Advanced Science

ADVANCED SCIENCE Open Access

www.advancedscience.com

RESEARCH ARTICLE OPEN ACCESS

Accurate Identification of Protein Binding Sites for All Drug Modalities Using ALLSites

Minjie Mou[1,2] | Mingkun Lu[2] | Zhimeng Zhou[2] | Yanlin Ren[2] | Xinyuan Yu[2] | Ziqi Pan[2] | Yuan Zhou[2] | Hao Yang[3] | Lingyan Zheng[2] | Shukai Gu[2] | Yang Zhang[3] | Wei Hu[1] | Fengcheng Li[4] | Haibin Dai[1] | Feng Zhu[1,2]

[1]Department of Pharmacy, The Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China | [2]College of Pharmaceutical Sciences, State Key Laboratory of Advanced Drug Delivery and Release Systems, Zhejiang University, Hangzhou, China | [3]School of Pharmacy, Hebei Medical University, Shijiazhuang, China | [4]Children's Hospital, Zhejiang University School of Medicine, National Clinical Research Center for Child Health, Hangzhou, China

Correspondence: Fengcheng Li (lifengcheng@zju.edu.cn) | Haibin Dai (haibindai@zju.edu.cn) | Feng Zhu (zhufeng@zju.edu.cn)

Received: 26 August 2025 | **Revised:** 7 December 2025 | **Accepted:** 14 December 2025

Keywords: binding site | drug modality | protein druggability | protein language model | transformer

ABSTRACT

Proteins interact with diverse molecular modalities, yet the incomplete identification of their binding sites has left the proteome-wide druggability largely underexplored. Although various computational methods have been developed for the prediction of protein binding sites, existing approaches are limited by their specificity to a single drug modality, dependence on high-quality structural data, or insufficient predictive accuracy. Here, a unified sequence-based framework, ALLSites, is constructed to identify proteome-wide binding sites across all drug modalities. Leveraging ESM-2 embeddings, ALLSites integrates a gated convolutional network with a transformer architecture to capture both global and local sequence features, effectively modeling residue interactions directly from sequence. This design bridges the gap between sequence-based and structure-based approaches, enabling ALLSites to achieve superior predictive performance across diverse drug modalities, including proteins, peptides, small molecules, carbohydrates, DNA, and RNA. It achieves state-of-the-art performance among sequence-based methods and matches the accuracy of the best structure-based tools. By enabling accurate and structure-free binding site prediction across all drug modalities, ALLSites is expected to expand the druggable proteome and provide a powerful resource for drug discovery.

1 | Introduction

Proteins play fundamental roles in cellular processes by interacting with a variety of molecular modalities [1–3]. Nevertheless, the druggability of proteins remains largely underexplored due to limited ligand-modulated proteins and incomplete mechanistic understanding, as evidenced by small molecules' ability to modulate less than 15% of the human proteome despite being the most prevalent drug modality [4–7]. Researchers then began to explore alternative drug modalities, including protein-, peptide-, nucleic acid-, and carbohydrate-based therapeutics, to modulate protein functions [8–12]. Therefore, the comprehensive identification of binding sites for all drug modalities is of vital importance, as it can greatly expand proteome druggability by redefining "undruggable" proteins under one modality as "druggable" under another [13–15]. The diversity and complexity of drug modalities pose great challenges in the experimental identification of protein binding sites [16–18]. In response, substantial efforts have been devoted to developing computational methods for predicting protein binding sites across various drug modalities [19–22].

Minjie Mou, Mingkun Lu and Zhimeng Zhou contributed equally to this work.

Computational methods generally fall into two categories according to their input requirements, namely, structure-based and sequence-based ones [23]. For example, methods such as DeepPPISP [24], DELPHI [25], and EnsemPPIS [26] utilize either protein structural or sequence information to identify protein-protein interaction (PPI) sites. Tools like PepBind [27] and PepSite [28] are specifically designed to predict peptide-protein interaction (PepPI) sites. These tools facilitate the design of protein- and peptide-based therapeutics. Several tools have been developed for predicting small-molecule-protein interaction (SMPI) sites. For example, P2Rank employs machine learning to identify potential small molecule binding pockets [29], while methods such as CAPSIF: V are tailored for predicting carbohydrate-protein interaction (CarbPI) sites [30]. In addition, various nucleic acid-binding site prediction tools have been developed to accelerate the understanding of biological processes and facilitate nucleic acid-based drug design [31]. Examples include DNAPred [32] and DNABind [33] for predicting DNA-protein interaction (DPI) sites, and NucleicNet for identifying RNA-protein interaction (RPI) sites [34].

However, existing approaches still face challenges in accurately identifying and distinguishing binding sites across all drug modalities at the proteome-wide scale. First, existing methods face issues of high dependency on precise structure or low predictive accuracy [26]. Specifically, the application of structure-based methods is limited by the low availability of high-resolution structures and high sensitivity to structural errors [35, 36]. In particular, many undruggable proteins lack experimentally determined structures, and the use of predicted structures may reduce the accuracy of structure-based methods [37, 38]. By contrast, sequence-based methods offer broader applicability but suffer from suboptimal predictive performance due to the neglect of residue interaction information [39]. Second, most existing methods are designed for modality-specific binding site prediction, and there remains no universal method capable of predicting binding sites for all drug modalities [40]. Currently, only a few methods support the prediction of binding sites for multiple modalities. For example, GraphBind is tailored for nucleic acid (DNA/RNA) binding sites [41], while PepBCL is effective only for peptides and performs poorly on DNA or RNA [42]. Overall, the insufficient proteome/modality coverage and poor performance of existing methods hinder their practical application. Therefore, it is highly demanded to develop a method capable of accurately predicting proteome-wide binding sites for all drug modalities.

In this study, considering that the key features of binding sites across various drug modalities is inherently encoded within the protein sequence, a unified sequence-based framework, named ALLSites, is constructed to identify proteome-wide binding sites for all drug modalities. Built upon the protein language model (PLM) ESM-2 for sequence embedding, ALLSites integrates a gated convolutional network (GatedCNN) with a transformer architecture to jointly learn the global sequence features and local contextual patterns. Moreover, it can model complex residue interactions directly from sequence data, thereby bridging the gap between sequence-based and structure-based methods. ALLSites achieves accurate protein binding site prediction for various drug modalities, including proteins, peptides, small molecules, carbohydrates, DNA, and RNA, while maintaining broad applicability. It exhibits state-of-the-art performance on these drug modalities, outperforming all sequence-based methods and achieving results comparable to the best structure-based method. The balance between accuracy and applicability makes ALLSites a valuable resource for advancing the understanding of proteome-wide druggability and accelerating the translation of various molecular modalities into clinical applications.

2 | Results and Discussion

2.1 | The Framework of ALLSites for Predicting Binding Sites of All Drug Modalities

To accurately identify protein binding sites for all drug modalities, a unified deep learning framework named ALLSites is designed based on the transformer architecture. The overall framework of ALLSites is illustrated in Figure 1a. First, the model takes the protein sequence as input and employs the powerful PLM, ESM-2, to generate residue-level embeddings. Next, these embeddings are fed into an encoder that incorporates a GatedCNN, as shown in Figure 1b. The function of the encoder is to extract local contextual patterns for each residue and integrate them to form the global sequence feature of the protein. Subsequently, the original residue embedding and the global sequence feature extracted by the encoder are passed into a modified transformer decoder, as shown in Figure 1c. This decoder incorporates a cross-attention mechanism and adapts the mask operation of the original transformer to ensure learning across the full length of the protein. The multi-head cross-attention mechanism enables ALLSites to capture interactions between each residue and other residues throughout the sequence. Finally, the embedding output by the decoder is fed into a classifier composed of fully connected layers (FCs), which predicts the probability of each residue being a binding site for different drug modalities. The ability of ALLSites to learn diverse residue features directly from protein sequence allows it to serve as a unified framework for identifying proteome-wide binding sites across all drug modalities, including proteins, peptides, small molecules, carbohydrates, DNA, and RNA.

2.2 | Evaluation of ALLSites in Identifying Binding Sites of Proteins/Peptides

Identifying PPI sites and PepPI sites is essential for the development of biologics such as protein- and peptide-based therapeutics [43, 44]. To comprehensively evaluate the performance of ALLSites in identifying binding sites of proteins and peptides, we employed four commonly used benchmark datasets for PPI sites and two benchmark datasets for PepPI sites.

2.2.1 | Performance Evaluation of ALLSites in Predicting PPI Sites

We first compared the performance of ALLSites and 13 competing methods on the PPI-Test70 dataset. The competing methods included four structure-based methods (IntPred [45], SPPIDER [46], DeepPPISP [24], and EGRET [39]) and nine sequence-based methods (ISIS [47], RF_PPI [48], PSIVER [49], SPRINGS [50],

