, they tested two well-known GNN designs: the Graph Convolutional Network (GCN) (Kipf and Welling, 2016) and the Graph Attention Network (GAT) (Veliˇckovi´c et al., 2017).

LaMPSite (Zhang and Xie, 2023) predicts ligand binding sites using protein sequences and ligand molecular graphs. This approach incorporates residue-level embeddings from the ESM-2 protein language model (Lin et al., 2023) for proteins and atom-level embeddings from a graph neural network for ligands. Additionally, LaMPSite employs a pooling module to aggregate interaction embeddings, simplifying them to generate a residue-specific score. Then it clusters residues using the protein contact map, ranking these clusters to pinpoint binding sites. Current clustering and filtering processes typically yield one binding site per prediction, which may limit the identification of multiple or cryptic binding sites.

Pseq2Sites (Seo et al., 2024) uses ProtTrans, a transformer-based model, to extract amino acid-level embeddings for protein sequence analysis. Subsequently, 1D-CNNs were utilized to extract local features from the resulting embedding sequence, followed by the application of methods employing position-based attention mechanisms to capture long-distance contextual information.

Seq-InSite (Hosseini et al., 2024) utilizes ProtT5 and MSA-transformer embeddings to predict protein interaction sites from sequence data. Its architecture employs ensemble learning techniques, integrating a Multi-Layer Perceptron (MLP) and a Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) network. While Seq-InSite predicts a broad range of protein interaction sites, including protein-ligand binding sites.

Overall, accurate prediction of protein-ligand binding sites is a crucial step in the drug discovery pipeline. Beyond theoretical predictions, these methods provide actionable insights that support drug target identification, lead optimization, and ligand design. Once protein binding sites are identified, these predictions lead to a variety of applications, including virtual screening (Kimber et al., 2021), studying off-target effects (Rao et al., 2023), predicting druggability scores (Raies et al., 2022), protein function prediction (Kulmanov and Hoehndorf, 2020), assessing mutation impacts (Sun et al., 2021), and pose prediction (Wang et al., 2022), among others.

3 Analysis and discussion

This section discusses four main topics: advancements in extracting features from protein sequences, the limitations of sequence-based methods with an analysis of the approaches listed in Table 1, the advantages of hybrid methods that combine sequence- and structure-based techniques, and a review of the datasets used for testing, as well as tools like AlphaFold that are employed for protein folding predictions. Each topic highlights critical aspects of the methodologies and their contributions to improving protein-ligand binding site predictions.

The models in Table 1 demonstrate a broad range of feature extraction techniques, spanning traditional evolution- and structure-based encodings to advanced protein language models (pLMs). 1D-CNNs are effective at extracting local motifs from protein sequences but may lose global context when motifs are spread across non-consecutive regions (Lee and Nam, 2022). PSSMs, a cornerstone of traditional methods, remain critical for capturing evolutionary information, with their removal causing significant performance drops (Chelur and Priyakumar, 2022). Relative solvent accessibility (RSA) and secondary structure elements add structural insights, but their impact on performance is less pronounced than that of embeddings or PSSMs (Chelur and Priyakumar, 2022). Secondary structure features and predicted dihedral angles provide structural context, with dihedral angles offering more fine-grained information; however, these features may also introduce noise (Cui et al., 2019). Protein language models, such as ProtT5-XL, offer significant advantages in terms of processing speed, generating embeddings for a human protein in as little as 0.12 s (Elnaggar et al., 2021). This efficiency is essential when analyzing extensive datasets with millions of sequences, allowing for high accuracy without reliance on traditional, computationally intensive evolutionary steps. ProtT5-XL embeddings, for example, deliver high accuracy and rich information, outperforming alternatives such as MSA-transformer embeddings in predictive tasks (Hosseini et al., 2024). Protein language models (pLMs) tend to be less effective for proteins that are rare or underrepresented in training datasets. However, pLMs perform best with well-represented proteins, and challenges remain in predicting binding sites for rare or novel proteins due to limited sequence data representation. As shown in Table 1, studies T5 GAT Ensemble, LaMPSite, Pseq2Sites, and Seq-InSite, which utilize pLMs extraction methods, demonstrate promising results compared to other studies listed in Table 1 that use traditional feature extraction methods.

