Journal of Cheminformatics

RESEARCH

Open Access

Protein-small molecule binding site prediction based on a pre-trained protein language model with contrastive learning

Jue Wang[1†], Yufan Liu[2†] and Boxue Tian[1*]

Abstract

Predicting protein-small molecule binding sites, the initial step in structure-guided drug design, remains challenging for proteins lacking experimentally derived ligand-bound structures. Here, we propose CLAPE-SMB, which integrates a pre-trained protein language model with contrastive learning to provide high accuracy predictions of small molecule binding sites that can accommodate proteins without a published crystal structure. We trained and tested CLAPE-SMB on the SJC dataset, a non-redundant dataset based on sc-PDB, JOINED, and COACH420, and achieved an MCC of 0.529. We also compiled the UniProtSMB dataset, which merges sites from similar proteins based on raw data from UniProtKB database, and achieved an MCC of 0.699 on the test set. In addition, CLAPE-SMB achieved an MCC of 0.815 on our intrinsically disordered protein (IDP) dataset that contains 336 non-redundant sequences. Case studies of DAPK1, RebH, and Nep1 support the potential of this binding site prediction tool to aid in drug design. The code and datasets are freely available at https://github.com/JueWangTHU/CLAPE-SMB.

Scientific contribution CLAPE-SMB combines a pre-trained protein language model with contrastive learning to accurately predict protein-small molecule binding sites, especially for proteins without experimental structures, such as IDPs. Trained across various datasets, this model shows strong adaptability, making it a valuable tool for advancing drug design and understanding protein-small molecule interactions.

Keywords Protein-small molecule binding site prediction, Protein language model, Contrastive learning, Intrinsically disordered proteins, Drug discovery

Introduction

Protein-small molecule interactions regulate almost all biological processes. Detailed characterization of small molecule binding interactions with proteins is essential for engineering or manipulating biological processes and designing targeted small molecule protein inhibitors or activators [1–5]. However, it is first necessary to determine the small molecule binding sites on a protein of interest. Several experimental methods have been established for binding site identification, including surface plasmon resonance (SPR), isothermal titration calorimetry (ITC), and hydrogen deuterium exchange mass spectrometry (HDX-MS) [6–8]. In addition, structural biology methods like X-ray crystallography, NMR

† Jue Wang and Yufan Liu have contributed equally to this work.

*Correspondence:
Boxue Tian
boxuetian@mail.tsinghua.edu.cn
[1] MOE Key Laboratory of Bioinformatics, State Key Laboratory of Molecular Oncology, Beijing Frontier Research Center for Biological Structure, School of Pharmaceutical Sciences, Tsinghua University, Beijing 100084, China
[2] Computer Science Program, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia

spectroscopy, and cryo-electron microscopy (Cryo-EM) that generate high-resolution three-dimensional structures of proteins in complex with their small molecule ligands have also helped advanced our understanding of protein-small molecule interactions [9]. While these experimental methods offer unparalleled precision and irreplaceable validation capabilities, they require substantial commitments of time, labor, and resources [10]. To focus investigation and reduce efforts required for screening, computational methods for predicting protein-small molecule binding sites have emerged as a complementary approach to the above experimental methods.

Protein-small molecule binding site prediction models can be generally classified as either structure-based or sequence-based. Structure-based models rely on the 3D structure of a protein as input. For example, ScanNet [11] utilized an interpretable geometric deep learning model and built representations of atoms and amino acids based on the spatio-chemical arrangement of their neighbours. P2Rank [12] uses Random Forest (RF) modeling to classify the points evenly spread on the Solvent Accessible Surface (SAS) of proteins. DeepSite [13] treats the protein structure as a 3D voxel and uses a deep convolutional neural network (DCNN) as the classifier. However, the preparation of protein structures can be time-consuming [14]. Alternatively, for proteins without an experimentally determined structure, AlphaFold [15] can be used to rapidly generate a protein structure for downstream binding site prediction, albeit with some potential loss of accuracy [16, 17].

Compared to structure-based binding site prediction, sequence-based models offer a simpler and more user-friendly approach that only require the protein sequence as input, typically in FASTA or plain text format [18–20]. However, these models often have relatively low predictive accuracy [21]. Notably, GraphBind [17] combines both sequence- and structure-based features, employing a graph neural network (GNN) for classification. DeepProSite [22] leverages ESMFold [23] and the ProtTrans T5 language model [24] to extract features from protein sequence, followed by prediction using a multilayer perceptron (MLP). However, a combination of both sequence- and structure-based features inevitably leads to redundancies in the input data and computational inefficiency [25–27].

To address the limitations in existing methods, in this study we propose Contrastive Learning And Pre-trained Encoder for Small Molecule Binding (CLAPE-SMB) based on our previous method, CLAPE, which is primarily used for predicting protein-DNA binding sites [21]. We used a pre-trained protein language model, Evolutional Scale Modeling-2 (ESM-2) [23], to extract sequence features, followed by an MLP for binding site classification. We also applied triplet center loss (TCL) as a contrastive learning technique [28] to improve prediction accuracy. CLAPE-SMB was first trained on the CHEN11 datasets [12, 29], and later showed high accuracy in binding site prediction using the COACH420 test sets [12, 30], on par with the best current models. Then, we integrated three benchmark datasets, sc-PDB, JOINED, and COACH420 [12, 30, 31] into a new dataset, i.e. SJC, on which CLAPE-SMB showed high accuracy. We also constructed the UniProtSMB dataset, in which we merged multiple sites from similar proteins. CLAPE-SMB showed high accuracy on this new dataset, thus illustrating the effects of data quality, as well as the combination of a pre-trained protein language model and contrastive learning on the accuracy of protein-small molecule binding site prediction. The model constructed here can facilitate basic research of small molecule function in biological systems, and also guide drug development.

Methods

Sequence embedding

ESM-2 [23], a protein language model pre-trained on a large number of protein sequences, has been used to obtain the initial representation of proteins. As a 1D-based protein sequence model, ESM-2 has been extensively used for predicting various properties of proteins. Deeper layers in ESM-2 tend to focus more on binding sites and contacts, which are high-level concepts related to protein folding. In contrast, secondary structures, which are lower- to mid-level concepts, are targeted more evenly across layers [32]. This suggests that ESM-2 can capture important aspects of protein folding, therefore generating high-accuracy predictions of various properties.

In this study, we used the esm2_t33_650M_UR50D version of ESM-2, which consists of 33 layers and a total of 650 million parameters without fine-tuning the ESM-2 layers. Protein sequences containing 21 tokens, including 20 standard amino acids and one special token 'X' to represent unknown residues, were fed into ESM-2 to obtain 1,280-dimensional embeddings for subsequent training steps.

Backbone model

The backbone model of CLAPE-SMB was an MLP consisting of 5 fully connected layers, an activation function, layer normalization, and Softmax function. The initial dimension was 1,280, and the output dimensions of the 5 layers were 1,024, 256, 128, 64, and 2, respectively. Rectified Linear Unit (ReLU), an activation function

applied after each layer, introduced nonlinearity into the model. We used dropout of 0.3 to prevent overfitting and enhance generalizability. Layer normalizations ensured stable training by maintaining consistent mean and variance. A Softmax function was used to process output of the last layer to get a mutually exclusive prediction score between 0 to 1, representing the predicted probability for small molecule binding site classification.

Loss function

In this study, the loss function is expressed as:

$$L = L_{focal} + \lambda L_{tc} \tag{1}$$

where L_{focal} and L_{tc} represent class-balanced focal loss [33, 34] and TCL [28] respectively. Both were used to address the issue of data imbalance, since the binding sites ratio was lower than 5%. We used a weight, λ, to balance the different parts of the loss function.

The class-balanced focal loss was originally introduced by Cui et al. [34] based on focal loss, which was first proposed by Lin et al. [33]. It can be expressed as follows:

$$L_{focal} = -\frac{1-\beta}{1-\beta^{n_y}} \sum_{i=1}^{C} \left(1 - p_i^t\right)^{\gamma} log\left(p_i^t\right) \tag{2}$$

In the formula, p_i^t is the probability of a particular classification. $\left(1 - p_i^t\right)^{\gamma}$ is a modulating factor, representing different emphasis we placed on different classes. γ is a hyperparameter called focusing parameter, which was set to 3 in our study after searching. $\frac{1-\beta}{1-\beta^{n_y}}$, noted as E_n, is the effective number of the class proposed by Cui et al.β was set to 0.999 in our study to obtain best model performance according to their work.

TCL is a supervised contrastive learning loss introduced by He et al. [28], which aims to distinguish small molecule binding sites (positive samples) and non-binding sites (negative samples) better. Imagine a protein sequence with M amino acids. In this context, we treat the entire sequence as a batch, where each amino acid is a sample within the batch. Each amino acid in the sequence has an associated feature vector that captures the characteristics of that amino acid. TCL maintains center points for both the positive and negative classes, which are initially chosen at random and optimized during training. TCL can be calculated as follows:

$$L_{tc} = \sum_{i=1}^{M} \max\left(D\left(f_i, c_{y^i}\right) + m - D\left(f_i, c_{1-y^i}\right), 0\right) \tag{3}$$

For each amino acid (i) in the batch, y^i represents this amino acid is positive or negative. $D\left(f_i, c_{y^i}\right)$ represents the Euclidean distance between predicted probability (f_i) and its cluster center (c_{y^i}) of class y^i in the embedding space. $D(f_i, c_{1-y^i})$, represents the Euclidean distance between predicted probability (f_i) and the opposite cluster center (c_{1-y^i}) of the opposite class of y^i. To be specific, if an amino acid is binding site, y^i is positive and $1 - y^i$ is negative. And m represents margin, a hyperparameters in TCL. And then each TCL of each amino acid was added to get final TCL of this batch. It is minimized when $D\left(f_i, c_{y^i}\right)$ is very small and $D\left(f_i, c_{1-y^i}\right)$ is very large. To be specific, $D\left(f_i, c_{1-y^i}\right)$ should be larger than $D\left(f_i, c_{y^i}\right)$ by at least a hyperparameter margin (m). This means that the binding residues are far from non-binding residues in the feature space. This separation in the feature space makes it easier for the model to distinguish between binding and non-binding residues, leading to more accurate predictions.

Evaluation metrics

Precision (Pre), recall (Rec), Matthews correlation coefficient (MCC), area under receiver operating characteristic (ROC) curve (AUROC) and area under precision-recall (PR) curve (AUPRC) are commonly-used evaluation metrics used in this classification task to measure the generalizability of our model and facilitate comparative analyses with previous models.

The three threshold-dependent evaluation metrics are expressed as:

$$Pre = \frac{TP}{TP+FP} \tag{4}$$

$$Rec = \frac{TP}{TP+FN} \tag{5}$$

$$MCC = \frac{TP \times TN - FN \times FP}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}} \tag{6}$$

In these formulas, TP represents true positive, indicating the number of residues that are correctly classified as small molecule binding sites. TN represents true negative, indicating the number of residues that are correctly classified as non-binding sites. FP represents false positive, indicating the number of residues that are incorrectly classified as small molecule binding sites. FN represents false negative, indicating the number of residues that are incorrectly classified as non-binding sites. Therefore, Pre represents the precision of our positive predictions. Rec is the ratio of small molecule binding sites successfully identified by our model. MCC and F1-score are both overall measures of prediction ability from both positive and negative aspects. Notably, we used MCC rather than F1-score due to the issue of data imbalance [35].

To better evaluate our model, we plotted ROC curve and PR curve to obtain an overall intuitive measurement.