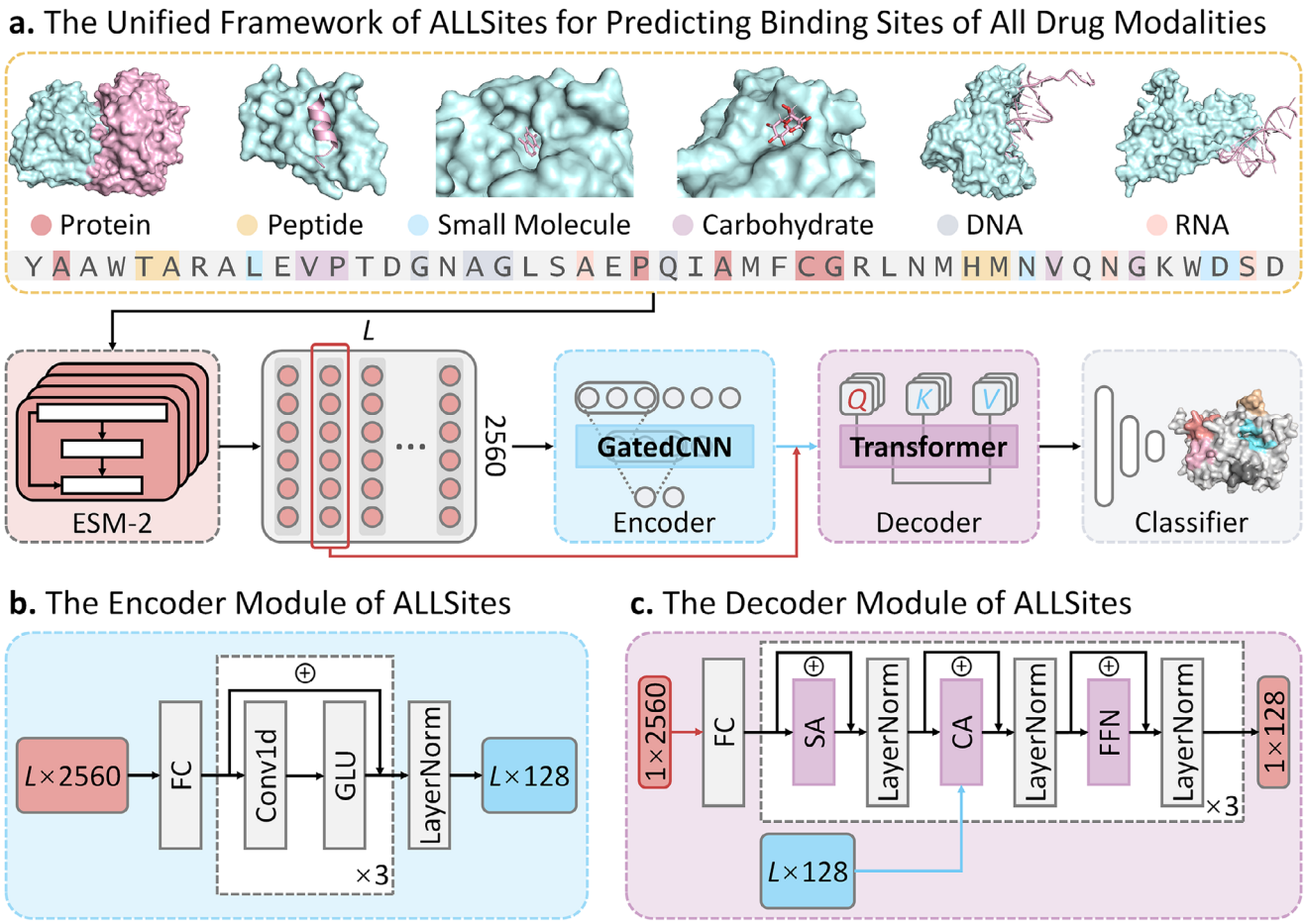

FIGURE 1 | The framework of ALLSites for predicting protein binding sites of all drug modalities. a). The overall framework of ALLSites. Using only the protein sequence as input, ALLSites enables binding site prediction for multiple drug modalities, including proteins, peptides, small molecules, carbohydrates, DNA, and RNA. The model architecture comprises three main modules, namely a protein feature encoder for processing ESM-2-derived representations, a cross-attention decoder, and a classification module. b). The architecture of the encoder module. The encoder is essentially a gated convolutional network (GatedCNN) designed to extract both local and global features from protein sequences. It consists of 1D convolutional layers (Conv1D) and the gated linear unit (GLU) activation function. c). The architecture of the decoder module. The decoder takes residue-level representation and global protein feature as input. It employs a cross-attention mechanism to learn residue interaction information. Both the encoder and decoder consist of three layers. FC, fully connected layers; SA, self-attention; CA, cross-attention; FFN, position-wise feed-forward network.

ProNA2020 [51], SCRIBER [52], DELPHI [25], DLPred [53], and EnsemPPIS [26]). Except for SPPIDER, ProNA2020, SCRIBER, and DLPred results, which were obtained using the web server, the results of the other methods were obtained by reproducing the source code or directly collected from the DeepPPISP literature [24], as they used the same PPI-Train352 as the training dataset. As shown in Table 1, ALLSites consistently performed best in terms of AUROC, AUPRC, F1, and MCC metrics, surpassing all sequence-based and structure-based methods. Particularly, compared to the second-best method, EnsemPPIS, ALLSites achieved improvements of 5.0% in AUROC and 8.1% in AUPRC. In terms of the F1 and MCC, ALLSites improved by 0.034 and 0.042, respectively. In another widely used PPI site prediction task (using PPI-Train9982 as the training set and PPI-Test355 as the test set), structural information was unavailable in the training data. Thus, the comparison was conducted exclusively among sequence-based methods. As shown in Table 2, ALLSites still exhibited superior performance compared to all sequence-based methods, achieving the highest scores in all metrics. ALLSites was 6.9% and 24.6% higher than EnsemPPIS in terms of AUROC and AUPRC, respectively, and was 0.081 and 0.096 higher in terms of F1 and MCC metrics, respectively.

In addition, ALLSites was evaluated on the PPI-Test60 and PPI-Test315 datasets. The model evaluated on these two test sets was the same model, which was trained on the PPI-Train335 dataset using the same training scheme. As presented in Figure 2a, in terms of the key metrics AUROC and MCC, ALLSites consistently outperformed all sequence-based methods by a large margin and achieved performance comparable to the best structure-based method, RGN [54]. By contrast, the second-best sequence-based method, EnsemPPIS, consistently underperformed compared to the structure-based methods GraphPPIS and RGN in terms of MCC. It was worth noting that, on the PPI-Test60 dataset, ALLSites showed slightly better performance than MaSIF-site [55], which used advanced geometric deep learning to learn surface features from protein structures. We also performed a calibration analysis of ALLSites on these two test datasets. As illustrated in Figure S1, the reliability diagram demonstrated well-calibrated predictions. Specifically, ALLSites achieved Brier scores of 0.153

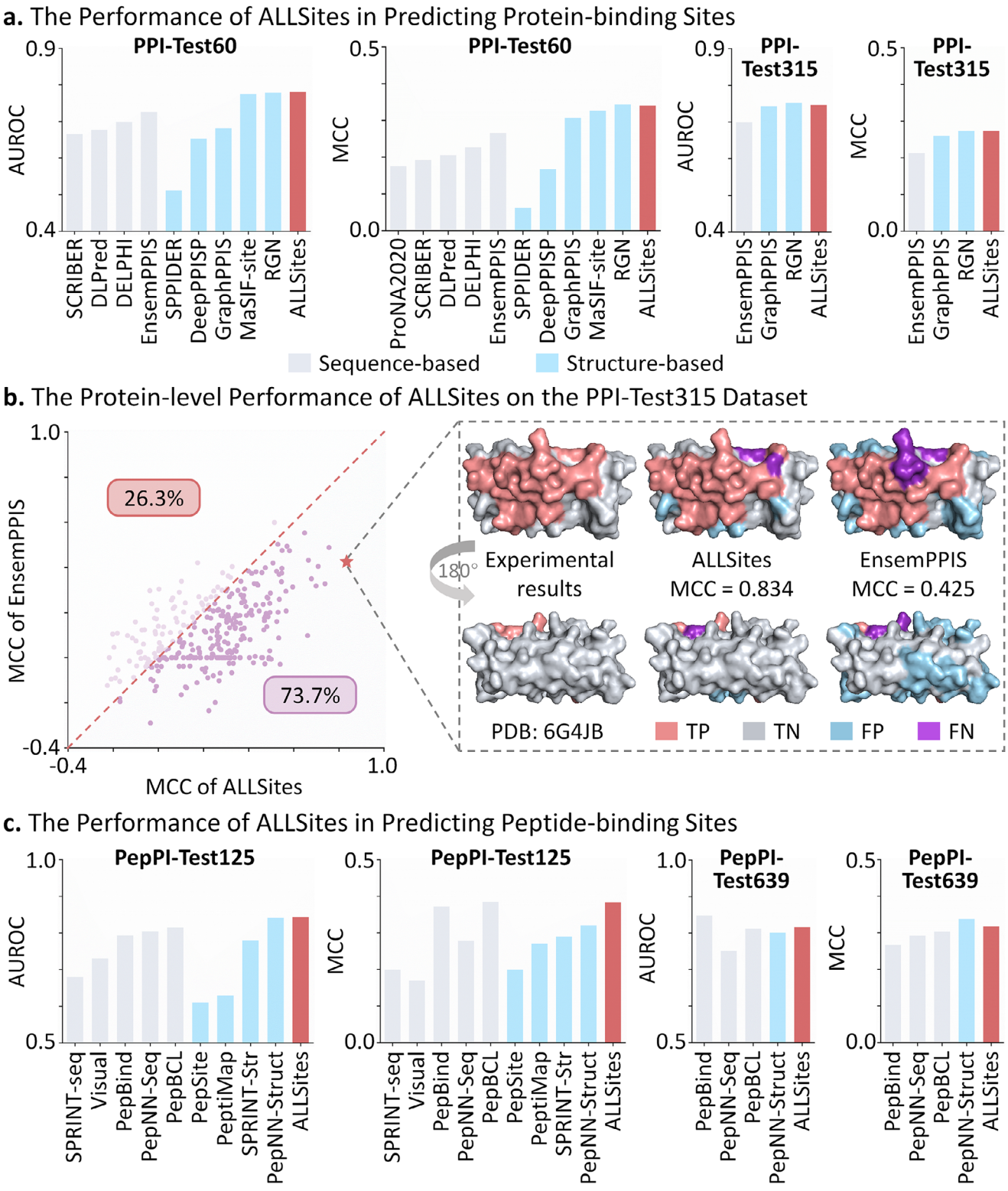

FIGURE 2 | The performance of ALLSites in identifying binding sites of proteins and peptides. a). The performance of ALLSites in predicting protein-binding sites. ALLSites is evaluated on the PPI-Test60 and PPI-Test315 datasets in terms of AUROC and MCC. Sequence-based methods are indicated by gray bars, and structure-based methods are indicated by blue bars. b). The protein-level performance of ALLSites on the PPI-Test315 dataset. The MCC metric is calculated for each protein based on ALLSites' predictions. A specific protein (PDB ID: 6G4JB) is presented to illustrate the predictions of ALLSites and EnsemPPIS alongside the corresponding experimental results. c). The performance of ALLSites in predicting peptide-binding sites. ALLSites is evaluated on the PepPI-Test125 and PepPI-Test639 datasets in terms of AUROC and MCC. Source data are provided in Tables S2 to S5.

TABLE 1 | Performance evaluation of ALLSites on the PPI-Test70 dataset. The best performance for each metric is highlighted in bold, and the second-best performance is underlined.