One key advantage of sequence-based methods is their computational efficiency. For instance, on the well-known COACH420 dataset, sequence-based protein-ligand binding site prediction methods achieved significantly faster execution times: Pseq2Sites completed predictions in 1.07 s, Birds in 3.97 s, DeepCSeqSite in 11.13 s, and HoTs in 51.84 s. In contrast, structure-based methods were considerably slower, with DeepPocket taking 894.28 s, DeepSurf 2436.76 s, and P2Rank 914.61 s (Seo et al., 2024). Although sequence-based methods are computationally efficient, they lack the spatial context needed to identify complex binding interactions, such as those involving residues across multiple protein chains. By analyzing each chain individually and then combining the results, traditional sequence-based methods often miss critical relationships, limiting their accuracy in predicting binding sites.

The studies in Table 1 highlight distinct characteristics of various models. For instance, SCRIBER incorporates over 1,000 input features and relies on feature elimination techniques to manage complexity, though it remains susceptible to overfitting. SCRIBER reported a Matthews correlation coefficient (MCC) (Chicco and Jurman, 2020) of 0.23. DELIA, on the other hand, is tailored for specific ligand types, which enhances predictive accuracy for those interactions but limits its applicability to general protein-ligand binding site prediction. DELIA achieved an average MCC of 0.469, which was derived from results across five different ligand types. Attention-based models like HoTS and Pseq2Sites excel at capturing both local interactions and long-range dependencies within sequences, making them effective for understanding complex sequence patterns. In the Pseq2Sites study, Pseq2Sites demonstrated a 96.8% success rate on the COACH test dataset, calculated as the number of correctly identified pockets divided by the total number of pockets. Additionally, the study reported success rates for other models, with HoTS achieving 14.3% and Birds reaching 70%, highlighting the comparative performance within the same evaluation framework. Seq-InSite achieved an MCC value of 0.462 on the Dset_448 dataset, which focuses on ligands that are not proteins. However, sequence-based models still struggle to fully capture inter-chain interactions, which are critical for predicting functional binding sites in multimeric proteins. Sequence-based approaches are generally less effective in identifying allosteric binding sites, which are often located far from the active site and can be missed without considering the protein's full 3D structure (Xia et al., 2024).

Hybrid approaches, which integrate both sequence-based and structural features, have emerged as powerful strategies to enhance the accuracy of protein function prediction tasks. The T5-GAT Ensemble, a hybrid model, combines sequence and structural features of proteins. While the sequence-based MLP model achieves an MCC of 0.54, the hybrid model improves this to 0.59 by incorporating structural features. Similarly, DELIA, tested on five ligand types, demonstrated that the hybrid architecture outperformed sequence-based models in MCC scores for all ligand types. Another method, LaMPSite, predicts ligand binding sites by utilizing both protein sequences and ligand molecular graphs. The ablation study for LaMPSite indicates a decrease in accuracy when the interaction module, which combines the benefits of both methods, is omitted. For this study, the reported success rate in terms of DCA (Distance Cutoff Accuracy) is 66.02%.

The choice of datasets in protein-ligand binding site prediction plays a crucial role in developing and evaluating computational models. To ensure fair testing, addressing data leakage is essential, especially the similarity between training and test datasets. For instance, LaMPSite excludes scPDB structures with more than 50% sequence identity or 0.9 ligand similarity and removes proteins from COACH420. Pseq2Sites takes additional steps by using unseen test datasets and filtering proteins with ≤40% structural similarity for unbiased evaluation. Studies like HoTS further promote fair analysis by reporting results at various similarity thresholds.

