



## OPEN ACCESS

### EDITED BY

Marco Masseroli,  
Polytechnic University of Milan, Italy

### REVIEWED BY

Emre Sefer,  
Özyeğin University, Türkiye  
Chenguang Zhao,  
St. Ambrose University, United States

### \*CORRESPONDENCE

Xiao Fan  
✉ xiaofan@ufl.edu

RECEIVED 08 December 2025

REVISED 27 February 2026

ACCEPTED 09 March 2026

PUBLISHED 25 March 2026

### CITATION

Zhang L, Li X, Song R, Song Q and  
Fan X (2026) PRIMED: predicting DNA  
binding residues by leveraging  
pre-trained protein language models.  
*Front. Artif. Intell.* 9:1763313.  
doi: 10.3389/frai.2026.1763313

### COPYRIGHT

© 2026 Zhang, Li, Song, Song and Fan.  
This is an open-access article distributed  
under the terms of the [Creative  
Commons Attribution License \(CC BY\)](#).  
The use, distribution or reproduction in  
other forums is permitted, provided the  
original author(s) and the copyright  
owner(s) are credited and that the  
original publication in this journal is  
cited, in accordance with accepted  
academic practice. No use, distribution  
or reproduction is permitted which does  
not comply with these terms.

# PRIMED: predicting DNA binding residues by leveraging pre-trained protein language models

Luoshu Zhang<sup>1</sup>, Xin Li<sup>1</sup>, Ruocen Song<sup>1</sup>, Qianqian Song<sup>2</sup> and  
Xiao Fan<sup>1\*</sup>

<sup>1</sup>J. Crayton Pruitt Family Department of Biomedical Engineering, University of Florida, Gainesville, FL, United States, <sup>2</sup>Department of Health Outcomes and Biomedical Informatics, University of Florida, Gainesville, FL, United States

**Introduction:** Protein-DNA interactions are central to gene regulation, genome stability, and disease mechanisms. Identifying DNA-binding residues (DBRs) is critical for structural modeling, protein engineering, and therapeutic design. Although experimental approaches provide valuable insights, they remain low-throughput and resource-intensive. Computational methods offer scalable alternatives by leveraging protein sequential and structural information to predict DBRs.

**Methods:** We present PRIMED (Protein Residue Inference using Multilayer perceptron for Enhanced DNA-binding predictions), a machine learning framework that integrates protein representations of distinct biochemical and structural properties from three protein language models: ESM-2, ESM-3, and ESM-C. These representations are concatenated and processed by a multilayer perceptron to perform DBR predictions.

**Results:** PRIMED demonstrated strong performance across three benchmark datasets: Test-46 and Test-129 from a previous study, CLAPE-DB, and Test-10 K, which we curated from UniProtKB/Swiss-Prot. The model achieves an area under the Receiver Operating Characteristic curve (AUC) of 0.92 and a Matthews Correlation Coefficient (MCC) of 0.64 on Test-46, as well as an AUC of 0.93 and MCC of 0.45 on Test-129. On Test-10 K, PRIMED demonstrates generalizability across proteins with varying DBR percentages, maintaining competitive performance relative to the runner-up method, CLAPE-DB.

**Discussion:** These results highlight the effectiveness of integrating diverse protein language model representations for accurate, transferable DBR predictions.

### KEYWORDS

DNA-binding protein, DNA-binding residue, protein language model, supervised machine learning, transfer learning

## 1 Introduction

DNA-binding proteins (DBPs) play a central role in numerous cellular processes, including gene regulation, replication, repair, recombination, and chromatin organization (Takeda et al., 1983). These proteins recognize specific DNA sequential (Qu et al., 2019) or structural motifs (Shanahan et al., 2004) through conserved DNA-binding domains (Lee et al., 1993), such as helix-turn-helix (HTH) (Aravind et al., 2005), zinc fingers (Jantz et al., 2004), leucine zippers (Landschulz et al., 1988), and homeodomains (Laughon, 1991). Dysregulation of DBPs is implicated in diverse diseases (Uryu et al., 2008), including cancer (Shiroma et al., 2020),

developmental syndromes (Micucci et al., 2015), and neurodegeneration (Chen-Plotkin et al., 2010). For example, DBP p53 activates DNA damage response genes (Li et al., 2012), while NF- $\kappa$ B (Hayden et al., 2006) and c-Myc (Buggins et al., 2001) regulate immune and proliferative pathways.

A more granular understanding of DNA-binding residues (DBRs) (Schleif, 1988), which are the specific amino acids that mediate direct contacts with DNA, is essential for modeling protein-DNA interactions (Si et al., 2015), engineering DNA-binding specificity (Bogdanove et al., 2018), and annotating functional sites in novel proteins (Nagarajan et al., 2013). Mutations in DBRs may disrupt regulatory networks (Lozada-Chávez et al., 2008), potentially leading to diseases (Pandey and Loscalzo, 2023). Residue-level annotations also support applications in synthetic biology (Gong et al., 2024) and drug design (Xiong et al., 2011).

Experimental techniques such as site-directed mutagenesis (Carter, 1986), electrophoretic mobility shift assays (EMSA) (Hellman and Fried, 2007), chromatin immunoprecipitation followed by sequencing (ChIP-seq) (Song et al., 2016), and DNA affinity purification sequencing (DAP-seq) (Kadonaga and Tjian, 1986) have been instrumental in characterizing DBRs. However, these approaches remain low-throughput, resource-intensive, and difficult to scale to proteome-level annotations (Rastogi et al., 2018). As of June 2023, the UniProt database held approximately 246 million protein sequences (UniProt Consortium, 2019), but fewer than 0.1% had experimentally validated annotations for DBRs (Zhu et al., 2024). In contrast, integrative estimates suggest that 18–36% of human proteins may possess DNA-binding functions (Peled et al., 2016), inferred from domain composition and transcriptional regulatory roles. This gap reflects both a limited experimental throughput and incomplete functional annotation for a significant portion of the human proteome. This observation underscores the urgent need for accurate computational approaches that can operate at the proteome scale (Gromiha and Nagarajan, 2013), enabling the high-throughput prediction of DBPs and their corresponding contact residues.

The evolution of protein digital representations in computational methods reflects a shift from handcrafted, simple features to learned, comprehensive ones. Early approaches relied on a small set, typically 20–30, of manually engineered features, such as physicochemical descriptors, amino acid composition, or position-specific scoring matrices (PSSMs) (Jones, 1999). While these features captured certain predictive patterns, they were insufficient to fully capture the high-dimensional, context-dependent complexity of protein sequences (Wang et al., 2019). Such limited representations failed to capture the nuanced interactions between residues, particularly in diverse, structurally complex proteins like DBPs (Ahmad and Sarai, 2005; Patro et al., 2012; Cetin and Sefer, 2025; Sefer, 2025). To address this, multiple sequence alignments (MSAs) were introduced to incorporate evolutionary context (Rao et al., 2021), offering a richer view of conservation and co-evolutionary patterns (Chatzou et al., 2016). For instance, TargetDBP leveraged MSAs to construct PSSMs that capture conserved sequence motifs critical for identifying DBPs (Hu et al., 2019). However, MSA generation is computationally expensive and sensitive to MSA depth (Chowdhury and Garai, 2017). The emergence of large protein language models (pLMs), such as the Evolutionary Scale Modeling (ESM) family (Lin et al., 2023; Hayes et al., 2025; ESM Team, 2024), marked a paradigm shift.

