

RESEARCH ARTICLE OPEN ACCESS

ProSiteHunter: A Unified Framework for Sequence-Based Prediction of Protein-Nucleic Acid and Protein-Protein Binding Sites

Dongliang Hou¹ | Qihang Zhen¹ | Zexin Lv¹ | Xinyue Cui¹ | Suhui Wang¹ | Minghua Hou¹ | Zhan Zhou²  | Xiaogen Zhou¹  | Guijun Zhang¹ 

¹College of Information Engineering, Zhejiang University of Technology, HangZhou, China | ²State Key Laboratory of Advanced Drug Delivery and Release Systems and Innovation Institute for AI in Medicine, College of Pharmaceutical Sciences, Zhejiang University, HangZhou, China

Correspondence: Zhan Zhou (zhazhou@zju.edu.cn) | Xiaogen Zhou (zxcg@zjut.edu.cn) | Guijun Zhang (zg@zjut.edu.cn)

Received: 15 December 2025 | **Revised:** 19 May 2026 | **Accepted:** 22 May 2026

Keywords: artificial intelligence | computational biology | computer science | language model | molecular recognition | protein function | protein ligand | sequence

ABSTRACT

Accurate identification of protein binding sites is essential for elucidating protein function, decoding molecular recognition, and guiding drug design. However, existing sequence-based approaches are often designed for specific binding-site types and therefore lack generality, whereas structure-based methods typically rely on high-quality structural models, limiting their applicability. Here, we present ProSiteHunter, a unified sequence-based framework for predicting protein binding sites spanning protein-DNA, protein-RNA, protein-protein, and antibody-antigen interfaces. ProSiteHunter integrates the fine-tuned protein language model SiteT5 with evolutionary, geometric, and statistical features extracted from sequences. These representations are further processed through a Multi-Source Feature Fusion (MSFF) module, which captures bidirectional semantics, local associations, and global dependencies to achieve a comprehensive characterization of binding sites, thereby substantially improving predictive accuracy and generalization capability. Across comprehensive benchmarks, ProSiteHunter achieved a 38.4% average improvement in the area under the precision-recall curve (PRAUC) for protein-DNA/RNA/protein tasks and a 15.1% PRAUC enhancement on the particularly challenging antibody-antigen task over state-of-the-art methods. Moreover, ProSiteHunter is capable of identifying local flexible sites that complement AlphaFold3 predictions and improving the accuracy of antibody-antigen interaction prediction. These results highlight ProSiteHunter as an efficient and unified approach for accurate and robust prediction of diverse protein binding sites.

1 | Introduction

The interactions between proteins and macromolecules are the core mechanisms of life activities, playing an important role in processes such as cell signal transduction, gene expression regulation, and immune response [1–3]. The binding sites represent critical regions that mediate interactions between these biomolecules. Accurate determination of these sites is essential

not only for elucidating protein function but also for providing key insights into drug target identification, mechanistic studies of biomolecular processes, and antibody design [4–6]. The investigation of protein-nucleic acid binding sites contributes to elucidating gene regulatory networks [7, 8] and the mechanisms of transcriptional regulation [9]. Similarly, the prediction of protein-protein binding sites can help uncover signaling pathways associated with diseases [1]. The identification of

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *Advanced Science* published by Wiley-VCH GmbH

antibody-antigen binding sites, as a distinct category of protein interactions, holds particular research significance. It provides a critical foundation for epitope-centric vaccine design [10] and, importantly, for the development of antibody-based therapeutics [11]. Traditional biological experimental methods (such as affinity purification-mass spectrometry [12], immunoprecipitation [13], and hybrid screening [14, 15]), while capable of precisely detecting interaction sites, are time-consuming and costly, and have stringent requirements for experimental conditions, making it difficult to meet the needs of large-scale research. Consequently, the development of efficient and accurate computational methods for predicting binding sites has emerged as a critical scientific challenge demanding urgent solutions.

Structure-based computational methods are mainstream approaches that are generally divided into two groups: graph neural network (GNN) based methods and geometric feature engineering-based methods. GNN-based methods model the spatial topological relationships between residues by constructing a graph representation of protein structures. Typical works include GraphBind [16], which integrates physicochemical features and hierarchical graph networks to predict nucleic acid binding residues, and EGPD [17], which constructs a graph convolutional network based on residue coordinates to predict protein-DNA binding sites. Geometric feature engineering methods focus on extracting key features from three-dimensional spatial information. Representative methods include PeSTo [18], which predicts protein interaction sites by characterizing protein surface shapes with atomic point clouds. While these methods have achieved remarkable success in predictive performance, several important challenges remain to be addressed for their broader practical application. A primary limitation is the strong dependency on high-quality structural data, where even minor conformational deviations can lead to prediction inaccuracies. Although advances in computational tools such as AlphaFold2 [19] have dramatically improved protein backbone prediction, accurately modeling flexible binding regions remains a major technical hurdle [20]. Furthermore, many functional interactions rely on dynamic conformational changes [21], suggesting that predictions derived from static structural snapshots may fail to capture authentic binding mechanisms [22]. These inherent constraints currently limit the generalizability and practical utility of structure-based binding site prediction methods.

Sequence-based prediction methods have emerged as a prominent and rapidly advancing frontier in bioinformatics, requiring only protein primary sequences as input. This approach demonstrates significantly greater robustness than structure-based methods, as it circumvents conformational prediction inaccuracies and inherent challenges posed by protein structural dynamics. Furthermore, sequence-based techniques substantially reduce computational resource requirements compared to structure-dependent approaches [23]. Notably, recent advances in large language models have dramatically improved the accuracy of sequence-based binding site prediction [24, 25], further establishing this paradigm as a pivotal direction for future research. Sequence-based computational methods can be generally classified into three categories based on the predicted binding site types: the first category includes protein-protein interaction (PPI) sites prediction methods. For example, ISPPRED-SEQ [26] integrates ESM-1b [27] embeddings with a convolutional neural

network to enable efficient prediction of protein interaction sites. In addition, Seq-InSite [28] enhances interaction site prediction by integrating ProtT5 and MSA Transformer language models, enabling more comprehensive representation learning of protein sequences; the second category involves protein-nucleic acid binding sites prediction methods, such as CLAPE [29], which employs contrastive learning to identify protein-nucleic acid binding sites; the third category comprises B-cell epitope prediction methods, including BepiPred-3.0 [30], which integrates the ESM-2 [24] language model for epitopes prediction, and SEMA2.0-1D [31], which optimizes epitopes feature learning for prediction by fine-tuning the ESM-2 [24] model.

While sequence-based binding site prediction methods have made remarkable progress, two critical challenges remain: limited task generalizability and inadequate modeling of sequence dependencies. Most existing approaches are tailored to specific types of interactions, such as protein-protein or protein-nucleic acid binding, using distinct feature sets and network architectures. Although this specialization enhances performance on dedicated tasks, it considerably restricts model generalization across different prediction scenarios. Furthermore, owing to constraints in local receptive fields and insufficient global context modeling, current methods often fail to capture long-range dependencies within sequences, a fundamental requirement for accurately identifying distal functional residues and achieving a comprehensive understanding of protein function mechanisms. Addressing these limitations would significantly advance the development of universal binding site predictors and provide a new paradigm for systematically deciphering the functional code of proteins.

In this work, we present ProSiteHunter, a universal sequence-based framework for predicting general binding sites between protein-nucleic acid and protein-protein interactions. The innovation of the proposed framework is manifested at three distinct levels. At the architectural design level, we have designed a multi-source feature fusion module that achieves spatial alignment and collaborative optimization of multimodal feature representation through three-track interaction. At the feature representation level, we integrated three complementary types of protein information: evolutionary features, statistical features, and geometric features to construct a multi-dimensional representation space of binding sites. At the sequence encoding level, we have developed SiteT5, a protein language model fine-tuned from ProtT5-XL-UniRef50 [25] by integrating evolutionary information. Compared to existing methods, ProSiteHunter not only exhibits broad applicability, capable of predicting protein-nucleic acid (DNA/RNA), protein-protein, and specialized protein (antibody-antigen) binding sites, but also demonstrates superior prediction accuracy and robustness. Cross-task validation further confirms that the integrated multi-source features and the unique architectural design are critical contributions to the model's performance.

2 | Results

2.1 | Overview of ProSiteHunter

ProSiteHunter leverages a unified deep learning framework based on an encoder-decoder architecture. It takes a single amino acid

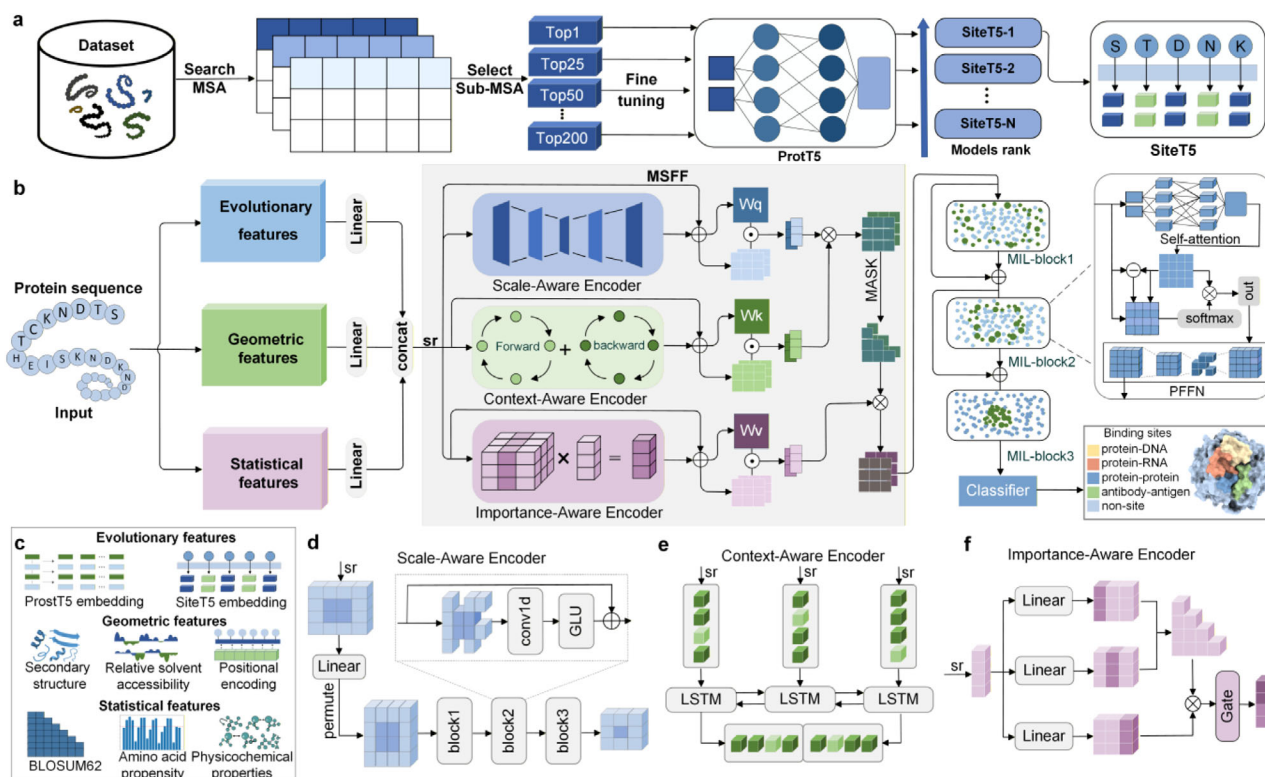


FIGURE 1 | Pipeline of ProSiteHunter. (a) Sub-MSAs are sampled from the MSA of each training sequence and used to fine-tune ProtT5 with LoRA. The optimal SiteT5 model is then selected to extract the SiteT5 embeddings shown in Figure 1c. (b) ProSiteHunter takes a protein sequence as input and constructs multi-level sequence representations, which are subsequently processed by the Multi-Source Feature Fusion (MSFF) module to capture both local patterns and global dependencies. The fused representations are further refined by the Multi-Layer Interaction Learning (MIL) module. Finally, a classifier outputs the probability of each residue being part of a binding site. Protein structures shown in the figure are schematic. (c) Feature extraction integrates three complementary feature types: evolutionary, geometric, and statistical, to construct a multidimensional representation of each site. All features are collectively referred to as sequence representations (sr). (d–f) Architectures of the Scale-Aware Encoder, Context-Aware Encoder, and Importance-Aware Encoder modules.

sequence as input and outputs residue-level binding propensity scores, indicating the likelihood of each residue being part of a binding site. Residues with predicted scores greater than 0.5 are classified as binding sites. We employ an information enhancement strategy through multi-level feature optimization. Our framework predicts binding sites for four major types of protein interactions: protein-DNA, protein-RNA, protein-protein, and antibody-antigen. Four specialized models are independently trained on type-specific datasets, enabling effective identification of distinctive structural features and binding patterns pertinent to each interaction type.