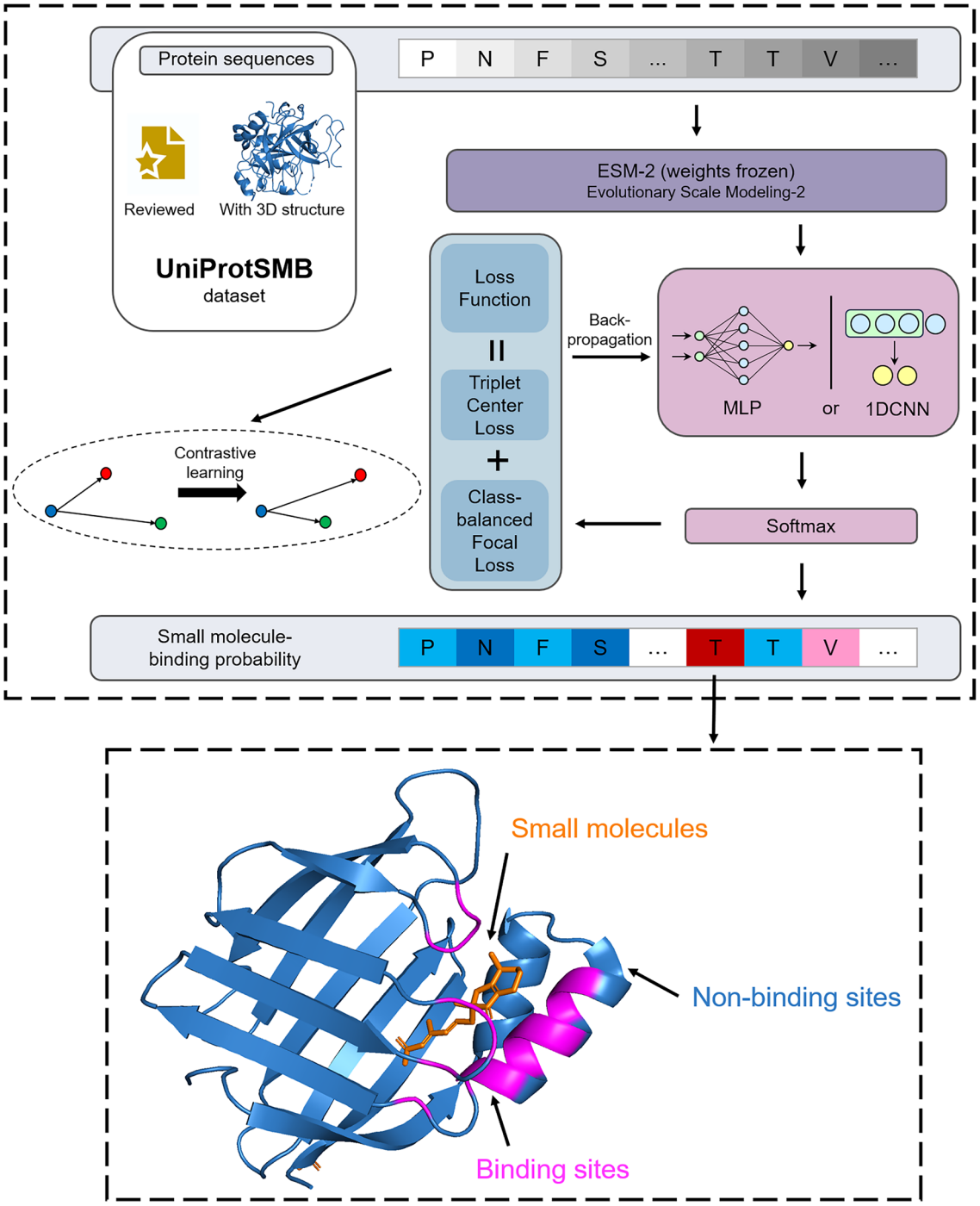

Fig. 1 Structure of the CLAPE-SMB model, comprising three primary components: a sequence embedding module based on the large, pre-trained protein language model, ESM-2, with its weights kept fixed during training; a backbone neural network employing either MLP or 1DCNN, both integrated with a Softmax function serving as the classification head; and a loss function module employing a combination of triplet center loss and class-balanced focal loss. The output is the probability of small molecule binding at each residue of in input protein sequence, with > 50% probability classified as a likely binding site

We calculated the area under these curves, i.e. AUROC and AUPRC, to obtain a quantitative result. Notably, in the presence of data imbalance, AUROC is less critical, while MCC plays a more decisive role [36, 37].

Results

The model architecture of CLAPE-SMB

The model architecture of CLAPE-SMB comprises three main modules (Fig. 1). In the first of these, protein sequence is fed into a pre-trained large language model, ESM-2, to extract informative features relevant to binding sites. Next, the extracted features are channeled into a 5-layer MLP, followed by a Softmax function to predict the probabilities of binding at each individual residue in the protein sequence. Finally, we included a loss function that to accommodate both contrastive loss [28] and class-balanced focal loss [33, 34] to address potential class imbalance issues in the training data. The back-propagation process solely updates the weights of the MLP, keeping the pre-trained ESM-2 model fixed, as previous studies have established that protein language models can automatically learn binding site information [38]. Fine-tuning a pre-trained model such as ESM-2 is computationally expensive and can sometimes lead to a phenomenon known as "catastrophic forgetting" [39, 40]. Keeping ESM-2 fixed can reduce the computational burden and overfitting in the model, thus highlighting the advantage of this design strategy. Additionally, CLAPE-SMB offers flexibility in the contrastive loss function. Alternative options such as InfoNCE [41], IOCRec [42] or LightGCL [43] are also viable choices for the loss function.

SJC dataset preparation

In this study, we employed four publicly available datasets, including sc-PDB, JOINED, COACH420, and CHEN11 [12, 29–31], to train and evaluate the performance of CLAPE-SMB (see Supplementary Table 1 for detailed statistics). Sc-PDB is an annotated repository of druggable binding sites within the Protein Data Bank (PDB), comprising 4,305 proteins with 72,579 small-molecule binding sites and 1,665,261 non-binding sites. The JOINED dataset, introduced for ligand-binding site prediction based on SAS points in the P2Rank study [12], is a relatively large compilation of several smaller datasets containing 560 proteins with 21,798 small-molecule binding sites and 156,286 non-binding sites. COACH420, a subset of the COACH test set processed in the P2Rank study, consists of 420 proteins with 5,376 small-molecule binding sites and 113,767 non-binding sites. CHEN11, containing 251 proteins, 3,592 small-molecule binding sites, and 56,930 non-binding sites, was first introduced in the LBS prediction benchmarking study [29].

The CHEN11 dataset was used in this study for direct comparison because it is the training set of GraphBind. The other three datasets sc-PDB, JOINED, and COACH420 were integrated into a new dataset named SJC (summarized in Table 1), which was taken from the first letter of each dataset name. To enhance model robustness, all protein sequences from the three datasets were processed using UCLUST [44], which can be used directly via the command line by following the provided instructions (https://www.drive5.com/usearch/manual/uclust_algo.html), with a sequence similarity cutoff of 50%. UCLUST was chosen because it offers several advantages over the commonly used CD-HIT program [45], including faster processing, lower memory consumption, better sensitivity, and the ability to cluster at lower identity thresholds while handling larger datasets [44]. This processing step ensured that our dataset did not contain sequences with more than 50% similarity, allowing for an accurate evaluation of the model's performance and preventing overly optimistic results. Then these non-redundant sequences were further divided into training (80%), validation (10%), and test (10%) sets.

UniProtSMB dataset preparation

To further evaluate our CLAPE-SMB model, we created the large UniProtSMB dataset (summarized in Table 1)

Table 1 Summary of protein-small molecule binding sites in SJC and UniProtSMB

Dataset		Small molecule binding residues	Non-binding residues	% of binding residues	Average length
SJC	Train	50,031	1,081,128	4.42	393
	Valid	5,974	136,029	4.21	403
	Test	6,153	138,259	4.26	410
UniProtSMB	Raw	91,187	3,158,304	2.81	415
	Train	46,294	1,636,512	2.75	424
	Valid	5,898	199,320	2.87	414
	Test	5,568	200,280	2.70	415

by collecting all experimentally supported proteins with 3D structures from the UniProtKB database [46] and identifying their small-molecule binding sites. In the UniProtKB dataset, for each protein with binding sites, the "Feature" section provides a table showing which amino acid on the protein sequence binds to which ligand. DNA, RNA, peptides, polymers, metal ions and Fe-S clusters were excluded because they are not small molecules. Through this process, we obtained a dataset of 7,828 sequences, i.e. the raw UniProtSMB dataset.

To identify potentially overlooked binding residues, the raw UniProtSMB dataset was further processed as following. We first clustered proteins with a sequence similarity cutoff of 50% using UCLUST. This threshold, commonly used alongside others like 30%, 90%, and 100% in datasets such as UniRef [47], was chosen to balance two key factors: (1) preventing overly similar sequences in test and training sets, which could lead to overoptimistic results, and (2) avoiding the risks of merging binding sites from lower thresholds with dissimilar sequences and compromised reliability. We also evaluated model performance using different similarity thresholds, with results aligning with our theoretical analysis that supports 50% as a suitable choice, as shown in Supplementary Table 2. We examine generalizability of our model by fixing the clustering threshold (e.g., 50%) and varying the similarity threshold for dataset splitting (e.g., lowering it to 30%). This approach ensured that sequences within the same cluster did not appear in both the training and test sets. Despite the low similarity between train and test data, the model's performance, as measured by MCC, showed only a slight decrease (seen in Supplementary Fig. 1), demonstrating the model's high generalizability.

Next, we conducted a multiple sequence alignment of all proteins within each cluster using MAFFT [48]. We then identified the longest sequence to serve as the center sequence, onto which we integrated the small molecule binding sites from all other sequences in the same cluster. This process resulted in a single representative protein sequence that contained all binding sites for each cluster. The final UniProtSMB dataset has 4,964 proteins, which was further divided into training (80%), validation (10%), and test (10%) sets. The training set comprised 3,972 proteins with 46,294 small-molecule binding sites and 1,636,512 non-binding sites. The validation set included 496 proteins with 5,898 small-molecule binding sites and 199,320 non-binding sites. The test set contained 496 proteins with 5,568 small-molecule binding sites and 200,280 non-binding sites. The overall process is summarized in Fig. 2.

Necessity of merging multiple binding sites of similar proteins

High sequence similarity often indicates structural similarity [49], and proteins with similar structures frequently contain the same (i.e., conserved) binding sites. However, datasets may contain homologous protein sequences with identical or highly similar amino acid sequences, but only annotate a portion of the sequence as the binding site based on interaction with a specific ligand, while labeling the remaining sequence as non-binding sites, regardless of possible binding with other small molecule ligands. To address this inconsistency, we clustered proteins with a sequence similarity cutoff of 50%, following methods in a previous study [50], then aligned and merged sequences into a single central sequence within each cluster. Figure 3 illustrates the rationale for this procedure using two specific examples. We labeled specific protein sequence fragments as binding sites according to their homologous binding sites though they were marked as non-binding sites in the original dataset. It is possible that these later-labeled binding sites could enhance the performance of our model and provide some assistance in inference.

To further verify that merging was necessary to ensure the predictive accuracy of our model, we analyzed the raw UniProtSMB dataset. Among 7,828 proteins, 4,241 exhibited over 50% similarity with others. Merging these proteins resulted in 1,377 unique sequences, with 383 sequences gained new binding sites. Therefore, we recommend the continuous updating of all existing protein-small molecule binding site datasets through experimental validation to guarantee accuracy.

Ligand analysis of binding sites and its impact on prediction accuracy

We analyzed the binding ligands of all proteins in UniProtSMB. We calculated the frequency of each small molecule and displayed the top 8 in Fig. 4a. The results showed that ATP ranked first, possibly because a large number of kinases require ATP as a substrate [51]. GTP and cAMP are related to GPCRs, the largest family of membrane proteins and one of the most promising drug targets due to their crucial role in numerous physiological processes [52]. Moreover, coenzyme such as NAD^+, $NADP^+$, FAD, and FMN also frequently appeared, playing a role in transmitting electrons, atoms, or functional groups in proteins [53]. Next, we categorized all small molecules into several main categories (Fig. 4c), from majority to least, including nucleotide and its derivatives, coenzyme and cofactors, amino acids, and carbohydrate, which account for 84.5% of all ligands. The remaining 15.5% mainly included marketed and clinically tested drug molecules, lipids, or low molecular weight acids and esters.