Unlike manually engineered features or MSA-derived profiles, pLMs trained on millions of proteins across diverse species implicitly capture structural, functional, and evolutionary information without

relying on curated annotations or alignment heuristics (Lin et al., 2023). By learning the contextual relationships between amino acids at scale, pLMs encode biochemical and biophysical properties in a data-driven manner, enabling downstream models (Qiu and Wei, 2023) to access rich information about the residue environment (Meier et al., 2021), secondary structure (Chowdhury et al., 2022), binding potential (Zhang and Liu, 2024), and other relevant properties. This not only reduces dependence on domain-specific preprocessing but also enables generalization across previously unseen protein families. As a result, pLM representations are particularly suited for tasks such as DBR prediction, where the scarcity of comprehensive DBR features presents significant challenges. Ultimately, pLM-derived features offer a scalable, unbiased, alignment-free, and biologically informed representation of protein sequences, enabling diverse and generalizable residue-level inference despite their limited direct interpretability (Li et al., 2024).

In this study, we introduce PRIMED (Protein Residue Inference using MLP for Enhanced DNA-binding predictions), a high-throughput machine learning framework for predicting DBRs by integrating complementary representations from multiple pre-trained pLMs (Lin et al., 2023; Hayes et al., 2025; ESM Team, 2024). PRIMED captures diverse biological signals relevant to sequence and structural context, enabling accurate and generalizable residue-level DNA-binding annotation across a broad range of proteins.

## 2 Materials and methods

### 2.1 Datasets for training and evaluation

We used the training datasets from CLAPE-DB (Liu and Tian, 2024), which are based on curated DBP datasets originally introduced by DBPred (Patiyal et al., 2022) and GraphBind (Xia et al., 2021). These datasets contain residue-level annotations derived from experimentally resolved protein-DNA complexes. DBRs were defined as those containing any atom located within 0.5 Å plus the sum of van der Waals radii from any DNA atom. For model evaluation, we used three non-redundant test sets: Test-46 (46 DBPs) and Test-129 (129 DBPs) from CLAPE-DB, and Test-10 K (12,067 DBPs) curated from UniProt/Swiss-Prot. CLAPE-DB test datasets were compiled from previous studies, including GraphBind (Xia et al., 2021), DBPred (Patiyal et al., 2022), BindN+ (Wang et al., 2010), ProNA2020 (Qiu et al., 2020), and PDNA-62 (Ahmad et al., 2004). Residue-level annotations in the test sets followed the same distance-based criterion as those in the training set.

Since most proteins in the previous datasets are from the PDB, we curated a third test dataset from UniProtKB/Swiss-Prot to further assess the model's generalizability at scale. We constructed this dataset by scanning the entire UniProt database for entries containing either feature annotations with 'DNA-BIND' or textual descriptions indicating DNA-binding activity. Residues not annotated as DNA-binding were treated as non-binding. The DBR definition is less stringent compared to the distance-based criteria used in the PDB (Huang et al., 2009), resulting in more continuous binding regions. This difference in annotation standards and diverse protein sequences provides an independent dataset for assessing model generalizability.

To reduce sequence redundancy and prevent information leakage, we performed identifier-based filtering using UniProt-PDB ID mappings (UniProt Consortium, 2015) to ensure that no overlapping proteins from

UniProt and PDB were included in our datasets, even though their sequences are not identical (Zaru et al., 2023). We then applied a second round of pairwise sequence similarity filtering within all datasets. Proteins were clustered at 80% sequence identity using MMseqs2 (Steinegger and Söding, 2017), and only a single representative from each protein cluster was retained. This multi-stage filtering process ensured a non-redundant and partition-independent protein corpus.

The resulting datasets consisted of a training set with 1,198 proteins, along with three test sets containing 46, 129, and 12,067 proteins, respectively. A summary of the training and testing dataset compositions, including the percentage of annotated DBRs, is provided in Table 1.

To further characterize the residue-level properties of the annotated datasets, we summarized the amino acid composition of DNA-binding residues and compared it with the overall amino acid composition across all residues. The composition in the training set is summarized in Table 2, and summaries for the Test-129, Test-46, and Test-10 K are provided in Supplementary Tables S1–S3. Arginine (R) and lysine (K) are consistently the most enriched residues among DBRs across datasets. Leucine (L) and glutamate (E) are consistently depleted among DBRs relative to their background frequencies.

## 2.2 Protein representations

Unlike traditional prediction methods that rely on handcrafted protein features derived from prior human knowledge, we leveraged general, comprehensive representations from pre-trained pLMs, which are free of such biases. These representations can capture novel, previously unseen protein features. In this study, we employed three transformer-based pLMs from the ESM family. These models were trained on millions of protein sequences to capture biologically relevant sequence information, without relying on task-specific objectives, making them well-suited for diverse downstream applications. Specifically, we used the ESM-2 (esm2\_t36\_3B\_UR50D) (Lin et al., 2023), the ESM-3 (ESM3\_OPEN\_SMALL) (Hayes et al., 2025), and the ESM-C (esm\_c\_600m) (ESM Team, 2024).

Sequences were tokenized and passed through the respective pLMs, and we further trained the amino acid representation for DBR predictions. The representation dimensions for ESM-2, ESM-3, and ESM-C are 2,560, 1,536, and 1,152, respectively. We also concatenated the per-residue representations derived from different ESM models, considering that each model captures complementary aspects of protein properties. Although ESM-2 and ESM-C are both sequence-based protein language models, they differ in architecture, training scale, and optimization strategy, leading to distinct residue-level representations. ESM-2 is trained at a larger scale and is effective at capturing long-range evolutionary and contextual dependencies across protein sequences (Lin et al., 2023). In contrast, ESM-C incorporates updated architectural and training refinements that emphasize efficient local contextual encoding (ESM Team, 2024). ESM-3 further complements these sequence-based representations by incorporating structure-aware pretraining objectives (Hayes et al., 2025), enabling it to encode information related to three-dimensional protein organization. Although ESM-2 and ESM-C are both sequence-based models, differences in their architectures and training regimes yield distinct residue-level representations, whereas ESM-3 provides complementary structure-aware features derived from its structure-focused pretraining. Together, these models capture complementary sequence- and structure-level signals relevant to DNA-binding residue prediction.

TABLE 1 Summary of training and testing dataset composition.

Subset name	Number of proteins	Number of residues	% binding residues
Training	1,198	470,074	6.35%
Test-46	46	10,876	8.87%
Test-129	129	37,515	5.97%
Test-10 K	12,067	4,874,438	12.98%

TABLE 2 Amino acid enrichment and depletion at DNA-binding residues in the training dataset.

Residue	DBR (%)	Background (%)	Enrichment (DBR/background)
R	16.41	5.95	2.76
K	13.16	6.65	1.98
Y	5.07	3.09	1.64
W	1.79	1.12	1.59
H	3.78	2.40	1.58
N	5.74	4.16	1.38
T	6.57	5.15	1.28
S	8.19	7.18	1.14
Q	4.72	4.56	1.04
G	6.33	6.43	0.98
F	3.34	3.68	0.91
M	1.80	2.20	0.82
P	3.10	5.42	0.57
D	2.98	5.31	0.56
A	4.05	7.38	0.55
I	2.79	5.22	0.53
V	3.14	6.09	0.52
C	0.65	1.41	0.46
E	3.05	7.12	0.43
L	3.33	9.48	0.35

DBR (%) denotes the percentage of each amino acid among DNA-binding residues, while Background (%) denotes the percentage among all residues in the dataset. The enrichment ratio was calculated as DBR% divided by Background%, and residues are sorted in descending order of this ratio.

## 2.3 Model architecture

We implemented an MLP for DBR predictions using per-residue pLM representation as features. The network comprised fully connected layers with ReLU activations, yielding binding probabilities for each residue through a sigmoid function. We used binary cross-entropy as the loss function, and a dropout rate of 0.5 in all hidden layers to reduce overfitting. Early stopping was applied based on validation loss, with a patience of five epochs and a minimum improvement threshold of  $10^{-4}$  in validation loss. All weights were initialized with default settings.