ProSiteHunter comprises two main stages. In the feature fusion stage, a Multi-Source Feature Fusion (MSFF) with three-track semantic parsing is constructed to realize the complementary representation of sequence information (Figure 1b illustrates the architectural optimization of the three networks, which share the same sequence representation (sr) as input.): (1) Scale-Aware Encoder module, which uses serial convolutional kernels to extract short-range inter-residue information in the local receptive field; (2) Context-Aware Encoder module is used for bidirectional semantic modeling, and directional sensitivity features are captured through forward and reverse sequence scanning; (3) The Importance-Aware Encoder module learns global dependencies and explicitly captures long-distance residue asso-

ciations. In particular, trainable weight matrices are employed to adaptively weight the features from the three tracks and project them into the Query (Q), Key (K), and Value (V) vectors for the cross-attention module, enabling dynamic alignment and fusion of the feature spaces. In the self-interaction stage, these representations are processed by a Multi-Level Interactive Learning (MIL) module composed of three blocks. Within this module, relevant information about binding sites is further refined through multi-head sub-attention with gating mechanisms. Nonlinear transformation and feature enhancement are then applied via a Position-wise Feed-Forward Network (PFFN). Finally, a classifier produces a probability score for each residue, indicating its likelihood of being a binding site.

In the process of feature extraction, we leverage two Protein Large Language Models (PLLM), SiteT5 and ProstT5, to encode key features of proteins from the perspectives of evolutionary conservation and geometric plausibility, thereby generating significant synergistic effects in binding site prediction tasks. Among them, SiteT5 is a specialized model we developed by fine-tuning ProtT5-XL-UniRef50[25], aiming to effectively capture the evolutionarily conserved features of functional sites (Figure 1a). For four different binding site prediction tasks, we searched for multiple sequence alignments (MSAs) from the protein sequences of the corresponding training sets and selected sub-MSAs as inputs,

enabling the model to learn the distinctive patterns and features of different protein families. Beyond the protein language model, we integrated three geometric features: symmetrically normalized position encoding, secondary structure, and relative solvent accessibility. These features quantify spatial exposure and structural context and provide a foundation for binding site localization. We further incorporated three statistical features: BLOSUM62 [32] matrix, physicochemical properties, and amino acid propensity to characterize residue-level attributes. See the ‘Sequence feature’ section in Methods and Table S1.

By organically integrating multi-source feature synergy with a dynamic fusion mechanism, ProSiteHunter achieves a significant enhancement in the accuracy of protein binding site identification, all while maintaining a lightweight model architecture. This capability is substantiated by its robust performance in two challenging scenarios: structural inaccuracies from AlphaFold3 predictions and conformational changes in proteins driven by dynamics (Figures 3e and 5e).

2.2 | Performance of Protein-Nucleic Acid Binding Sites Prediction

To fairly compare the predictive performance of ProSiteHunter on protein-nucleic acid binding sites, both we and the benchmark methods trained and tested our models using two different datasets from GraphBind [16]: one for protein-DNA and another for protein-RNA binding sites. In order to mitigate potential data leakage, a temporal split was applied by assigning all protein chains released before January 6, 2016 to the training set, and those released on or after that date to the test set. Furthermore, to address the class imbalance between positive and negative samples, protein chains were clustered. Within each cluster, binding site annotations were transferred to the longest chain to enhance the representativeness of the data [16]. For further details, see the ‘Data set’ section in Methods.

We benchmarked our method against state-of-the-art predictors, including DRNApred [33], iDRNA-ITF [34], CLAPE-DB [29], and CLAPE-RB [29]. The predictions from DRNApred [33] and iDRNA-ITF [34] were generated using their respective web servers, while those for CLAPE-DB [29] and CLAPE-RB [29] were reproduced based on their publicly available source code. The performance was evaluated using a comprehensive set of standard metrics, comprising the area under the receiver operating characteristic curve (ROCAUC), F1-score, and Matthew’s correlation coefficient (MCC). It is worth noting that in order to avoid the bias caused by the imbalance of positive and negative sample data in the test set (see Table S2), we introduced the area under the precision-recall curve (PRAUC) to further evaluate ProSiteHunter performance.

As illustrated in Figure 2a, ProSiteHunter demonstrates significantly superior performance compared to other benchmarked methods. Specifically, it achieved ROCAUC values of 0.917 and 0.852 for predicting protein-DNA and protein-RNA binding sites, respectively, surpassing the second-best method, CLAPE-DB [29], by 6.0% and CLAPE-RB [29] by 8.1%. Notably, ProSiteHunter also showed substantial advantages in other core metrics such as PRAUC, F1-score, and MCC (Figures 2b,c). For protein-DNA

binding site prediction, it improved by 26.1% in PRAUC, 34.5% in F1-score, and 32.4% in MCC compared to CLAPE-DB [29]. Similarly, for protein-RNA prediction, the corresponding improvements over CLAPE-RB [29] reached 78.8% in PRAUC, 33.5% in F1-score, and 41.7% in MCC (Tables S3 and S4). These results collectively indicate that ProSiteHunter maintains exceptional predictive accuracy even under highly imbalanced classification conditions, where the site-to-non-site ratio approaches 1:10 (Table S2), highlighting its robust classification capability.

In the sequence-wise comparative analysis, ProSiteHunter consistently outperformed CLAPE-DB [29] and CLAPE-RB [29], achieving higher AUC values on 64.3% and 67.5% of test sequences in protein-DNA and protein-RNA binding site prediction, respectively (Figure 2d). A similar trend was observed in comparisons with iDRNA-ITF [34] and DRNApred [33], where ProSiteHunter also demonstrated clear advantages (Figure S1). We hypothesize that this robust performance stems from our three-track feature fusion architecture, which synergistically integrates the Scale-Aware Encoder module, Context-Aware Encoder module, and Importance-Aware Encoder module. This architectural design, as further validated by the ablation studies in Figure 4c, effectively combines site and non-site information across diverse representation spaces, thereby enhancing both discriminative power and generalization capacity. In contrast, approaches relying on single-network architectures may not sufficiently capture the complex and hierarchical determinants of binding sites, which could partially explain their comparatively limited predictive performance.

To evaluate the discriminatory ability of ProSiteHunter between binding sites and non-binding sites, we analyzed the predictive score distributions for both positive and negative samples across protein-DNA and protein-RNA tasks (Figures 2e,f). ProSiteHunter demonstrated a pronounced advantage over state-of-the-art methods such as iDRNA-ITF [34], CLAPE-DB [29], and CLAPE-RB [29]. Specifically, it assigned high prediction scores (mostly above 0.6) to binding sites, indicating strong discriminative confidence for positive samples. In comparison, other methods yielded predominantly lower scores (below 0.5) for true binding residues, reflecting limited sensitivity in identifying positive instances. These results confirm the superior capability of ProSiteHunter in recognizing binding sites and its robustness in class-imbalanced prediction scenarios.

ProSiteHunter demonstrates robust and consistent superiority over existing other methods across three fundamental aspects: comprehensive evaluation metrics, sequence-wise predictive accuracy, and discriminative capability in distinguishing binding from non-binding residues. We attribute this advantage to two core designs: the SiteT5 module, a protein language model-based encoder that captures rich evolutionary and biochemical information, and a three-track fusion architecture that integrates multi-scale and hierarchical sequence semantics. Ablation experiments (Figures 4c,f) further confirm that both components contribute significantly to the model’s enhanced performance. The SiteT5 encoder provides highly discriminative residue-level features, while the fusion mechanism enables effective integration of multi-source contextual information. Collectively, these designs endow ProSiteHunter with strengthened generalization capacity and improved robustness in capturing

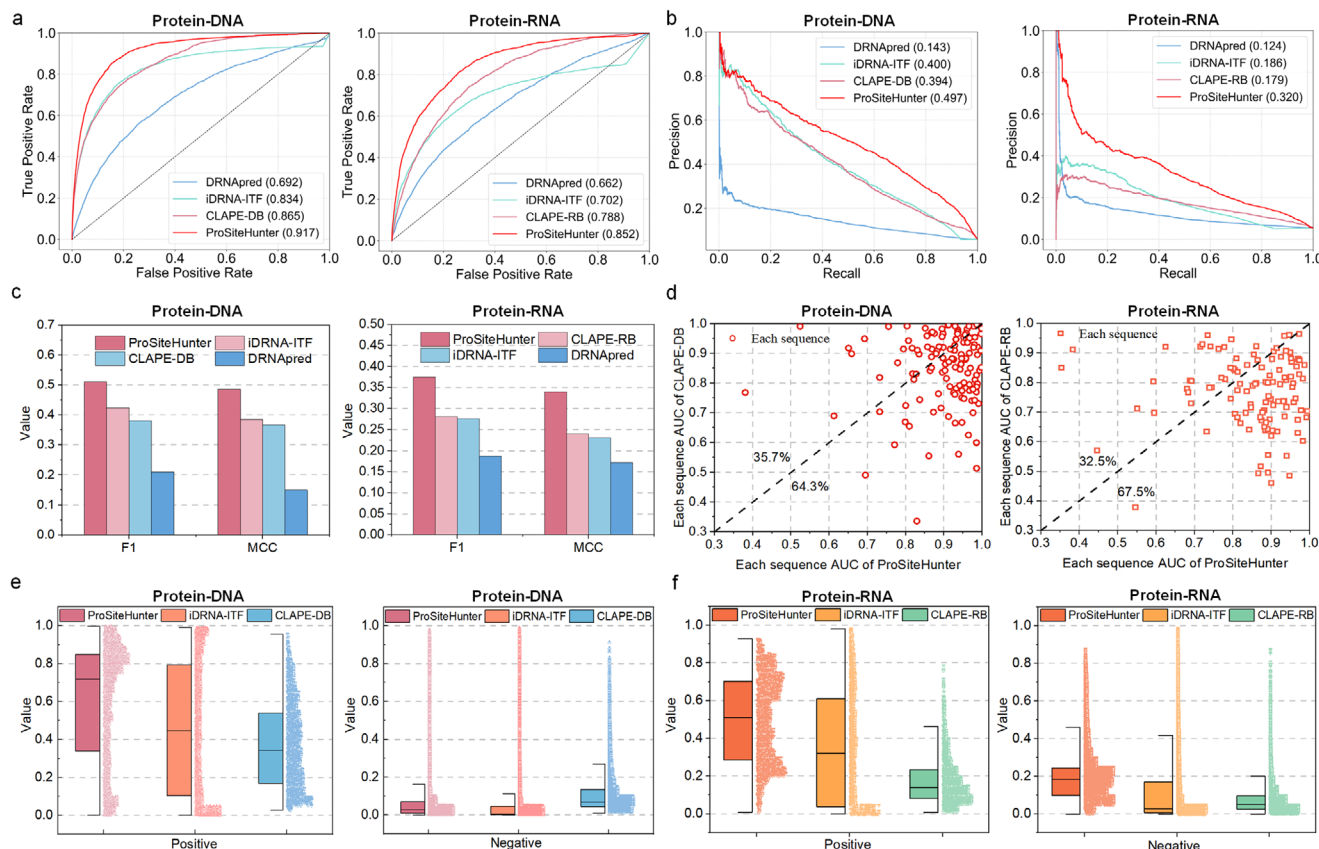


FIGURE 2 | Performance of ProSiteHunter in protein-nucleic acid binding sites prediction. (a) ROCAUC curve of ProSiteHunter on the protein-nucleic acid prediction task. (b) PRAUC curves of ProSiteHunter on protein-nucleic acid prediction tasks. (c) The F1 and MCC performance comparison of the four methods: ProSiteHunter, CLAPE-DB/RB, iDRNA-ITF, and DRNApred. (d) Scatter plots display the sequence-wise AUC values of ProSiteHunter vs. CLAPE-DB for protein-DNA binding and CLAPE-RB for protein-RNA binding. (e) Half-box plots comparing the classification performance of ProSiteHunter, CLAPE-DB, and iDRNA-ITF on positive and negative samples in protein-DNA binding site prediction. (f) Half-box plots comparing the classification performance of ProSiteHunter, CLAPE-RB, and iDRNA-ITF on positive and negative samples in protein-RNA binding site prediction.

the complex sequence features of protein-nucleic acid binding sites.

2.3 | Performance of Protein-Protein Binding Sites and Antibody-Antigen Binding Sites Prediction

To evaluate the applicability of the ProSiteHunter framework for predicting protein-protein interaction sites, we maintained its original architecture and performed training and validation using the Seq-InSite [28] dataset, which contains 13518 protein sequences with less than 25% sequence identity. Further details are described in the ‘Data set’ section of Methods. Benchmark results showed that ProSiteHunter maintained strong predictive performance on this task. Given the importance of antibody-antigen interactions in pharmaceutical applications, we further assessed the performance of ProSiteHunter on antibody-antigen binding sites prediction. To ensure fair comparison, we trained and tested the model on the SEMA [35] antibody-antigen conformational epitope dataset (as epitope labels are masked in SEMA2.0 [31]). This dataset assigns protein chains published before January 1, 2020, to the training set and protein chains published on or after that date to the test set. Further details

are described in the ‘Data set’ section of Methods. Benchmark results demonstrated that ProSiteHunter consistently outperformed other methods.