Protein folding software such as AlphaFold can facilitate hybrid approaches, certain limitations persist. AlphaFold2 (AF2) relies on patterns extracted from known protein folds rather than understanding the physical and chemical basis of proteins (Agarwal and McShan, 2024). The experimentally determined 3D structural dataset is limited to fewer than 300,000 structures, compared to the billions of protein sequences available in public repositories. AlphaFold 3 (AF3) builds on the evoformer architecture from AF2, incorporating a diffusion network that refines a cloud of atoms iteratively to generate highly accurate protein structures. AF3 can predict heme-binding sites; however, its reliance on structurally similar proteins in its training data limits its effectiveness for less-represented or novel protein sequences (Kondo and Takano, 2024). AF3 struggles to accurately predict ligand-binding poses, particularly for complex ligands such as peptides, ions, and non-standard molecules (He et al., 2024). Additionally, the lack of support for user-defined ligands and a broader range of ligand types further restricts AF3's applicability in practical drug discovery efforts. Single changes in the sequence (e.g., point mutations) can significantly alter a protein's function or cause misfolding. AF2 is not trained to predict the effects of mutations on protein structure or stability (Agarwal and McShan, 2024; Pak et al., 2023). AF3 has limitations in stereochemistry, hallucinations, dynamic behavior, and accuracy for specific targets (Abramson et al., 2024). The Predicted Local Distance Difference Test (pLDDT) serves as a confidence metric in AF2 and AF3 for evaluating the reliability of protein structure predictions. However, high pLDDT values or low Predicted Aligned Error (PAE) scores do not necessarily ensure alignment with experimental structures (Carugo, 2023; Buel and Walters, 2022).

Overall, the paper highlights the strengths and limitations of both 3D and 1D approaches, concluding in the discussion section that hybrid methodologies represent a promising direction for future research.

4 Future directions

Future advancements in protein binding site prediction are likely to focus on integrating sequence-based and structure-based data to improve model accuracy, particularly for complex binding sites that depend on 3D spatial context. Hybrid models that combine these two types of data show promise in addressing limitations of sequence-only methods, such as identifying distant allosteric sites or inter-chain interactions. Another promising direction involves the development of transformer-based models specifically tailored for protein-ligand interactions, utilizing advanced embeddings to capture intricate sequence patterns and dependencies. Recently, GPT-based (Brown et al., 2020) studies have emerged in protein engineering, harnessing protein sequence data and the capabilities of large language models (LLMs) rooted in natural language processing (NLP). These advancements emphasize the need for a deeper understanding of protein sequence data, improving its representation, and designing deep learning architectures to align with these enhancements. Reviews like ours, which focus on sequence-based protein structures, are expected to make valuable contributions to the development of these tools. To further advance the field, it will be critical to enhance the adaptability of protein language models (pLMs) for underrepresented or rare proteins. This could be achieved by expanding training datasets or developing adaptive embedding methods. Additionally, collaboration across computational, experimental, and industrial fields will be essential for validating and refining these models. Such efforts aim to improve generalizability and optimize predictive tools for specific therapeutic targets, ultimately accelerating advancements in computational drug discovery.

5 Conclusion

The prediction of protein-ligand binding sites is crucial for advancing drug discovery and development, as it enables the identification of potential drug targets and the design of more effective therapeutics. Accurate prediction methods can significantly streamline the drug discovery process, reducing the time and cost associated with experimental validation. Our study reviews various sequence-based approaches for predicting protein-ligand binding sites using machine learning techniques in computational drug discovery. Our examination explores the models, focusing on their embedding methods and deep learning architectures, and discusses the challenges and future directions associated with sequence-based methods. Our study aims to serve as a comprehensive guide for sequence-based prediction of protein-ligand binding sites, providing a thorough understanding of the existing literature within a single paper.