Class	Method	ACC	AUROC	AUPRC	F1	MCC
Structure-based	IntPred [a]	0.672	–	–	0.332	0.165
	SPPIDER [c]	0.667	0.518	0.235	0.273	0.063
	DeepPPISP [b]	0.655	0.671	0.320	0.397	0.206
	EGRET [b]	0.715	<u>0.719</u>	<u>0.405</u>	0.438	0.270
Sequence-based	ISIS [a]	0.622	–	0.240	0.267	0.097
	RF_PPI [a]	0.598	–	0.210	0.258	0.118
	PSIVER [a]	0.653	–	0.250	0.328	0.138
	SPRINGS [b]	0.631	–	0.280	0.350	0.181
	ProNA2020 [c]	**0.741**	–	–	0.258	0.106
	SCRIBER [c]	0.616	0.635	0.307	0.370	0.159
	DELPHI [b]	0.667	0.690	0.360	0.418	0.236
	DLPred [c]	0.680	0.697	0.380	0.416	0.235
	EnsemPPIS [b]	<u>0.732</u>	<u>0.719</u>	<u>0.405</u>	<u>0.440</u>	<u>0.277</u>
	ALLSites	0.720	**0.755**	**0.438**	**0.474**	**0.319**

[a]Results reported by DeepPPISP.
[b]Results generated by reproducing the source code.
[c]Results obtained by using the web server. ProNA2020 only makes binary predictions, and its AUROC and AUPRC are not calculated.

TABLE 2 | Performance evaluation of ALLSites on the PPI-Test355 dataset. All the comparison methods use only protein sequences. The best performance for each metric is highlighted in bold, and the second-best performance is underlined.

Method	ACC	AUROC	AUPRC	F1	MCC
SPRINGS [a]	0.811	0.608	0.178	0.211	0.103
DLPred [b]	0.835	0.724	0.272	0.308	0.214
SCRIBER [b]	0.838	0.719	0.275	0.322	0.230
DELPHI [a]	<u>0.848</u>	0.746	0.326	0.364	0.278
EnsemPPIS [a]	0.821	<u>0.770</u>	<u>0.354</u>	<u>0.385</u>	<u>0.291</u>
ALLSites	**0.850**	**0.823**	**0.441**	**0.466**	**0.387**

[a]Results generated by reproducing the source code.
[b]Results obtained by using the web server.

and 0.163 on PPI-Test60 and PPI-Test315, respectively. Both Brier scores were below the commonly accepted threshold of 0.25, suggesting that their predicted probabilities were reasonably reliable [56].

In fact, the protein representation module in ALLSites could leverage newer PLMs beyond ESM-2. A notable example is ESM-C, a parallel model family specifically designed to generate biologically meaningful protein representations alongside the ESM3 generative models [57]. Therefore, it is necessary to compare the performance of ALLSites when using either ESM-2 (ALLSites-ESM2) or ESM-C (ALLSites-ESMC) for protein representation. Due to hardware constraints, we selected ESM-C 600M to represent protein sequences and evaluated ALLSites on PPI-Test70, PPI-Test315, and PPI-Test60. It generated a 1,152-dimensional embedding for each protein residue. As shown in Table S1, replacing ESM-2 with ESM-C 600M led to a slight improvement in AUROC and AUPRC on PPI-Test70, but resulted in decreased performance across other metrics, including ACC, F1, and MCC. Notably, on both PPI-Test60 and PPI-Test315, the use of ESM-C caused a consistent and substantial drop in nearly all evaluation metrics, particularly MCC, which decreased by 0.037 and 0.032, respectively. These findings indicated that, for the task of PPI site prediction, adopting the newer ESM-C did not further enhance the predictive performance of ALLSites. This conclusion aligned with a recent large-scale benchmarking study of PLMs [58], reinforcing that the choice of ESM-2 as the protein representation backbone in ALLSites was well justified.

Furthermore, we additionally assessed the protein-level performance of ALLSites on the PPI-Test315 and PPI-Test60 datasets. As shown in Figure 2b, compared with the current best sequence-based method EnsemPPIS, ALLSites achieved higher MCC scores on 73.7% of the 315 proteins in the PPI-Test315 dataset. For a selected case protein (PDB ID: 6G4JB), ALLSites produced fewer false positives and false negatives, resulting in an MCC of 0.834, whereas EnsemPPIS achieved an MCC of only 0.425. Similar results were observed on the PPI-Test60 dataset. As shown in Figure S2a, ALLSites outperformed EnsemPPIS in terms of MCC on 81.7% of the proteins. For the case protein (PDB ID: 4EMJB), EnsemPPIS predicted a large number of false positive PPI sites, leading to an MCC less than half of that achieved by ALLSites. However, in the PPI-Test60 test set, ALLSites still yielded lower MCC values than EnsemPPIS for 18.3% of the proteins. Taking a case protein as an example (PDB ID: 4HLUA), the MCC value of ALLSites' prediction was only 0.095, compared to 0.333 for

EnsemPPIS. Although ALLSites produced fewer false-positive PPI sites, its extremely low number of true-positive predictions resulted in a substantially lower MCC value.

In terms of the AUROC metric, ALLSites also achieved higher AUROC values than EnsemPPIS on the majority of proteins, as shown in Figure S2b. In summary, these results indicated that ALLSites achieved SOTA predictive performance in PPI site prediction at both the residue and protein levels.

2.2.2 | Performance Evaluation of ALLSites in Predicting PepPI Sites

The performance evaluation of ALLSites in PepPI site prediction was conducted on two tasks. The first task involved training on PepPI-Train1154 and testing on PepPI-Test125, while the second task used PepPI-Train640 for training and PepPI-Test639 for testing. A total of nine PepPI site prediction methods were compared with ALLSites, including five sequence-based methods (SPRINT-seq [59], Visual [60], PepBind [27], PepNN-Seq [61], and PepBCL [42]) and four structure-based methods (PepSite [28], PeptiMap [62], SPRINT-Str [63], and PepNN-Struct [61]). Since the same training dataset and model training strategy were employed, the performance metrics of competing methods were obtained from the PepBCL literature [42].

As shown in Figure 2c, ALLSites exhibited excellent predictive performance in both tasks. Specifically, on the PepPI-Test125 dataset, ALLSites outperformed all sequence-based (gray bars) and structure-based (blue bars) methods in both AUROC and MCC metrics. In terms of AUROC, ALLSites surpassed all sequence-based methods and slightly exceeded the best structure-based method, PepNN-Struct. Regarding the MCC metric, ALLSites outperformed all structure-based methods, with only a 0.002 margin below the best-performing method, PepBCL. On PepPI-Test639, PepBind and PepNN-Struct achieved the highest scores in AUROC and MCC, respectively, while ALLSites consistently ranked second, highlighting its strong robustness. These findings demonstrated that ALLSites was capable of accurately identifying PepPI sites on proteins, offering a valuable tool for advancing peptide drug discovery and design.

2.3 | Evaluation of ALLSites in Identifying Binding Sites of Small Molecules/Carbohydrates

2.3.1 | Performance Evaluation of ALLSites in Predicting SMPI Sites

Small molecules represent the most prevalent molecular modality among approved therapeutics, making the identification of potential SMPI sites on proteins crucial for the development of novel small-molecule drugs [64]. To evaluate the performance of ALLSites in predicting SMPI sites, a new benchmark dataset of SMPI sites was constructed based on the sc-PDB database [65], which comprised 2 324 non-redundant proteins with low pairwise sequence identity. These proteins were randomly partitioned into a training set (SMPI-Train1628), a validation set (SMPI-Valid348), and a test set (SMPI-Test348). We compared the performance of ALLSites and P2Rank [29]. (a widely used structure-based method) on the SMPI-Test348 dataset. Following the instructions described in the original publication, the performance of P2Rank was calculated using the top-ranked predicted pocket. It should be noted that P2Rank's predictive performance was obtained by directly loading its pre-trained model parameters for inference. Consequently, the comparison with ALLSites may be inherently unfair due to potential differences in the training data used. Nevertheless, from a practical application standpoint, evaluating ALLSites against the pre-trained P2Rank remains meaningful. As illustrated in the left panel of Figure 3a, ALLSites outperformed P2Rank across all five metrics, including accuracy, recall, precision, F1, and MCC. Particularly, ALLSites achieved F1, MCC, and recall scores of 0.601, 0.560, and 0.593, respectively, achieving improvements of 0.151, 0.136, and 0.232 over P2Rank.

Similarly, we assessed the protein-level performance of ALLSites on the SMPI-Test348 dataset. As illustrated in Figure 3b, among the 348 unique proteins, ALLSites achieved higher MCC values than P2Rank for 66.4% of the cases. As demonstrated by two representative proteins (PDB ID: 2QA1A and 3OUMA), ALLSites achieved MCC scores of 0.755 and 0.717, markedly surpassing those of P2Rank. Obviously, the binding sites predicted by ALLSites were much closer to the true binding pockets, whereas P2Rank exhibited a substantially higher number of false negative predictions (purple residues). Notably, ALLSites achieved superior performance compared to the well-established P2Rank algorithm without using any structural information, highlighting its powerful capability in extracting binding site features from sequence alone.

2.3.2 | Performance Evaluation of ALLSites in Predicting CarbPI Sites

Research on the interactions between carbohydrates and proteins has already led to several approved drugs [66–68]. A considerable portion of carbohydrate molecules falls within the category of small molecules. However, the distinctive chemical properties of carbohydrates, especially their rich hydroxyl groups, generate binding sites that are fundamentally different from those of typical small molecules [69, 70]. Therefore, it is essential to evaluate the performance of ALLSites in identifying CarbPI sites.