In the protein binding sites prediction task (Figures 3a,b), ProSiteHunter demonstrated significantly superior performance over state-of-the-art methods, including PITHIA [36], ISPRE-SEQ [26], and Seq-InSite [28]. Relative to the best-performing method, ProSiteHunter achieved notable improvements across multiple evaluation metrics, with increases of 2.67% in ROCAUC, 10.25% in PRAUC, 2.49% in F1-score, and 4.37% in MCC. Additionally, Table S5 presents detailed comparisons with a broader set of competing methods [37–39], further confirming the overall superiority of the model. Similarly, we conducted additional evaluations on the benchmark dataset Test70, and the results show that ProSiteHunter maintains strong and stable performance. Compared with the best-performing method, it achieves a 10.78% improvement in PRAUC (see Table S6 and Figure S2), demonstrating its robust generalization capability. Furthermore, in the antibody-antigen binding sites prediction task, ProSiteHunter consistently outperformed the top-performing method, CALIBER [40] (Figures 3a,b), achieving improvements of 4.10%, 15.10%, 11.91% and 15.74% in ROCAUC, PRAUC, F1-score, and

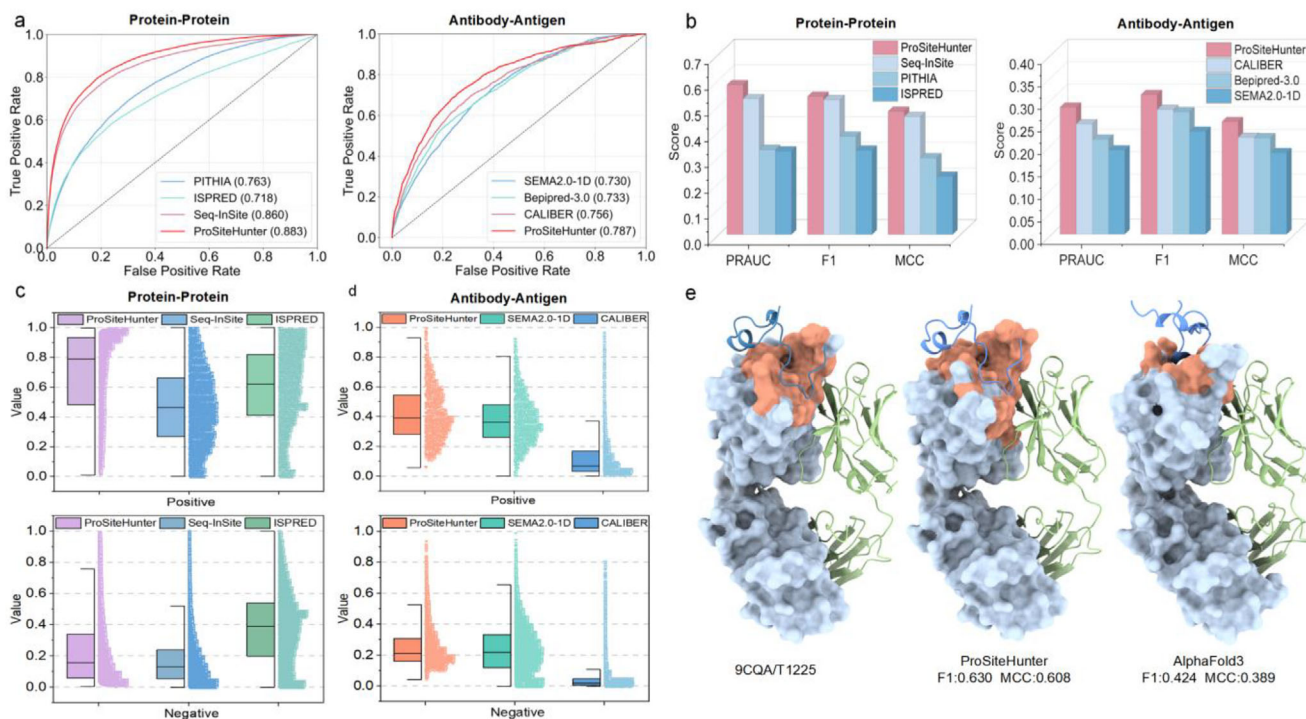


FIGURE 3 | Performance of ProSiteHunter in protein-protein binding sites, antibody-antigen binding sites prediction. (a) ROCAUC curves of ProSiteHunter for protein-protein binding sites and antibody-antigen binding sites prediction tasks. (b) Performance comparison of ProSiteHunter and other methods in terms of PRAUC, F1, and MCC for protein-protein binding sites and antibody-antigen binding sites prediction tasks. (c) Half-box plots display the classification performance of ProSiteHunter vs. Seq-InSite and ISPRED on positive and negative samples in the protein binding sites prediction task. (d) Half-box plots display the classification performance of ProSiteHunter vs. SEMA2.0-1D and CALIBER on positive and negative samples in the antibody-antigen binding sites prediction task. (e) Comparison of binding site predictions for the CASP16 target T1225 (PDB ID: 9CQA): the experimental structure (left, the light blue portion represents the antibody heavy chain region, which binds to the dark blue respiratory syncytial virus glycoprotein protein. The orange region indicates the true binding sites), predictions from ProSiteHunter (middle; orange region for predicted binding sites), and predictions from AlphaFold3 (right; orange region for predicted binding sites).

MCC, respectively (Table S7). Collectively, these results highlight that ProSiteHunter delivers high predictive accuracy while maintaining strong generalization capability across diverse prediction tasks, underscoring its effectiveness as a unified framework for binding site prediction.

In addition, we tested the classification performance of ProSiteHunter against state-of-the-art methods (ISPRED-SEQ [26] and Seq-InSite [28]) (Figure 3c). ProSiteHunter demonstrates a clear advantage in positive sample identification, with a median score of 0.788, which is substantially higher than that of Seq-InSite [28] (0.462) and ISPRED-SEQ [26] (0.620). From the distribution of predicted scores, ProSiteHunter's outputs for positive samples are predominantly concentrated in the high-confidence range (0.8–1.0). In contrast, the predictions of Seq-InSite [28] are more dispersed and mainly fall within the low-to-medium confidence interval (0.2–0.8). Although ISPRED-SEQ [26] assigns relatively high scores to a subset of samples, its overall distribution still tends toward a moderate confidence range (0.4–1.0). These observations indicate that ProSiteHunter produces more stable and higher-confidence predictions for positive samples, reflecting its stronger discriminative capability. In more challenging antibody-antigen binding sites prediction tasks, ProSiteHunter's classification performance is still superior to the comparison methods. (Figure 3d). In contrast, other methods are overly conservative in predicting sites and are not sensitive to the distinc-

tion between positive and negative samples. These performance differences likely arise from the reliance of these methods on single-modal sequence features, without incorporating complementary information such as protein spatial topology, long-range interactions, or evolutionary conservation. Consequently, they face challenges in establishing precise decision boundaries when classifying positive and negative samples.

Figure 3e presents the structural model of the complex between the central conserved domain of respiratory syncytial virus glycoprotein (RSV G) and monoclonal antibody 1G1, serving as a representative case study that demonstrates the capability of our approach in resolving functional epitope-paratope interfaces. RSV is a major pathogen causing severe lower respiratory tract infections in infants, young children, and the elderly. Its attachment glycoprotein facilitates viral entry into host cells and modulates the host immune response by binding to the chemokine receptor CX3CR1 [41]. In the native protein structure shown on the left side of Figure 3e, the light blue portion represents the antibody heavy chain region, which binds to the dark blue RSV G protein. The orange region indicates the actual binding sites. This protein, designated as T1225 in CASP16. We used ProSiteHunter to predict binding sites on antibody heavy chains, achieving an F1 score of 0.630 and an MCC of 0.608. Compared with AlphaFold3, which yielded an F1 score of 0.424 and an MCC of 0.389, this represents improvements of 48.6% and

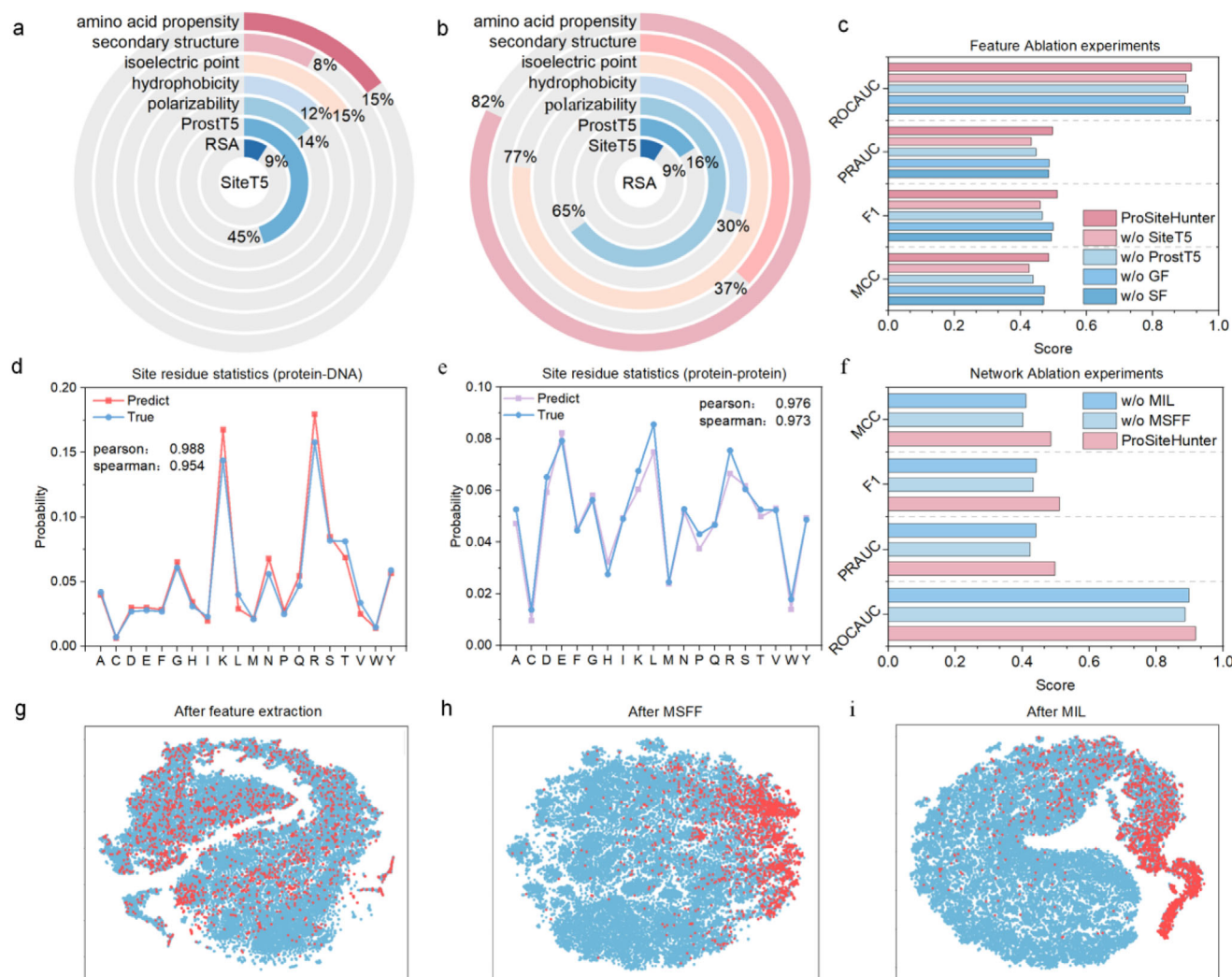


FIGURE 4 | Ablation studies and interpretability analysis. (a) Correlation analysis between SiteT5 embedding and a set of representative features, including RSA, ProstT5, polarizability, hydrophobicity, isoelectric point, secondary structure, and amino acid propensity. (b) Correlation analysis between RSA embedding and a set of representative features, including SiteT5, ProstT5, polarizability, hydrophobicity, isoelectric point, secondary structure, and amino acid propensity. (c) Characteristic ablation experiments. (d) Amino acid propensity distribution of real protein-DNA binding sites and ProSiteHunter predicted binding sites. (e) Amino acid propensity distribution of real protein-protein binding sites and ProSiteHunter predicted binding sites. (f) Network ablation experiments. (g–i), The t-SNE graph is used to map the high-dimensional tensor after feature extraction, MSFF module, and MIL module into a two-dimensional space to describe the relationship between protein-DNA binding sites (red) and non-sites (blue).

56.3%, respectively. Structure-based methods often predict incorrect binding site regions when the predicted protein structure deviates from experimental results. In contrast, ProSiteHunter delivers superior and more reliable performance in identifying the majority of binding sites, establishing it as a highly effective complement to AlphaFold3 [42].

Our analysis confirms the consistent superiority of ProSiteHunter over benchmark methods across all four sub-tasks (protein-DNA, protein-RNA, protein-protein, and antibody-antigen), despite significant variations in intrinsic prediction difficulty. Notably, the task of antibody-antigen binding sites prediction is more challenging than protein-DNA binding sites prediction. We propose that the observed discrepancy arises from distinct binding mechanisms: antibodies primarily interact with epitopes via their complementarity-determining regions (CDRs), whose conformational diversity contributes to increased structural heterogeneity

within epitopes. This observation aligns with the documented difficulties of both AlphaFold2 [19] and AlphaFold3 [42] in modeling antibody-antigen complexes. Moreover, these findings suggest that a hybrid strategy integrating sequence-based epitope prediction with the structural modeling capabilities of AlphaFold2/3 may be a worthwhile avenue to explore for improving the accuracy of antibody-antigen complex determination.