Author contributions

OV: Conceptualization, Data curation, Formal Analysis, Investigation, Resources, Writing–original draft, Writing–review and editing. LJ: Project administration, Supervision, Validation, Writing–original draft, Writing–review and editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

The authors would like to sincerely thank Dr. Lurong Pan for her valuable guidance and support throughout this study. We also extend our gratitude to Recep Tayyip Erdogan University.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., et al. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 630, 493–500. doi:10.1038/s41586-024-07487-w

Agarwal, V., and McShan, A. C. (2024). The power and pitfalls of AlphaFold2 for structure prediction beyond rigid globular proteins. *Nat. Chem. Biol.* 20 (8), 950–959. doi:10.1038/s41589-024-01638-w

Aggarwal, R., Gupta, A., Chelur, V., Jawahar, C., and Priyakumar, U. D. (2021). DeepPocket: ligand binding site detection and segmentation using 3D convolutional neural networks. *J. Chem. Inf. Model.* 62 (21), 5069–5079. doi:10.1021/acs.jcim.1c00799

Alfaro, J. A., Bohländer, P., Dai, M., Filius, M., Howard, C. J., Van Kooten, X. F., et al. (2021). The emerging landscape of single-molecule protein sequencing technologies. *Nat. methods* 18 (6), 604–617. doi:10.1038/s41592-021-01143-1

Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M., and Church, G. M. (2019). Unified rational protein engineering with sequence-based deep representation learning. *Nat. methods* 16 (12), 1315–1322. doi:10.1038/s41592-019-0598-1

Asgari, E., and Mofrad, M. R. (2015). Continuous distributed representation of biological sequences for deep proteomics and genomics. *PloS one* 10 (11), e0141287. doi:10.1371/journal.pone.0141287

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., and Bairoch, A. (2007). "UniProtKB/Swiss-Prot: the manually annotated section of the UniProt KnowledgeBase," in *Plant bioinformatics: methods and protocols* (Springer), 89–112.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Adv. neural Inf. Process. Syst.* 33, 1877–1901. doi:10.48550/arXiv.2005.14165

Buchan, D. W., Minneci, F., Nugent, T. C., Bryson, K., and Jones, D. T. (2013). Scalable web services for the PSIPRED protein analysis workbench. *Nucleic acids Res.* 41 (W1), W349–W357. doi:10.1093/nar/gkt381

Buel, G. R., and Walters, K. J. (2022). Can AlphaFold2 predict the impact of missense mutations on structure? *Nat. Struct. and Mol. Biol.* 29 (1), 1–2. doi:10.1038/s41594-021-00714-2

Carugo, O. (2023). pLDDT values in AlphaFold2 protein models are unrelated to globular protein local flexibility. *Crystals* 13 (11), 1560. doi:10.3390/cryst13111560

Chelur, V. R., and Priyakumar, U. D. (2022). Birds-binding residue detection from protein sequences using deep resnets. *J. Chem. Inf. Model.* 62 (8), 1809–1818. doi:10.1021/acs.jcim.1c00972

Cheng, J., Randall, A. Z., Sweredoski, M. J., and Baldi, P. (2005). SCRATCH: a protein structure and structural feature prediction server. *Nucleic acids Res.* 33 (Suppl. l_2), W72–W76. doi:10.1093/nar/gki396

Chicco, D., and Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics* 21, 6–13. doi:10.1186/s12864-019-6413-7

Cramer, J. S. (2002). *The origins of logistic regression, tinbergen Institute working paper, no. 2002-119/4*. doi:10.2139/ssrn.360300

Cui, Y., Dong, Q., Hong, D., and Wang, X. (2019). Predicting protein-ligand binding residues with deep convolutional neural networks. *BMC Bioinforma.* 20, 93–12. doi:10.1186/s12859-019-2672-1

Desaphy, J., Bret, G., Rognan, D., and Kellenberger, E. (2015). sc-PDB: a 3D-database of ligandable binding sites—10 years on. *Nucleic acids Res.* 43 (D1), D399–D404. doi:10.1093/nar/gku928

Dosztányi, Z., Mészáros, B., and Simon, I. (2009). ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics* 25 (20), 2745–2746. doi:10.1093/bioinformatics/btp518