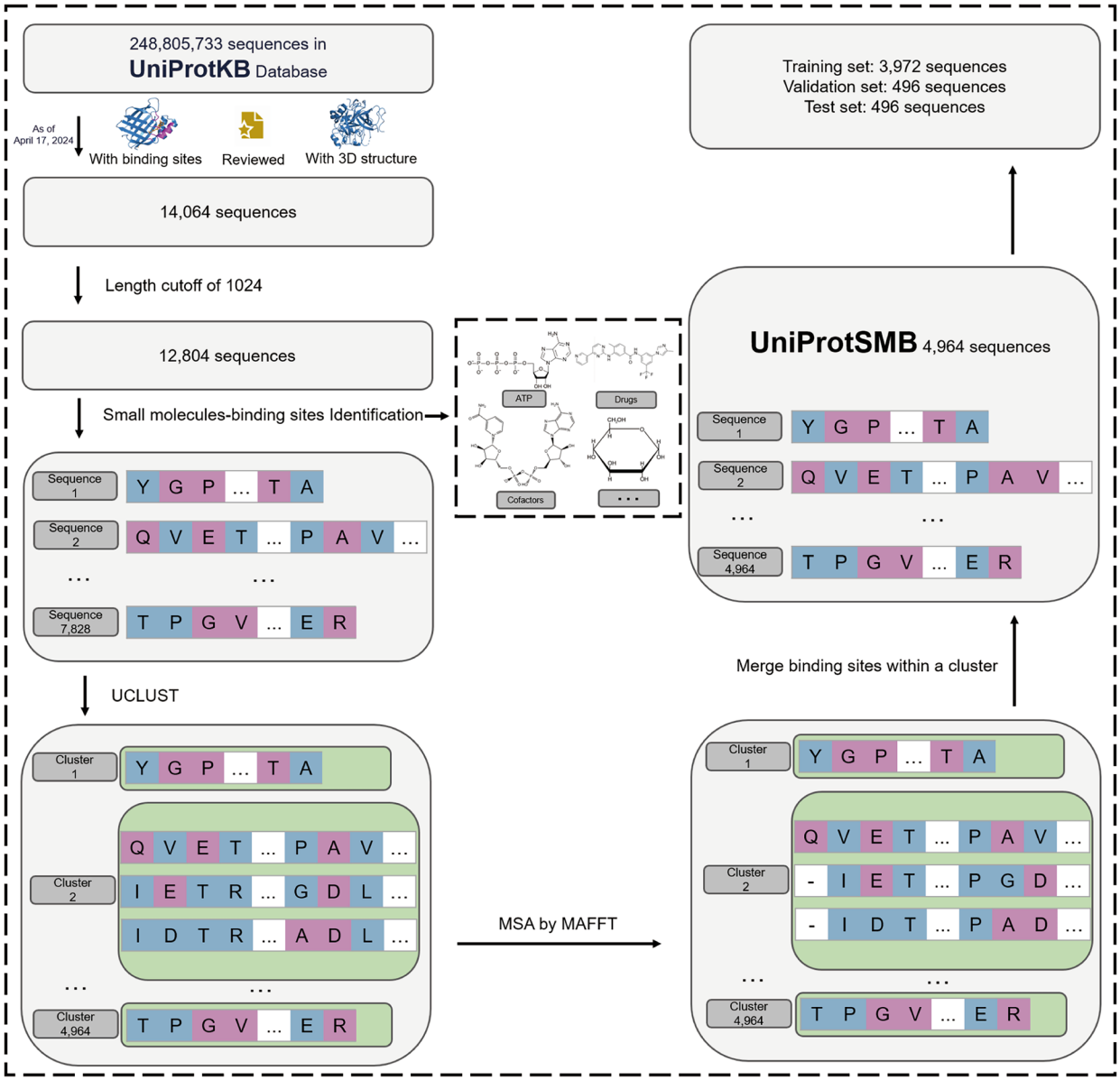

Fig. 2 Pipeline for assembly and curation of the UniProtSMB dataset. First, we collected 14,064 experimentally supported proteins with 3D structures and small molecule binding sites from among 248,805,733 total proteins in the UniProtKB database as of April 17, 2024. After removing proteins longer than 1,024 amino acids, we examined UniProtKB annotations to collect binding site information, including the relevant residues, drugs, cofactors, ATP, and other small molecule ligands. A total of 7,828 small molecules binding protein sequences were collected in this step. Residues involved in binding (pink) or not involved in small molecule binding (blue) were labeled in the sequence of each protein. We then clustered proteins with a sequence similarity cutoff of 50% using UCLUST, which resulted in 4,964 sequence clusters. All proteins within each cluster were subsequently aligned by MAFFT and all binding sites in each cluster were merged onto the longest sequence in that cluster, resulting in a final total set of 4,964 proteins. Finally, the resulting UniProtSMB dataset was divided into a training set (3,972 proteins), a validation set (496 proteins) and a test set (496 proteins)

To evaluate the ability of CLAPE-SMB in predicting specific type of small molecule binding sites, we organized four major types of specific small molecules binding site subsets: nucleotide and its derivatives, coenzyme and cofactors, amino acids, and carbohydrate. In this process, it was common to find proteins that lacked specific binding sites for certain small molecules, such as nucleotides. We removed these protein sequences from the corresponding dataset. Therefore, these subsets were much smaller than UniProtSMB. However, evaluation results

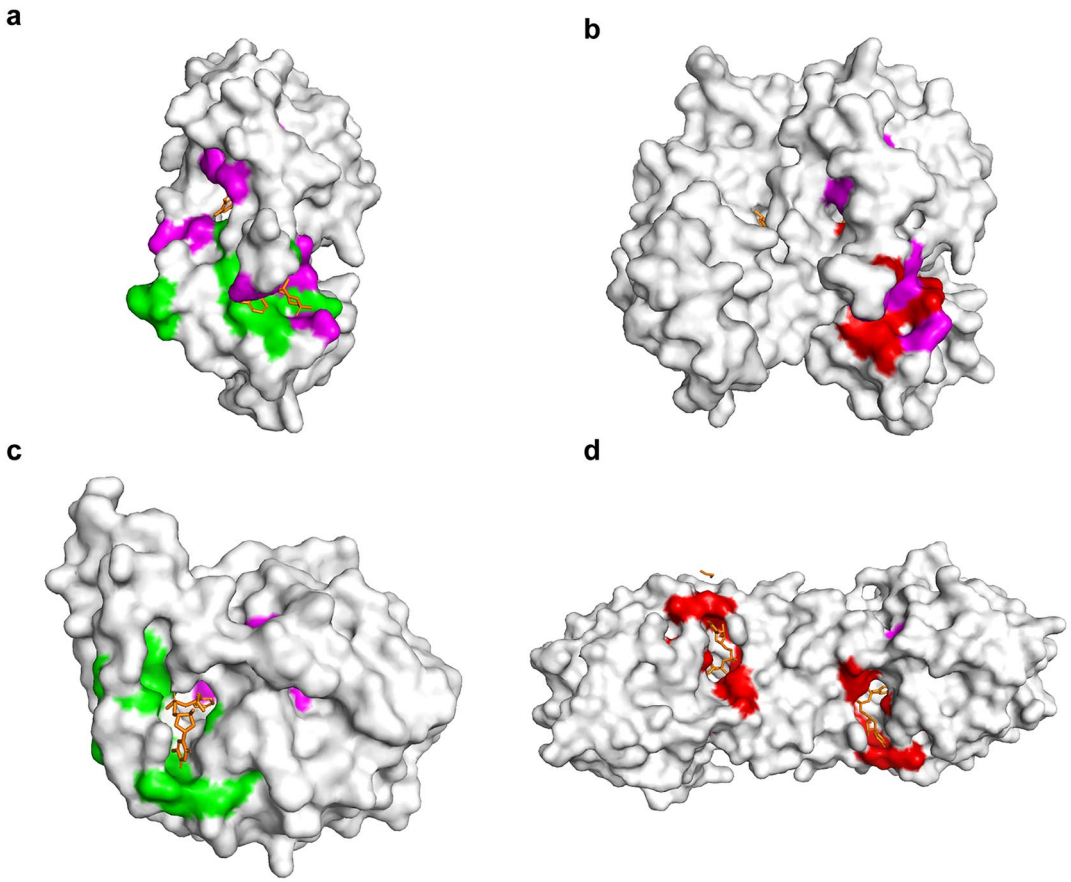

Fig. 3 Example illustration of binding site merging results. **a** P55222 and **b** P0ACJ8 share a sequence similarity of 66.8%. Magenta residues represent homologous binding sites between these sequences that were marked as small molecule binding sites for both proteins in UniProtKB. Green residues in **a** are homologous with red residues in **b** which were marked as non-binding sites for P55222 but binding sites for P0ACJ8. Thus, merged proteins include such binding sites that were annotated on one member of cluster but not others, so that the final merged protein of each cluster contains all known binding sites in that cluster. An additional example of merging small molecule binding sites within a cluster is shown for **c** P95748 and **d** Q9ZGH6 which share a sequence similarity of 60.9%

showed that CLAPE-SMB performed as well as, or even better on these smaller subsets (Fig. 4b).

We conducted a simple amino acid analysis (Fig. 4d–g) on these datasets to explore the reasons behind this phenomenon. On one hand, nucleotide and its derivatives subset (Fig. 4d), amino acids subset (Fig. 4f), and carbohydrate subset (Fig. 4g) showed concentrated distribution of amino acids at binding sites, making model prediction easier and more accurate. In fact, CLAPE-SMB achieved excellent MCC of 0.744 and 0.680 on nucleotide and its derivatives subset (2,099 sequences) and amino acids subset (600 sequences), respectively, compared to an MCC of 0.699 on the large UniProtSMB dataset (4,964 sequences). However, MCC on carbohydrate subset was only 0.561 because the sample size is simply too small (165 sequences). On the other hand, coenzyme and cofactors subset did not show clear pattern of binding site distribution,

(See figure on next page.)

Fig. 4 Ligand analysis of binding sites and its impact on model performance. **a** The top 8 small molecules that bind to proteins: ATP, GTP, NAD^+, $NADP^+$, FAD, cAMP, S-Adenosyl methionine (SAM), and FMN. **b** Performance metrics (Recall, Precision, MCC, and AUROC) of CLAPE-SMB on different datasets including different types of small molecule binding sites, including (**c**) four main categories of small molecules: nucleotides and their derivatives, coenzymes and cofactors, amino acids, and carbohydrate. Furthermore, proportions of each amino acid at binding and non-binding sites in the ground truth data on different datasets: nucleotide and its derivatives (**d**), coenzyme and cofactors (**e**), amino acids (**f**), and carbohydrate (**g**)