2.4 | Ablation Studies and Interpretability Analysis

To validate the rationale of feature selection and the effectiveness of the network architecture in ProSiteHunter, we first performed feature correlation analyses and ablation studies (including both feature and network ablations). To further examine the model's ability to identify binding sites and capture underlying interaction

patterns, we conducted amino acid propensity analyses within binding regions and integrated them with t-SNE [43] visualization. These analyses demonstrated that the selected features make significant contributions to binding site prediction, and that the designed architecture excels in feature integration and pattern learning, enabling effective discrimination between binding and non-binding residues. Collectively, these results enhance the interpretability of the model and substantiate its effectiveness in both binding site identification and pattern recognition.

To assess the interdependence among different features, we conducted a systematic correlation analysis using cosine similarity as the evaluation metric. In particular, we focused on SiteT5 and relative solvent accessibility (RSA), as SiteT5 represents our evolution-based language model, whereas RSA is a classical structural descriptor closely associated with binding site formation (Figures 4a,b). ProST5 [44] and SiteT5 showed a moderate correlation (45%), reflecting that although they share similarities, they have different focuses due to different fine-tuning strategies: ProST5 [44] relies on Foldseek [45] 3Di sequences for structure guidance, while SiteT5 focuses on evolutionary information. SiteT5 generally showed low correlations with traditional physicochemical features (isoelectric point, polarizability, etc.) (< 20%), demonstrating the importance of a multi-feature fusion strategy. RSA showed weak correlations with SiteT5 (9%) and ProST5 [44] (16%), highlighting the limitations of language models in capturing structural features and the necessity of incorporating RSA into sequence prediction. Isoelectric point (77%), polarizability (65%), and amino acid propensity (82%) were highly correlated with RSA, further highlighting the close connection between solvent accessibility and functional residues. Overall, correlation analyses reveal the complementary roles of each indispensable feature category: language models provide semantic and evolutionary constraints, physicochemical descriptors reveal residue-level properties, and geometric descriptors capture the spatial environment. Their synergistic integration is therefore crucial for accurate binding site prediction.

Subsequently, we conducted systematic feature ablation experiments (Figure 4c and Figure S3) to evaluate the contributions of different feature categories. Compared with the full ProSiteHunter model, removing SiteT5 embedding, ProST5 embedding, Geometric features (GF, including secondary structure, Positional encoding, Relative solvent accessibility), and statistical features (SF, including BLOSUM62 Matrix, Amino acid propensity, Physicochemical properties) individually led to performance degradation to varying degrees. Specifically, removing SF and GF resulted in only minor decreases, with ROCAUC reduced by 0.23% and 2.18%, respectively, while the declines in F1, MCC, and PRAUC were all within 3.5%. In contrast, removing ProST5 embedding led to more noticeable performance drops, with ROCAUC decreasing by 1.27% and F1, MCC, and PRAUC declining by 8.79%, 9.79%, and 10.07%, respectively. A more pronounced degradation was observed when SiteT5 embedding was removed, where ROCAUC decreased by 1.76% and F1, MCC, and PRAUC dropped by 10.21%, 12.43%, and 12.92%, respectively. Overall, these findings indicate that geometric and statistical features primarily serve as auxiliary information that complements model performance, whereas the SiteT5 language model and ProST5 semantic branch constitute the core representational foundation of the model, playing a critical role in binding site prediction.

To systematically evaluate the contributions of key components in ProSiteHunter, we constructed two ablation models by removing the multi-source feature fusion module (MSFF) and the multi-level interaction learning module (MIL), respectively, and compared them with the full model (Figure 4f). The results show that removing either module leads to a significant performance decline. Specifically, excluding MSFF results in decreases of 3.38%, 14.89%, 15.26%, and 17.28% in ROCAUC, PRAUC, F1, and MCC, respectively, while removing MIL leads to reductions of 2.18%, 11.27%, 13.50%, and 15.43% in the same metrics. Notably, removing MSFF leads to a more pronounced performance degradation, demonstrating that the three-track semantic feature fusion mechanism is a key source of the model's performance gains. However, MSFF alone is insufficient to fully exploit its representational capacity; MIL further enhances the learned multi-source representations through additional interaction modeling. Therefore, the two modules do not operate independently, but instead form a mutually reinforcing, synergistic mechanism that jointly improves the overall performance of the model in binding site prediction tasks. To further evaluate the assignment strategy of the Scale-Aware Encoder, Context-Aware Encoder, and Importance-Aware Encoder within the Q/K/V projections, we conducted systematic permutation experiments across different Q/K/V configurations (Figure S4 and Table S8). The results show that ProSiteHunter with the original Q/K/V assignment achieves the best performance across all evaluation metrics. Notably, it outperforms the second-best method by approximately 3.4% in PRAUC and 4.1% in MCC, indicating that the proposed Q/K/V assignment strategy more effectively integrates multi-scale and multi-semantic information, thereby enhancing the model's ability to discriminate complex binding sites.

To evaluate the generalization ability of ProSiteHunter across different sequence length intervals, we conducted a systematic analysis on four tasks (protein-DNA, protein-RNA, protein-protein, and antibody-antigen binding site prediction), as shown in Figure S5 and Tables S9 and S10. Overall, the model maintains relatively stable performance across all length ranges, with the smallest variation observed in the protein-DNA task (range = 0.018, standard deviation = 0.0055); for the protein-RNA, protein-protein, and antibody-antigen tasks, the ranges of variation are 0.057, 0.081, and 0.077, with corresponding standard deviations of 0.0195, 0.0273, and 0.0323, respectively. Although moderate fluctuations are observed, no clear monotonic trend with respect to sequence length is evident, indicating that the model does not exhibit systematic bias toward a specific length range. Furthermore, experiments with multiple random seeds (Tables S11–S14) consistently show stable performance under different initialization conditions, further confirming the model's robustness and reproducibility. To further assess computational efficiency, we conducted a computational cost analysis across different sequence lengths. All experiments were performed under single-sequence inference (batch size = 1) on an NVIDIA GeForce RTX 3090 GPU. As shown in Figure S6 and Table S15, both inference time and GPU memory usage increase approximately linearly with sequence length, even for longer sequences (500–1000 residues), the inference time remains within 11.78 s, and the GPU memory usage is approximately 123.8 MB. These results indicate that the proposed model achieves a favorable balance between predictive performance and computational efficiency, with low

resource requirements and fast inference speed, demonstrating strong practical applicability.

To evaluate model interpretability (Figures 4g–i), we selected the test set from the protein–DNA binding site prediction task, totaling 37 515 residues, including 2240 binding sites (positive samples) and 35 275 non-binding sites (negative samples). In the visualizations, red points denote binding sites and blue points denote non-binding sites. At the feature extraction stage, binding sites were randomly scattered among non-binding sites. With the incorporation of the multi-source feature fusion module, binding sites progressively clustered toward the right region, and this trend became more pronounced after further processing by the interactive learning module. These results demonstrate that ProSiteHunter effectively distinguishes positive from negative samples, underscoring its advantage in binding site discrimination. The respective interpretability experiments for protein–RNA, protein–protein, and antibody–antigen binding sites are presented in Figure S7.

As a unified framework, ProSiteHunter demonstrates versatile capability by achieving robust performance in both protein–DNA and protein–protein binding mode prediction. As shown in Figure 4d, the predicted amino acid frequency distribution of protein–DNA binding sites (red line) closely matches the ground truth (blue line) (Pearson = 0.988; Spearman = 0.954). Similarly, in protein–protein interaction prediction (Figure 4e), the predicted distribution (purple line) aligns well with the experimental data (blue line) (Pearson = 0.976; Spearman = 0.973). These findings indicate that ProSiteHunter accurately captures specific binding patterns, owing to the deep representation power of its multi-source feature fusion module and the superior generalization enabled by its hierarchical feature extraction and dynamic fusion mechanisms. Collectively, our work suggests that ProSiteHunter could become a widely applicable tool for the study of diverse protein–macromolecule interactions. Further analysis (Figure S8) revealed two notable insights: first, protein–DNA and protein–RNA interactions share similar binding patterns, while protein–protein and antibody–antigen interactions also exhibit overlapping binding features; second, a higher proportion of loops in the binding-site region generally confers greater structural flexibility, thereby increasing the difficulty of accurate binding site prediction.

2.5 | ProSiteHunter Captures Diverse Binding Modes and Dynamic Conformational Changes

To systematically evaluate the predictive performance of ProSiteHunter in different biological macromolecule interactions, we selected four representative proteins from the test set: PDB ID: 5H58 (protein–DNA binding, Figure 5a), PDB ID: 6D12 (protein–RNA binding, Figure 5b), PDB ID: 7MJS (antibody–antigen binding, Figure 5c), and PDB ID: 3E1Z (protein–protein binding, Figure 5d). These cases exemplify key binding site characteristics: the interacting residues form tight structural clusters despite being discretely distributed in the linear sequence. The conformational dynamics of hen egg-white lysozyme (PDB: 4GN4/1J1X) are of particular interest, as conformational changes in its loop regions enable the formation of two distinct protein complexes (4GN4 and 1J1X) (Figure 5e). The experimental results show that

ProSiteHunter is able to effectively identify this change in the binding sites region caused by protein conformational dynamics. The experimental results demonstrate ProSiteHunter’s capability for accurate identification of binding site alterations induced by protein conformational dynamics. This capability not only confirms its potential as a general-purpose prediction framework across diverse protein classes but also highlights its applicability in profiling dynamic binding sites.

5H58 is a complex of CprB of *Streptomyces coelicolor* and its biological longitudinal sequence [46]. CprB can specifically bind to operons (operons belong to DNA sequences, which are specific functional regions encoded by DNA in the genome of prokaryotes), and participates in the fine regulation of these physiological processes by regulating downstream gene expression [46]. ProSiteHunter accurately predicted most of the DNA binding sites on the protein (out of a total of 15 binding sites, 13 of which were successfully predicted. The prediction success rate was as high as 86.7%), and the AUC score reached 0.984, indicating that it has extremely high discriminant performance. As shown in Figure 5a, the green and yellow models are the DNA double helix structure, the blue protein part is the predicted non-binding site, and the orange part is the predicted binding site region.

6D12 is a complex of the C-terminal xRRM domain of human Larp7 bound to the 7SK stem-loop 4 RNA [47]. Larp7 contains a C-terminal atypical RNA recognition motif (xRRM) that specifically binds to 7SK and P-TEFb assembly, and xRRM deletion may drive gastric cancer progression through epigenetic dysregulation, such as aberrant transcriptional elongation [47]. As shown in Figure 5b, the green part is the RNA model (forming a double helix structure, similar to DNA, but composed of adenine, guanine, cytosine, and uracil, and is a type of RNA), the blue part is the non-binding site area predicted using ProSiteHunter, and the orange area is the predicted binding site area, which tightly wraps RNA, with an AUC score of up to 0.956.

As shown in Figure 5c, in the antibody–antigen complex 7MJS (PDB ID) [48], partial loops and α -helical regions of the antigen (blue) form contacts with the antibody heavy chain (purple). Remarkably, ProSiteHunter accurately predicted these binding interface sites (orange), achieving high prediction performance in this case (AUC = 0.965). Antibody–antigen binding interfaces are often highly structurally diverse and flexible, making prediction challenging. However, ProSiteHunter reliably identified the antibody–antigen binding sites in this case, further validating its versatility and effectiveness.

Figure 5d (PDB ID: 3E1Z) shows the dimeric crystal structure of the parasitic protease inhibitor chagasin in complex with papain. The blue papain subunit contains multiple highly flexible loops that mediate binding to chagasin [49]. Although the conformational flexibility of these loops complicates binding site prediction and leads to some mislocalized predictions, ProSiteHunter successfully identified the majority of true binding sites, achieving an AUC-ROC value of 0.935 and demonstrating robust discriminatory performance.