The benchmark dataset used to assess ALLSites' ability in predicting CarbPI sites was derived from a previous study and consisted of a training set (CarbPI-Train517), a validation set (CarbPI-Valid129), and a test set (CarbPI-Test162) [30]. Three structure-based methods were employed for comparison with ALLSites, including one general small-molecule binding site prediction tool (FTMap [71]) and two carbohydrate-specific binding site prediction tools (CAPSIF: V^{30} and CAPSIF: G^{30}). Since the same training and evaluation protocols were applied, the performance metrics of these competing methods were directly obtained from the original literature [30]. To ensure consistency and comparability, we adopted the same evaluation metrics as the original study, calculating the average metric values across all proteins in the test set. As shown in the right panel of Figure 3a, ALLSites outperformed all three methods in terms of average DICE, precision, recall, and MCC. The performance of ALLSites was substantially higher than that of FTMap, with average DICE

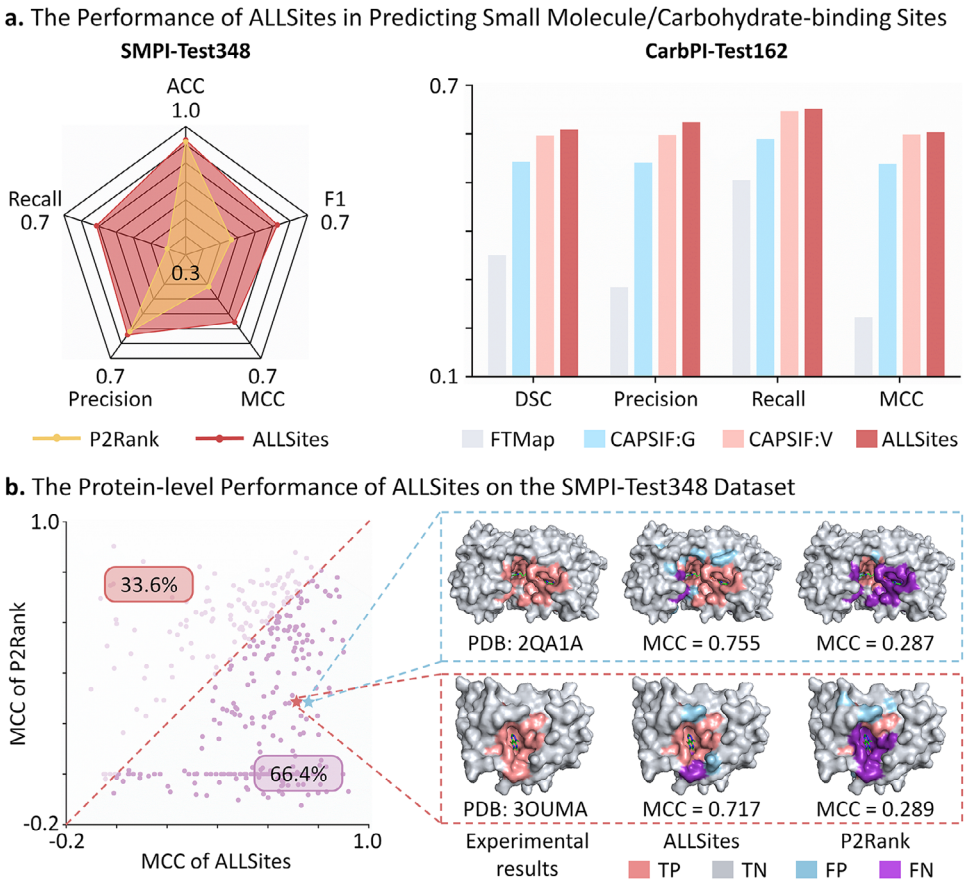

FIGURE 3 | The performance of ALLSites in identifying binding sites of small molecules and carbohydrates. a). The residue-level performance of ALLSites in predicting SMPI sites and CarbPI sites. ALLSites is evaluated on the SMPI-Test348 and CarbPI-Test162 datasets. All the compared methods are based on structural information. Source data are provided in Tables S6 and S7. b). The protein-level performance of ALLSites on the SMPI-Test348 dataset. The MCC metric is calculated for each protein based on ALLSites' predictions. Two specific proteins (PDB ID: 2QA1A and PDB ID: 3OUMA) are presented to illustrate the predictions of ALLSites and P2Rank alongside the corresponding experimental results.

and MCC values improved by 0.258 and 0.381, respectively. The inferior performance of FTMap can be attributed to its design for general small-molecule pocket prediction, as it is not specifically tailored for identifying CarbPI sites. These results further confirm the fact that carbohydrate-binding sites are distinctly different from conventional small-molecule binding sites. Moreover, ALLSites is highly portable and can be adapted to other small-molecule drug modalities, such as the prediction of covalent binding sites [72].

2.4 | Evaluation of ALLSites in Identifying Binding Sites of Nucleic Acids

2.4.1 | Performance Evaluation of ALLSites in Predicting DPI and RPI Sites

Nucleic acid-protein interactions play crucial roles in numerous essential cellular processes, such as DNA replication, transcription, and translation [73, 74]. Elucidating the molecular mechanisms underlying these interactions, including the characterization of DNA- and RNA-binding sites on proteins, can facilitate the development of drugs to treat diseases caused by aberrant regulation between proteins and nucleic acids [75–77]. To assess ALLSites' ability in identifying nucleic acid binding sites, we utilized two benchmark datasets for DPI sites and one for RPI sites.

In the task for DPI site prediction, three sequence-based methods (SVMnuc [78], NCBRPred [79], and DNAPred [32]) and four structure-based methods (COACH-D [80], NucBind [78], DNABind [33], and GraphBind [41]) were selected for performance comparison with ALLSites. The results for GraphBind were obtained by reproducing the source code, while the performances of the other methods were retrieved using their respective web servers. A single model trained on the DPI-Train573 dataset was applied to evaluate prediction performance on both the DPI-Test129 and DPI-Test181 test sets. As shown in Figure 4a, ALLSites significantly outperformed all sequence-based methods and the majority of structure-based methods on both AUROC and MCC metrics across the two test sets. Compared to the best-performing sequence-based method, ALLSites achieved improvements of 9.7% and 12.6% in AUROC on DPI-Test129 and DPI-Test181, respectively, along with MCC improvements of 0.148 and 0.150. Although ALLSites showed slightly lower MCC scores than the best-performing structure-based method, GraphBind, it achieved comparable AUROC performance on both test sets.

In the benchmarking task for RPI site prediction, the performance of ALLSites was compared with two sequence-based meth-

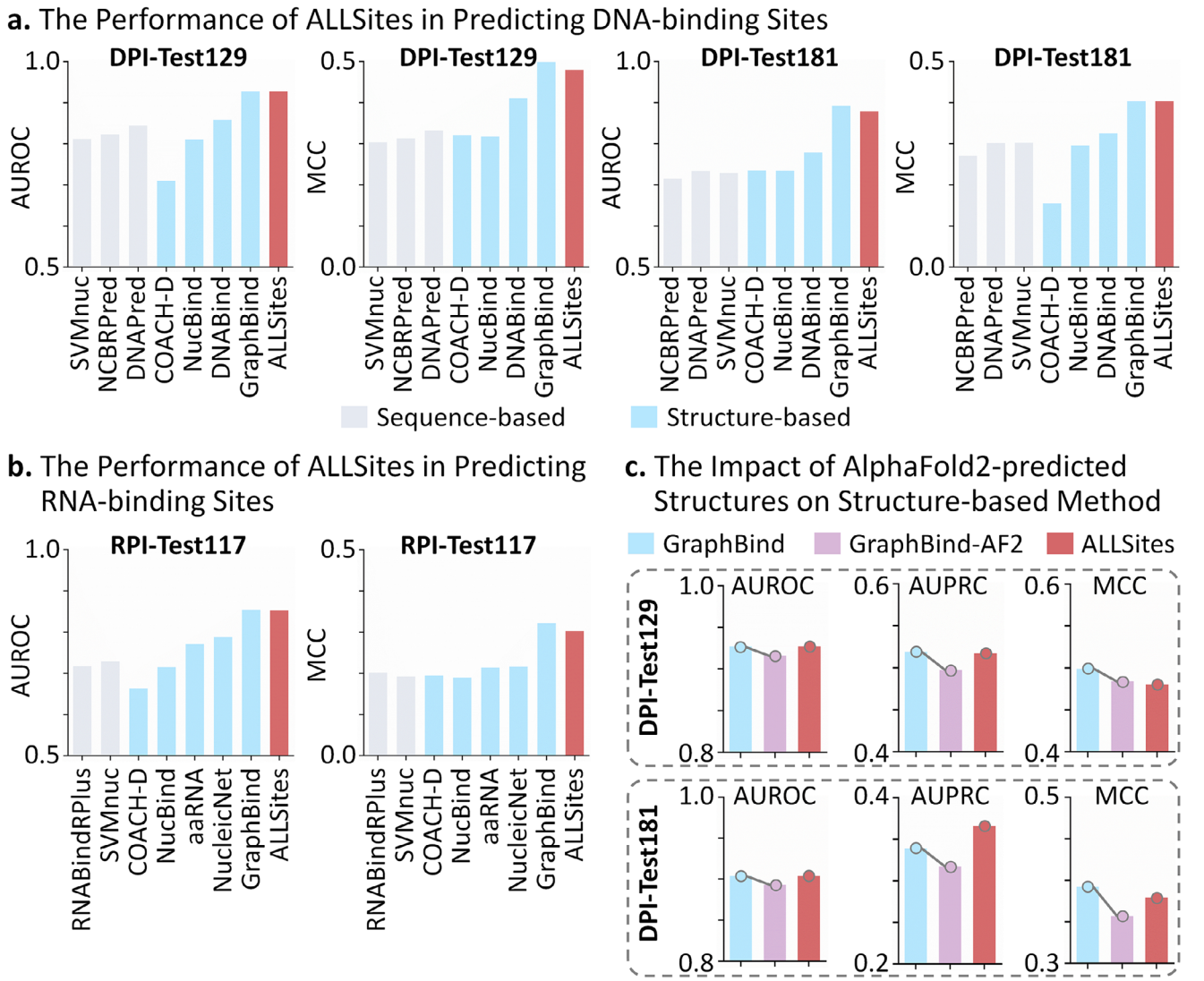

FIGURE 4 | The performance of ALLSites in identifying binding sites of nucleic acids. a). The performance of ALLSites in predicting DNA-binding sites. ALLSites is evaluated on the DPI-Test129 and DPI-Test181 datasets in terms of AUROC and MCC. Sequence-based methods are indicated by gray bars, and structure-based methods are indicated by blue bars. b). The performance of ALLSites in predicting RNA-binding sites. ALLSites is evaluated on the RPI-Test117 dataset in terms of AUROC and MCC. c). The impact of AlphaFold2-predicted structures on the structure-based method. The impact of predicted structures on GraphBind's performance was evaluated using the DPI-Test129 and DPI-Test181 datasets. Blue and purple bars represent the performance of GraphBind using experimentally determined structures and predicted structures (GraphBind-AF2), respectively. Source data are provided in Tables S8 to S14.

ods (RNABindRPlus [81] and SVMnuc [78]) and five structure-based methods (COACH-D [80], NucBind [78], aaRNA [82], NucleicNet [34], and GraphBind [41]). The results for NucleicNet and GraphBind were obtained by reproducing the provided source code, while the performance of the other methods was obtained using their web servers. As shown in Figure 4b, consistent with the results observed in the DPI site prediction, ALLSites outperformed all sequence-based methods and the majority of structure-based approaches. Specifically, ALLSites ranked second in both AUROC and MCC metrics, with AUROC values comparable to those of the best-performing structure-based method, GraphBind. These findings provide strong evidence that ALLSites can accurately identify nucleic acid binding sites solely from protein sequences.