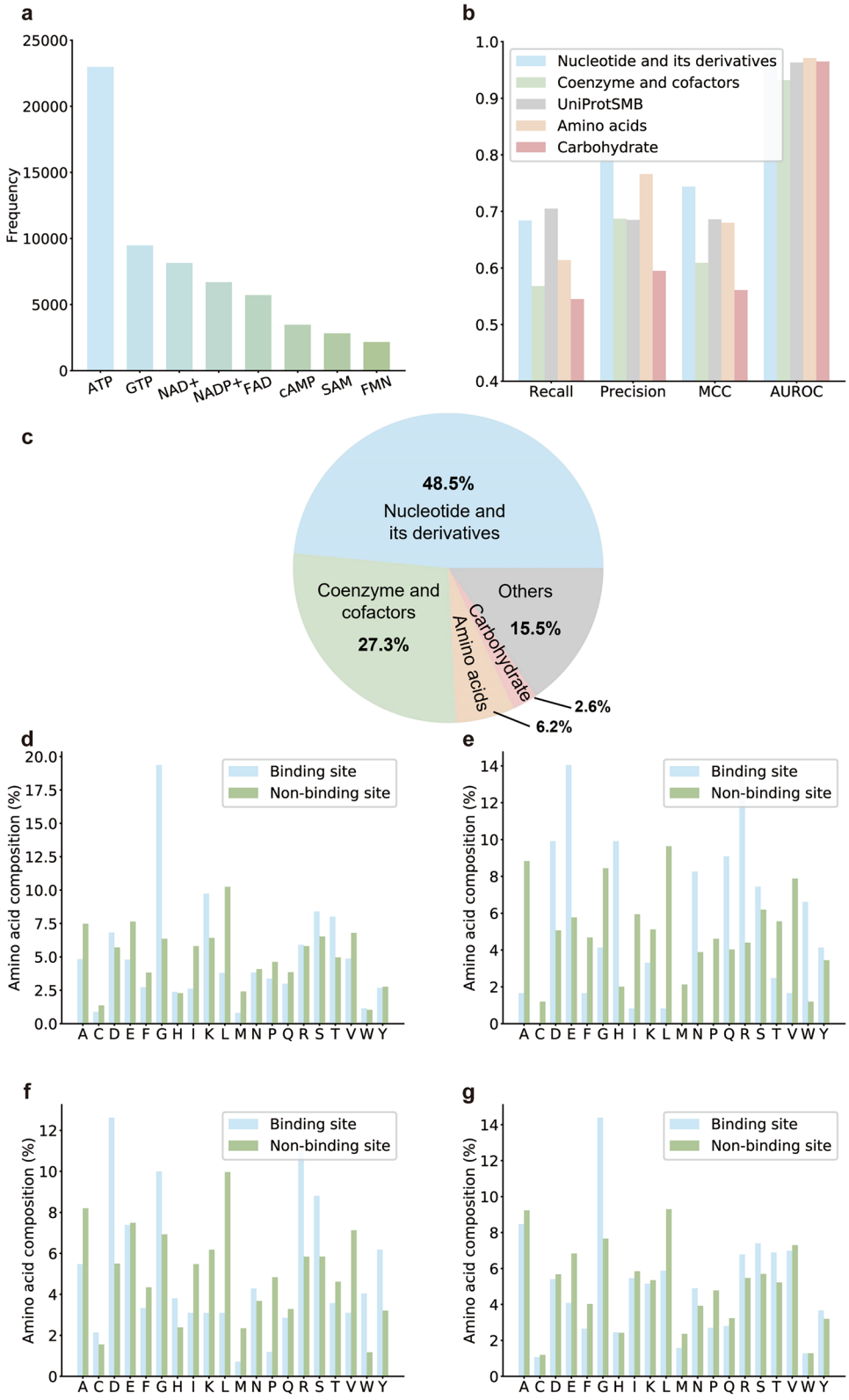

Fig. 4 (See legend on previous page.)

Table 2 Comparison of CLAPE-SMB with other models trained on CHEN11 and tested on COACH420

Model	Rec	Pre	MCC	AUROC
P2Rank	0.888	0.079	0.224	N/A
GraphBind	0.477	0.223	0.303	0.889
CLAPE-SMB	0.395 ± 0.066	0.412 ± 0.057	0.371 ± 0.009	0.876 ± 0.002

Table 3 Comparison of CLAPE-SMB with other models on the SJC

Model	Rec	Pre	MCC	AUROC
P2Rank	0.660	0.180	0.293	N/A
GraphBind	0.568 ± 0.024	0.462 ± 0.011	0.486 ± 0.005	0.906 ± 0.003
DeepProSite	0.458 ± 0.022	0.644 ± 0.011	0.524 ± 0.015	0.926 ± 0.002
CLAPE-SMB	0.456 ± 0.006	0.651 ± 0.016	0.529 ± 0.004	0.915 ± 0.002

and therefore CLAPE-SMB did not perform quite ideal (MCC: 0.609) on it, even though this subset was of moderate size (1,173 sequences).

Improved accuracy of small molecule binding site prediction with CLAPE-SMB

As GraphBind [17] also employed the CHEN11 dataset for training and COACH420 for testing, we adopted the same evaluation scheme to facilitate direct comparison with our model (Table 2). Interestingly, CLAPE-SMB achieved a precision of 0.412, representing an unexpectedly large increase of 84.8% over that of GraphBind. However, CLAPE-SMB exhibited a marginally lower recall, 17.2% lower than GraphBind, but still provided a 22.4% higher MCC (0.371). By contrast, P2Rank [12] achieved a markedly higher recall (0.888) than CLAPE-SMB, but considerably lower precision (0.079), resulting in a modest MCC (0.224), most likely arising from an overabundance of positive predictions that could potentially include an unknown number of misidentifications.

We also compared CLAPE-SMB with P2Rank [12], GraphBind [17], and DeepProSite [22] on the SJC and UniProtSMB datasets. As shown in Table 3, when trained and tested on the SJC dataset, CLAPE-SMB achieved a recall of 0.456, a precision of 0.651, an MCC of 0.529, and an AUROC of 0.915. Figure 5 provides an evaluation of CLAPE-SMB's performance on the SJC dataset with a default random seed of 42. On the UniProtSMB test set, CLAPE-SMB exhibited a recall of 0.673, precision of 0.743, MCC of 0.699, and AUROC of 0.960 (Table 4). We performed a statistical test comparing MCC and AUROC between CLAPE-SMB and other models (Supplementary Fig. 2). CLAPE-SMB consistently outperforms GraphBind on both metrics, showing significant improvements ($p < 0.01$) in MCC and AUROC. Although DeepProSite surpasses CLAPE-SMB in AUROC, CLAPE-SMB achieves a slightly higher MCC on the SJC dataset, though not statistically significant ($p \geq 0.05$). On the UniProtSMB dataset, CLAPE-SMB demonstrates a marked improvement, significantly exceeding DeepProSite ($p < 0.001$). This difference may be due to the nature of the datasets: SJC sequences have clear PDB IDs with defined structures, whereas sequences in UniProtSMB are derived from the UniProtKB database and do not correspond one-to-one with PDB structures, potentially leading to inaccuracies in structural information. Since DeepProSite can utilize structural information for prediction, it may have an advantage in cases where accurate structures are available, potentially leading to slightly better performance than our model, but still comparable.

All experiments of CLAPE-SMB on the chen11, SJC, and UniProtSMB datasets were rigorously tested multiple times using different random seeds (6, 17, 35, 42) to ensure reliability. For further reference, detailed results for UniProtSMB are provided in Supplementary Table 3. Additionally, we conducted ninefold cross-validation on UniProtSMB to assess model robustness, with results shown in Supplementary Table 4. The average performance across the folds was Rec: 0.662, Pre: 0.760, MCC: 0.702, and AUROC: 0.961, with small standard deviation, underscoring the model's consistency across different experimental setups.

Influence of model architecture on prediction accuracy

To ensure the selection of the most appropriate protein language model for CLAPE-SMB, we compared predictive accuracy between two widely used models, ESM-2 [23] and ProtBert [24], with a default random seed 42. We found that ESM-2 outperformed ProtBert on both test sets (Table 5), achieving higher MCC (0.535) and AUROC (0.917) than ProtBert (MCC: 0.352, AUROC: 0.856) on the SJC test set; and a striking 44.3% increase in MCC (0.704) over that of ProtBert on the UniProtSMB test set. These results suggest that the higher-dimensional embedding space of ESM-2, with 1,280-dimensional embeddings compared to ProtBERT's 1,024, captures more complex protein sequence information, enabling the extraction of a greater number of informative features necessary to distinguish binding from non-binding residues. Furthermore, the larger number of parameters in ESM-2 also contributes to its enhanced performance.

We then compared accuracy among three candidate backbone models (MLP, CNN, and Transformer) on the test set of SJC dataset. Among them, MLP achieved the highest accuracy (Fig. 5c), showing a slight advantage in

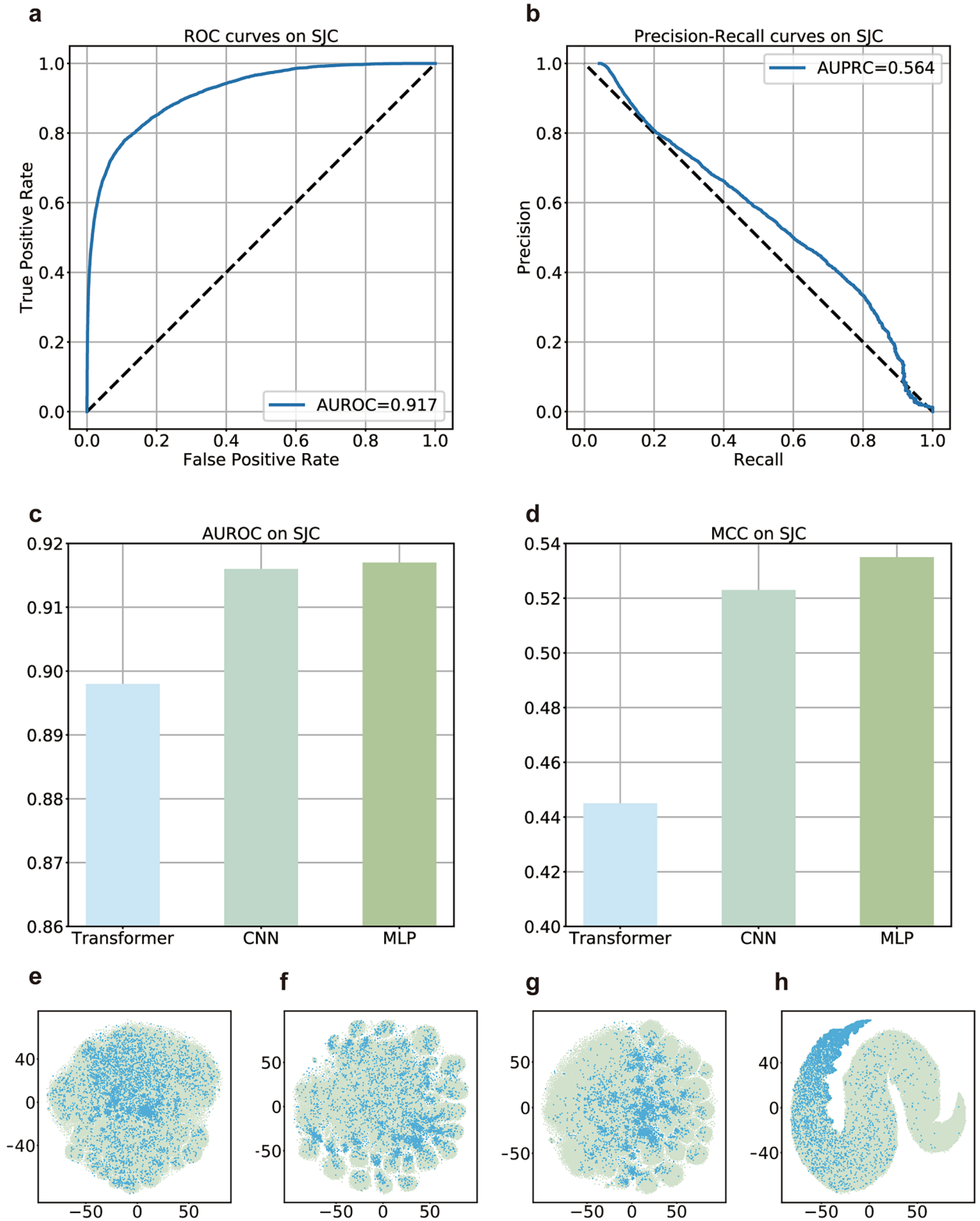

Fig. 5 Evaluation of CLAPE-SMB performance on the SJC dataset with a default random seed, 42. **a** ROC curves of CLAPE-SMB predictions on the SJC test set (AUROC = 0.917). **b** Precision-recall curves of CLAPE-SMB on the SJC test set (AUPRC = 0.564). Comparison of the Transformer, 1DCNN, and MLP backbone models showing that **c** MLP provides higher MCC than CNN; **d** MLP had a slightly higher AUROC than CNN. Plot of t-SNE dimensionality reduction of outputs from **e** the first layer of a randomly initialized MLP, (**f**) ESM-2, and **g** the first layer and **h** fourth layer of CLAPE-SMB. Green, non-binding sites; blue, small molecule binding sites