As shown in Figure 5e, we present a comparative analysis of two distinct complex structures of hen egg-white lysozyme

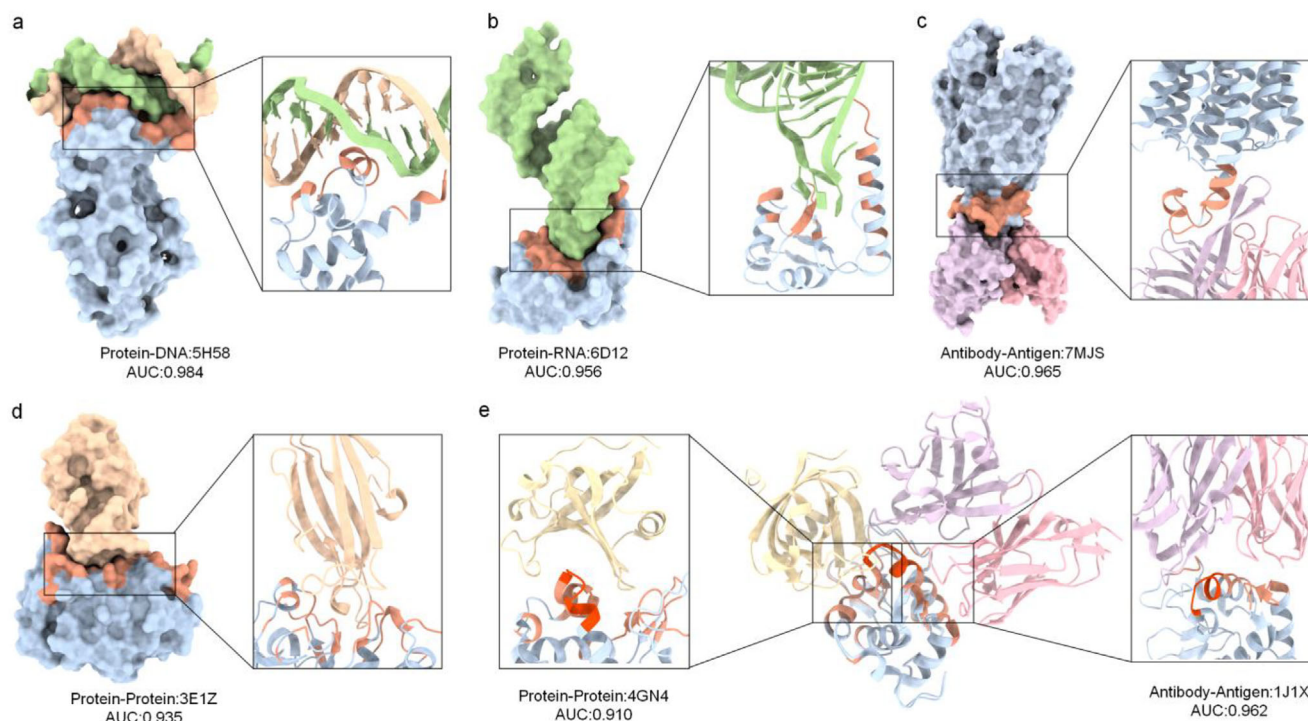


FIGURE 5 | Special case of five different binding sites. (a) The Protein (PDB ID:5H58) represents CprB (blue) from streptomyces coelicolor in complex with its cognate operator sequences (green and yellow). The orange region is the predicted binding sites. (b) The Protein (PDB ID:6D12) represents the crystal structure of the human Larp7 (blue) C-terminal xRRM domain bound to the 7SK stem-loop 4 RNA (green). The orange region indicates the predicted binding sites. (c) The Protein (PDB ID:7MJS) represents the complex of major promoter superfamily domain protein 2A (blue) and LPC-18:3 (pink and purple). The orange region indicates the predicted binding sites. (d) The Protein (PDB ID:3E1Z) represents the crystal structure of the parasite protease inhibitor chagasin (yellow) in complex with papain (blue). The orange region indicates the predicted binding sites. (e) The Protein (PDB ID:4GN4) represents AM2EP06 (yellow) bound to hen egg white lysozyme (blue). The Protein (PDB ID:1J1X) represents the crystal structure of the HyHEL-10 Fv mutant LS93A (purple, pink) bound to hen egg white lysozyme (blue). The predicted binding sites are highlighted in orange and red, and the red region indicates the overlap between the two binding sites.

(HEWL): the OBody AM2EP06-HEWL complex (PDB ID:4GN4 [50]) and the HyHEL-10 Fv mutant LS93A-HEWL complex (PDB ID:1J1X [51]). HEWL interacts with different molecular partners (proteins or antibodies) through conformational changes in its loop regions, thereby altering its binding interface. ProSiteHunter accurately identified the binding sites of HEWL in two complexes, achieving AUC values of 0.910 and 0.962 in 4GN4 and 1J1X, respectively. These findings highlight ProSiteHunter's ability to capture binding site variations caused by conformational flexibility, suggesting that structural dynamics play a key role in regulating binding specificity. This conformation-driven plasticity may represent one of the fundamental mechanisms underlying protein functional diversity.

These results establish ProSiteHunter's capability to accurately predict binding sites from sequence information alone. Notably, even its spurious predictions are informative, as they predominantly cluster spatially near genuine binding sites. This pattern strongly suggests that the model is implicitly learning the structural topology and evolutionary and physical constraints governing residue proximity directly from one-dimensional sequences (As illustrated in Figure S9). More remarkably, ProSiteHunter successfully predicts binding site remodeling induced by conformational changes from a single sequence (Figure 5e), highlighting its potential to illuminate allosteric mechanisms and dynamic protein functions.

2.6 | ProSiteHunter Enhances Downstream Prediction of Antibody-Antigen Interactions

Antibody-antigen interactions are a class of highly specific protein recognition processes that play a critical role in immune responses and biomedical applications [52]. The molecular basis of this interaction primarily relies on the precise recognition between the epitopes and the antibody complementarity-determining regions. To validate the applicability of ProSiteHunter in the field of immune recognition, this study specifically designed targeted downstream experiments: the epitopes information predicted by ProSiteHunter was integrated into MultisAAI [53], an antibody-antigen interaction prediction framework developed in-house in our laboratory. The dataset constructed by MultisAAI [53] includes 7580 antibody antigen pairs with a 1:1 ratio of positive to negative samples. We randomly selected 80% of the positive and negative samples as the training set and 20% as the test set. For the specific construction process of the dataset, please refer to MultisAAI [53]. Depending on whether to add sites features or not, we trained two models with the following results.

As shown in Figures 6a,b, evaluation results based on ROCAUC, F1 score, precision, and recall indicate that incorporating epitope prediction features improves model performance by 5.74%, 7.20%, 7.48%, and 6.94%, respectively. Further comparative analysis (Figures 6c,d) shows that this enhancement not only substantially

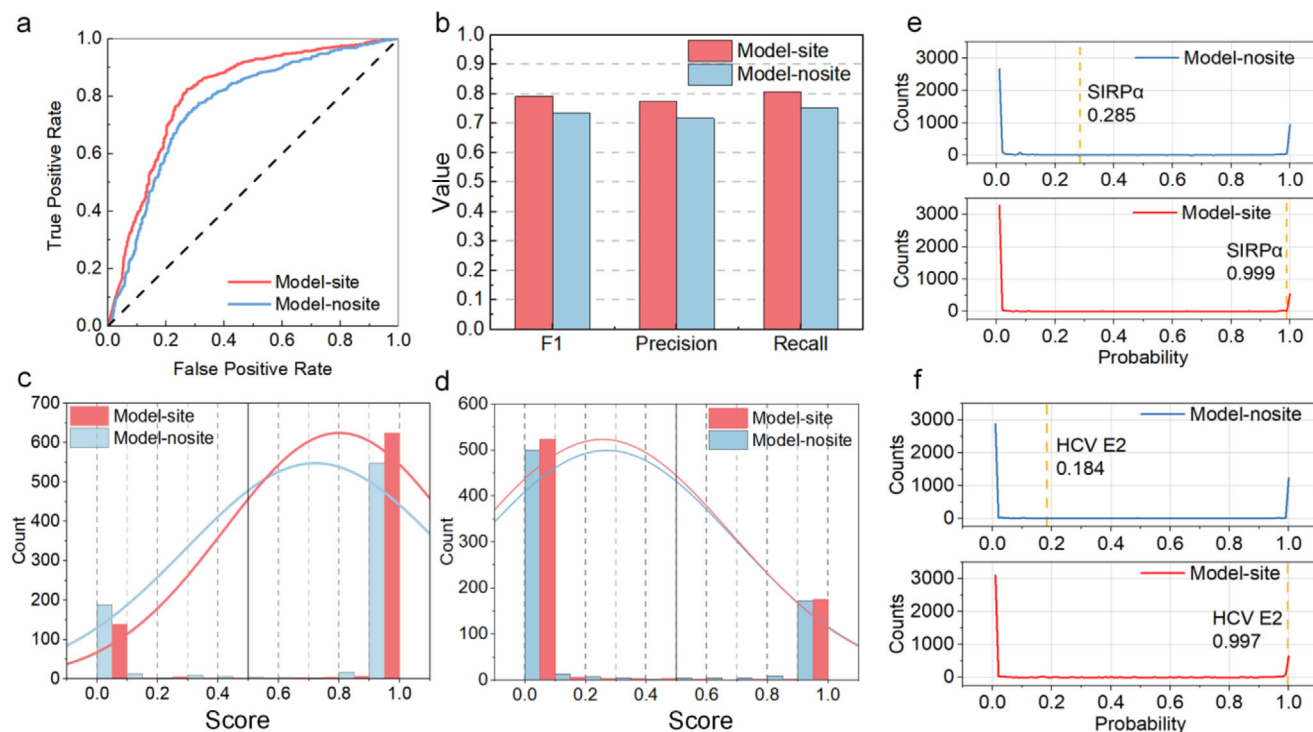


FIGURE 6 | ProSiteHunter enhances antibody-antigen interaction prediction performance. (a) Comparison of ROCAUC curves before and after the addition of predicted sites information. (b) Describes the performance of the two models on F1, Precision, and Recall. (c) Distribution plot illustrates the classification performance of positive samples before and after incorporating binding site information. (d) Distribution plot shows the classification performance of negative samples before and after the incorporation of binding site information. (e) The distribution of the interaction probability of Signal Regulatory Protein α (SIRP α) with all antibodies in the SAbDab database. (f) The distribution of the interaction probability of Hepatitis C Virus (HCV) E2 with all antibodies in the SAbDab database.

increases the identification rate of positive samples (by 14.1%, Figure 6c) but also strengthens the discrimination of negative samples (by 5.2%, Figure 6d). Together, these findings underscore the critical role of epitope information in predicting antibody-antigen interactions. ProSiteHunter effectively pinpoints most epitopes, and the site information it provides further enhances both the reliability and interpretability of interaction predictions.

To further validate the enhanced effect of integrating ProSiteHunter-predicted sites into MultisAAI [53], we analyzed two representative case studies. First, Signal Regulatory Protein α (SIRP α) is an inhibitory receptor expressed on the surface of myeloid cells, such as macrophages and neutrophils. By binding to the “don’t eat me” signal CD47, it inhibits the tumor-killing activity of myeloid cells. Kick-Off Fab115, an anti-SIRP α antibody, can effectively block this pathway [54]. As shown in Figure 6e, incorporating site information increased the interaction probability between SIRP α and Kick-Off Fab115 from 0.285 to 0.999. Second, Hepatitis C Virus (HCV) infects approximately 1%–2% of the global population and is a leading cause of liver failure and hepatocellular carcinoma. The development of an effective HCV vaccine requires targeting conserved epitopes recognized by broadly neutralizing antibodies (AR3B) [55]. With ProSiteHunter’s precise epitope prediction, the probability of interaction between genotype 6a E2 protein and AR3B increased from 0.184 to 0.997 (As shown in Figure 6f). Together, these cases illustrate that incorporating ProSiteHunter’s epitopes information can substantially improve antibody-antigen

interaction modeling and screening, thereby providing valuable support for antibody design and related studies.

Overall, ProSiteHunter demonstrates excellent adaptability and transferability in antibody-antigen recognition tasks. The epitope information predicted by the model significantly improves the discriminative performance of downstream models, achieving comprehensive improvements in overall metrics while demonstrating higher sensitivity and reliability in distinguishing positive from negative samples. Further case studies confirm this, demonstrating that the key epitope regions captured by ProSiteHunter are highly consistent with the true binding interface and can accurately characterize the distribution characteristics of antigen binding sites. These results demonstrate that ProSiteHunter not only has powerful epitope mapping capabilities but also provides high-quality semantic representations to support the screening of antibody-antigen interactions.

3 | Discussion

In this work, we present ProSiteHunter, a unified sequence-driven framework designed to address the long-standing challenge of limited generalizability in protein binding site prediction. By integrating a fine-tuned protein language model with a multi-source feature fusion architecture, ProSiteHunter can accurately identify binding interfaces across diverse biomolecular interactions, including protein-DNA, protein-RNA, protein-protein,

and antibody-antigen systems, solely from sequence information. Comprehensive evaluations demonstrate that ProSiteHunter exhibits both robustness and broad applicability, consistently outperforming state-of-the-art specialized predictors across multiple benchmark datasets, thereby providing a powerful tool for sequence-level studies of binding mechanisms and functional annotation.

ProSiteHunter further demonstrates its potential in modeling flexible regions and capturing dynamic conformational effects (Figures 3e and 5e). This method can identify local flexible sites and binding site differences caused by dynamic conformational changes, providing a feasible strategy for analyzing flexible epitopes and dynamic structures. These results show that ProSiteHunter is highly complementary to structure-based methods, and the combination of the two is expected to provide a more comprehensive analytical perspective in complex systems, thereby improving the reliability and applicability of binding site predictions and helping to reveal protein-macromolecule interaction mechanisms that are difficult to elucidate with other methods.

As training data continues to grow and diversify, ProSiteHunter's generalization capabilities are expected to further enhance. Not only will it maintain stable performance across a wider range of protein systems, but it will also naturally extend to the prediction of protein-small molecule ligand and protein-ion binding sites, further expanding its application. Furthermore, by combining ligand features with sequence information, ProSiteHunter is expected to achieve even higher accuracy and robustness in binding mode prediction. These improvements will provide a more reliable technical path for identifying functional binding sites, potentially making it a powerful tool for affinity prediction, molecular docking, and mutation effect analysis. This will facilitate a deeper understanding of protein binding mechanisms and hold great promise for applications in areas such as antibody design and antiviral vaccine development.