PLM selection and hyperparameter tuning were performed using a five-fold cross-validation approach. Model evaluation reflects differences in protein representation and architecture. The evaluation metric was the mean area under the Receiver Operating Characteristic

(ROC) curve (AUC) across folds. The hyperparameters examined were the number of hidden layers and the dimensionality of each hidden layer. After selecting the optimal hyperparameters, the final model was retrained on the entire training set.

The MLP model computed the output probability  $\hat{y}$  through a sequence of nonlinear transformations applied to the input vector  $x$ . For an MLP with  $n$  hidden layers, the feedforward computation was expressed as:

$$h^{(0)} = x$$

$$h^{(i)} = \text{Dropout}\left(\text{ReLU}\left(W^{(i)}h^{(i-1)} + b^{(i)}\right)\right), \text{ for } i = 1, \dots, n$$

$$\hat{y} = \sigma\left(W^{(n+1)}h^{(n)} + b^{(n+1)}\right)$$

Here,  $W^{(i)}$  and  $b^{(i)}$  denote the weight matrices and bias vectors of the  $i$ th layer;  $h^{(i)}$  represents the hidden state;  $\text{ReLU}(\cdot)$  is the rectified linear unit activation function;  $\text{Dropout}(\cdot)$  randomly sets elements to zero with probability  $p$ ; and  $\sigma(\cdot)$  is the sigmoid activation function used to compute the final probability score.

## 2.4 Evaluation

Model performance was evaluated at the residue level using the following metrics: AUC, Matthew's correlation coefficient (MCC), sensitivity, specificity, and F1 score. AUC evaluates how well the model's predicted scores distinguish between binding and non-binding residues, while MCC provides a balanced assessment of binary classification performance, particularly under class imbalance. Let TP, TN, FP, and FN denote the numbers of true positives, true negatives, false positives, and false negatives, respectively. Sensitivity is defined as

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

and specificity is defined as

$$\text{Specificity} = \frac{TN}{TN + FP}$$

The F1 score is the harmonic mean of precision and recall, given by

$$F1 = \frac{2TP}{2TP + FP + FN}$$

The MCC summarizes binary classification performance by jointly considering all four confusion matrix terms:

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

At the same time, the optimal classification threshold was determined to maximize MCC during the final training phase. Evaluation was conducted on the test sets (Test-46, Test-129, and Test-10 K) (see Section 2.1) and compared with several existing methods, including

CLAPE-DB (Liu and Tian, 2024), DRNAPred (Yan and Kurgan, 2017), DNAPred (Zhu et al., 2019), SVMnuc (Su et al., 2019), NCBRPred (Zhang et al., 2021), and DBPred (Patiyal et al., 2022). Their evaluation metrics on Test-46 and Test-129 were extracted from the CLAPE-DB study. We also collected the predictions of CLAPE-DB on the Test-10 K dataset and compared them with those of our model to assess the generalization performance.

To assess statistical significance, we performed a paired two-sided  $t$ -test over ten iterations. In each iteration, AUC scores were computed on 50% of randomly selected test proteins.  $p$ -values  $< 0.05$  were considered statistically significant.

## 2.5 Feature space visualization and correlation with solvent accessibility

Solvent accessibility has long been recognized as a critical structural determinant for identifying DBRs. Previous methods, such as BindN+ (Wang et al., 2010), DISPLAR (Tjong and Zhou, 2007), DNABind (Liu and Hu, 2013), and PDNAsite (Zhou et al., 2016), have all incorporated predicted or experimentally derived solvent accessibility as a key predictor, demonstrating that surface-exposed residues are more likely to interact with DNA. Building on this insight, we examined whether amino acid representations derived from pre-trained pLMs implicitly encode such structural information. To evaluate this, we calculated the Pearson correlation coefficient between the low-dimensional representation of amino acids and solvent accessibility. Per-residue representations were projected into two dimensions using the Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018). Solvent accessibility was estimated using the Dictionary of Secondary Structure of Proteins (DSSP) algorithm (Kabsch and Sander, 1983), which takes the three-dimensional atomic coordinates from PDB structures as input. We then normalized the solvent accessibility values using min-max scaling to obtain relative solvent accessibility (RSA) values for each protein, which range from 0 (completely buried) to 1 (fully exposed). We visualized the first two UMAP dimensions of Test-129, coloring residues by their RSA values. The protein sequences in Test-46 are not always consistent with their corresponding PDB entries; therefore, they were not included in this analysis.

## 3 Results

### 3.1 Model architecture and hyperparameter selection

We first trained MLPs using protein features from three pLMs: ESM-2, ESM-3, and ESM-C. For each model, we performed five-fold cross-validation on the training dataset to evaluate performance and tune the model's hyperparameters. Detailed hyperparameter tuning results for the three models are provided in [Supplementary Tables S4–S6](#). We then tested whether their combined representations offer complementary information.

As shown in [Table 3](#), the best-performing architecture consisted of two hidden layers with 256 and 64 units, achieving the highest AUC of 0.960. While single-layer networks achieved comparable but slightly lower performance (AUC in between 0.955 and 0.957), adding a third layer did not further improve accuracy and instead reduced generalization. These results indicate that shallow models may underfit

complex patterns, whereas overly deep networks tend to overfit the training data. [Supplementary Figure S1](#) shows cross-entropy loss converging across all five folds. Early stopping was triggered consistently between epochs 5 and 10, confirming that the model converged efficiently without performance degradation. The training loss curves for the three individual pLMs are provided in [Supplementary Figures S2–S4](#).

TABLE 3 The hyperparameter selection for the MLP architecture of PRIMED was evaluated using five-fold cross-validation on the training set.

First layer	Second layer	Third layer	Validation AUC
8	-	-	0.957
64	-	-	0.956
128	-	-	0.956
256	-	-	0.957
512	-	-	0.955
1,024	-	-	0.955
128	64	-	0.958
<b>256</b>	<b>64</b>	-	<b>0.960</b>
512	64	-	0.952
512	128	-	0.953
1,024	64	-	0.944
1,024	128	-	0.944
1,024	256	-	0.942
1,536	256	-	0.939
1,024	128	8	0.956
1,024	128	64	0.920
1,024	256	8	0.955
1,024	256	64	0.954
1,024	256	128	0.930

Shown are the averaged AUC across folds for each configuration. The best-performing setting is highlighted in bold.

As summarized in [Table 4](#), all models demonstrated strong predictive capability, supporting the effectiveness of the representations learned by the pre-trained pLMs. The PRIMED model, which concatenates ESM-2, ESM-3, and ESM-C representations, yielded the most robust predictions on the validation sets (AUC = 0.960, MCC = 0.676). Statistical analyses summarized in [Supplementary Table S7](#) show that PRIMED achieves significantly higher AUC than ESM-2 and ESM-3, while showing no significant difference compared with ESM-C. In terms of MCC, PRIMED is significantly better than ESM-3 and ESM-C but not significantly different from ESM-2. This performance gain suggests that concatenating representations from multiple foundation models captures complementary contextual information, thereby enhancing the model's discriminative power for classifying DBPs. The overall workflow of the PRIMED model is illustrated in [Figure 1](#).