Table 4 Comparison of CLAPE-SMB with other models on the UniProtSMB

Model	Rec	Pre	MCC	AUROC
P2Rank	0.632	0.124	0.236	N/A
GraphBind	0.565 ± 0.020	0.430 ± 0.007	0.473 ± 0.007	0.932 ± 0.003
DeepProSite	0.490 ± 0.013	0.756 ± 0.005	0.598 ± 0.006	0.965 ± 0.001
CLAPE-SMB	0.673 ± 0.031	0.743 ± 0.040	0.699 ± 0.004	0.960 ± 0.001

Table 5 Comparison of prediction accuracy between ProtBert and ESM-2 as the feature extractor of CLAPE-SMB

Feature extractor	Dataset	MCC	AUROC	AUPRC
ProtBert	SJC	0.352	0.856	0.354
ESM-2	SJC	0.535	0.917	0.564
ProtBert	UniProtSMB	0.488	0.922	0.498
ESM-2	UniProtSMB	0.704	0.961	0.750

Table 6 Model performance on the SJC using different loss functions

Loss functions	MCC	AUROC
TCL	0.528 ± 0.003	0.915 ± 0.003
Focal Loss	0.516 ± 0.003	0.912 ± 0.002
Focal Loss + TCL	0.529 ± 0.004	0.915 ± 0.002

MCC over CNN, although both models had comparable AUROC values [54–56]. To visually assess the model's ability to differentiate binding sites from non-binding sites, we employed t-distributed Stochastic Neighbor Embedding (t-SNE) [57] to reduce the dimensionality of outputs from ESM-2 and the first layer of MLP (Fig. 5f, g), as well as subsequent MLP layers (Supplementary Fig. 3). Additionally, we included the results of a randomly initialized MLP for comparison. The results demonstrated that ESM-2 could distinguish binding from non-binding residues to some extent, while an untrained MLP could not. Conversely, CLAPE-SMB could effectively distinguish between binding and non-binding classes, and its performance improved with increasing layer depth. While the Transformer model [58] has served as the foundation for other protein language models, it performed considerably worse than MLP or CNN as our backbone module, which might be attributable to the general strength of MLP in comprehensively integrating diverse features learned by the protein language models for this specific application.

Influence of loss function and its hyperparameters on prediction accuracy

We next compared model performance using three different loss functions: TCL [28] alone, class-balanced focal loss [33, 34] alone, and TCL in combination with focal loss (Table 6). Focal loss yielded a modest 0.2% increase in MCC, while TCL boosted MCC by 2.5%. These findings aligned well with the t-SNE plots of embeddings (Supplementary Fig. 4). While both focal loss and TCL individually could separate binding from non-binding sites, their combined application resulted in tighter clustering of binding sites, indicating improved model performance. Representative embeddings from the second layer are displayed in Supplementary Fig. 4. In addition to TCL, the unsupervised InfoNCE loss [41] could also be applied. We evaluated the model performance using InfoNCE, and the results showed that TCL slightly outperforms InfoNCE (Supplementary Table 5).

The hyperparameters of the loss function can also influence model performance [59, 60]. In this study, we considered four main hyperparameters, including γ in class-balanced focal loss, margin and learning rate of TCL, and weight, λ. Optimization of hyperparameters was completed on both the SJC (Fig. 6) and UniProtSMB (Supplementary Fig. 5) datasets. We initially set γ to 1, 5, 10, and 20 and witnessed a dramatic drop in AUROC between 5 and 10. Further fine-tuning between values of 1–5 revealed that $\gamma = 3$ conferred the highest MCC and a relatively high AUROC. After further optimization, the margin was set to 4 and λ was set to 0.2. The learning rate of TCL was set to 0.3, which was consistent with previous studies that suggested a higher learning rate to optimize cluster centers [61]. Similarly, γ was set to 3, margin was set to 5, the learning rate of TCL was set to 0.01, and λ was set to 0.2 on the UniProtSMB dataset after hyperparameter optimization.

CLAPE-SMB can rationally distinguish binding and non-binding sites based on residue features

To determine whether CLAPE-SMB predictions reflected rational, interpretable differences in residue properties, we next analyzed the composition and properties of binding and non-binding amino acid residues. First, we analyzed the frequency of each amino acid type within binding and non-binding sites in SJC (Fig. 7a) and UniProtSMB (Supplementary Fig. 6a). Both results suggested that glycine (G), serine (S), and threonine (T) had a higher propensity to participate in small-molecule binding, while leucine (L) exhibited the lowest propensity. The relatively high frequency of glycine binding sites aligns well with its small size and flexibility that facilitate induced fit by small molecule ligands, as reported in other studies [62, 63]. Additionally, serine and threonine

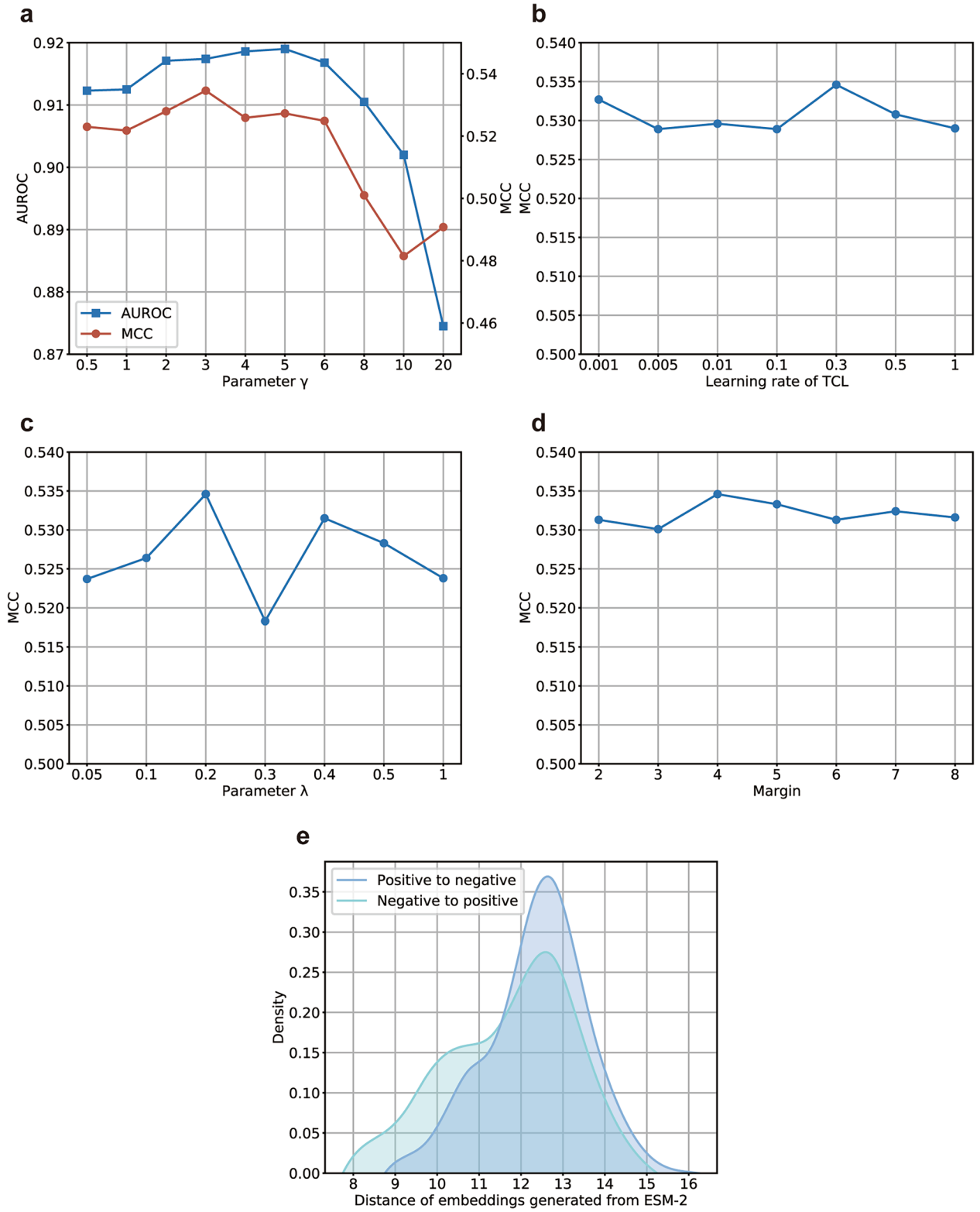

Fig. 6 Hyperparameter optimization for the loss function on the SJC dataset. **a** Parameter γ was set to 3 rather than 5 due to the highest MCC (which has a stronger effect in analyses of imbalanced data) and a relatively high AUROC. **b** Learning rate of TCL was set to 0.3 after fine-tuning. **c** Weight (λ) of TCL and focal loss was set to 0.2. **d** Margin of TCL is set to 4, which gives a maximum MCC. **e** Distribution of the maximum Euclidean distance between a given embedding generated from ESM-2 and the embeddings from the opposite class

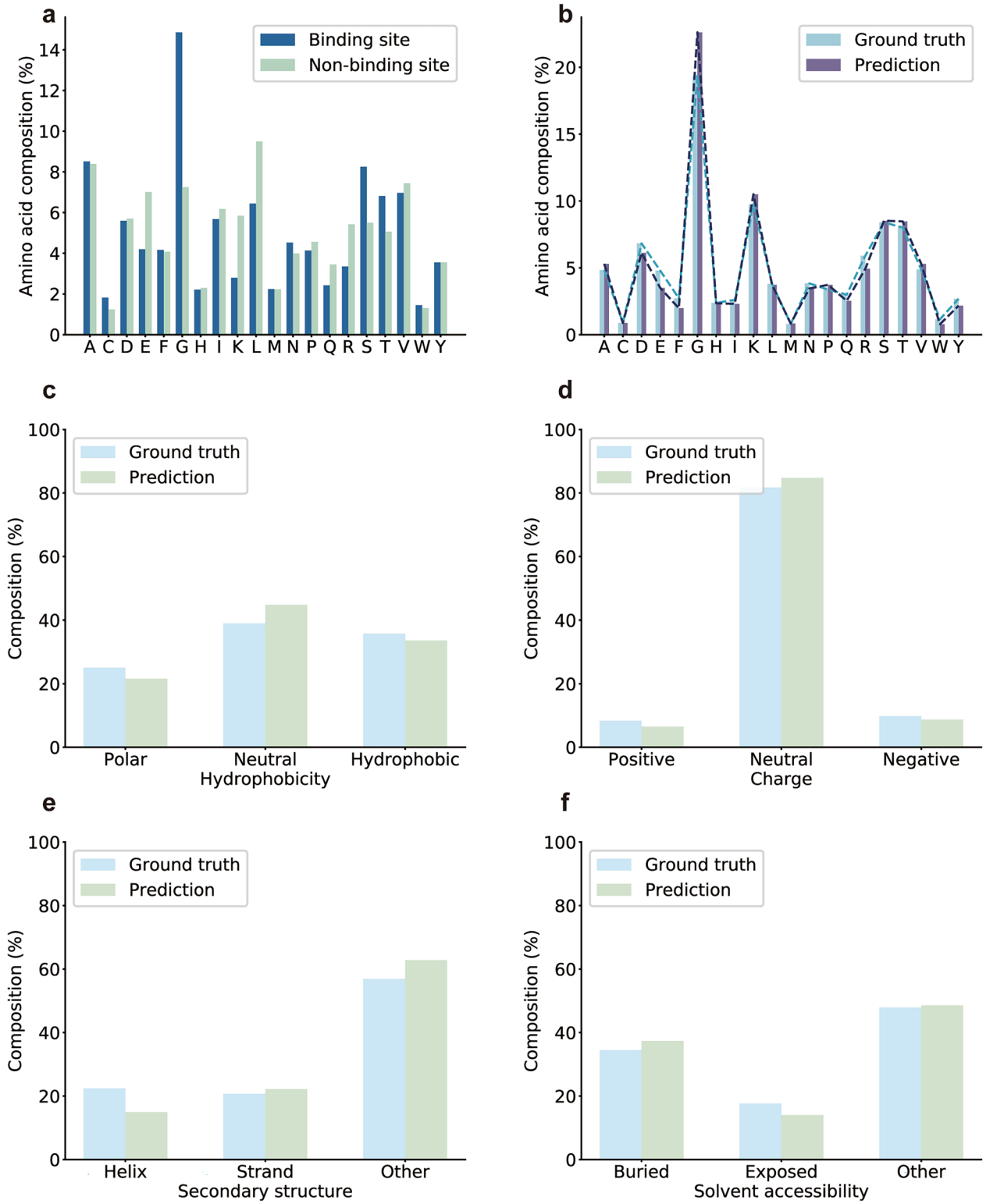

Fig. 7 Amino acid analysis on the SJC dataset. **a** Proportions of each amino acid at binding and non-binding sites in ground truth data. **b** Comparison of the proportions of amino acids between predicted and ground truth binding sites. Comparison of amino acid properties, including **c** hydrophobicity, **d** charge, **e** secondary structure, and **f** solvent accessibility between ground truth and predicted small molecule binding residues

possess hydroxyl groups in their side chains that enable hydrogen bonding interactions with small molecules [64]. These findings suggest interpretability in CLAPE-SMB of predictions.