4 | Methods

4.1 | Data Set

Protein-nucleic acid binding sites dataset: The DNA/RNA binding proteins were collected from the BioLiP database, released on December 5, 2018. If the minimum atomic distance between the target residue and the nucleic acid molecule is less than 0.5 Å plus the sum of the van der Waals radii of the two closest atoms, it is defined as a binding residue. Based on the release date, protein chains published before January 6, 2016, were assigned to the original training set (comprising 6731 DNA-binding protein chains and 6426 RNA-binding protein chains), while the remaining protein chains were allocated to the original test set (comprising 2843 DNA-binding protein chains and 1267 RNA-binding protein chains). The prediction of DNA/RNA binding residues faces a data imbalance issue, where the number of DNA/RNA binding residues is significantly smaller than that of non-binding residues; thus, data augmentation was applied to the original training set. Chains with sequence identities > 0.8 and TM scores > 0.5 were clustered, and the annotations of

protein chains in the same cluster were transferred to the chain with the largest number of residues. After the migration and combined annotations, CD-HIT was used to remove redundant protein chains to reduce the sequence identity in the training set to less than 30%. Ultimately, 573 DNA-binding protein chains and 495 RNA-binding protein chains were obtained as the training set. Protein chains from the original DNA/RNA binding test set that shared over 30% sequence identity with any chains in the DNA/RNA binding training set were removed, resulting in 129 DNA-binding proteins and 117 RNA-binding proteins for the DNA and RNA binding test set. Please refer to GraphBind [16] for the detailed processing procedure.

Protein-Protein interaction site dataset was constructed as follows: we further performed redundancy reduction based on the dataset originally constructed in Seq-InSite [28], and obtained the final dataset used in this study. Specifically, as described in Seq-InSite [28], 22 654 protein sequences were extracted from the latest version of the PiSite [56] database, and sequences that did not contain interacting residues or were shorter than 50 amino acids were removed, ultimately retaining 14 203 sequences. Following the Seq-InSite [28] pipeline, PSI-CD-HIT [57] (sequence identity threshold set to 25%) was applied to these 14 203 sequences to remove entries with high sequence similarity to the Test70 dataset, yielding 14 031 sequences, referred to as “trainD-set_without70_Pid_Pseq_label_V12Mar2023”. Considering that this study uses both the Test70 and Test355 benchmark datasets, and to avoid potential data leakage, we further performed redundancy reduction on these 14 031 sequences using MMseqs [58] with a sequence identity threshold of 25%, a minimum sequence coverage of 90%, and coverage defined as the alignment length divided by the length of the shorter sequence, resulting in 13518 sequences as the final training set. To remain consistent with Seq-InSite [28] as much as possible, we adopted the same data splitting strategy, dividing the dataset into 99% training and 1% validation sets.

Antibody-antigen conformational epitopes Dataset: The process of generating the conformational epitopes dataset for this dataset includes the following steps: The ANARCI [59] tool is used to screen published protein structural sequences in the PDB database, including the heavy and light chains of Fab. Heavy/light Fab pairs are identified by calculating the distance between heavy and light chain subunit residues, and only those heavy and light chains that are in direct contact at a distance of 4.5 Å within a non-CDR region are considered heavy/light pairs. The CDR loop is defined using the Chothia number based on the notes in the ANARCI [59] tool. A protein subunit that is not annotated as an antibody but has at least five residues interacting with antibody residues with L1/L2/L3 or H1/H2/H3 CDR antibody loops within a 4.5 Å radius is considered an antigen. Antigenic residues with a distance radius of less than 4.5 Å are defined as epitopes. According to the structure publication date, the dataset is divided into training set and test set. This test set includes structures first published in the PDB database after January 1, 2020, with no sequences of > 70% available homologues prior to this date. The test set had a total of 101 antigen sequences, while the remaining antigens were divided into a training set (713 items) and a validation set (70 items). See SEMA [35] for details.

4.2 | Sequence Features

Protein language model: (1) SiteT5 is an enhanced protein language model fine-tuned based on ProtT5-XL-UniRef50 specifically for protein binding sites prediction tasks. ProtT5-XL-UniRef50²⁵ is a large-scale protein language model with T5-3B architecture obtained from self-supervised pre-training of 45 million protein sequences on the UniRef50 dataset. In order to improve the performance of site prediction, we used UniRef30 for multiple sequence alignment (MSA) generation, which preserves a wider range of sequence diversity, helps to capture deeper evolutionary signals, and improves the recognition of conserved regions. During the MSA generation process, we employed the HHblits [60] tool with stringent filtering parameters (E value threshold 0.000001, coverage 25%) to ensure the quality of alignment and retain the top 50 high-quality MSA results for each sequence (see Table S16 for details). During the fine-tuning phase, in order to retain as much of the original information learned in the ProtT5 pre-trained model as possible, we fine-tuned only the last four layers of its 24-layer decoder. The LoRA (Low Rank Adaptation) strategy was adopted during fine-tuning to ensure efficient parameter updates while avoiding destabilizing the pre-trained representation. The final total number of fine-tuned parameters was 655 360. We searched for the corresponding MSAs using the protein sequences in their respective training sets for four different binding sites prediction tasks, and input these MSAs as important information into the model for 30 epochs to enable the model to learn specific patterns and characteristics of the protein family. To prevent overfitting, we used protein data with less than 30% similarity as a validation set, and ultimately obtained four high-performance fine-tuned models dedicated to different binding sites prediction tasks. Each of these amino acids is expressed using a tensor of 1024 dimensions. (2) ProstT5 [44] (Protein Structure-Sequence T5) uses the 3Di coding scheme proposed by Foldseek [45] to convert the three-dimensional spatial structure information of proteins into one-dimensional discrete labeled sequences, so as to realize the bidirectional conversion between protein sequences and three-dimensional structures. The model is based on ProtT5-XL-UniRef50 and fine-tuned with 17 M high-quality protein sequence-structure pairs (3D structure prediction from AlphaFoldDB [61]). Each of these amino acids is expressed using a tensor of 1024 dimensions.

BLOSUM62: It is based on evolutionarily conserved statistical properties and is used to measure the probability of substitution between different amino acids in protein sequence alignment. To avoid over-reliance on highly similar sequences, BLOSUM62 clusters sequences with a similarity of $\geq 62\%$ (known as ‘62% clustering’). Conserved properties reflect the critical role of these regions in protein function or structure, so these conserved regions can be better described by BLOSUM62. It is of great significance for site identification.

Physicochemical properties: To provide a more specific description of amino acids, we introduce five physicochemical properties related to the site prediction task. Among these, the steric hindrance parameter measures the spatial volume or steric hindrance of the amino acid side chains, which affects spatial compatibility in protein interactions. The polarizability parameter describes the ease with which the electron cloud of the amino acid side chain deforms in an electric field, which is related to

the solvent accessibility of proteins and influences intermolecular interactions. The volume parameter reflects the van der Waals volume occupied by amino acid side chains, providing insight into steric constraints in protein structures. The hydrophobicity of amino acids, reflecting their tendency to exclude water and aggregate in non-polar environments, can be leveraged to identify residues likely exposed on the protein surface. The isoelectric point parameter is used to describe the pH value at which the net charge of an amino acid or protein is zero, indirectly indicating the charge characteristics of the amino acid. A sequence of length L corresponds to a tensor representation of its physicochemical properties as $(L, 5)$.

Relative solvent accessibility characteristic is the ratio of the solvent accessible surface area of an amino acid residue in a protein structure to the maximum accessible surface area of that amino acid in a fully stretched ‘free’ state (e.g., Gly-X-Gly tripeptide conformation). This feature allows for a clear distinction between surface residues and buried residues, as the sites are present on the protein surface, and the addition of this feature narrows the scope of binding sites for prediction tasks. In addition, we believe that proteins are flexible and more likely to interact with other macromolecules. Therefore, secondary structure information is also necessary, so we have selected eight more detailed secondary structures to describe. These two characteristics were predicted by the NetSurfP-3.0 [62] tool, which inputs the entire amino acid sequence to predict the probability of which secondary structure each amino acid belongs to in the sequence, as well as relative solvent accessibility information.

Amino acid propensity: Through a systematic statistical analysis of regions marked as binding sites in the training set, we found significant differences in amino acid composition between different types of protein binding sites. The binding sites of proteins to DNA/RNA exhibited highly similar amino acid distribution characteristics, with the positively charged hydrophilic amino acids lysine (K) and arginine (R) showing considerable enrichment, with their occurrence frequencies reaching 2.6 times and 3 times the average of twenty amino acids, respectively. Further analysis revealed that aspartic acid (D), glutamic acid (E), lysine (K), arginine (R), and leucine (L) were more abundant at protein-protein interaction sites, while antibody-antigen binding sites also displayed a high proportion of glycine (G), serine (S), and threonine (T). It is noteworthy that among these amino acids with a high proportion, leucine (L) is the only strong hydrophobic amino acid; the remaining are all hydrophilic, and more than half are charged amino acids (D/E/K/R), suggesting that electrostatic interactions and hydrophilicity may play an important role in sites recognition.

After that, we have added a symmetry-normalized position code that measures the degree of symmetry of amino acids in the sequence, with outputs ranging from $[0, 1]$. It can be used in protein sequence analysis to enhance a model’s understanding of the global structure.

$$N = \frac{\min(i, L - i)}{L} \cdot 2 \quad (1)$$

i represents the index of residues, while L denotes the total length of the protein sequence.

4.3 | Network Architecture

We adopt an encoder-decoder architecture, which primarily comprises two components: a Multi-Source Feature Fusion (MSFF) module and a Multi-Level Interaction Learning (MIL) module. Below, we provide a detailed description of these modules. While various types of features are introduced to characterize protein sequences, a key challenge lies in effectively processing these heterogeneous features. To address this, we designed the MSFF module, which consists of three parts: a Scale-Aware Encoder module, a Context-Aware Encoder module, and an Importance-Aware Encoder module. The tensors representing different patterns learned by these three networks are then mapped as Queries (Q), Keys (K), and Values (V) and fed into a cross-attention layer for integrative information fusion. In order to simplify the representation of features, we refer to all other features, except for the language model features SiteT5 and ProstT5, as the protein composite features (pcf).

Let us first introduce the Scale-Aware Encoder module. To better learn the relationships between local residues, we designed the Scale - Aware Encoder module. Conv1D mainly captures the contextual representations of residues with local biases and learns global protein features by assembling the local features of all residues. By aggregating the residue representations through convolutional kernels of different sizes, specifically 3, 5, and 7, we can learn information across different receptive fields.

$$f_0 = \text{concat} [f_{pcf}, \text{Linear}(f_{\text{SiteT5}}), \text{Linear}(f_{\text{ProstT5}})] \quad (2)$$

$$f_{i+1} = \text{GLU}(\text{Cov1D}(\text{Dropout}(f_i))) + f_i, i = 0, 1, 2, 3 \quad (3)$$

$$f_{\text{end1}} = \text{LayerNorm}(f_1 \oplus f_2 \oplus f_3) \quad (4)$$

Among these, f_{pcf} represents a 43-dimensional feature, f_{SiteT5} is a 1024-dimensional feature that is transformed into a 102-dimensional feature after passing through a linear layer, and f_{ProstT5} is a 1024-dimensional feature that is reduced to a 102-dimensional feature after a linear layer. f_0 contains features obtained from multi-scale convolution, while $f_i, i = 0, 1, 2, 3$ represents feature tensors after applying different convolution kernels (3, 5, 7). $\oplus$ describes the serial relationship, and f_{end1} is a 64-dimensional tensor.

The Context-Aware Encoder module simultaneously includes a forward LSTM (which processes sequences from past to future) and a backward LSTM (which processes sequences from future to past), ultimately concatenating the hidden states from both directions. This module is capable of capturing past (historical) information while also leveraging future (subsequent) information, selectively memorizing important data (such as features at critical time steps), and ignoring irrelevant data. It is particularly well-suited for handling long sequence data, which is why we have incorporated this module.

$$f_{\text{end2}} = \text{BiLSTM}(\text{concat} [f_{pcf}, \text{Linear}(f_{\text{SiteT5}}), \text{Linear}(f_{\text{ProstT5}})]) \quad (5)$$

To effectively capture the interaction information between long-distance residue pairs in protein sequences and highlight the important residue tensors that contribute significantly to structure or function, we have employed an Importance-Aware Encoder module. The attention module is proficient at capturing long-range information; however, it is unable to dynamically adjust the focus on important features. Through gated units, high-weight residue signals are preserved and enhanced via gating (such as potential residue sites).

$$A_0 = f_0 \quad (6)$$

$$G_{i+1} = \text{sigmoid}(\text{Linear}(\text{concat} [A_i, A_{i+1}, (A_i - A_{i+1})])) \quad (7)$$

$$A_{i+1} = \left(\left(\frac{W_q^i \cdot \text{Linear}(A_i) \cdot (W_k^i \cdot \text{Linear}(A_i))^T}{\sqrt{d_k^i}} \right) W_v^i \cdot \text{Linear}(A_i) \right) \quad (8)$$

$$D_{i+1} = \text{LayerNorm}(A_{i+1} \cdot G_{i+1} + A_i \cdot (1 - G_{i+1})), i = 0, 1 \quad (9)$$

$$f_{\text{end3}} = \text{LayerNorm}(\text{Dropout}(D_1 \oplus D_2)) \quad (10)$$

where A_0 is the feature of the input, which is clearly described in multiscale convolution. A_1, A_2 is the tensor after passing through the 8-head self-attention network, G_1, G_2 is the weight generated by the gated network is a 1-dimensional tensor, D_1, D_2 is the tensor output after the gated weights are multiplied by points, $\oplus$ describes the concatenation relationship, and f_{end3} is a 64-dimensional tensor.