Dunbar, J. B., Jr, Smith, R. D., Damm-Ganamet, K. L., Ahmed, A., Esposito, E. X., Delproposto, J., et al. (2013). CSAR data set release 2012: ligands, affinities, complexes, and docking decoys. *J. Chem. Inf. Model.* 53 (8), 1842–1852. doi:10.1021/ci4000486

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., et al. (2021). Prottrans: toward understanding the language of life through self-supervised learning. *IEEE Trans. pattern analysis Mach. Intell.* 44 (10), 7112–7127. doi:10.1109/tpami.2021.3095381

Faraggi, E., Zhou, Y., and Kloczkowski, A. (2014). Accurate single-sequence prediction of solvent accessible surface area using local and global features. *Proteins Struct. Funct. Bioinforma.* 82 (11), 3170–3176. doi:10.1002/prot.24682

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., et al. (2014). Pfam: the protein families database. *Nucleic acids Res.* 42 (D1), D222–D230. doi:10.1093/nar/gkt1223

Gagliardi, L., Raffo, A., Fugacci, U., Biasotti, S., Rocchia, W., Huang, H., et al. (2022). SHREC 2022: protein–ligand binding site recognition. *Comput. and Graph.* 107, 20–31. doi:10.1016/j.cag.2022.07.005

Gamouh, H., Hoksza, D., and Novotny, M. (2023). *Hybrid protein-ligand binding residue prediction with protein language models: does the structure matter?* bioRxiv. 2023.08. 11.553028.

Guo, Y., Wu, J., Ma, H., Wang, S., and Huang, J. (2021). Comprehensive study on enhancing low-quality position-specific scoring matrix with deep learning for accurate protein structure property prediction: using bagging multiple sequence alignment learning. *J. Comput. Biol.* 28 (4), 346–361. doi:10.1089/cmb.2020.0416

Gupta, R., Srivastava, D., Sahu, M., Tiwari, S., Ambasta, R. K., and Kumar, P. (2021). Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Mol. Divers.* 25, 1315–1360. doi:10.1007/s11030-021-10217-3

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, June 27 2016–June 30 2016, 770–778.

He, Xh., Li, J. R., Shen, S. Y., and Xu, H. E. (2024). AlphaFold3 versus experimental structures: assessment of the accuracy in ligand-bound G protein-coupled receptors. *Acta Pharmacol. Sin*. doi:10.1038/s41401-024-01429-y

Heinzinger, M., Elnaggar, A., Wang, Y., Dallago, C., Nechaev, D., Matthes, F., et al. (2019). Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinforma.* 20, 723. doi:10.1186/s12859-019-3220-8

Higurashi, M., Ishida, T., and Kinoshita, K. (2009). PiSite: a database of protein interaction sites using multiple binding states in the PDB. *Nucleic acids Res.* 37 (Suppl. l_1), D360–D364. doi:10.1093/nar/gkn659

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput. MIT-Press* 9, 1735–1780. doi:10.1162/neco.1997.9.8.1735

Hoksza, D., and Gamouh, H. (2022). "Exploration of protein sequence embeddings for protein-ligand binding site detection," in 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Las Vegas, NV, USA, 06-08 December 2022 (IEEE), 3356–3361.

Hosseini, S., Golding, G. B., and Ilie, L. (2024). Seq-InSite: sequence supersedes structure for protein interaction site prediction. *Bioinformatics* 40 (1), btad738. doi:10.1093/bioinformatics/btad738

Ibtehaz, N., and Kihara, D. (2023). "Application of sequence embedding in protein sequence-based predictions," in *Machine learning in bioinformatics of protein sequences: algorithms, databases and resources for modern protein bioinformatics* (World Scientific), 31–55.