We then focused strictly on binding sites and compared the distribution of predicted amino acid against the ground truth (Fig. 7b, Supplementary Fig. 6b), which revealed similar distribution of binding amino acids to the ground truth, with forward and reverse Kullback–Leibler (KL) divergence of 0.0208 and 0.0206, respectively. By contrast, a prediction of uniform distribution for each amino acid resulted in much larger forward and reverse KL divergence of 0.165 and 0.164, respectively. Next, we analyzed the amino acid properties of hydrophobicity, charge, secondary structure, and solvent accessibility (Fig. 7c–f). Again, CLAPE-SMB predicted a property distribution among binding residues similar to the ground truth, suggesting high predictive accuracy. In addition, t-SNE dimensionality reduction of the binding site property distribution in the first layer output of MLP showed that CLAPE-SMB could correctly cluster binding residues according to their hydrophobicity and charge (Supplementary Fig. 7). These results suggested that CLAPE-SMB can rationally distinguish binding and non-binding sites based on residue features.

Comparison of CLAPE-SMB with DeepProSite and heuristic analyses in three protein case studies

Given the above results showing high accuracy and interpretability of our CLAPE-SMB model, we next evaluated its performance in case study. We compared CLAPE-SMB performance with that of DeepProSite, the best sequence-based model currently available for predicting small molecule binding sites in three protein case studies from the SJC test set. The first protein was death-associated protein kinase 1 (DAPK1; PDB ID: 1JKK, chain A, or 1JKK_A), which participates in programmed cell death and has been proposed as a potential target for treating cancer [65, 66]. The second protein, tryptophan 7-halogenase RebH (RebH; PDB ID: 2OAL, chain B, or 2OAL_B), functions in chlorinating tryptophan residues, often in the synthesis bioactive compounds [67–69]. The third protein selected as a case study was ribosomal RNA small subunit methyltransferase (Nep1; PDB ID: 3O7B, chain A, or 3O7B_A), which is responsible for catalyzing methyl group addition to the small subunit of ribosomal RNA and which has also been proposed as a candidate antimicrobial drug target [70, 71].

The ground truth binding sites, sites predicted by CLAPE-SMB, and DeepProSite predictions for DAPK1 and RebH are shown in Fig. 8a-f. CLAPE-SMB successfully predicted over half of the known binding sites with relatively few false positives (e.g., some incorrect predictions near the actual binding pockets). While DeepProSite made some correct predictions, it also generated a comparatively higher number of false positives, suggesting higher recall but lower precision. However, it warrants mention that these potentially misleading predictions might also inspire the discovery of previously unrecognized novel allosteric binding sites.

Interestingly, in the third case study of Nep1, in which CLAPE-SMB achieved its highest accuracy in predicting bona fide binding sites, we compared CLAPE-SMB predictions with potential small-molecule binding sites identified through heuristic analysis of the experimentally derived structure (Fig. 8g). CLAPE-SMB accurately identified 19 binding residues (out of 22), suggesting its potential to reduce time and labor in structural analyses. The ground truth, CLAPE-SMB predictions, and DeepProSite predictions of Nep1 small molecule binding sites are shown in Fig. 8h-j, emphasizing the high accuracy provided by CLAPE-SMB in predicting protein-small molecule binding interactions.

CLAPE-SMB accurately predicts small molecule binding sites in intrinsically disordered proteins

Intrinsically disordered proteins (IDPs) lack stable tertiary structures, making their study and functional characterization challenging. Experimental determination of small molecule binding sites in IDPs is difficult due to their inherent structural flexibility and dynamic nature [72, 73]. Structure-based prediction models struggle with accuracy because IDPs lack stable conformations, and AlphaFold's predicted structures may not capture the functional dynamics of these regions.

However, CLAPE-SMB successfully predicted small molecule binding sites of IDPs with high accuracy. In the UniProtKB database, a protein is classified as an IDP if it contains at least one intrinsically disordered region (IDR). For our study, we randomly selected 526 experimentally supported IDPs from the UniProtKB database and annotated their small molecule binding sites as described in Fig. 2. After deduplication based on a cutoff of 50% sequence similarity between training sets of UniProtSMB and IDP dataset, we obtained 336 non-redundant sequences. CLAPE-SMB was trained on UniProtSMB and then tested on this IDP dataset, achieving an MCC of 0.815. Despite the high MCC, further analysis revealed that almost all binding sites are located on non-IDRs: only 5 out of 336 IDPs have small molecule binding sites on IDRs, which was not satisfactory. This also underscores the necessity for traditional experimental methods to precisely determine the structure of binding sites.

Therefore, we continued to verify whether CLAPE-SMB can accurately predict binding sites of IDRs. We

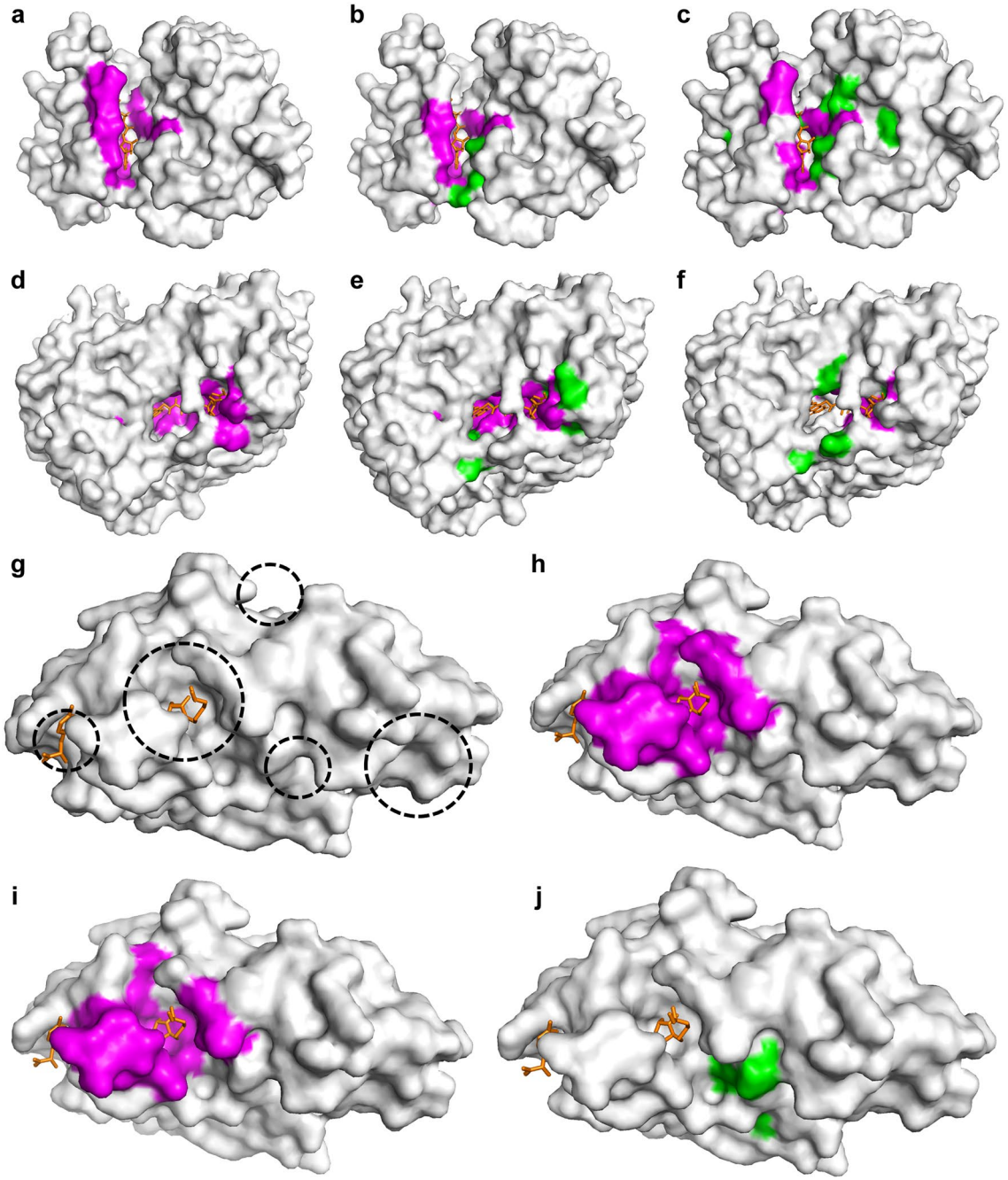

Fig. 8 Case study of binding site prediction for 1JKK_A, 2OAL_B, and 3O7B_A proteins. **a**-**c** Small molecule binding sites on 1JKK_A in **a** ground truth data, **b** CLAPE-SMB predictions, and **c** predicted by DeepProSite. **d**-**f** Small-molecule binding sites of 2OAL_B in (**d**) ground truth data and predicted by (**e**) CLAPE-SMB or (**f**) DeepProSite. **g** Putative small molecule binding pockets of 3O7B_A predicted by heuristic analysis. **h** Ground truth of small molecule binding sites on 3O7B_A. (**i**-**j**) Small-molecule binding sites of 3O7B_A predicted by (**i**) CLAPE-SMB and **j** DeepProSite. Magenta residues in (**a**, **d**, and **h**) indicate binding sites in ground truth data. Magenta residues in **b**, **c**, **e**, **f**, **i**, and **j** indicate correct predictions and green residues indicate false positive predictions. Orange sticks are small molecule ligands

identified 34 protein sequences with small molecule binding sites on IDRs in the UniProtKB database. These sequences are non-redundant and have less than 50% similarity to any sequence in the UniProtSMB training set. Using these 34 sequences as a test set, CLAPE-SMB achieved a high MCC of 0.730. Next, we performed two case studies of IDPs, i.e. bifunctional protein GlmU (GlmU, UniProtKB ID: A1TUE2) and cobalamin