Three networks are then mapped as Queries (Q), Keys (K), and Values (V), and fed into a cross-attention layer for integrative information fusion.

$$f_{\text{end}} = \left(\left(\frac{W_q^f \cdot f_{\text{end1}} \cdot (W_k^f \cdot f_{\text{end2}})^T}{\sqrt{d_k^f}} \right) W_v^f \cdot f_{\text{end3}} \right) \quad (11)$$

f_{end} is the tensor after the 8-head cross-attention module.

The decoder is designed as a Multi-Level Interaction Learning (MIL) module composed of three identical stacked sub-blocks. Each sub-block consists of a gated multi-head self-attention unit and a position-wise feed-forward network (PFFN), enabling the MIL to iteratively aggregate and enhance critical feature information. The integrated features are subsequently passed through a multilayer perceptron (MLP) to produce the final prediction output.

$$f_{\text{end}} = F_0 \quad (12)$$

$$F_{i+1} = \left(\left(\frac{W_q^{F_i} \cdot F_i \cdot (W_k^{F_i} \cdot F_i)^T}{\sqrt{d_k^{F_i}}} \right) W_v^{F_i} \cdot F_i \right) \quad (13)$$

$$G_{F_{i+1}} = \text{sigmoid}(\text{Linear}(\text{concat}[F_i, F_{i+1}, (F_{i+1} - F_i)]) \quad (14)$$

$$D_{F_{i+1}} = \text{LayerNorm}(F_{i+1} \cdot G_{F_{i+1}} + F_{i+1} \cdot (1 - G_{F_{i+1}})) \quad (15)$$

$$D_{\text{PFFN}_{i+1}} = \text{Cov1D}(\text{Dropout}(\text{RELU}(\text{Cov1D}(D_{F_{i+1}})))) \quad (16)$$

$$D_{\text{PFFN}_{i+1}} = F_{i+1} \quad (17)$$

$$S = \text{MLP}(\text{softmax}(\text{Norm}(D_{\text{PFFN}_3})) \cdot D_{\text{PFFN}_3}) \quad (18)$$

F_0 is the tensor after the 8-head cross-attention module, F_i , $i = 0, 1, 2$ is the tensor after the 8-head self-attention, $W_q^{F_i}$ represents the weight matrix, $G_{F_{i+1}}$ is the gating parameter, and $D_{F_{i+1}}$ is the tensor output after the gating layer. $D_{\text{PFFN}_{i+1}}$ is the tensor output from the position-wise feed-forward network, and S is the result output by the MLP.

4.4 | Network Design Intent

Multi-Source Feature Fusion module (MSFF): In this tri-track semantic parsing structure, the Query (Q), Key (K), and Value (V) components serve distinct informational roles: Q determines what the model should focus on, K provides the contextual semantics to be referenced, and V defines the information content to be aggregated and updated in the final representation. Therefore, their corresponding subnetworks are designed to be semantically complementary and functionally aligned. Specifically, the Scale-Aware Encoder module captures local sequence patterns and short-range residue dependencies with different receptive fields, providing high-resolution and position-sensitive features. Hence, it is most suitable as Q (Query), initiating attention queries toward relevant contexts. The Context-Aware Encoder module bidirectional contextual dependencies, offering a smooth and consistent global semantic reference, thus serving as K (Key) to provide globally aligned semantic coordinates for each query. Meanwhile, the Importance-Aware Encoder module performs global aggregation and adaptive weighting, suppressing irrelevant noise while amplifying functionally important signals; it is therefore assigned as V (Value), providing the information that will be integrated during fusion. After mapping to Q, K, and V spaces, these representations are fed into a cross-attention module, where attention weights are computed to dynamically align heterogeneous feature spaces and achieve context-sensitive fusion. This hierarchical information flow (from local feature extraction (Q) to global semantic alignment (K) and final information aggregation (V)) ensures explicit semantic propagation and functional complementarity within the model. Reversing this correspondence would disrupt the intrinsic logic of the attention mechanism, weakening both the interpretability and the effectiveness of the feature fusion process.

Multi-Level Interactive Learning (MIL): Although static feature fusion can integrate multi-modal semantics, it remains insufficient for capturing the dynamic coupling and cooperative effects among residues. To address this limitation, the MIL module adopts a hierarchical three-block architecture, where each block consists of an 8-head self-attention mechanism, a gating mechanism, and a position-wise feed-forward network (PFFN). Within each block, the multi-head self-attention mechanism models

residue relationships across multiple semantic subspaces in parallel, thereby implicitly capturing multi-scale dependencies ranging from local to long-range interactions and enabling dynamic context aggregation. The gating mechanism adaptively modulates attention weights according to feature relevance, suppressing noisy activations while emphasizing functionally critical residues. The subsequent PFFN performs a nonlinear transformation and semantic enhancement on the aggregated features to improve discriminative capacity and generalization. Through progressive three-layer interaction, the MIL module achieves hierarchical semantic evolution (from local dependency learning to global cooperative modeling), enabling ProSiteHunter to capture high-order residue interactions directly from sequence data and deliver superior accuracy, robustness, and biological interpretability in binding-site prediction.

4.5 | Network Hyperparameters

All network parameters are as follows: In the multi-source feature fusion module, sequential features are processed by three distinct networks, the Context-Aware Encoder consists of two LSTM layers, each with 32 units, and a dropout rate of 0.3 is applied between layers to prevent overfitting. As it is implemented as a bidirectional LSTM, the outputs concatenate the hidden states from both the forward and backward directions, resulting in a final tensor with a dimensionality of (L, 64). For the Scale-Aware Encoder module, we have configured three layers of 1D convolutional layers with kernel sizes of 3, 5, and 7. Each convolutional layer leverages residual connections to enhance network stability, producing a tensor of dimension (L, 64). The Importance-Aware Encoder module comprises two concatenated layers of 8-head self-attention, with an inter-layer dropout probability of 0.1. Additionally, gated parameters generated through softmax are set to dynamically adjust in between layers, resulting in a tensor of dimension (L, 64). The features outputted by the three networks are multiplied by their corresponding weights to be reused as Q, K, and V for attention learning. Finally, the MLP includes three hidden layers of fully connected networks.

4.6 | Training

ProSiteHunter is implemented in Python in Pytorch and trained for up to 30 epochs. Since it is a common sequence-based framework, it consumes less resources and the training batch size is 1. The Radam optimizer is used to minimize the loss.

To address the class imbalance problem, we employed the binary Focal Loss function, which assigns different weights to positive and negative samples. Specifically, a higher weight $\alpha_1 = 0.75$ is assigned to the positive class ($y = 1$), while a lower weight $\alpha_0 = 0.25$ is assigned to the negative class ($y = 0$). The loss function is defined as:

$$\text{Loss}(p_i, y_i) = -\alpha_1 y_i (1 - p_i)^{\gamma} \log(p_i) - \alpha_0 (1 - y_i) p_i^{\gamma} \log(1 - p_i), i = 0, 1 \quad (19)$$

where $y_i = 0, 1$ denotes the ground-truth label of residue i , and p_i is the predicted probability for the positive class. The focusing

parameter γ is set to 2, which reduces the contribution of well-classified samples and emphasizes harder examples. The model is optimized by minimizing the overall Focal Loss during training. In addition, we conducted a comparative analysis of BCE Loss and Focal Loss (Figure S10), and the results further support the effectiveness of using Focal Loss in our framework.

The learning rate was initially set to 0.0005 and gradually decreased during training following an exponential decay schedule. The model was trained on an NVIDIA GeForce RTX 3090 GPU.

Author Contributions

G.Z. conceived and supervised the research. Z.Z. and X.Z. helped supervise the research. G.Z., D.H., and X.Z. designed the experiment. G.Z., D.H., Q.Z., Z.L., and S.W. collected the data and performed the experiment. G.Z., D.H., Q.Z., Z.L., and M.H. analyzed the data. G.Z., D.H., and X.C. wrote the manuscript, and all authors read and approved the final manuscript.

Acknowledgements

The authors thank members of the Guijun Zhang lab for discussion and feedback. Computational resources were provided by the College of Information Engineering at Zhejiang University of Technology. This work was supported by the National Key R&D Program of China [2022ZD0115103, G.Z.], the National Nature Science Foundation of China [62173304, G.Z.; 62203389, X.Z.; 32370712, Z.Z.], the “Pioneer” and “Leading Goose” R&D Program of Zhejiang [2025C01190, G.Z.], the Zhejiang Province High-level Talent Special Support Program [2023R5248, G.Z.], and Fundamental Research Funds for the Provincial Universities of Zhejiang [RF-C2024006, X.Z.].

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study and the code of ProSiteHunter are openly available in GitHub at <https://github.com/ibio-zjut/ProSiteHunter>. The online server of ProSiteHunter is made freely available at <http://zhanglab-bioinf.com/ProSiteHunter/>.

References

1. M. Fessenden, “Protein Maps Chart the Causes of Disease,” *Nature* 549 (2017): 293–295, <https://doi.org/10.1038/549293a>.
2. M. Vidal, M. E. Cusick, and A.-L. Barabási, “Interactome Networks and Human Disease,” *Cell* 144 (2011): 986–998, <https://doi.org/10.1016/j.cell.2011.02.016>.
3. C. V. Robinson, A. Sali, and W. Baumeister, “The Molecular Sociology of the Cell,” *Nature* 450 (2007): 973–982, <https://doi.org/10.1038/nature06523>.
4. R. Qi and Q. Zou, “Trends and Potential of Machine Learning and Deep Learning in Drug Study at Single-Cell Level,” *Research* 6 (2023): 0050, <https://doi.org/10.34133/research.0050>.
5. J. Wang, S. Lisanza, D. Juergens, et al., “Scaffolding Protein Functional Sites Using Deep Learning,” *Science* 377 (2022): 387–394, <https://doi.org/10.1126/science.abn2100>.
6. A. Hamdi and P. Colas, “Yeast Two-Hybrid Methods and Their Applications in Drug Discovery,” *Trends in Pharmacological Sciences* 33 (2012): 109–118, <https://doi.org/10.1016/j.tips.2011.10.008>.

7. L. A. Gallagher, E. Velazquez, S. B. Peterson, et al., “Genome-Wide Protein–DNA Interaction Site Mapping in Bacteria Using a Double-Stranded DNA-Specific Cytosine Deaminase,” *Nature Microbiology* 7 (2022): 844–855, <https://doi.org/10.1038/s41564-022-01133-9>.
8. G. D. Stormo and Y. Zhao, “Determining the Specificity of Protein–DNA Interactions,” *Nature Reviews Genetics* 11 (2010): 751–760, <https://doi.org/10.1038/nrg2845>.
9. Z. Wang, D. Xie, D. Wu, et al., “Robust Enzyme Discovery and Engineering With Deep Learning Using CataPro,” *Nature Communications* 16 (2025): 2736, <https://doi.org/10.1038/s41467-025-58038-4>.
10. M. Elavarasu and S. K., “Rational Design of an Epitope-Centric Vaccine Against *Pseudomonas Aeruginosa* Using Pangenomic Insights and Immunoinformatics Approach,” *Frontiers in Immunology* 16 (2025): 1617251, <https://doi.org/10.3389/fimmu.2025.1617251>.
11. J. Mucci, S. J. Carmona, R. Volcovich, et al., “Next-Generation ELISA Diagnostic Assay for Chagas Disease Based on the Combination of Short Peptidic Epitopes,” *PLOS Neglected Tropical Diseases* 11 (2017): 0005972, <https://doi.org/10.1371/journal.pntd.0005972>.
12. X. Liu, K. Salokas, R. G. Weldatsadi, L. Gawrylski, and M. Varjosalo, “Combined Proximity Labeling and Affinity Purification–Mass Spectrometry Workflow for Mapping and Visualizing Protein Interaction Networks,” *Nature Protocols* 15 (2020): 3182–3211, <https://doi.org/10.1038/s41596-020-0365-x>.
13. M. S. Sung, J.-H. Jung, C. Jeong, T.-Y. Yoon, and J.-H. Park, “Single-Molecule Co-Immunoprecipitation Reveals Functional Inheritance of EGFRs in Extracellular Vesicles,” *Small* 14 (2018): 1802358, <https://doi.org/10.1002/smll.201802358>.
14. H.-W. Tang, K. Spirohn, Y. Hu, et al., “Next-Generation Large-Scale Binary Protein Interaction Network for *Drosophila Melanogaster*,” *Nature Communications* 14 (2023): 2162, <https://doi.org/10.1038/s41467-023-37876-0>.
15. L. Giot, J. S. Bader, C. Brouwer, et al., “A Protein Interaction Map of *Drosophila Melanogaster*,” *Science* 302 (2003): 1727–1736, <https://doi.org/10.1126/science.1090289>.
16. 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): e51–e51, <https://doi.org/10.1093/nar/gkab044>.
17. M. Zheng, G. Sun, X. Li, and Y. Fan, “EGPDI: Identifying Protein–DNA Binding Sites Based on Multi-View Graph Embedding Fusion,” *Briefings in Bioinformatics* 25 (2024): bbae330, <https://doi.org/10.1093/bib/bbae330>.
18. L. F. Krapp, L. A. Abriata, F. Cortés Rodríguez, and M. Dal Peraro, “PeSTo: Parameter-Free Geometric Deep Learning for Accurate Prediction of Protein Binding Interfaces,” *Nature Communications* 14 (2023): 2175, <https://doi.org/10.1038/s41467-023-37701-8>.
19. J. Jumper, R. Evans, A. Pritzel, et al., “Highly Accurate Protein Structure Prediction With AlphaFold,” *Nature* 596 (2021): 583–589, <https://doi.org/10.1038/s41586-021-03819-2>.
20. M. Akdel, D. E. V. Pires, E. P. Pardo, et al., “A Structural Biology Community Assessment of AlphaFold2 Applications,” *Nature Structural & Molecular Biology* 29 (2022): 1056–1067, <https://doi.org/10.1038/s41594-022-00849-w>.
21. J. Yu, C. Yan, T. Dodd, et al., “Dynamic Conformational Switching Underlies TFIIH Function in Transcription and DNA Repair and Impacts Genetic Diseases,” *Nature Communications* 14 (2023): 2758, <https://doi.org/10.1038/s41467-023-38416-6>.
22. M. Lazou, O. Khan, T. Nguyen, et al., “Predicting Multiple Conformations of Ligand Binding Sites in Proteins Suggests That AlphaFold2 May Remember Too Much,” *Proceedings of the National Academy of Sciences* 121 (2024): 2412719121, <https://doi.org/10.1073/pnas.2412719121>.
23. J. Zhang and L. Kurgan, “Review and Comparative Assessment of Sequence-Based Predictors of Protein-Binding Residues,” *Briefings in Bioinformatics* 19 (2018): 821–837, <https://doi.org/10.1093/bib/bbx022>.