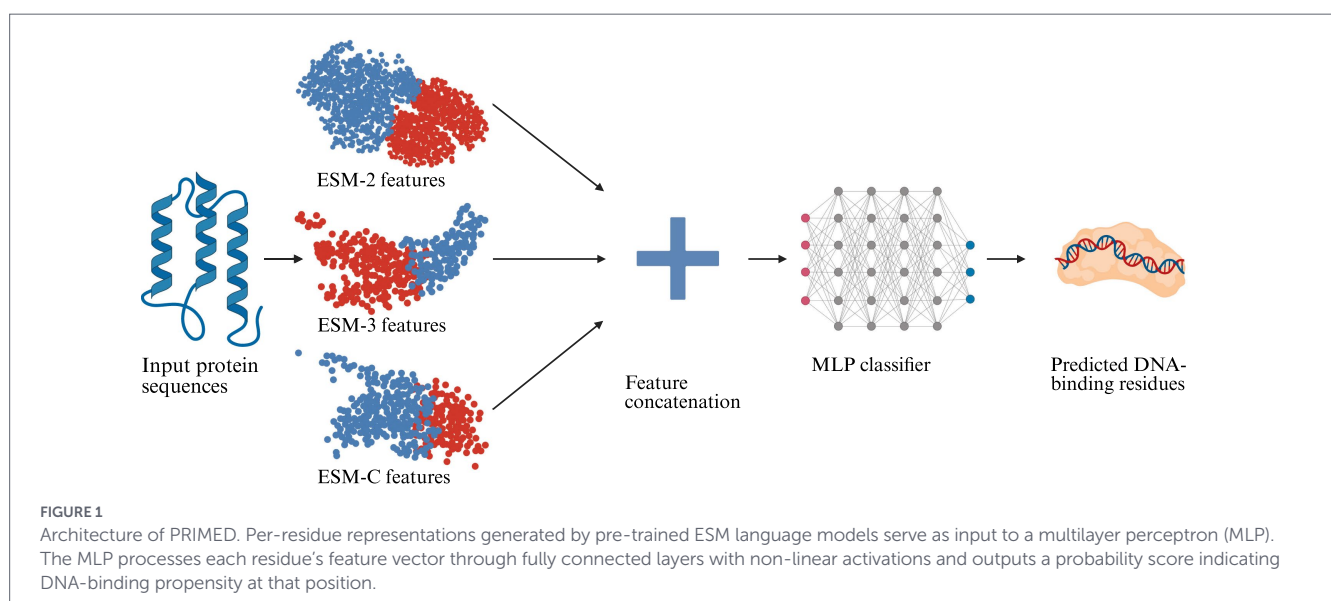
### 3.2 Prediction performance on benchmark datasets

[Figure 2](#) illustrates the distribution of predicted scores, showing clear separation between DNA-binding (positive) and non-binding (negative) residues across Test-46 and Test-129 datasets. The kernel density estimates reveal that negative residues are predominantly concentrated near zero, while positive residues show a broader distribution, with higher scores extending toward and beyond the threshold region. The threshold of 0.813, optimized on the training set, effectively discriminates between the two classes on the test sets,

TABLE 4 DBR prediction performance using different ESM-based model representations.

Model	Validation AUC	Validation MCC
ESM-2	0.954	0.635
ESM-3	0.957	0.522
ESM-C	0.960	0.649
PRIMED	0.960	0.676

Reported values are the mean AUC and MCC across five folds based on the best-performing model parameters.



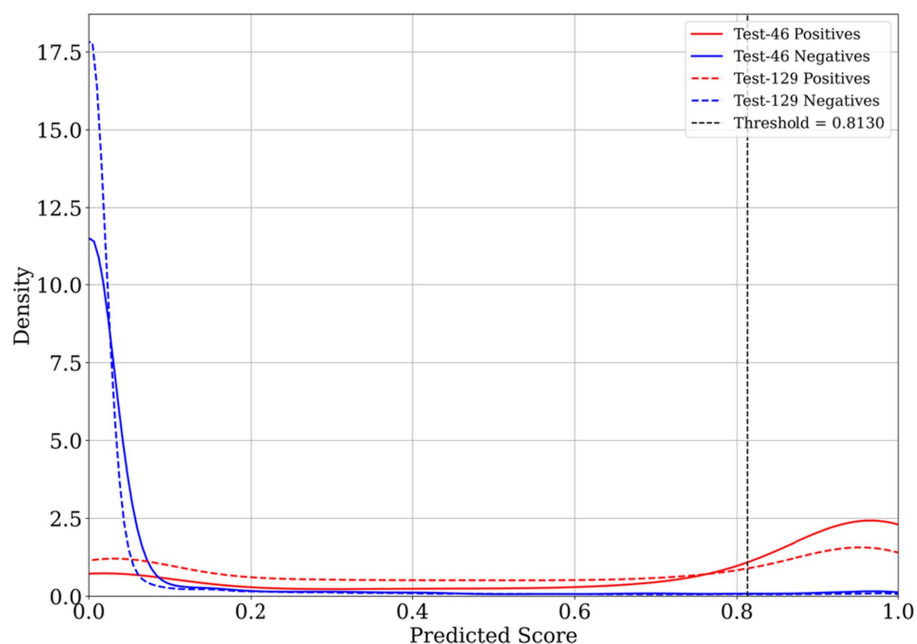


FIGURE 2

Distribution of predicted residue-level binding scores from PRIMED on Test-46 and Test-129 datasets. Kernel density estimates are shown for binding residues (red lines) and non-binding residues (blue lines), with solid lines representing Test-46 data and dashed lines representing Test-129 data. The vertical dashed line indicates the threshold for binarization as 0.813.

TABLE 5 Residue-level classification performance of PRIMED and other existing methods on two benchmark datasets.

Dataset	Method	Sensitivity	Specificity	F1	AUC	MCC
Test-46	PRIMED	0.631	<b>0.975</b>	<b>0.668</b>	<b>0.919</b>	<b>0.631</b>
	CLAPE-DB	<b>0.747</b>	0.835	0.434	0.871*	0.401*
	DRNAPred	0.677	0.692	0.291	0.755	0.226
	DNAPred	0.671	0.655	0.254	0.730	0.194
	SVMnuc	0.668	0.666	0.250	0.715	0.192
	NCBRPred	0.677	0.674	0.265	0.713	0.207
	DBPred	0.708	0.784	0.362	0.794	0.320
Test-129	PRIMED	0.375	<b>0.985</b>	<b>0.465</b>	<b>0.925</b>	<b>0.455</b>
	CLAPE-DB	<b>0.464</b>	0.955	0.427	0.881*	0.389*
	DRNAPred	0.233	0.937	0.210	0.693	0.155
	DNAPred	0.396	0.954	0.373	0.845	0.332
	SVMnuc	0.316	0.966	0.341	0.812	0.304
	NCBRPred	0.312	0.969	0.347	0.823	0.313

Reported are the sensitivity, specificity, F1 score, AUC, and MCC values for each model. \* indicate that a statistical test was performed, and PRIMED significantly outperformed the corresponding method. The highest value for each metric within a dataset is shown in bold.

demonstrating the model's robust classification capability across different test sets.