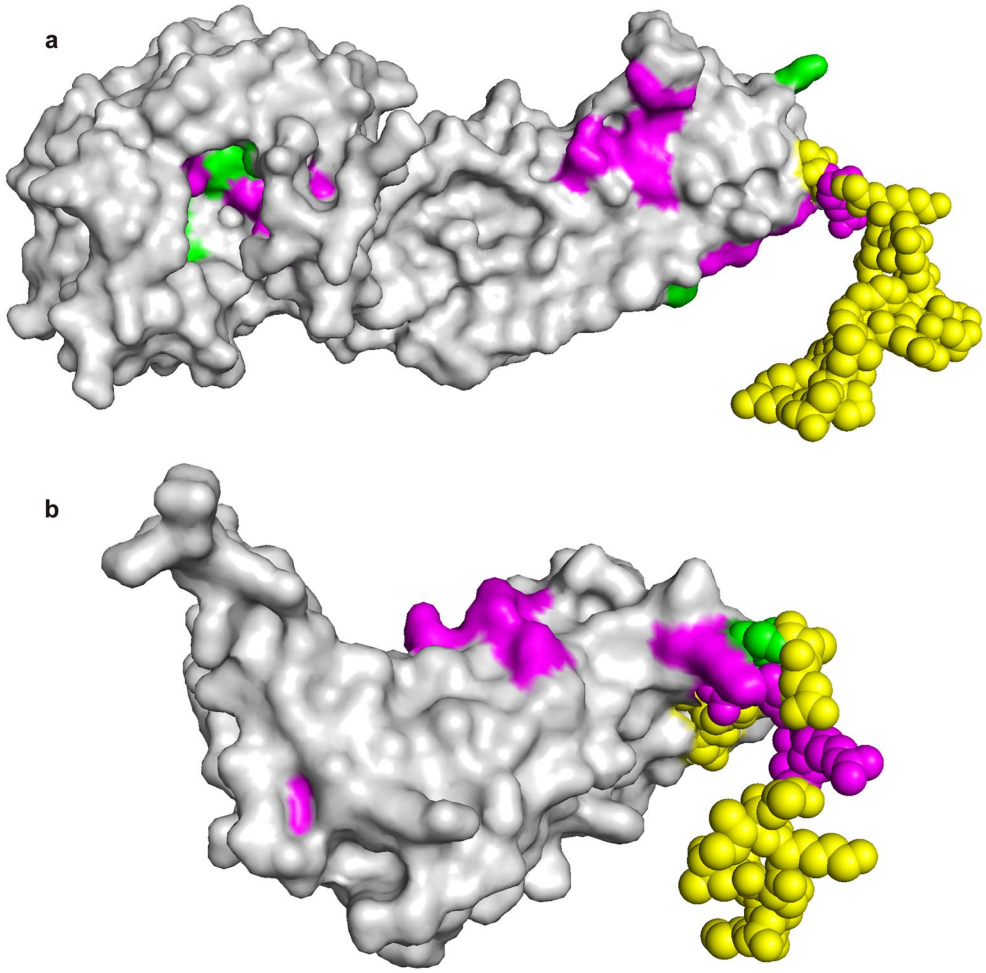

Fig. 9 Two case studies of IDPs: **a** A1TUE2 and **b** Q1LJ80. Yellow spheres represent IDRs, and gray surfaces represent non-IDRs. CLAPE-SMB accurately predicted most of small molecule binding sites (magenta residues) with a few false positive predictions (green residues). The structure and conformation around IDRs rarely form binding pockets, indicating a significant difference between AlphaFold predictions and the in vivo situation

adenosyltransferase (CobO, UniProtKB ID: Q1LJ80) (Fig. 9). GlmU catalyzes the last two sequential reactions in the de novo biosynthetic pathway for UDP-N-acetylglucosamine [74]. The C-terminal domain is IDR, catalyzing the transfer of acetyl group from acetyl coenzyme A. CobO catalyzes the conversion of cob(II) alamin to adenosylcob(III)alamin in the presence of ATP, which binds to the disordered N-terminal region of CobO [75]. Because most IDPs do not have available PDB structures, we used structures predicted by AlphaFold instead. The results showed that CLAPE-SMB successfully identified the majority of small

Table 7 Computational efficiency comparison of CLAPE-SMB with other models on the UniProtSMB

	CLAPE-SMB	DeepProSite	GraphBind	P2Rank
GPU	NVIDIA A100-PCIE-40 GB			
CPU	Intel(R) Xeon(R) Silver 4210R CPU @ 2.40 GHz			
Data preparation time	0.91 s / sequence	2.56 s / sequence	~4 min / sequence	N/A
Training time	~15 min / epoch	~3 min / epoch	~200 min / epoch	N/A
Inference time	191 sequence / s	51 sequence / s	~3 s / sequence	~10 s / sequence

molecule binding sites on IDRs. This work underscores the potential of sequence-based prediction models in accurately predicting small molecule binding sites in IDPs, facilitating further research and drug discovery.

Computational efficiency of CLAPE-SMB

We evaluated the computational efficiency of CLAPE-SMB and compared it with DeepProSite [22], GraphBind [17], and P2Rank [12] across data preparation, training, and inference time (Table 7). These comparisons were conducted on an NVIDIA A100-PCIE-40GB GPU and an Intel Xeon Silver 4210R CPU. Specifically, data preparation and inference times were tested on the CPU, while training time was measured on the GPU. The result shows that while CLAPE-SMB required slightly longer training times (~15 min/epoch) compared to DeepProSite (~3 min/epoch), it demonstrated significantly faster data preparation and inference speeds. CLAPE-SMB required only 0.91 s per sequence for data preparation, substantially less than DeepProSite (2.56 s/sequence) and GraphBind (~4 min/sequence). In terms of inference, CLAPE-SMB achieved 191 sequences per second, considerably faster than DeepProSite (51 sequences/second), GraphBind (~3 s/sequence), and P2Rank (~10 s/sequence). This highlights the computational efficiency of CLAPE-SMB, as it can make residue-level binding site predictions from a 1D amino acid sequence within a second, emphasizing its potential for large-scale applications. This evaluation highlights the computational efficiency of CLAPE-SMB, which is valuable for high-throughput applications.

Discussion

To overcome challenges in the accurate prediction of protein-small molecule interactions that are ubiquitous in biological systems and essential for drug development and research, it is necessary to ensure that the extracted protein information is suitable as data input and the imbalanced data between small molecule binding sites and non-binding sites in protein sequences is properly handled. In our data, binding sites represented less than 5% of the data (Table 1 and Supplementary Table 1). To address these challenges in our current approach, we leveraged ESM-2 to generate protein embeddings and employed contrastive learning to deal with the imbalanced data. Our CLAPE-SMB model demonstrates strong performance on both UniProtSMB and SJC test set, as well as the COACH420 benchmark set. In addition, our findings highlight the impact of dataset quality on model performance. To provide the highest possible quality input data, we assembled the UniProtSMB dataset by merging binding sites of similar sequences, which resulted in markedly higher accuracy binding site prediction by CLAPE-SMB using the UniProtSMB compared to that obtained with SJC and COACH420.

Protein language models are highly proficient in extracting accurate sequence information and thus show potential for a wide array of applications [76]. To assess the contribution of ESM-2 in the performance of CLAPE-SMB, we evaluated a simpler model using a single-layer MLP to process the protein embedding step in place of the 5-layer MLP. This simpler model also showed relatively strong performance on SJC dataset, with an MCC of 0.300 and an AUROC of 0.882 (Supplementary Table 6). This strong performance despite using fewer layers illustrates the strong capability of ESM-2 to extract informative features from protein sequence relevant to binding site prediction [23, 77, 78]. Intriguingly, we observed that hyperparameter margin was inconsistent with the maximum Euclidean distance distribution of 1,280-dimensional embeddings (Fig. 6d-e, Supplementary Fig. 5c-d). For instance, in the UniProtSMB training set, the distances between negative to positive residues, and positive to negative residues were distributed from 8 to 15 and 9 to 16, respectively. However, the best margin selected after hyperparameter optimization was 5, much smaller than expected [21]. This discovery implied that certain dimensions within the 1,280-dimensional features embedded by ESM-2 might not obviously contribute to binding site prediction. Exploring these less informative dimensions could be a promising avenue for future research for better understanding ESM-2. Additionally, with the recent release of ESM-3 [79], incorporating this model presents a valuable opportunity for further improving predictive performance and enhancing our understanding of protein features.

An interesting question is whether the inclusion of sequences in ESM-2's training set affects the performance of CLAPE-SMB. The esm2_t33_650M_UR50D model was trained on the UniRef50 dataset, which contained approximately 6.6 million protein sequences [23, 47]. Approximately 20%-40% of the sequences in our experiments are new to the esm2_t33_650M_UR50D model (Supplementary Table 7). While we did not specifically exclude sequences from UniRef50 in our experiments, we did assess the model's performance on the UniProtSMB dataset for sequences both included in and excluded from UniRef50. As shown in Supplementary Table 8, the model actually performed slightly better on sequences excluded from UniRef50, with higher MCC and AUROC values. Therefore, we believe that whether a sequence is part of ESM-2's training set is not a significant factor in the model's performance.

One limitation of our study is the potential presence of mislabeled data in both SJC and UniProtSMB. All possible mislabeled data can be divided into two categories:

binding sites marked as non-binding sites, and non-binding sites marked as binding sites. In our study, those marked as binding sites in our dataset are unlikely to be false, as they are all collected from published literature. However, bona fide binding sites might be incorrectly labeled as non-binding residues due to inherent limitations in our current knowledge regarding the full scope of all possible small-molecule interactions for any given protein. This situation may persist even with the merging approach we employed in compiling UniProtSMB. Therefore, some structurally plausible predicted binding sites but are labeled as non-binding sites in dataset could indicate previously undiscovered or unvalidated binding sites, potentially including novel drug targets. Experimental validation and subsequent dataset updates could lead to the identification of new drug targets and improve future model performance. Currently, CLAPE-SMB predicts all potential small-molecule binding sites on a protein. Future work will focus on integrating information for specific small molecules as an input condition, potentially by incorporating language models for SMILES strings or graph neural networks for structural formula processing to ultimately enable binding site for specific molecules.

Conclusion

We propose CLAPE-SMB, which integrates a pre-trained protein language model with contrastive learning to overcome current challenges in predicting small molecule binding sites of proteins, and also guide drug development. On COACH420 benchmark test set, CLAPE-SMB outperformed other current methods. We integrated three benchmark datasets—sc-PDB, JOINED, and COACH420—into a new dataset, SJC, on which CLAPE-SMB performed well. We also constructed the UniProtSMB dataset, in which we merged multiple sites from similar proteins. CLAPE-SMB showed high accuracy on it as well. Analysis of composition and properties of amino acid residues, and case studies of DAPK1, RebH, and Nep1 are also done to provide extra proof to CLAPE-SMB.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-024-00920-2.

Supplementary material 1.

Acknowledgements
We thank Jingrui Zhong for helpful discussion, and Hedi Chen, Xiaochun Zhang, and Lin Guo for necessary support.

Author contributions
J.W conducted most of the experiments and was involved in virtually every step of this work. Y.L wrote the original code for CLAPE-SMB and revised the manuscript. J.W and B.T wrote the manuscript.

Funding
This work was supported by the Beijing Frontier Research Center for Biological Structure (No. 041500002), Tsinghua University Initiative Scientific Research Program (No.20231080030), the Tsinghua-Peking University Center for Life Sciences (No.20111770319).

Availability of data and materials
The datasets and codes of CLAPE-SMB are freely available at https://github.com/JueWangTHU/CLAPE-SMB.

Declarations

Competing interests
The authors declare no competing interests.