2.4.2 | Performance Evaluation of Structure-Based Method Using Predicted Structures

Although ALLSites consistently shows slightly lower performance than the best-performing structure-based method, GraphBind, in the tasks for both DPI and RPI site prediction, it offers broader applicability across the proteome due to its reliance solely on sequence information. This advantage is significant because structure-based methods are limited by the low availability of high-resolution structures and their high sensitivity to structural errors. Specifically, only about 35% of human proteins have experimentally determined crystal structures, and in many cases, these structures cover only a fragment of the full sequence [35]. Moreover, although advanced protein structure prediction tools

(such as AlphaFold2 [83] and RoseTTaFold [84]) can partially alleviate the scarcity of structural data, the inherent deviations between predicted and native structures often substantially degrade the performance of structure-based prediction methods.

To elucidate the limitations of structure-based methods, we evaluated the impact of AlphaFold2-predicted structures on GraphBind using the DPI-Test129 and DPI-Test181 datasets. As shown in Figure 4c, GraphBind exhibited a clear performance decline across all three metrics (AUROC, AUPRC, and MCC) when predicted structures were used instead of experimentally resolved ones. On the DPI-Test129 dataset, GraphBind's AUROC and AUPRC dropped below those of ALLSites, with MCC values becoming comparable to ALLSites. On DPI-Test181, the use of predicted structures led to GraphBind underperforming ALLSites across all three metrics. These results confirm that structure-based methods are highly sensitive to structural errors, and even state-of-the-art structure prediction tools cannot fully compensate for the limitations imposed by the scarcity of experimentally resolved structures. Source data for all benchmark results were provided in Tables S2–S14.

ALLSites exhibits a very fast inference speed. On a single NVIDIA V100 GPU, it can screen the entire human proteome (comprising 20 420 reviewed human proteins from the UniProt database) within 16 h. On average, it takes 2.81 s per protein and only 0.0075 s per residue. In conclusion, given its reliance solely on protein sequence and its fast inference speed, ALLSites is well-suited for proteome-wide mapping of binding sites across all drug modalities.

3 | Materials and Methods

3.1 | Dataset Collection and Data Processing

3.1.1 | Benchmark Datasets for PPI Site and PepPI Site Prediction

In this study, four commonly used benchmark datasets for PPI site prediction and two benchmark datasets for PepPI site prediction were used to assess the performance of ALLSites in identifying binding sites of proteins and peptides.

The first PPI site benchmark (PPI-Train352 and PPI-Test70) was sourced from DeepPPISP [24], originally curated from the PDB [85] through a six-step data filtering process [49]. Both datasets comprised proteins with less than 25% sequence homology, ensuring low redundancy in model training and evaluation. A surface residue was annotated as a PPI site if its absolute solvent accessibility decreased by at least 1.0 $Å^2$ upon protein binding [86]. A subset of 50 proteins was randomly selected from PPI-Train352 to form a hold-out validation set. The second PPI site benchmark (PPI-Train9982 and PPI-Test355) was collected by DELPHI [25]. The PPI-Test355 dataset was built based on the BioLip database [87] and comprised 355 non-redundant proteins with pairwise sequence similarity below 25%. Residues were annotated as PPI sites if the distance between any two atoms from different chains was less than 0.5 Å plus the sum of their van der Waals radii. The PPI-Train9982 dataset was collected from a previous study [88], with all proteins exhibiting less than 25% sequence similarity to those in the PPI-Test355 set. A total of 1 110 sequences were randomly selected from PPI-Train9982 to form a validation set, while the remaining sequences were used for model training. The PPI-Train9982 dataset lacks structural annotations and is therefore unsuitable for training structure-based PPI site prediction methods. The third PPI site benchmark (PPI-Train335 and PPI-Test60) was constructed by GraphPPIS [36]. The sequences in both datasets also exhibited less than 25% sequence similarity. To ensure a fair comparison, the PPI-Train335 and PPI-Test60 datasets were identical to those employed in the GraphPPIS study. Furthermore, the model trained on PPI-Train335 was also evaluated on the PPI-Test315 dataset [36], which was a previously published dataset comprising proteins with less than 25% sequence identity with those in PPI-Train335.

The two PepPI benchmark datasets were directly adopted from a previous study [42]. The first benchmark (PepPI-Train1154 and PepPI-Test125) was originally introduced in the SPRINT-Str study [63]. The second benchmark (PepPI-Train640 and PepPI-Test639) was derived from a previous work as well [27]. Both benchmarks underwent a similar preprocessing pipeline, and sequence identity between the training and test sets was reduced to a maximum of 30% using the BLAST-Clust in the BLAST package to ensure reliable performance evaluation [89].

3.1.2 | Benchmark Datasets for SMPI Site and CarbPI Site Prediction

In this study, to evaluate the performance of ALLSites in identifying small-molecule binding sites, we constructed a new benchmark dataset for SMPI site prediction. First, 17 594 small-molecule-protein complex structures were downloaded from the sc-PDB database [65]. According to the sc-PDB definition, a protein residue was considered as a binding site if any of its atoms was located within 6.5 Å of a ligand atom. Following the protocol adopted in a previous study [90], binding site annotations from multiple PDB entries of the same protein were mapped to their corresponding UniProt sequences, resulting in 4,993 unique protein sequences. To avoid bias in performance evaluation, the sequence identity was reduced to 30% using the BLASTClust algorithm. This yielded a non-redundant set of 2 324 proteins. From this set, 348 proteins were randomly selected to form the independent test dataset (SMPI-Test348), while the remaining 1 976 proteins constituted the training set. Furthermore, an additional 348 proteins were randomly selected from SMPI-Train1976 to form a validation set (SMPI-Valid348), leaving 1 628 proteins as the final training set (SMPI-Train1628) for model training.

The benchmark dataset for CarbPI site prediction was obtained from the CAPSIF study [30]. It comprises 517 proteins for training (Carb-Train517), 129 for validation (Carb-Valid129), and 162 for an independent test (Carb-Test162). All protein structures in this benchmark had a resolution lower than 3.0 Å, and sequence identity between any two proteins was below 30%. A residue was defined as a CarbPI site if any of its heavy atoms were within 4.2 Å of a heavy atom in the bound carbohydrate.

3.1.3 | Benchmark Datasets for DPI Site and RPI Site Prediction

Two benchmark datasets for DPI site prediction and one benchmark dataset for RPI site were collected to evaluate the performance of ALLSites in identifying binding sites of nucleic acids.

The first benchmark dataset for DPI site prediction (DPI-Train573 and DPI-Test129) was adopted from the GraphBind study [41]. The DNA-binding proteins were initially curated from the BioLip database and processed through a series of filtering steps [87], resulting in 573 proteins in the training set (DPI-Train573) and 129 in the independent test set (DPI-Test129). Sequence identity between proteins in the training and test sets was below 30%. A residue was defined as a DPI site if the shortest atomic distance between it and the DNA molecule was less than 0.5 Å plus the sum of the van der Waals radii of the two closest atoms. Another independent test dataset for DPI site prediction, DPI-Test181, was collected from a previous study [91]. The proteins in DPI-Test181 shared less than 30% sequence identity with those in the DPI-Train573 training set, enabling an unbiased evaluation. Therefore, the model trained on DPI-Train573 was also assessed on DPI-Test181 to validate its generalizability.

The benchmark dataset for RPI site prediction (RPI-Train495 and RPI-Test117) was also obtained from the GraphBind study [41]. The data preprocessing pipeline and binding site definition criteria were kept consistent with those applied to the DPI-Train573 and DPI-Test129 datasets. Sequence identity between proteins in RPI-Train495 (495 proteins) and RPI-Test117 (117 proteins) was also below 30%, ensuring minimal sequence redundancy and a reliable evaluation.

Notably, class imbalance is prevalent across all benchmark datasets, where the number of non-binding sites exceeds that of binding sites. Dataset statistics for all benchmark tasks are summarized in Table S15, which includes protein counts, numbers of binding and non-binding residues, and the fraction of binding residues among all residues.

3.2 | Protein Representation

Protein sequences were represented using ESM-2, a transformer-based protein language model pre-trained on 65 million protein sequences with 3B parameters [92]. ESM-2 leverages large-scale self-supervised pre-training to extract semantic knowledge at a molecular level, enabling the inference of deep embeddings that align with biological semantics. This language model was selected for its ability to capture evolutionary information and complex structural patterns within protein sequences without requiring explicit structural data. For each protein sequence, residue-level features were extracted using ESM-2, with each amino acid represented as a fixed-dimensional vector of 2 560 features. This provided a rich, context-aware representation for every residue position, which was crucial for accurate binding site prediction.

3.3 | Model Architecture of ALLSites

ALLSites is a novel deep learning framework for protein binding site prediction, featuring an encoder-decoder architecture enhanced with a cross-attention mechanism. As illustrated in Figure 1a, its framework comprises three key components: (1) a protein feature encoder, (2) a cross-attention decoder, and (3) a classification module. By capturing both local and global protein features, as well as residue interaction features, ALLSites can identify potential binding sites across all drug modalities.

3.3.1 | Protein Feature Encoder

The protein feature encoder extracts meaningful representations from protein sequences. The structure of the encoder is depicted in Figure 1b. Initially, the encoder maps the input protein features through a fully connected layer to obtain a hidden representation of dimension d_{hid}. The mapped features are then processed through a series of 1D convolutional layers (Conv1D) with gated linear units (GLU) activation functions [93]. Specifically, the encoder contains n convolutional layers, each with kernel size k and padding $(k-1)/2$ to maintain the sequence length. Each convolutional layer produces an output of dimension $2 \times d_{hid}$, which is then processed through the GLU activation function to obtain an output of dimension d_{hid}.

To facilitate gradient flow through the network, we employed residual connections that combined the input and output of each convolutional layer with a scaling factor. Layer normalization is applied to the final output to stabilize the training process. The protein encoder's computation is defined by Equation (1).

$$\boldsymbol{H}_p = \text{Encoder}(\boldsymbol{P}) \tag{1}$$

where $\boldsymbol{P} \in \mathbb{R}^{B \times L_p \times d_p}$ $\boldsymbol{P} \in \mathbb{R}^{B \times L_p \times d_p}$ represents the protein features with batch size B, sequence length L_p, and input feature dimension d_p, and $\boldsymbol{H}_p \in \mathbb{R}^{B \times L_p \times d_{hid}}$ is the encoded protein representation.