Jiang, M., Wei, Z., Zhang, S., Wang, S., Wang, X., and Li, Z. (2019). Frsite: protein drug binding site prediction based on faster r–cnn. *J. Mol. Graph. Model.* 93, 107454. doi:10.1016/j.jmgm.2019.107454

Jing, X., Dong, Q., Hong, D., and Lu, R. (2019). Amino acid encoding methods for protein sequences: a comprehensive review and assessment. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 17 (6), 1918–1931. doi:10.1109/tcbb.2019.2911677

Jones, D. T., Singh, T., Kosciolek, T., and Tetchner, S. (2015). MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics* 31 (7), 999–1006. doi:10.1093/bioinformatics/btu791

Joo, K., Lee, S. J., and Lee, J. (2012). Sann: solvent accessibility prediction of proteins by nearest neighbor method. *Proteins Struct. Funct. Bioinforma.* 80 (7), 1791–1797. doi:10.1002/prot.24074

Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., and Kanehisa, M. (2007). AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.* 36 (Suppl. l_1), D202–D205. doi:10.1093/nar/gkm998

Kimber, T. B., Chen, Y., and Volkamer, A. (2021). Deep learning in virtual screening: recent applications and developments. *Int. J. Mol. Sci.* 22 (9), 4435. doi:10.3390/ijms22094435

Kipf, T. N., and Welling, M. (2016). *Semi-supervised classification with graph convolutional networks*. arXiv preprint arXiv:1609.02907.

Kondo, H. X., and Takano, Y. (2024). Structure comparison of heme-binding sites in heme protein predicted by AlphaFold3 and AlphaFold2. *Chem. Lett.* 53 (8), upae148. doi:10.1093/chemle/upae148

Krause, B., Lu, L., Murray, I., and Renals, S. (2016). *Multiplicative LSTM for sequence modelling*. arXiv preprint arXiv:1609.07959.

Krivák, R., and Hoksza, D. (2018). P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *J. cheminformatics* 10, 39–12. doi:10.1186/s13321-018-0285-8

Kulmanov, M., and Hoehndorf, R. (2020). DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics* 36 (2), 422–429. doi:10.1093/bioinformatics/btz595

Laine, E., Eismann, S., Elofsson, A., and Grudinin, S. (2021). Protein sequence-to-structure learning: is this the end (-to-end revolution)? *Proteins Struct. Funct. Bioinforma.* 89 (12), 1770–1786. doi:10.1002/prot.26235

LeCun, Y., and Bengio, Y. (1995). Convolutional networks for images, speech, and time series. *Handb. brain theory neural Netw.* 3361 (10), 1995.

Lee, I., and Nam, H. (2022). Sequence-based prediction of protein binding regions and drug–target interactions. *J. cheminformatics* 14 (1), 5. doi:10.1186/s13321-022-00584-w

Li, P., Cao, B., Tu, S., and Xu, L. (2022). "Recurpocket: recurrent lmser network with gating mechanism for protein binding site detection," in 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Las Vegas, NV, USA, 06-08 December 2022 (IEEE), 334–339.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 379 (6637), 1123–1130. doi:10.1126/science.ade2574

Liu, Y., Li, P., Tu, S., and Xu, L. (2023). Refinepocket: an attention-enhanced and mask-guided deep learning approach for protein binding site prediction. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* 20, 3314–3321. doi:10.1109/tcbb.2023.3265640

Liu, Z., Li, Y., Han, L., Li, J., Liu, J., Zhao, Z., et al. (2015). PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics* 31 (3), 405–412. doi:10.1093/bioinformatics/btu626

Maveyraud, L., and Mourey, L. (2020). Protein X-ray crystallography and drug discovery. *Molecules* 25 (5), 1030. doi:10.3390/molecules25051030

Miciaccia, M., Belviso, B. D., Iaselli, M., Cingolani, G., Ferorelli, S., Cappellari, M., et al. (2021). Three-dimensional structure of human cyclooxygenase (h COX)-1. *Sci. Rep.* 11 (1), 4312. doi:10.1038/s41598-021-83438-z

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). *Efficient estimation of word representations in vector space*. *arXiv preprint arXiv:1301.3781*.