Received: 4 June 2024 Accepted: 20 October 2024
Published online: 06 November 2024

References

1. Burslem GM, Crews CM (2017) Small-molecule modulation of protein homeostasis. Chem Rev 117(17):11269–11301
2. Schenone M, Dančík V, Wagner BK, Clemons PA (2013) Target identification and mechanism of action in chemical biology and drug discovery. Nat Chem Biol 9(4):232–240
3. Sneha P, Doss CGP (2016) Molecular dynamics: new frontier in personalized medicine. Adv Protein Chem Struct Biol 102:181–224
4. Xiao B, Sanders MJ, Carmena D, Bright NJ, Haire LF, Underwood E et al (2013) Structural basis of AMPK regulation by small molecule activators. Nat Commun 4(1):3017
5. Zhang J, Yang PL, Gray NS (2009) Targeting cancer with small molecule kinase inhibitors. Nat Rev Cancer 9(1):28–39
6. Gal M, Bloch I, Shechter N, Romanenko O, Shir M, O. (2016) Efficient isothermal titration calorimetry technique identifies direct interaction of small molecule inhibitors with the target protein. Comb Chem High Throughput Screen 19(1):4–13
7. Kennedy AE, Sheffield KS, Eibl JK, Murphy MB, Vohra R, Scott JA et al (2016) A surface plasmon resonance spectroscopy method for characterizing small-molecule binding to nerve growth factor. J Biomol Screen 21(1):96–100
8. Masson GR, Jenkins ML, Burke JE (2017) An overview of hydrogen deuterium exchange mass spectrometry (HDX-MS) in drug discovery. Expert Opin Drug Discov 12(10):981–994
9. Merk A, Bartesaghi A, Banerjee S, Falconieri V, Rao P, Davis MI et al (2016) Breaking cryo-EM resolution barriers to facilitate drug discovery. Cell 165(7):1698–1707
10. Guvench O, MacKerell AD Jr (2009) Computational evaluation of protein–small molecule binding. Curr Opin Struct Biol 19(1):56–61
11. Tubiana J, Schneidman-Duhovny D, Wolfson HJ (2022) ScanNet: an interpretable geometric deep learning model for structure-based protein binding site prediction. Nat Methods 19(6):730–739
12. Krivák R, Hoksza D (2018) P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. J Cheminform 10:1–12
13. Jiménez J, Doerr S, Martínez-Rosell G, Rose AS, De Fabritiis G (2017) DeepSite: protein-binding site predictor using 3D-convolutional neural networks. Bioinformatics 33(19):3036–3042
14. Hu J, Yan C (2009) A tool for calculating binding-site residues on proteins from PDB structures. BMC Struct Biol 9:1–6
15. Wang S, Sun S, Li Z, Zhang R, Xu J (2017) Accurate de novo prediction of protein contact map by ultra-deep learning model. PLoS Comput Biol 13(1):e1005324

16. Villegas-Morcillo A, Makrodimitris S, van Ham RC, Gomez AM, Sanchez V, Reinders MJ (2021) Unsupervised protein embeddings outperform hand-crafted sequence and structure features at predicting molecular function. Bioinformatics 37(2):162–170
17. Xia Y, Xia C-Q, Pan X, Shen H-B (2021) GraphBind: protein structural context embedded rules learned by hierarchical graph neural networks for recognizing nucleic-acid-binding residues. Nucleic Acids Res 49(9):e51–e51
18. Pai PP, Dattatreya RK, Mondal S (2017) Ensemble architecture for prediction of enzyme-ligand binding residues using evolutionary information. Mol Inform 36(11):1700021
19. Macari G, Toti D, Polticelli F (2019) Computational methods and tools for binding site recognition between proteins and small molecules: from classical geometrical approaches to modern machine learning strategies. J Comput Aided Mol Des 33(10):887–903
20. Zhou X, Zheng W, Li Y, Pearce R, Zhang C, Bell EW et al (2022) I-TASSER-MTD: a deep-learning-based platform for multi-domain protein structure and function prediction. Nat Protoc 17(10):2326–2353
21. Liu Y, Tian B (2024) Protein–DNA binding sites prediction based on pre-trained protein language model and contrastive learning. Brief Bioinform 25(1):bbad488
22. Fang Y, Jiang Y, Wei L, Ma Q, Ren Z, Yuan Q et al (2023) DeepProSite: structure-aware protein binding site prediction using ESMFold and pretrained language model. Bioinformatics 39(12):btad718
23. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W et al (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379(6637):1123–1130
24. Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L et al (2021) Prottrans: toward understanding the language of life through self-supervised learning. IEEE Trans Pattern Anal Mach Intell 44(10):7112–7127
25. Marks DS, Hopf TA, Sander C (2012) Protein structure prediction from sequence variation. Nat Biotechnol 30(11):1072–1080
26. Shenoy R, S., & Jayaram, B. (2010) Proteins: sequence to structure and function-current status. Curr Protein Pept Sci 11(7):498–514
27. Whisstock JC, Lesk AM (2003) Prediction of protein function from protein sequence and structure. Q Rev Biophys 36(3):307–340
28. He X, Zhou Y, Zhou Z, Bai S, Bai X. Triplet-center loss for multi-view 3d object retrieval. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018 (pp. 1945–1954)
29. Chen K, Mizianty MJ, Gao J, Kurgan L (2011) A critical comparative assessment of predictions of protein-binding sites for biologically relevant organic compounds. Structure 19(5):613–621
30. Yang J, Roy A, Zhang Y (2013) Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. Bioinformatics 29(20):2588–2595
31. Desaphy J, Bret G, Rognan D, Kellenberger E (2015) sc-PDB: a 3D-database of ligandable binding sites—10 years on. Nucleic Acids Res 43(D1):D399–D404
32. Vig J, Madani A, Varshney LR, Xiong C, Socher R, Rajani NF (2020) Bertology meets biology: interpreting attention in protein language models. *arXiv preprint* arXiv:2006.15222.
33. Lin T-Y, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In 2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017 (pp. 2980–2988)
34. Cui Y, Jia M, Lin T-Y, Song Y, Belongie S. Class-balanced loss based on effective number of samples. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019 (pp. 9268–9277)
35. Chicco D, Jurman G (2020) The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 21:1–13
36. Boughorbel S, Jarray F, El-Anbari M (2017) Optimal classifier for imbalanced data using Matthews correlation coefficient metric. PLoS ONE 12(6):e0177678
37. Chicco D, Tötsch N, Jurman G (2021) The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. BioData Min 14:1–22
38. Unsal S, Atas H, Albayrak M, Turhan K, Acar AC, Doğan T (2022) Learning functional properties of proteins with language models. Nat Mach Intell 4(3):227–245
39. French RM (1999) Catastrophic forgetting in connectionist networks. Trends Cogn Sci 3(4):128–135
40. Kirkpatrick J, Pascanu R, Rabinowitz N, Veness J, Desjardins G, Rusu AA et al (2017) Overcoming catastrophic forgetting in neural networks. Proc Natl Acad Sci U S A 114(13):3521–3526
41. Oord AVD, Li Y, Vinyals O (2018) Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748
42. Li X, Sun A, Zhao M, Yu J, Zhu K, Jin D, et al. Multi-intention oriented contrastive learning for sequential recommendation. In ACM International Conference on Web Search and Data Mining, 2023 (pp. 411–419)
43. Cai X, Huang C, Xia L, Ren X. (2023). LightGCL: simple yet effective graph contrastive learning for *recommendation. arXiv preprint* arXiv:2302.08191
44. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26(19):2460–2461
45. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22(13):1658–1659
46. UniProt: the universal protein knowledgebase in 2021 (2021). Nucleic Acids Res, 49(D1), D480–D489.
47. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 23(10):1282–1288
48. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30(4):772–780
49. Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. EMBO J 5(4):823–826
50. Yang A-S, Honig B (2000) An integrated approach to the analysis and modeling of protein sequences and structures. II. On the relationship between sequence and structural similarity for proteins that are not obviously related in sequence. J Mol Biol 301(3):679–689
51. Fischer EH, Krebs EG (1955) Conversion of phosphorylase b to phosphorylase a in muscle extracts. J Biol Chem 216(1):121–132
52. Pierce KL, Premont RT, Lefkowitz RJ (2002) Seven-transmembrane receptors. Nat Rev Mol Cell Biol 3(9):639–650
53. Walker JE (1992) The NADH: ubiquinone oxidoreductase (complex I) of respiratory chains. Q Rev Biophys 25(3):253–324
54. Carrington AM, Fieguth PW, Qazi H, Holzinger A, Chen HH, Mayr F et al (2020) A new concordant partial AUC and partial c statistic for imbalanced data in the evaluation of machine learning algorithms. BMC Med Inform Decis Mak 20:1–12
55. Bekkar M, Djemaa HK, Alitouche TA (2013) Evaluation measures for models assessment over imbalanced data sets. J Inf Eng Appl 3(10):2224
56. Weng CG, Poon J. A new evaluation measure for imbalanced datasets. In Australasian data mining conference, 2008 (pp. 27–32)
57. Van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9(11):12
58. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN et al (2017) Attention is all you need. Adv Neural Inf Process Syst 30:415
59. Yang L, Shami A (2020) On hyperparameter optimization of machine learning algorithms: theory and practice. Neurocomputing 415:295–316
60. Wu J, Chen X-Y, Zhang H, Xiong L-D, Lei H, Deng S-H (2019) Hyperparameter optimization for machine learning models based on Bayesian optimization. J Electron Sci Technol 17(1):26–40
61. Wen Y, Zhang K, Li Z, Qiao Y (2016) A discriminative feature learning approach for deep face recognition. In: Wen Y (ed) European conference on computer vision. Springer, Cham, pp 499–515
62. Katsoulidis AP, Antypov D, Whitehead GF, Carrington EJ, Adams DJ, Berry NG et al (2019) Chemical control of structure and guest uptake by a conformationally mobile porous material. Nature 565(7738):213–217
63. Roskoski R Jr (2015) A historical overview of protein kinases and their targeted small molecule inhibitors. Pharmacol Res 100:1–23
64. Ippolito JA, Alexander RS, Christianson DW (1990) Hydrogen bond stereochemistry in protein structure and function. J Mol Biol 215(3):457–471
65. Chen D, Zhou XZ, Lee TH (2019) Death-associated protein kinase 1 as a promising drug target in cancer and Alzheimer's disease. Recent Pat Anticancer Drug Discov 14(2):144–157
66. Singh P, Ravanan P, Talwar P (2016) Death associated protein kinase 1 (DAPK1): a regulator of apoptosis and autophagy. Front Mol Neurosci 9:46

67. Yeh E, Blasiak LC, Koglin A, Drennan CL, Walsh CT (2007) Chlorination by a long-lived intermediate in the mechanism of flavin-dependent halogenases. Biochemistry 46(5):1284–1292
68. Yeh E, Garneau S, Walsh CT (2005) Robust in vitro activity of RebF and RebH, a two-component reductase/halogenase, generating 7-chlorotryptophan during rebeccamycin biosynthesis. Proc Natl Acad Sci U S A 102(11):3960–3965
69. Sánchez C, Butovich IA, Braña AF, Rohr J, Méndez C, Salas JA (2002) The biosynthetic gene cluster for the antitumor rebeccamycin: characterization and generation of indolocarbazole derivatives. Chem Biol 9(4):519–531
70. Sergiev PV, Aleksashin NA, Chugunova AA, Polikanov YS, Dontsova OA (2018) Structural and evolutionary insights into ribosomal RNA methylation. Nat Chem Biol 14(3):226–235
71. Wurm JP, Meyer B, Bahr U, Held M, Frolow O, Kötter P et al (2010) The ribosome assembly factor Nep1 responsible for Bowen-Conradi syndrome is a pseudouridine-N1-specific methyltransferase. Nucleic Acids Res 38(7):2387–2398
72. Uversky VN, Oldfield CJ, Dunker AK (2008) Intrinsically disordered proteins in human diseases: introducing the D2 concept. Annu Rev Biophys 37:215–246
73. Wright PE, Dyson HJ (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J Mol Biol 293(2):321–331
74. Olsen LR, Roderick SL (2001) Structure of the Escherichia coli GlmU pyrophosphorylase and acetyltransferase active sites. Biochemistry 40(7):1913–1921
75. Li Z, Kitanishi K, Twahir UT, Cracan V, Chapman D, Warncke K et al (2017) Cofactor editing by the G-protein metallochaperone domain regulates the radical B12 enzyme IcmF. J Biol Chem 292(10):3977–3987
76. Liu W, Wang Z, You R, Xie C, Wei H, Xiong Y et al (2024) PLMSearch: Protein language model powers accurate and fast sequence search for remote homology. Nat Commun 15(1):2775
77. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W et al (2022) Language models of protein sequences at the scale of evolution enable accurate structure prediction. BioRxiv 2022:500902
78. Du Z, Ding X, Hsu W, Munir A, Xu Y, Li Y (2024) pLM4ACE: a protein language model based predictor for antihypertensive peptide screening. Food Chem 431:137162
79. Hayes T, Rao R, Akin H, Sofroniew NJ, Oktay D, Lin Z, et al (2024) Simulating 500 million years of evolution with a language model. *bioRxiv*, 2024.2007. 2001.600583.

Publisher's Note