3.3.2 | Cross-Attention Decoder

The decoder processes the encoded global protein features while also attending to the residue interactions. As shown in Figure 1c, the decoder architecture comprises multiple decoder layers, each featuring a cross-attention mechanism and a position-wise feed-forward network, both enhanced by residual connections and layer normalization [94]. In the decoder layer, the local protein features first undergo self-attention to capture internal relationships within the local sequence. The cross-attention mechanism enables the model to focus on residue interactions potentially involved in binding interactions. Subsequently, the position-wise feed-forward network, comprising two convolutional layers with a ReLU activation in between, enhances the model's representational capacity.

The cross-attention mechanism is a critical component that enables the model to capture long-range dependencies between

the potential binding sites and other protein residues. Our implementation follows the multi-head attention paradigm, where the attention is computed in parallel across multiple representation subspaces. For each attention head, we compute query ($\boldsymbol{Q}$), key ($\boldsymbol{K}$), and value ($\boldsymbol{V}$) from the input features. As shown in Equation (2), the attention scores are calculated as scaled dot products between query and keys, followed by softmax normalization.

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{2}$$

where d_k is the dimensionality of the key. The outputs from all attention heads are concatenated and projected to obtain the final attention output. The decoder can be expressed as Equation (3).

$$\boldsymbol{H}_l, \boldsymbol{A} = \text{Decoder}\left(\boldsymbol{L}, \boldsymbol{H}_p, \boldsymbol{M}_l, \boldsymbol{M}_p\right) \tag{3}$$

where $\boldsymbol{L} \in \mathbb{R}^{B\times L_l \times d_l}$ represents the local features of a specific residue, $\boldsymbol{M}_l$ and $\boldsymbol{M}_p$ are the mask matrices for local and global protein features, respectively, $\boldsymbol{H}_l \in \mathbb{R}^{B\times d_{hid}}$ is the processed local interaction representation, and $\boldsymbol{A}$ represents the attention weights.

3.3.3 | Classification Module

After obtaining the processed local interaction representation, we employ a significance-weighted aggregation strategy to combine information from all local features. The norm of each local feature representation is used to compute a softmax-normalized significance score, which is then used to weight its contribution to the final representation. $\boldsymbol{H}_{\text{agg}}$. The computation formula is shown in Equation (4).

$$\boldsymbol{H}_{\text{agg}} = \sum_{j=1}^{L_l} \frac{\exp(|\left|\boldsymbol{H}_l^j\right||_2)}{\sum_{k=1}^{L_l}\exp(|\left|\boldsymbol{H}_l^k\right||_2)} \cdot \boldsymbol{H}_l^j \tag{4}$$

The aggregated representation is passed through a series of FCs with the ReLU activation to predict the binding probability.

$$\boldsymbol{z} = \text{FC}_3\left(\text{FC}_2\left(\text{FC}_1\left(\boldsymbol{H}_{\text{agg}}\right)\right)\right) \tag{5}$$

where $\boldsymbol{z} \in \mathbb{R}^{B\times 2}$ is the logits for binary classification of binding site or non-binding site.

3.4 | Model Training and Implementation

3.4.1 | Training Procedure

ALLSites was trained on the binding site classification task using the weighted cross-entropy loss function. Since the number of binding residues in all prediction tasks was substantially greater than that of non-binding residues, assigning a higher loss weight to binding residues could enhance the model's ability to predict them accurately. In this study, the loss weight for each task was treated as a hyperparameter and optimized accordingly. The final loss weight values used for all tasks were provided in Table S15. The weight loss was set prior to model training and remained fixed throughout the training process. The computation formula of weighted cross-entropy loss was shown in Equation (6).

$$\text{Loss} = -\sum_{i=1}^{N} w * y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i) \tag{6}$$

where y_i is the true label of residue i, $\hat{y}_i$ is the predicted probability of residue i being a binding site, and N is the total number of all residues, w is the loss weight of a true binding site.

To enhance training stability and convergence, we employed the RAdam optimizer combined with the Lookahead optimization technique [95]. The regularization methods, including dropout and weight decay, were applied to improve the capacity of generalization. For each drug modality's binding site, ALLSites was trained separately using the corresponding binding site dataset. In each prediction task, unless otherwise specified, ALLSites adopted the same training scheme as the competing methods.

Specifically, for PPI-Test70, PPI-Test355, SMPI-Test348, and CarbPI-Test162, models were trained on their respective training sets, evaluated on the corresponding validation sets for hyperparameter tuning, and the best-performing models were finally evaluated on the test sets. For PPI-Test60 and PPI-Test315, the five-fold cross-validation was first performed on the PPI-Train335 dataset to determine the best hyperparameters. Finally, the model was retrained on the entire PPI-Train335 dataset using these best hyperparameters, and the model performance was evaluated on PPI-Test60 and PPI-Test315, respectively. The model training schemes for the two PepPI site prediction tasks were consistent with the original studies from which the datasets were sourced. For PepPI-Train1154, the ten-fold cross-validation was conducted to identify the best hyperparameters. Subsequently, the model was retrained on the entire PepPI-Train1154 dataset using the best hyperparameters and evaluated on the PepPI-Test125 dataset. For PepPI-Train640, a random subset of 128 proteins was first held out from the training set to form a validation set, used for hyperparameter selection. The performance of the best-performing model was then evaluated on the PepPI-Test639 test set. For DPI-Train573 and RPI-Train495, following the training scheme described in GraphBind [41], the training sets were randomly split into training and validation subsets in an 8:2 ratio. This splitting and training process was repeated ten times to generate ten independent models. The performance of each model was evaluated on its respective test set (DPI-Test129, DPI-Test181, or RPI-Test117). The reported performance of ALLSites on these test sets represented the average performance across ten models. An early stopping strategy with a patience of ten epochs was employed in all tasks to mitigate overfitting.

3.4.2 | Model Implementation

ALLSites was configured with a series of settings. The hidden dimension (d_{hid}) was 128; the number of encoder layers was 3; the number of decoder layers was 3; the number of attention heads was 8; the hidden dimension in the position-wise feed-forward network was 256; the kernel size of Conv1D was 7; and the weight decay was set to 1E-4. Additionally, the four most influential hyperparameters (including batch size, learning rate,

dropout rate, and loss weight) were optimized based on the model's predictive performance on the validation dataset. Given the large dataset size arising from treating each residue as an individual sample, ALLSites supported distributed training to accelerate model training. ALLSites was implemented in Python 3.10 and Pytorch 1.12.0 (http://pytorch.org/). All models were developed on the platform with Intel(R) Xeon(R) Gold 6132 CPU @ 2.60GHz, NVIDIA(R) Tesla(R) V100 32GB GPU, and 263GB RAM on CentOS Linux release 7.9.2009 (Core).

3.5 | Evaluation Metrics

Several evaluation metrics were used to evaluate the model performance, including accuracy (ACC), precision, recall, area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), F1 score, and Matthews correlation coefficient (MCC). Due to the class imbalance inherent in binding site datasets, the MCC metric is a particularly important metric, as it provides a robust assessment that accounts for imbalance in the data [96]. For the two test sets, PPI-Test60 and PPI-Test315, we further performed a calibration analysis of ALLSites based on the Brier score. It has been reported that the Brier score ranges from 0 to 1, and a Brier score below 0.25 indicates that the model's predictions are reliable [56]. Additionally, the Dice similarity coefficient (DSC) was calculated for the CarbPI site prediction task [97]. Unlike the other tasks, performance evaluation for CarbPI involved calculating metrics for each individual protein and then averaging these results across the entire test set. All the metrics were calculated using Scikit-learn [98]. The formulas for computing these evaluation metrics were provided in Supporting Information

Table S16 explicitly documents the sources of the evaluation performance for all baseline methods across the assessed tasks. In general, for those methods whose results were obtained by reproducing their source code, we employed the same training and evaluation data splits as well as identical training protocols used by ALLSites. For certain methods, results were retrieved from their respective web servers primarily because their source code was not publicly available for retraining. For other methods, reported performance values were directly extracted from the original literature, as these studies utilized the same training data and training schemes as ALLSites, thereby ensuring a fair comparison.

3.6 | Statistical Analysis

In this work, the benchmark results of ALLSites on DPI-Test129, DPI-Test181, and RPI-Test117 datasets were obtained from ten independent runs. The results were presented as mean ± standard deviation (SD).

Author Contributions

F.Z., H.B.D., F.C.L., and M.J.M. conceived the idea, designed the research, and wrote the manuscript. M.J.M., M.K.L. and Z.M.Z. constructed the model and performed benchmark evaluation. Y.L.R., X.Y.Y., Z.Q.P. and Y.Z. performed data analysis. H.Y., L.Y.Z., S.K.G. and Y.Z. collected the data. All authors have approved the latest version of the manuscript.

Acknowledgements

Natural Science Foundation of Zhejiang (RG25H300001); National Natural Science Foundation of China (82373790, 22220102001, 82504916); National Key R&D Programs of China (2024YFA1307503); Project for Building Research Capacity in Digital and Intelligent Health (HJ2024011003); Information Technology Center of Zhejiang University.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

All benchmark datasets are available on GitHub (https://github.com/idrblab/ALLSites/). Source data for benchmark evaluation results are fully provided in the Supplementary Information.

References

1. T. Rolland, M. Tasan, B. Charloteaux, et al., "A Proteome-Scale Map of the Human Interactome Network," *Cell* 159 (2014): 1212–1226.

2. L. F. Krapp, L. A. Abriata, F. Cortés Rodriguez, and M. Dal Peraro, "PeSTo: Parameter-Free Geometric Deep Learning for Accurate Prediction of Protein Binding Interfaces," *Nature Communications* 14 (2023): 2175.

3. H. Cui, A. Tejada-Lapuerta, M. Brbic, et al., "Towards Multimodal Foundation Models in Molecular Cell Biology," *Nature* 640 (2025): 623–633.

4. J. N. Spradlin, E. Zhang, and D. K. Nomura, "Reimagining Druggability Using Chemoproteomic Platforms," *Accounts of Chemical Research* 54 (2021): 1801–1813.

5. S. Ackloo, A. A. Antolin, J. M. Bartolome, et al., "Target 2035 – an Update on Private Sector Contributions," *RSC Medicinal Chemistry* 14 (2023): 1002–1011.

6. B. Zdrazil, E. Felix, F. Hunter, et al., "The ChEMBL Database in 2023: A Drug Discovery Platform Spanning Multiple Bioactivity Data Types and Time Periods," *Nucleic Acids Research* 52 (2024): D1180–D1192.