As shown in Table 5, PRIMED achieves the strongest performance among all sequence-based predictors across both benchmark datasets, consistently attaining the highest AUC and MCC values. On Test-46, PRIMED yields an AUC of 0.919 and an MCC of 0.631, outperforming the second-best model, CLAPE-DB, which scored 0.871 in AUC and 0.401 in MCC. These results correspond to relative improvements of 5.5 and 57%, respectively. Notably, CLAPE-DB achieves higher sensitivity (0.747) than PRIMED (0.631), but this comes at the expense of reduced specificity (0.835 versus 0.975). In contrast, PRIMED

maintains a more balanced performance, achieving both high specificity and competitive sensitivity, which translates into more balanced identification of true binding residues while minimizing false positives. The advantages of PRIMED are similarly evident on Test-129. It records an AUC of 0.925 and an MCC of 0.455, compared to 0.881 and 0.389 from CLAPE-DB. This reflects additional gains of 5.0% in AUC and 17% in MCC. Although PRIMED's sensitivity on Test-129 is modest at 0.375, it achieves the highest specificity among all evaluated methods at 0.985, exceeding that of CLAPE-DB (0.955), SVMnuc (0.966), and NCBRPred (0.969). These findings underscore PRIMED's strength in precisely distinguishing DBRs from non-binding residues,

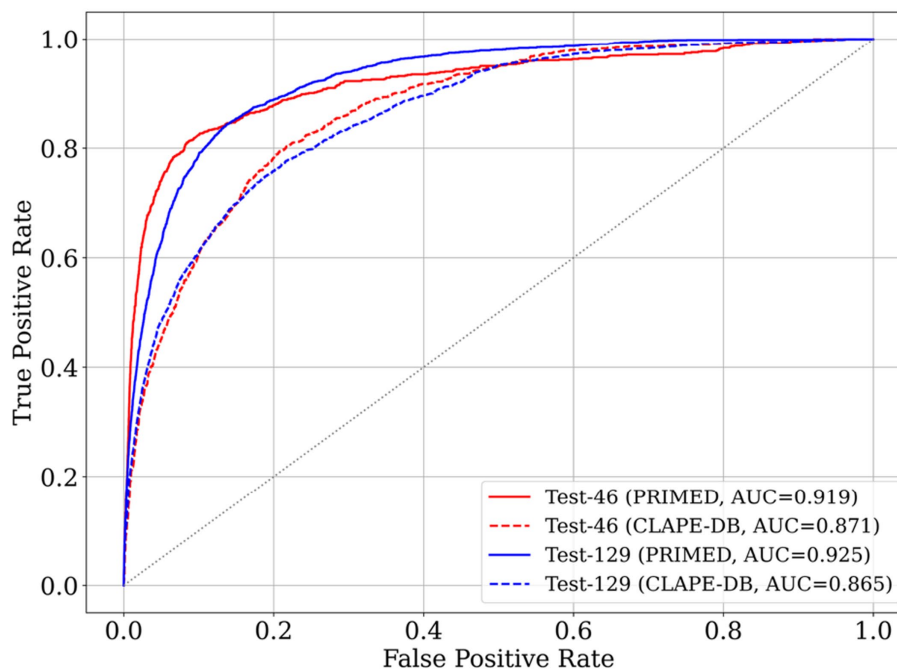


FIGURE 3  
ROC curves for PRIMED and CLAPE-DB on Test-46 and Test-129.

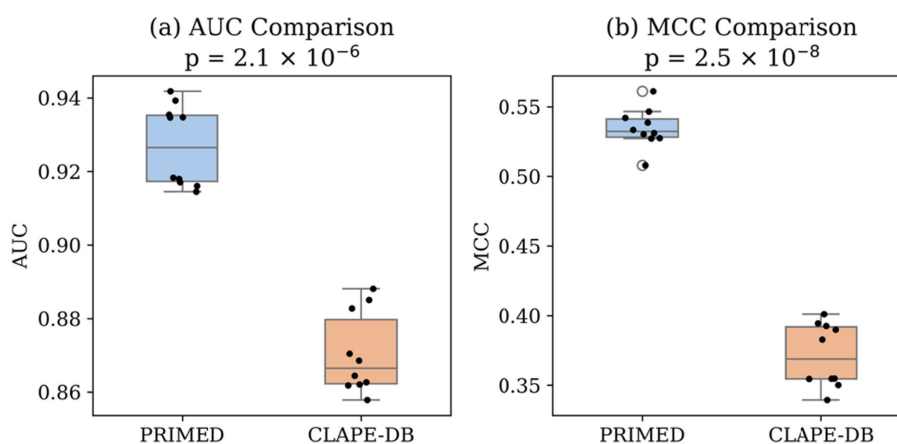


FIGURE 4  
(a) Boxplot of AUC scores for ten sampled protein subsets, comparing PRIMED to CLAPE-DB. (b) Boxplot of MCC scores from the same protein subsets. Each dot represents the per-subset prediction performance (sampled at 50% of the combined Test-46 and Test-129 benchmark datasets).