24. 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, <https://doi.org/10.1126/science.ade2574>.
25. A. Elnaggar, M. Heinzinger, C. Dallago, et al., "ProtTrans: Toward Understanding the Language of Life through Self-Supervised Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44 (2022): 7112–7127, <https://doi.org/10.1109/TPAMI.2021.3095381>.
26. M. Manfredi, C. Savojardo, P. L. Martelli, and R. Casadio, "ISPRED-SEQ: Deep Neural Networks and Embeddings for Predicting Interaction Sites in Protein Sequences," *Journal of Molecular Biology* 435 (2023): 167963, <https://doi.org/10.1016/j.jmb.2023.167963>.
27. A. Rives, J. Meier, T. Sercu, et al., "Biological Structure and Function Emerge From Scaling Unsupervised Learning to 250 Million Protein Sequences," *Proceedings of the National Academy of Sciences* 118 (2021): 2016239118, <https://doi.org/10.1073/pnas.2016239118>.
28. S. Hosseini, G. B. Golding, and L. Ilie, "Seq-InSite: Sequence Supersedes Structure for Protein Interaction Site Prediction," *Bioinformatics* 40 (2024): btad738, <https://doi.org/10.1093/bioinformatics/btad738>.
29. Y. Liu and B. Tian, "Protein–DNA Binding Sites Prediction Based on Pre-Trained Protein Language Model and Contrastive Learning," *Briefings in Bioinformatics* 25 (2024): bbad488, <https://doi.org/10.1093/bib/bbad488>.
30. J. N. Clifford, M. H. Høie, S. Deleuran, B. Peters, M. Nielsen, and P. Marcatili, "BepiPred -3.0: Improved B-Cell Epitope Prediction Using Protein Language Models," *Protein Science* 31 (2022): 4497, <https://doi.org/10.1002/pro.4497>.
31. N. V. Ivanisenko, T. I. Shashkova, A. Shevtsov, M. Sindeeva, D. Umerenkov, and O. Kardymon, "SEMA 2.0: Web-Platform for B-Cell Conformational Epitopes Prediction Using Artificial Intelligence," *Nucleic Acids Research* 52 (2024): W533–W539, <https://doi.org/10.1093/nar/gkae386>.
32. S. Henikoff and J. G. Henikoff, "Amino Acid Substitution Matrices From Protein Blocks," *Proceedings of the National Academy of Sciences* 89 (1992): 10915–10919, <https://doi.org/10.1073/pnas.89.22.10915>.
33. J. Yan and L. Kurgan, "DRNApred, Fast Sequence-Based Method That Accurately Predicts and Discriminates DNA- and RNA-Binding Residues," *Nucleic Acids Research* 45 (2017): e84–e84, <https://doi.org/10.1093/nar/gkx059>.
34. N. Wang, K. Yan, J. Zhang, and B. Liu, "iDRNA-ITF: Identifying DNA- and RNA-Binding Residues in Proteins Based on Induction and Transfer Framework," *Briefings in Bioinformatics* 23 (2022): bbac236, <https://doi.org/10.1093/bib/bbac236>.
35. T. I. Shashkova, D. Umerenkov, M. Salnikov, et al., "SEMA: Antigen B-Cell Conformational Epitope Prediction Using Deep Transfer Learning," *Frontiers in Immunology* 13 (2022): 960985, <https://doi.org/10.3389/fimmu.2022.960985>.
36. S. Hosseini and L. Ilie, "PITHIA: Protein Interaction Site Prediction Using Multiple Sequence Alignments and Attention," *International Journal of Molecular Sciences* 23 (2022): 12814, <https://doi.org/10.3390/ijms232112814>.
37. J. Hu, M. Dong, Y.-X. Tang, and G.-J. Zhang, "Improving Protein-Protein Interaction Site Prediction Using Deep Residual Neural Network," *Analytical Biochemistry* 670 (2023): 115132, <https://doi.org/10.1016/j.ab.2023.115132>.
38. 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, <https://doi.org/10.1093/bioinformatics/btq302>.
39. K. Dhole, G. Singh, P. P. Pai, and S. Mondal, "Sequence-Based Prediction of Protein–Protein Interaction Sites With L1-Logreg Classifier," *Journal of Theoretical Biology* 348 (2014): 47–54, <https://doi.org/10.1016/j.jtbi.2014.01.028>.
40. S. Israeli and Y. Louzoun, "Single-Residue Linear and Conformational B Cell Epitopes Prediction Using Random and ESM-2 Based Projections," *Briefings in Bioinformatics* 25 (2024): bbae084, <https://doi.org/10.1093/bib/bbae084>.
41. M. G. Juarez, S. M. O'Rourke, J. V. Dzimiński, et al., "Structures of Respiratory Syncytial Virus G Bound to Broadly Reactive Antibodies Provide Insights Into Vaccine Design," *Scientific Reports* 15 (2025): 8666, <https://doi.org/10.1038/s41598-025-92886-w>.
42. J. Abramson, J. Adler, J. Dunger, et al., "Accurate Structure Prediction of Biomolecular Interactions With AlphaFold 3," *Nature* 630 (2024): 493–500, <https://doi.org/10.1038/s41586-024-07487-w>.
43. L. V. Maaten and G. Hinton, "Visualizing Data Using T-SNE," *Journal of Machine Learning Research* 9 (2008): 2579–2605, <http://jmlr.org/papers/v9/vandemaaten08a.html>.
44. M. Heinzinger, K. Weissenow, J. G. Sanchez, et al., "Bilingual Language Model for Protein Sequence and Structure," *NAR Genomics and Bioinformatics* 6 (2024): lqae150, <https://doi.org/10.1093/nargab/lqae150>.
45. M. van Kempen, S. S. Kim, C. Tumescheit, et al., "Fast and Accurate Protein Structure Search With Foldseek," *Nature Biotechnology* 42 (2024): 243–246, <https://doi.org/10.1038/s41587-023-01773-0>.
46. H. Bhukya, A. K. Jana, N. Sengupta, and R. Anand, "Structural and Dynamics Studies of the TetR Family Protein, CprB From Streptomyces Coelicolor in Complex With Its Biological Operator Sequence," *Journal of Structural Biology* 198 (2017): 134–146, <https://doi.org/10.1016/j.jsb.2017.03.006>.
47. C. D. Eichhorn, Y. Yang, L. Repeta, and J. Feigon, "Structural Basis for Recognition of Human 7SK Long Noncoding RNA by the La-Related Protein Larpp7," *Proceedings of the National Academy of Sciences* 115 (2018): E6457–E6466, <https://doi.org/10.1073/pnas.1806276115>.
48. R. J. Cater, G. L. Chua, S. K. Erramilli, et al., "Structural Basis of Omega-3 Fatty Acid Transport Across the Blood–Brain Barrier," *Nature* 595 (2021): 315–319, <https://doi.org/10.1038/s41586-021-03650-9>.
49. I. Redzynia, A. Ljunggren, A. Bujacz, M. Abrahamson, M. Jaskolski, and G. Bujacz, "Crystal Structure of the Parasite Inhibitor Chagasin in Complex With Papain Allows Identification of Structural Requirements for Broad Reactivity and Specificity Determinants for Target Proteases," *The FEBS Journal* 276 (2009): 793–806, <https://doi.org/10.1111/j.1742-4658.2008.06824.x>.
50. J. D. Steemson, M. Baake, J. Rakonjac, V. L. Arcus, and M. T. Liddament, "Tracking Molecular Recognition at the Atomic Level With a New Protein Scaffold Based on the OB-Fold," *PLoS ONE* 9 (2014): 86050, <https://doi.org/10.1371/journal.pone.0086050>.
51. A. Yokota, K. Tsumoto, M. Shiroishi, H. Kondo, and I. Kumagai, "The Role of Hydrogen Bonding via Interfacial Water Molecules in Antigen–Antibody Complexation," *Journal of Biological Chemistry* 278 (2003): 5410–5418, <https://doi.org/10.1074/jbc.M210182200>.
52. D. M. Mason, S. Friedensohn, C. R. Weber, et al., "Optimization of Therapeutic Antibodies by Predicting Antigen Specificity From Antibody Sequence via Deep Learning," *Nature Biomedical Engineering* 5 (2021): 600–612, <https://doi.org/10.1038/s41551-021-00699-9>.
53. Z. Lv, D. Hou, M. Hou, et al., "MultiSAAI: Sequence-Informed Antibody–Antigen Interactions Prediction Using Multi-Scale Deep Learning," *BioRxiv* (2025), <https://doi.org/10.1101/2025.05.29.656915>.
54. J. Sim, J. T. Sockolosky, E. Sangalang, et al., "Discovery of High Affinity, Pan-Allelic, and Pan-Mammalian Reactive Antibodies Against the Myeloid Checkpoint Receptor SIRPα," *mAbs* 11 (2019): 1036–1052, <https://doi.org/10.1080/19420862.2019.1624123>.
55. N. Tzarum, E. Giang, L. Kong, et al., "Genetic and Structural Insights Into Broad Neutralization of Hepatitis C Virus by Human V_H1-69 Antibodies," *Science Advances* 5 (2019): aav1882, <https://doi.org/10.1126/sciadv.aav1882>.
56. M. Higurashi, T. Ishida, and K. Kinoshita, "PiSite: A Database of Protein Interaction Sites Using Multiple Binding States in the PDB,"

Nucleic Acids Research 37 (2009): D360–D364, <https://doi.org/10.1093/nar/gkn659>.

57. L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, “CD-HIT: Accelerated for Clustering the Next-Generation Sequencing Data,” *Bioinformatics* 28 (2012): 3150–3152, <https://doi.org/10.1093/bioinformatics/bts565>.

58. M. Hauser, M. Steinegger, and J. Söding, “MMseqs Software Suite for Fast and Deep Clustering and Searching of Large Protein Sequence Sets,” *Bioinformatics* 32 (2016): 1323–1330, <https://doi.org/10.1093/bioinformatics/btw006>.

59. J. Dunbar and C. M. Deane, “ANARCI: Antigen Receptor Numbering and Receptor Classification,” *Bioinformatics* 32 (2015): 298–300, <https://doi.org/10.1093/bioinformatics/btv552>.

60. M. Remmert, A. Biegert, A. Hauser, and J. Söding, “HHblits: Lightning-Fast Iterative Protein Sequence Searching by HMM-HMM Alignment,” *Nature Methods* 9 (2012): 173–175, <https://doi.org/10.1038/nmeth.1818>.

61. M. Varadi, S. Anyango, M. Deshpande, et al., “AlphaFold Protein Structure Database: Massively Expanding the Structural Coverage of Protein-Sequence Space With High-Accuracy Models,” *Nucleic Acids Research* 50 (2021): D439–D444, <https://doi.org/10.1093/nar/gkab1061>.

62. M. H. Høie, E. N. Kiehl, B. Petersen, et al., “NetSurfP-3.0: Accurate and Fast Prediction of Protein Structural Features by Protein Language Models and Deep Learning,” *Nucleic Acids Research* 50 (2022): W510–W515, <https://doi.org/10.1093/nar/gkac439>.

Supporting Information

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

Supporting File: advs75931-sup-0001-SuppMat.docx.