Pak, M. A., Markhieva, K. A., Novikova, M. S., Petrov, D. S., Vorobyev, I. S., Maksimova, E. S., et al. (2023). Using AlphaFold to predict the impact of single mutations on protein stability and function. *Plos one* 18 (3), e0282689. doi:10.1371/journal.pone.0282689

Raies, A., Tulodziecka, E., Stainer, J., Middleton, L., Dhindsa, R. S., Hill, P., et al. (2022). DrugnomeAI is an ensemble machine-learning framework for predicting druggability of candidate drug targets. *Commun. Biol.* 5 (1), 1291. doi:10.1038/s42003-022-04245-4

Raj, S. S., and Chandra, S. V. (2024). Significance of sequence features in classification of protein–protein interactions using machine learning. *Protein J.* 43 (1), 72–83. doi:10.1007/s10930-023-10168-8

Rao, M., McDuffie, E., and Sachs, C. (2023). Artificial intelligence/machine learning-driven small molecule repurposing via off-target prediction and transcriptomics. *Toxics* 11 (10), 875. doi:10.3390/toxics11100875

Rao, R. M., Liu, J., Verkuil, R., Meier, J., Canny, J. F., Abbeel, P., et al. (2021). "MSA transformer," in *International conference on machine learning* (PMLR), 8844–8856. https://proceedings.mlr.press/v139/rao21a.html.

Remmert, M., Biegert, A., Hauser, A., and Söding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. methods* 9 (2), 173–175. doi:10.1038/nmeth.1818

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., et al. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci.* 118 (15), e2016239118. doi:10.1073/pnas.2016239118

Sadybekov, A. V., and Katritch, V. (2023). Computational approaches streamlining drug discovery. *Nature* 616 (7958), 673–685. doi:10.1038/s41586-023-05905-z

Sarzynska-Wawer, J., Wawer, A., Pawlak, A., Szymanowska, J., Stefaniak, I., Jarkiewicz, M., et al. (2021). Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Res.* 304, 114135. doi:10.1016/j.psychres.2021.114135

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. (2008). The graph neural network model. *IEEE Trans. neural Netw.* 20 (1), 61–80. doi:10.1109/TNN.2008.2005605

Schrodinger, L. (2015). *The PyMOL molecular graphics system*, 8. Version 1.

Schuster, M., and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* 45 (11), 2673–2681. doi:10.1109/78.650093

Seo, S., Choi, J., Choi, S., Lee, J., Park, C., and Park, S. (2024). Pseq2Sites: enhancing protein sequence-based ligand binding-site prediction accuracy via the deep convolutional network and attention mechanism. *Eng. Appl. Artif. Intell.* 127, 107257. doi:10.1016/j.engappai.2023.107257

Stank, A., Kokh, D. B., Fuller, J. C., and Wade, R. C. (2016). Protein binding pocket dynamics. *Accounts Chem. Res.* 49 (5), 809–815. doi:10.1021/acs.accounts.5b00516

Stepniewska-Dziubinska, M. M., Zielenkiewicz, P., and Siedlecki, P. (2020). Improving detection of protein-ligand binding sites with 3D segmentation. *Sci. Rep.* 10 (1), 5035. doi:10.1038/s41598-020-61860-z

Sun, T., Chen, Y., Wen, Y., Zhu, Z., and Li, M. (2021). PremPLI: a machine learning model for predicting the effects of missense mutations on protein-ligand interactions. *Commun. Biol.* 4 (1), 1311. doi:10.1038/s42003-021-02826-3

Sunseri, J., and Koes, D. R. (2020). Libmolgrid: graphics processing unit accelerated molecular gridding for deep learning applications. *J. Chem. Inf. Model.* 60 (3), 1079–1084. doi:10.1021/acs.jcim.9b01145

Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R., and Wu, C. H. (2007). UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* 23 (10), 1282–1288. doi:10.1093/bioinformatics/btm098

Tran, C., Khadkikar, S., and Porollo, A. (2023). Survey of protein sequence embedding models. *Int. J. Mol. Sci.* 24 (4), 3775. doi:10.3390/ijms24043775

UniProt Consortium (2015). UniProt: a hub for protein information. *Nucleic acids Res.* 43 (D1), D204–D212. doi:10.1093/nar/gku989

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Aidan, N., et al. (2017). Attention is all you need. *Adv. neural Inf. Process. Syst.* 30. doi:10.48550/arXiv.1706.03762

Veliˇckovi´c, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. (2017). *Graph attention networks*. arXiv preprint arXiv:1710.10903.