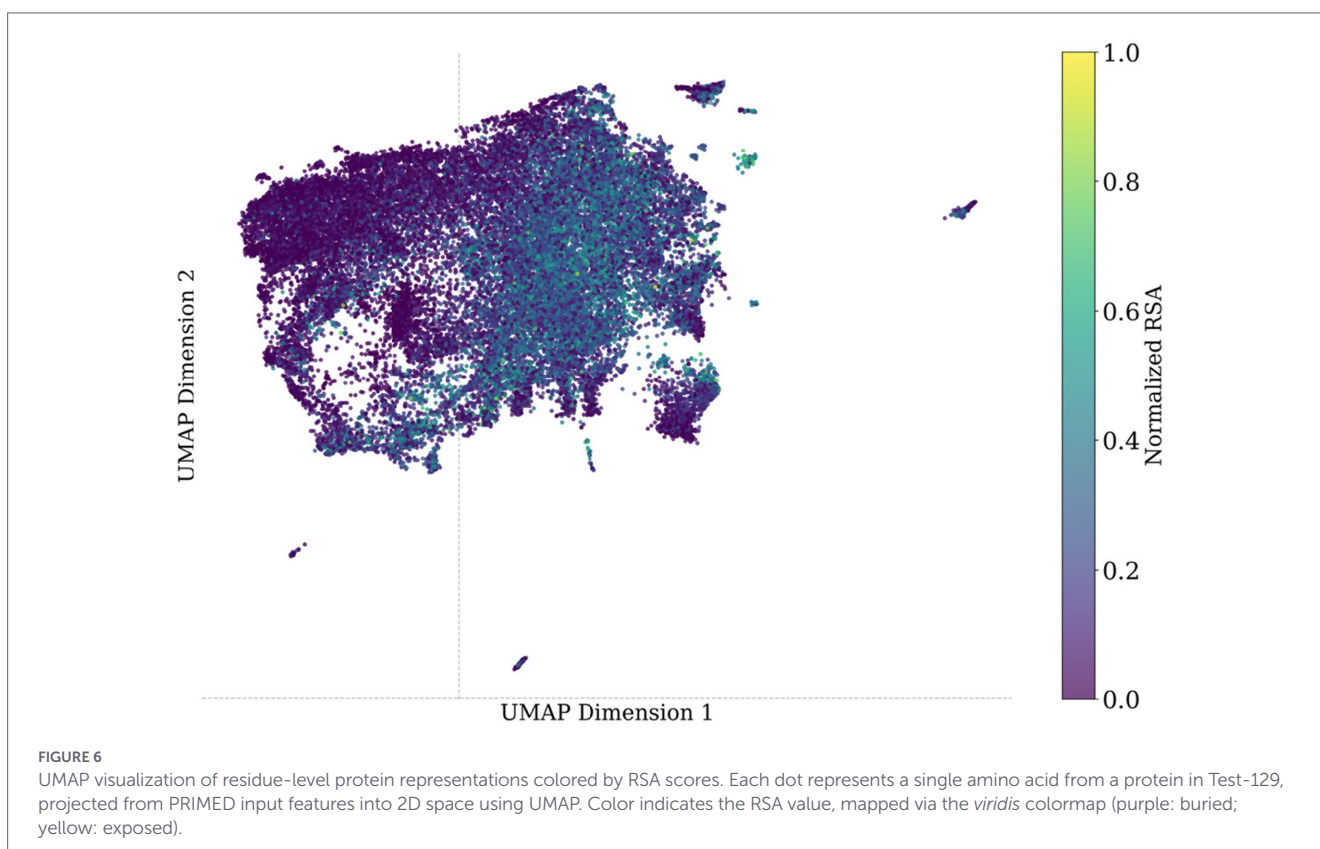
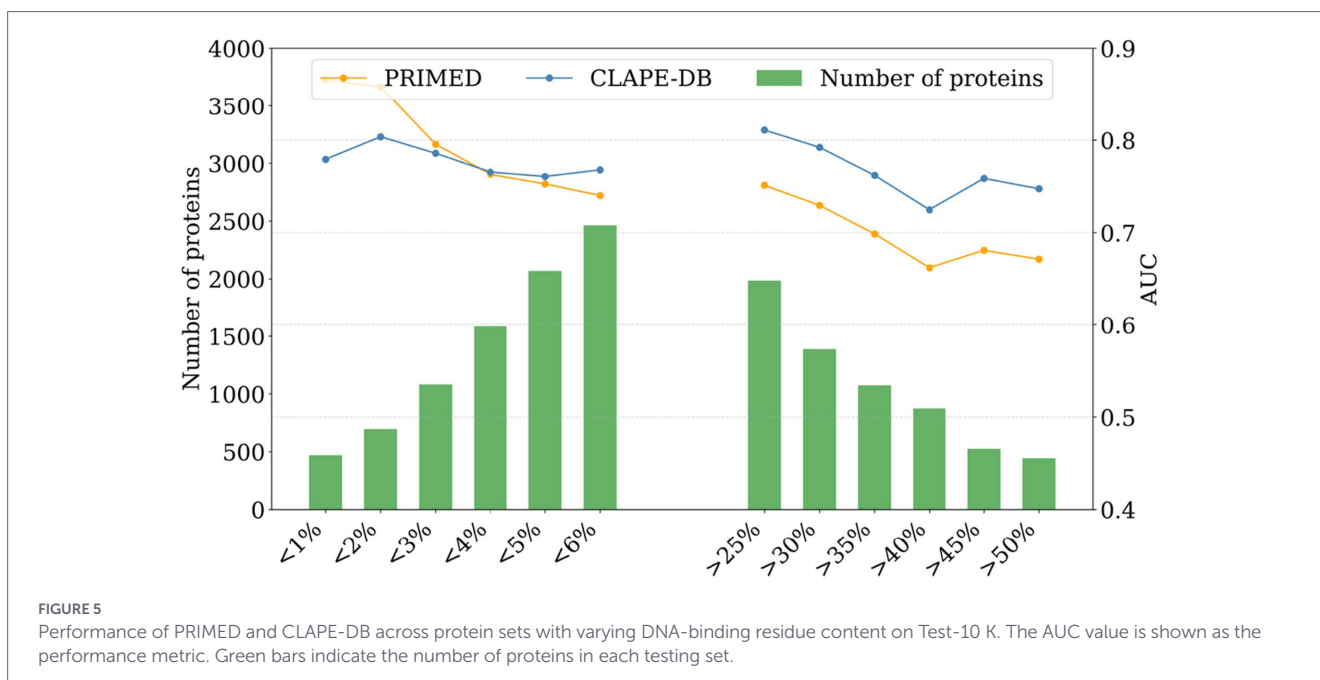
a property that is especially valuable in large-scale annotation tasks where controlling the false positive rate is critical. These results demonstrate that PRIMED offers a robust and generalizable solution for DBR prediction. Its consistently strong performance across both datasets, especially in terms of specificity and composite metrics such as MCC, highlights its practical advantages over existing methods.

As shown in Figure 3, the ROC curves provide a clear visualization of the performance difference between PRIMED and CLAPE-DB. Across both Test-46 and Test-129, PRIMED maintains a consistently higher true positive rate over nearly the entire false positive rate spectrum. PRIMED demonstrates a sharper rise in sensitivity within the low-FPR region (FPR < 0.1), a region that is especially important for high-confidence residue-level annotation. This behavior indicates that PRIMED is more effective at capturing true DBRs while

minimizing false positives, complementing the quantitative improvements reported in Table 5 and reinforcing its reliability across diverse protein sequences.

Additionally, we performed a statistical analysis, combining the two test datasets. As shown in Figure 4, PRIMED significantly outperformed CLAPE-DB across both metrics. The average AUC improvement was statistically significant ( $p$ -value =  $2.1 \times 10^{-6}$ ), and the MCC comparison also showed a highly significant difference ( $p$ -value =  $2.5 \times 10^{-8}$ ). These results confirm that the improvements observed in overall model performance (Section 3.2) are robust and statistically reliable.

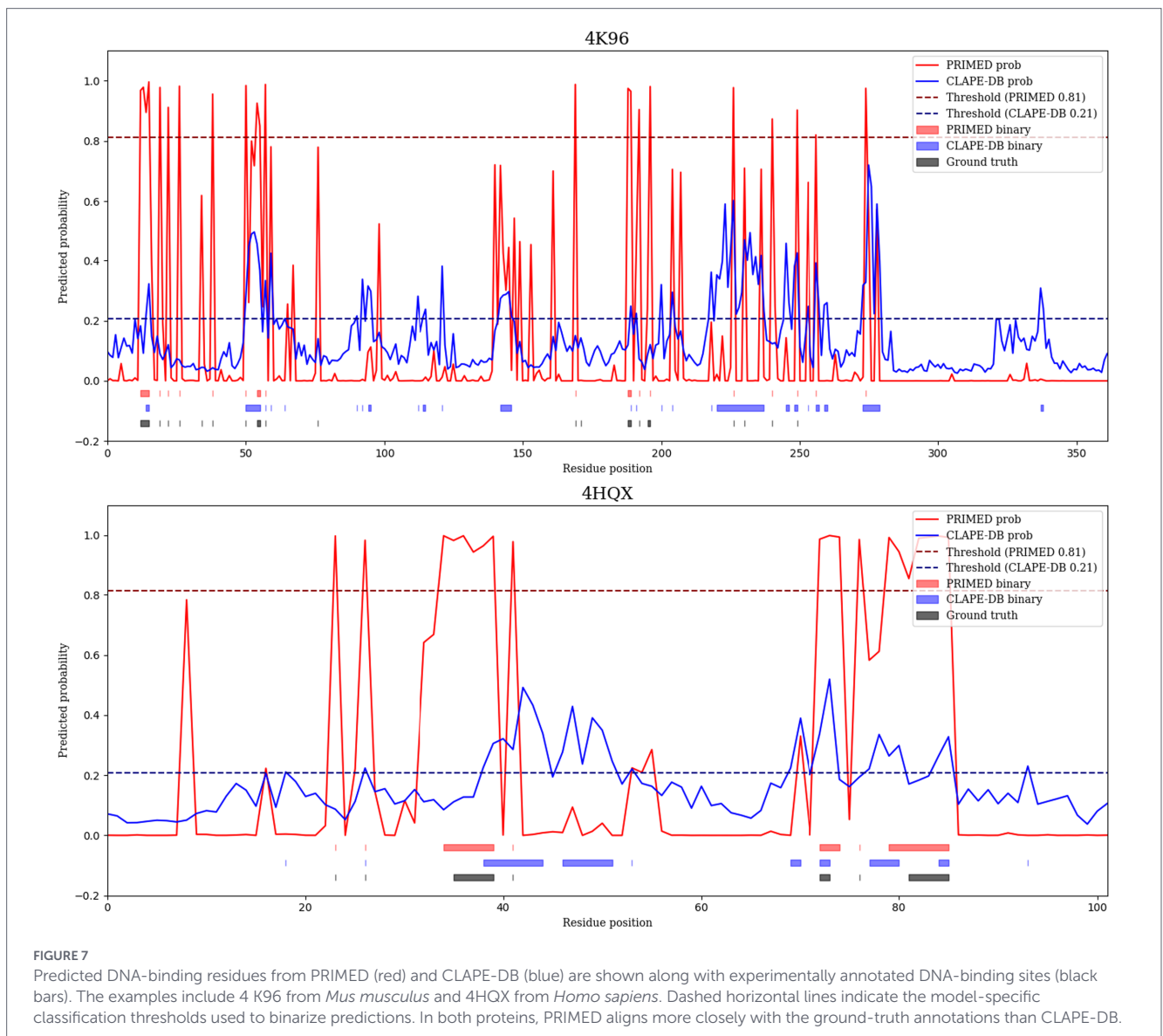
On the datasets curated from the UniProt database (Test-10 K), we evaluated the model generalizability of PRIMED and CLAPE-DB across diverse evaluation sets with different DBR