7. Y. Zhou, Y. Zhang, D. Zhao, et al., "TTD: Therapeutic Target Database Describing Target Druggability Information," *Nucleic Acids Research* 52 (2024): D1465–D1477.

8. X. X. Li and T. M. Woodruff, "The Complement System: Biology, Pathology, and Therapeutic Interventions," *Pharmacological Reviews* 77 (2025): 100079.

9. S. Linciano, Y. Mazzocato, Z. Romanyuk, et al., "Screening Macrocyclic Peptide Libraries by Yeast Display Allows Control of Selection Process and Affinity Ranking," *Nature Communications* 16 (2025): 5367.

10. R. Roche, B. Moussad, M. H. Shuvo, S. Tarafder, and D. Bhattacharya, "EquiPNAS: Improved Protein–Nucleic Acid Binding Site Prediction Using Protein-Language-Model-Informed Equivariant Deep Graph Neural Networks," *Nucleic Acids Research* 52 (2024): 27.

11. B. A. H. Smith and C. R. Bertozzi, "The Clinical Impact of Glycobiology: Targeting Selectins, Siglecs and Mammalian Glycans," *Nature Reviews Drug Discovery* 20 (2021): 217–243.

12. J. Mao, S. Guan, Y. Chen, et al., "Application of a Deep Generative Model Produces Novel and Diverse Functional Peptides Against Microbial Resistance," *Computational and Structural Biotechnology Journal* 21 (2023): 463–471.

13. A. E. Wakefield, D. Kozakov, and S. Vajda, "Mapping the Binding Sites of Challenging Drug Targets," *Current Opinion in Structural Biology* 75 (2022): 102396.

14. L. Qian, X. Lin, X. Gao, et al., "The Dawn of a New Era: Targeting the "Undruggables" with Antibody-Based Therapeutics," *Chemical Reviews* 123 (2023): 7782–7853.

15. M. Lazou, D. Kozakov, D. Joseph-McCarthy, and S. Vajda, "Which Cryptic Sites are Feasible Drug Targets?" *Drug Discovery Today* 29 (2024): 104197.

16. X. He, L. Zhao, Y. Tian, et al., "Highly Accurate Carbohydrate-Binding Site Prediction with DeepGlycanSite," *Nature Communications* 15 (2024): 5163.

17. Q. C. Zhang, D. Petrey, L. Deng, et al., "Structure-Based Prediction Of Protein–Protein Interactions on a Genome-Wide Scale," *Nature* 490 (2012): 556–560.

18. Z. Peng and L. Kurgan, "High-Throughput Prediction of RNA, DNA and Protein Binding Regions Mediated by Intrinsic Disorder," *Nucleic Acids Research* 43 (2015): 121.

19. D. Xiong, et al., "A Structurally Informed Human Protein-Protein Interactome Reveals Proteome-Wide Perturbations Caused by Disease Mutations," *Nature Biotechnology* 43 (2024): 1510–1524.

20. P. Li and Z. P. Liu, "GeoBind: Segmentation of Nucleic Acid Binding Interface on Protein Surface with Geometric Deep Learning," *Nucleic Acids Research* 51 (2023): 60.

21. P. Schmidtke and X. Barril, "Understanding and Predicting Druggability. A High-Throughput Method for Detection of Drug Binding Sites," *Journal of Medicinal Chemistry* 53 (2010): 5858–5867.

22. M. Mou, Z. Zhang, Z. Pan, and F. Zhu, "Deep Learning for Predicting Biomolecular Binding Sites of Proteins," *Research* 8 (2025): 0615.

23. S. Li, S. Wu, L. Wang, F. Li, H. Jiang, and F. Bai, "Recent Advances in Predicting Protein–Protein Interactions with the Aid of Artificial Intelligence Algorithms," *Current Opinion in Structural Biology* 73 (2022): 102344.

24. M. Zeng, F. Zhang, F.-X. Wu, Y. Li, J. Wang, and M. Li, "Protein–Protein Interaction Site Prediction Through Combining Local and Global Features with Deep Neural Networks," *Bioinformatics* 36 (2020): 1114–1120.

25. Y. Li, G. B. Golding, and L. Ilie, "DELPHI: Accurate Deep Ensemble Model for Protein Interaction Sites Prediction," *Bioinformatics* 37 (2021): 896–904.

26. M. Mou, et al., "A Transformer-Based Ensemble Framework for the Prediction of Protein-Protein Interaction Sites," *Research* 6 (2023): 0240.

27. Z. Zhao, Z. Peng, and J. Yang, "Improving Sequence-Based Prediction of Protein–Peptide Binding Residues by Introducing Intrinsic Disorder and a Consensus Method," *Journal of Chemical Information and Modeling* 58 (2018): 1459–1468.

28. E. Petsalaki, A. Stark, E. García-Urdiales, and R. B. Russell, "Accurate Prediction of Peptide Binding Sites on Protein Surfaces," *PLoS Computational Biology* 5 (2009): 1000335.

29. R. Krivák and D. Hoksza, "P2Rank: Machine Learning Based Tool for Rapid and Accurate Prediction of Ligand Binding Sites From Protein Structure," *Journal of Cheminformatics* 10 (2018): 39.

30. S. W. Canner, S. Shanker, and J. J. Gray, "Structure-Based Neural Network Protein-Carbohydrate Interaction Predictions at the Residue Level," *Frontiers in Bioinformatics* 3 (2023): 1186531.

31. S. Basu, J. Yu, D. Kihara, and L. Kurgan, "Twenty Years of Advances in Prediction of Nucleic Acid-Binding Residues in Protein Sequences," *Briefings in Bioinformatics* 26 (2024): bbaf016.

32. Y. H. Zhu, J. Hu, X. N. Song, and D. J. Yu, "DNAPred: Accurate Identification of DNA-Binding Sites From Protein Sequence by Ensembled Hyperplane-Distance-Based Support Vector Machines," *Journal of Chemical Information and Modeling* 59 (2019): 3057–3071.

33. R. Liu and J. Hu, "DNABind: A Hybrid Algorithm for Structure-Based Prediction of DNA-Binding Residues by Combining Machine Learning-and Template-Based Approaches," *Proteins: Structure, Function, and Bioinformatics* 81 (2013): 1885–1899.

34. J. H. Lam, Y. Li, L. Zhu, et al., "A Deep Learning Framework to Predict Binding Preference of RNA Constituents on Protein Surface," *Nature Communications* 10 (2019): 4941.

35. K. Tunyasuvunakool, J. Adler, Z. Wu, et al., "Highly Accurate Protein Structure Prediction for the Human Proteome," *Nature* 596 (2021): 590–596.

36. Q. Yuan, J. Chen, H. Zhao, Y. Zhou, and Y. Yang, "Structure-Aware Protein–Protein Interaction Site Prediction Using Deep Graph Convolutional Network," *Bioinformatics* 38 (2021): 125–132.

37. Q. Sun, H. Wang, J. Xie, et al., "Computer-Aided Drug Discovery for Undruggable Targets," *Chemical Reviews* 125 (2025): 6309–6365.

38. Z. Yang, X. Zeng, Y. Zhao, and R. Chen, "AlphaFold2 and Its Applications in the Fields of Biology and Medicine," *Signal Transduction and Targeted Therapy* 8 (2023): 115.

39. S. Mahbub and M. S. Bayzid, "EGRET: Edge Aggregated Graph Attention Networks and Transfer Learning Improve Protein-Protein Interaction Site Prediction," *Briefings in Bioinformatics* 23 (2022): bbab578.

40. J. Tubiana, D. Schneidman-Duhovny, and H. J. Wolfson, "ScanNet: An Interpretable Geometric Deep Learning Model for Structure-Based Protein Binding Site Prediction," *Nature Methods* 19 (2022): 730–739.

41. Y. Xia, C. Q. Xia, X. Pan, and H. B. Shen, "GraphBind: Protein Structural Context Embedded Rules Learned by Hierarchical Graph Neural Networks for Recognizing Nucleic-Acid-Binding Residues," *Nucleic Acids Research* 49 (2021): 51.

42. R. Wang, J. Jin, Q. Zou, K. Nakai, and L. Wei, "Predicting Protein–Peptide Binding Residues via Interpretable Deep Learning," *Bioinformatics* 38 (2022): 3351–3360.

43. J. Dapkūnas, A. Timinskas, K. Olechnovič, M. Tomkuvienė, and Č. Venclovas, "PPI3D: A Web Server for Searching, Analyzing and Modeling Protein–Protein, Protein–Peptide and Protein–Nucleic Acid Interactions," *Nucleic Acids Research* 52 (2024): W264–W271.

44. J. Wang, X. Wang, Y. Chu, et al., "Exploring the Conformational Ensembles of Protein–Protein Complex with Transformer-Based Generative Model," *Journal of Chemical Theory and Computation* 20 (2024): 4469–4480.

45. T. C. Northey, A. Barešić, and A. C. R. Martin, "IntPred: A Structure-Based Predictor of Protein–Protein Interaction Sites," *Bioinformatics* 34 (2018): 223–229.

46. A. Porollo and J. Meller, "Prediction-Based Fingerprints of Protein–Protein Interactions," *Proteins: Structure, Function, and Bioinformatics* 66 (2007): 630–645.

47. Y. Ofran and B. Rost, "ISIS: Interaction Sites Identified From Sequence," *Bioinformatics* 23 (2007): e13–e16.

48. Q. Hou, P. F. G. De Geest, W. F. Vranken, J. Heringa, and K. A. Feenstra, "Seeing the Trees Through the Forest: Sequence-Based Homo- and Heteromeric Protein-Protein Interaction Sites Prediction Using Random Forest," *Bioinformatics* 33 (2017): 1479–1487.

49. Y. Murakami and K. Mizuguchi, "Applying the Naïve Bayes Classifier with kernel Density Estimation to the Prediction of Protein–Protein Interaction Sites," *Bioinformatics* 26 (2010): 1841–1848.

50. G. Singh, K. D. Dhole, P. Pai, and S. K. Mondal, "SPRINGS: Prediction of Protein-Protein Interaction Sites Using Artificial Neural Networks," *Journal of Proteomics and Computational Biology* 1 (2014): 7.

51. J. Qiu, M. Bernhofer, M. Heinzinger, et al., "ProNA2020 Predicts Protein–DNA, Protein–RNA, and Protein–Protein Binding Proteins and Residues From Sequence," *Journal of Molecular Biology* 432 (2020): 2428–2443.