Villegas-Morcillo, A., Gomez, A. M., and Sanchez, V. (2022). An analysis of protein language model embeddings for fold prediction. *Briefings Bioinforma.* 23 (3), bbac142. doi:10.1093/bib/bbac142

Wang, C., Chen, Y., Zhang, Y., Li, K., Lin, M., Pan, F., et al. (2022). A reinforcement learning approach for protein–ligand binding pose prediction. *BMC Bioinforma.* 23 (1), 368. doi:10.1186/s12859-022-04912-7

Wang, M., Zheng, D., Ye, Z., Gan, Q., Li, M., Song, X., et al. (2019). *Deep graph library: a graph-centric, highly-performant package for graph neural networks*. arXiv preprint arXiv:1909.01315.

Wu, S., and Zhang, Y. (2008). ANGLOR: a composite machine-learning algorithm for protein backbone torsion angle prediction. *PloS one* 3 (10), e3400. doi:10.1371/journal.pone.0003400

Xia, C. Q., Pan, X., and Shen, H.-B. (2020). Protein–ligand binding residue prediction enhancement through hybrid deep heterogeneous learning of sequence and structure data. *Bioinformatics* 36 (10), 3018–3027. doi:10.1093/bioinformatics/btaa110

Xia, Y., Pan, X., and Shen, H.-B. (2024). A comprehensive survey on protein-ligand binding site prediction. *Curr. Opin. Struct. Biol.* 86, 102793. doi:10.1016/j.sbi.2024.102793

Yan, R., Xu, D., Yang, J., Walker, S., and Zhang, Y. (2013). A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction. *Sci. Rep.* 3 (1), 2619. doi:10.1038/srep02619

Yan, X., Lu, Y., Li, Z., Wei, Q., Gao, X., Wang, S., et al. (2022). PointSite: a point cloud segmentation tool for identification of protein ligand binding atoms. *J. Chem. Inf. Model.* 62 (11), 2835–2845. doi:10.1021/acs.jcim.1c01512

Yang, J., Roy, A., and Zhang, Y. (2013a). Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics* 29 (20), 2588–2595. doi:10.1093/bioinformatics/btt447

Yang, J., Roy, A., and Zhang, Y. (2013b). Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics* 29 (20), 2588–2595. doi:10.1093/bioinformatics/btt447

Zhang, C., Zhang, X., Freddolino, P. L., and Zhang, Y. (2024). BioLiP2: an updated structure database for biologically relevant ligand–protein interactions. *Nucleic Acids Res.* 52 (D1), D404–D412. doi:10.1093/nar/gkad630

Zhang, C., Zheng, W., Mortuza, S., Li, Y., and Zhang, Y. (2020). DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. *Bioinformatics* 36 (7), 2105–2112. doi:10.1093/bioinformatics/btz863

Zhang, J., and Kurgan, L. (2019). SCRIBER: accurate and partner type-specific prediction of protein-binding residues from proteins sequences. *Bioinformatics* 35 (14), i343–i353. doi:10.1093/bioinformatics/btz324

Zhang, J., and Liu, B. (2019). A review on the recent developments of sequence-based protein feature extraction methods. *Curr. Bioinforma.* 14 (3), 190–199. doi:10.2174/1574893614666181212102749

Zhang, S., and Xie, L. (2023). *Protein Language model-powered 3D ligand binding site prediction from protein sequence*. arXiv preprint arXiv:2312.03016.

Zhao, J., Cao, Y., and Zhang, L. (2020). Exploring the computational methods for protein-ligand binding site prediction. *Comput. Struct. Biotechnol. J.* 18, 417–426. doi:10.1016/j.csbj.2020.02.008