percentages (e.g., <1, <2%, ..., >25%, ..., >50%). Overall, both models exhibit reduced AUC values on the Test-10 K dataset compared to the PDB-based test datasets. Such a phenomenon reflects different DBR ascertainment and/or sequence diversity in the UniProt database. Figure 5 shows that PRIMED consistently outperforms CLAPE-DB for proteins with low DBR content (<6%), highlighting its high specificity. In proteins with high DBR content (>25%), CLAPE-DB shows stronger performance, consistent with

its high sensitivity. These results suggest the importance of harmonizing the definition of DBR and improving the generalizability of computational models.

In addition, PRIMED applies a threshold optimized for MCC during training (0.813), which balances sensitivity and specificity under class imbalance. However, this threshold can be adjusted depending on application needs. For example, lowering it to increase sensitivity at the cost of specificity may be desirable in contexts where



missing DBRs is more detrimental than including false positives. This flexibility allows PRIMED to adapt to different prioritization strategies in downstream tasks. Reducing PRIMED's threshold to 0.5 improves sensitivity at the expense of classification balance, resulting in an MCC of 0.311 on the full Test-10 K dataset. In comparison, CLAPE-DB achieves an MCC of 0.353 at its predefined threshold of 0.207, while PRIMED attains an MCC of 0.298 at its own optimized threshold of 0.813. This illustrates the inherent trade-off between sensitivity and specificity introduced by threshold adjustments.

### 3.3 Interpretation of pLM representations

The UMAP plot in Figure 6 reveals a separation between solvent-exposed (high RSA, colored yellow) and buried (low RSA, colored blue) residues. UMAP1 exhibited a moderate correlation with RSA (Pearson's  $r = 0.41$ ), suggesting that the first UMAP component partially reflects variation in residue-level structural exposure. This correlation confirms that pre-trained protein language models capture biologically meaningful structural patterns associated with protein interactions.

### 3.4 Case studies

To illustrate the performance of our model, we visualized residue-level DBR predictions for two representative protein-DNA complexes: PDB IDs 4 K96 (Gao et al., 2013) and 4HGX (Davies et al., 2012). 4 K96 represents the crystal structure of the murine cyclic GMP-AMP synthase bound to double-stranded DNA, a key innate immune sensor that triggers type I interferon responses. The protein-DNA interface in 4 K96 is biologically critical for the activation of DNA sensing and signaling. 4HGX depicts the crystal structure of human PDGF-BB in complex with a modified DNA aptamer, highlighting a hormone-DNA interaction interface relevant to regulatory signaling pathways. For each protein, we compared DNA-binding scores predicted by PRIMED and CLAPE-DB against the ground truth.

As shown in Figure 7, the predicted probability profiles produced by PRIMED (red curves) more accurately align with real DNA-binding regions (indicated by black ground truth bars), while CLAPE-DB (blue curves) often overpredicts binding regions. Notably, our model captures DBRs in multiple regions that CLAPE-DB fails to identify.

## 4 Discussion

In this study, we developed a machine learning framework named PRIMED that integrates concatenated protein representations from three pLMs: ESM-2, ESM-3, and ESM-C, to predict DBRs. Our model, using a simple MLP classifier, achieved significant improvements in prediction performance and robustness across three independent benchmark datasets (Test-46, Test-129, Test-10 K). The ability to generalize across datasets suggests that our methodology captures transferable biological signals, surpassing recent benchmarks and reinforcing the shift toward self-supervised learning in bioinformatics.

Our findings align with the growing body of work leveraging pLMs to predict protein structure and function. In addition, UMAP-based projection demonstrated the biochemical context carried by the pLM representations. These results underscore the value of pLMs, offering expressive and scalable representations for fine-grained protein structural and functional predictions.

This work also benefits from the use of high-quality benchmark datasets, rigorous sample-redundancy removal, and consistent separation between the training and evaluation datasets. However, the lack of a unified definition of DBRs remains a major challenge in the field. Methods trained under one definition often fail to generalize to DBRs defined differently. Moreover, residues lacking DBR annotations are typically treated as non-binding, leading to inflated false negatives. We therefore advocate establishing a gold-standard definition of DBRs and developing more carefully curated benchmarks for training and testing computational methods.

Our results open new possibilities for high-throughput annotation of DBRs across diverse organisms. In conclusion, this study presents a scalable, sequence-based framework that achieves state-of-the-art accuracy in predicting DBRs using pre-trained pLM representations as features. By leveraging the pLMs and an MLP architecture, our approach balances interpretability, efficiency, and predictive power. These findings support the broader utility of pLM-derived protein representations in biological sequence analysis and highlight promising avenues for future work in protein structural and functional predictions.

## Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: <https://github.com/luoshu-zhang/PRIMED/tree/main/datasets>.

## Author contributions

LZ: Data curation, Formal analysis, Software, Validation, Visualization, Writing – original draft. XL: Software, Writing – review

& editing. RS: Writing – review & editing. QS: Writing – review & editing. XF: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Visualization, Writing – original draft.

## Funding

The author(s) declared that financial support was received for this work and/or its publication. XF and QS are supported by the National Institutes of Health (R00HG011490, DP2LM014811, and R35GM151089).

## Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Generative AI statement

The author(s) declared that Generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

## Publisher's note

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

## Supplementary material

The Supplementary material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/frai.2026.1763313/full#supplementary-material>

## References

Ahmad, S., Gromiha, M. M., and Sarai, A. (2004). Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics* 20, 477–486. doi: 10.1093/bioinformatics/btg432

Ahmad, S., and Sarai, A. (2005). PSSM-based prediction of DNA binding sites in proteins. *BMC Bioinformatics* 6, 1–6. doi: 10.1186/1471-2105-6-33

Aravind, L., Anantharaman, V., Balaji, S., Babu, M. M., and Iyer, L. M. (2005). The many faces of the helix-turn-helix domain: transcription regulation and beyond. *FEMS Microbiol. Rev.* 29, 231–262. doi: 10.1016/j.femsre.2004.12.008

Bogdanove, A. J., Bohm, A., Miller, J. C., Morgan, R. D., and Stoddard, B. L. (2018). Engineering altered protein–DNA recognition specificity. *Nucleic Acids Res.* 46, 4845–4871. doi: 10.1093/nar/gky289