52. J. Zhang and L. Kurgan, "SCRIBER: Accurate and Partner Type-Specific Prediction of Protein-Binding Residues From Proteins Sequences," *Bioinformatics* 35 (2019): i343–i353.

53. B. Zhang, J. Li, L. Quan, Y. Chen, and Q. Lü, "Sequence-Based Prediction of Protein-Protein Interaction Sites by Simplified Long Short-Term Memory Network," *Neurocomputing* 357 (2019): 86–100.

54. S. Wang, W. Chen, P. Han, X. Li, and T. Song, "RGN: Residue-Based Graph Attention and Convolutional Network for Protein–Protein Interaction Site Prediction," *Journal of Chemical Information and Modeling* 62 (2022): 5961–5974.

55. P. Gainza, F. Sverrisson, F. Monti, et al., "Deciphering Interaction Fingerprints From Protein Molecular Surfaces Using Geometric Deep Learning," *Nature Methods* 17 (2020): 184–192.

56. H. Haider, B. Hoehn, S. Davis, and R. J. A. Greiner, "Effective Ways to Build and Evaluate Individual Survival Distributions," *J Machine Learn Res* 21 (2020): 1–63.

57. T. Hayes, R. Rao, H. Akin, et al., "Simulating 500 Million Years of Evolution with a Language Model," *Science* 387 (2025): 850–858.

58. Z. Gao, et al., "PFMBench: Protein Foundation Model Benchmark," *arXiv preprint arXiv* 2506 (2025): 14796.

59. G. Taherzadeh, Y. Yang, T. Zhang, A. W. Liew, and Y. Zhou, "Sequence-Based Prediction of Protein–Peptide Binding Sites Using Support Vector Machine," *Journal of Computational Chemistry* 37 (2016): 1223–1229.

60. W. Wardah, A. Dehzangi, G. Taherzadeh, et al., "Predicting Protein-Peptide Binding Sites with a Deep Convolutional Neural Network," *Journal of Theoretical Biology* 496 (2020): 110278.

61. O. Abdin, S. Nim, H. Wen, and P. M. Kim, "PepNN: A Deep Attention Model for the Identification of Peptide Binding Sites," *Communications Biology* 5 (2022): 503.

62. A. Lavi, C. H. Ngan, D. Movshovitz-Attias, et al., "Detection of Peptide-Binding Sites on Protein Surfaces: The First Step Toward the Modeling and Targeting of Peptide-Mediated Interactions," *Proteins: Structure, Function, and Bioinformatics* 81 (2013): 2096–2105.

63. G. Taherzadeh, Y. Zhou, A. W. Liew, and Y. Yang, "Structure-Based Prediction of Protein– Peptide Binding Regions Using Random Forest," *Bioinformatics* 34 (2018): 477–484.

64. L. Zhong, Y. Li, L. Xiong, et al., "Small Molecules in Targeted Cancer Therapy: Advances, Challenges, and Future Perspectives," *Signal Transduction and Targeted Therapy* 6 (2021): 201.

65. J. Desaphy, G. Bret, D. Rognan, and E. Kellenberger, "sc-PDB: A 3D-Database of Ligandable Binding Sites—10 Years on," *Nucleic Acids Research* 43 (2015): D399–D404.

66. É. Bokor, S. Kun, D. Goyard, et al., "C-Glycopyranosyl Arenes and Hetarenes: Synthetic Methods and Bioactivity Focused on Antidiabetic Potential," *Chemical Reviews* 117 (2017): 1687–1764.

67. M. Buerke, A. S. Weyrich, Z. Zheng, F. C. Gaeta, M. J. Forrest, and A. M. Lefer, "Sialyl Lewisx-Containing Oligosaccharide Attenuates Myocardial Reperfusion Injury in Cats," *Journal of Clinical Investigation* 93 (1994): 1140–1148.

68. M. Von Itzstein, "The War Against Influenza: Discovery and Development of Sialidase Inhibitors," *Nature Reviews Drug Discovery* 6 (2007): 967–974.

69. M. E. Griffin and L. C. Hsieh-Wilson, "Tools for Mammalian Glycoscience Research," *Cell* 185 (2022): 2657–2677.

70. B. Ernst and J. L. Magnani, "From Carbohydrate Leads to Glycomimetic Drugs," *Nature Reviews Drug Discovery* 8 (2009): 661–677.

71. D. Kozakov, L. E. Grove, D. R. Hall, et al., "The FTMap Family of Web Servers for Determining and Characterizing Ligand-Binding Hot Spots of Proteins," *Nature Protocols* 10 (2015): 733–755.

72. H. Du, et al., "Proteome-Wide Profiling of the Covalent-Druggable Cysteines With a Structure-Based Deep Graph Learning Network," *Research* 2022 (2022): 9873564.

73. S. Djebali, C. A. Davis, A. Merkel, et al., "Landscape of Transcription in Human Cells," *Nature* 489 (2012): 101–108.

74. F. Cozzolino, I. Iacobucci, V. Monaco, and M. Monti, "Protein–DNA/RNA Interactions: An Overview of Investigation Methods in the -Omics Era," *Journal of Proteome Research* 20 (2021): 3018–3030.

75. Y. Tao, Q. Zhang, H. Wang, X. Yang, and H. Mu, "Alternative Splicing and Related RNA Binding Proteins in Human Health and Disease," *Signal Transduction and Targeted Therapy* 9 (2024): 26.

76. M. R. Cookson, "RNA-Binding Proteins Implicated in Neurodegenerative Diseases," *WIREs RNA* 8 (2017): 1397.

77. D. Chen, X. Gu, Y. Nurzat, et al., "Writers, Readers, and Erasers RNA Modifications and Drug Resistance in Cancer," *Molecular Cancer* 23 (2024): 178.

78. H. Su, M. Liu, S. Sun, Z. Peng, and J. Yang, "Improving the Prediction of Protein–Nucleic Acids Binding Residues via Multiple Sequence Profiles and the Consensus of Complementary Methods," *Bioinformatics* 35 (2019): 930–936.

79. J. Zhang, Q. Chen, and B. Liu, "NCBRPred: Predicting Nucleic Acid Binding Residues in Proteins Based on Multilabel Learning," *Briefings in bioinformatics* 22 (2021): bbaa397.

80. Q. Wu, Z. Peng, Y. Zhang, and J. Yang, "COACH-D: Improved Protein–Ligand Binding Sites Prediction with Refined Ligand-Binding Poses Through Molecular Docking," *Nucleic Acids Research* 46 (2018): W438–W442.

81. R. R. Walia, L. C. Xue, K. Wilkins, Y. El-Manzalawy, D. Dobbs, and V. Honavar, "RNABindRPlus: A Predictor that Combines Machine Learning and Sequence Homology-Based Methods to Improve the Reliability of Predicted RNA-Binding Residues in Proteins," *PLoS ONE* 9 (2014): 97725.

82. S. Li, K. Yamashita, K. M. Amada, and D. M. Standley, "Quantifying Sequence and Structural Features of Protein–RNA Interactions," *Nucleic Acids Research* 42 (2014): 10086–10098.

83. J. Jumper, R. Evans, A. Pritzel, et al., "Highly Accurate Protein Structure Prediction With AlphaFold," *Nature* 596 (2021): 583–589.

84. M. Baek, F. DiMaio, I. Anishchenko, et al., "Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network," *Science* 373 (2021): 871–876.

85. S. K. Burley, C. Bhikadiya, C. Bi, et al., "RCSB Protein Data Bank (RCSB.org): Delivery of Experimentally-Determined PDB Structures Alongside One Million Computed Structure Models of Proteins From Artificial Intelligence/Machine Learning," *Nucleic Acids Research* 51 (2023): D488.

86. S. Jones and J. M. Thornton, "Analysis of Protein-Protein Interaction Sites Using Surface Patches 1 1Edited by G.Von Heijne," *Journal of Molecular Biology* 272 (1997): 121–132.

87. J. Yang, A. Roy, and Y. Zhang, "BioLiP: A Semi-Manually Curated Database for Biologically Relevant Ligand–Protein Interactions," *Nucleic Acids Research* 41 (2013): D1096–D1103.

88. J. Zhang, Z. Ma, and L. Kurgan, "Comprehensive Review and Empirical Analysis of Hallmarks of DNA-, RNA- and Protein-Binding Residues in Protein Chains," *Briefings in Bioinformatics* 20 (2019): 1250–1268.

89. S. Altschul, "Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs," *Nucleic Acids Research* 25 (1997): 3389–3402.

90. I. Lee and H. Nam, "Sequence-Based Prediction of Protein Binding Regions and Drug–Target Interactions," *Journal of Cheminformatics* 14 (2022): 5.

91. Q. Yuan, et al., "AlphaFold2-aware protein-DNA Binding Site Prediction Using Graph Transformer," *Briefings in Bioinformatics* 23 (2022): bbab564.

92. Z. Lin, H. Akin, R. Rao, et al., "Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model," *Science* 379 (2023): 1123–1130.

93. T. Li, X. M. Zhao, and L. Li, "Co-VAE: Drug-Target Binding Affinity Prediction by Co-Regularized Variational Autoencoders," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44 (2022): 8861–8873.

94. J. Mao, T. Sui, K.-H. Cho, K. T. No, J. Wang, and D. Shan, "IUPAC-GPT: An IUPAC-Based Large-Scale Molecular Pre-Trained Model for Property Prediction and Molecule Generation," *Molecular Diversity* (2025): 1–9.

95. L. Chen, X. Tan, D. Wang, et al., "TransformerCPI: Improving Compound–Protein Interaction Prediction by Sequence-Based Deep Learning with Self-Attention Mechanism and Label Reversal Experiments," *Bioinformatics* 36 (2020): 4406–4414.

96. Z. Xu, D. Shen, Y. Kou, and T. Nie, "A Synthetic Minority Oversampling Technique Based on Gaussian Mixture Model Filtering for Imbalanced Data Classification," *IEEE Transactions on Neural Networks and Learning Systems* 35 (2024): 3740–3753.

97. M. Arshad, C. Wang, M. W. Us Sima, et al., "BioAug-Net: A Bioimage Sensor-Driven Attention-Augmented Segmentation Framework with Physiological Coupling for Early Prostate Cancer Detection in T2-Weighted MRI," *BioData Mining* 18 (2025): 49.

98. F. Pedregosa, et al., "Scikit-Learn: Machine Learning in Python," *J Machine Learn Res* 12 (2011): 2825–2830.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.

Supporting File: advs73503-sup-0001-SuppMat.docx.