- Buggins, A. G., Milojkovic, D., Arno, M. J., Lea, N. C., Mufti, G. J., Thomas, N. S. B., et al. (2001). Microenvironment produced by acute myeloid leukemia cells prevents T cell activation and proliferation by inhibition of NF- $\kappa$ B, c-Myc, and pRb pathways. *J. Immunol.* 167, 6021–6030. doi: 10.4049/jimmunol.167.10.6021
- Carter, P. (1986). Site-directed mutagenesis. *Biochem. J.* 237, 1–7. doi: 10.1042/bj2370001
- Cetin, S., and Sefer, E. (2025). A graphlet-based explanation generator for graph neural networks over biological datasets. *Curr. Bioinforma.* 20, 840–851. doi: 10.2174/40115748936355418250114104026
- Chatzou, M., Magis, C., Chang, J.-M., Kemena, C., Bussotti, G., Erb, I., et al. (2016). Multiple sequence alignment modeling: methods and applications. *Brief. Bioinform.* 17, 1009–1023. doi: 10.1093/bib/bbv099
- Chen-Plotkin, A. S., Lee, V. M.-Y., and Trojanowski, J. Q. (2010). TAR DNA-binding protein 43 in neurodegenerative disease. *Nat. Rev. Neurol.* 6, 211–220. doi: 10.1038/nrneurol.2010.18
- Chowdhury, R., Bouatta, N., Biswas, S., Floristean, C., Kharkar, A., Roy, K., et al. (2022). Single-sequence protein structure prediction using a language model and deep learning. *Nat. Biotechnol.* 40, 1617–1623. doi: 10.1038/s41587-022-01432-w
- Chowdhury, B., and Garai, G. (2017). A review on multiple sequence alignment from the perspective of genetic algorithm. *Genomics* 109, 419–431. doi: 10.1016/j.ygeno.2017.06.007
- Davies, D. R., Gelinas, A. D., Zhang, C., Rohloff, J. C., Carter, J. D., O'Connell, D., et al. (2012). Unique motifs and hydrophobic interactions shape the binding of modified DNA ligands to protein targets. *Proc. Natl. Acad. Sci.* 109, 19971–19976. doi: 10.1073/pnas.1213933109
- ESM Team. (2024). ESM Cambrian: Revealing the Mysteries of Proteins with Unsupervised Learning in, EvolutionaryScale Website.
- Gao, P., Ascano, M., Wu, Y., Barchet, W., Gaffney, B. L., Zillinger, T., et al. (2013). Cyclic [G (2', 5') pA (3', 5') p] is the metazoan second messenger produced by DNA-activated cyclic GMP-AMP synthase. *Cell* 153, 1094–1107. doi: 10.1016/j.cell.2013.04.046
- Gong, X., Zhang, J., Gan, Q., Teng, Y., Hou, J., Lyu, Y., et al. (2024). Advancing microbial production through artificial intelligence-aided biology. *Biotechnol. Adv.* 74:108399. doi: 10.1016/j.biotechadv.2024.108399
- Gromiha, M. M., and Nagarajan, R. (2013). Computational approaches for predicting the binding sites and understanding the recognition mechanism of protein–DNA complexes. *Adv. Protein Chem. Struct. Biol.* 91, 65–99. doi: 10.1016/B978-0-12-411637-5.00003-2
- Hayden, M., West, A., and Ghosh, S. (2006). NF- $\kappa$ B and the immune response. *Oncogene* 25, 6758–6780. doi: 10.1038/sj.onc.1209943
- Hayes, T., Rao, R., Akin, H., Sofroniew, N. J., Oktay, D., Lin, Z., et al. (2025). Simulating 500 million years of evolution with a language model. *Science* 387:eads0018. doi: 10.1126/science.ads0018
- Hellman, L. M., and Fried, M. G. (2007). Electrophoretic mobility shift assay (EMSA) for detecting protein–nucleic acid interactions. *Nat. Protoc.* 2, 1849–1861. doi: 10.1038/nprot.2007.249
- Hu, J., Zhou, X.-G., Zhu, Y.-H., Yu, D.-J., and Zhang, G.-J. (2019). TargetDBP: accurate DNA-binding protein prediction via sequence-based multi-view feature learning. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 17, 1419–1429. doi: 10.1109/TCBB.2019.2893634
- Huang, Y.-F., Huang, C.-C., Liu, Y.-C., Oyang, Y.-J., and Huang, C.-K. (2009). DNA-binding residues and binding mode prediction with binding-mechanism concerned models. *BMC Genomics* 10:S23. doi: 10.1186/1471-2164-10-S3-S23
- Jantz, D., Amann, B. T., Gatto, G. J., and Berg, J. M. (2004). The design of functional DNA-binding proteins based on zinc finger domains. *Chem. Rev.* 104, 789–800. doi: 10.1021/cr020603o
- Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* 292, 195–202. doi: 10.1006/jmbi.1999.3091
- Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22, 2577–2637. doi: 10.1002/bip.360221211
- Kadonaga, J. T., and Tjian, R. (1986). Affinity purification of sequence-specific DNA binding proteins. *Proc. Natl. Acad. Sci.* 83, 5889–5893. doi: 10.1073/pnas.83.16.5889
- Landschulz, W. H., Johnson, P. F., and McKnight, S. L. (1988). The leucine zipper: a hypothetical structure common to a new class of DNA binding proteins. *Science* 240, 1759–1764. doi: 10.1126/science.3289117
- Laughon, A. (1991). DNA binding specificity of homeodomains. *Biochemistry* 30, 11357–11367. doi: 10.1021/bi00112a001
- Lee, M. S., Kliewer, S. A., Provencal, J., Wright, P. E., and Evans, R. M. (1993). Structure of the retinoid X receptor  $\alpha$  DNA binding domain: a helix required for homodimeric DNA binding. *Science* 260, 1117–1121. doi: 10.1126/science.8388124
- Li, F.-Z., Amini, A. P., Yue, Y., Yang, K. K., and Lu, A. X. (2024). Feature reuse and scaling: understanding transfer learning with protein language models. *bioRxiv*. [Preprint]. doi: 10.1101/2024.02.05.578959
- Li, M., He, Y., Dubois, W., Wu, X., Shi, J., and Huang, J. (2012). Distinct regulatory mechanisms and functions for p53-activated and p53-repressed DNA damage response genes in embryonic stem cells. *Mol. Cell* 46, 30–42. doi: 10.1016/j.molcel.2012.01.020
- 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, 1123–1130. doi: 10.1126/science.ade2574
- Liu, R., and Hu, J. (2013). DNABind: a hybrid algorithm for structure-based prediction of DNA-binding residues by combining machine learning-and template-based approaches. *Proteins* 81, 1885–1899. doi: 10.1002/prot.24330
- Liu, Y., and Tian, B. (2024). Protein–DNA binding sites prediction based on pre-trained protein language model and contrastive learning. *Brief. Bioinform.* 25:bbad488. doi: 10.1093/bib/bbad488
- Lozada-Chávez, I., Angarica, V. E., Collado-Vides, J., and Contreras-Moreira, B. (2008). The role of DNA-binding specificity in the evolution of bacterial regulatory networks. *J. Mol. Biol.* 379, 627–643. doi: 10.1016/j.jmb.2008.04.008
- McInnes, L., Healy, J., and Melville, J. (2018). Umap: uniform manifold approximation and projection for dimension reduction. *arXiv*. [Preprint] arXiv:1802.03426. doi: 10.48550/arXiv.1802.03426
- Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T., and Rives, A. (2021). Language models enable zero-shot prediction of the effects of mutations on protein function, advances in neural information processing systems 34, 29287–29303.
- Micucci, J. A., Sperry, E. D., and Martin, D. M. (2015). Chromodomain helicase DNA-binding proteins in stem cells and human developmental diseases. *Stem Cells Dev.* 24, 917–926. doi: 10.1089/scd.2014.0544
- Nagarajan, R., Ahmad, S., and Michael Gromiha, M. (2013). Novel approach for selecting the best predictor for identifying the binding sites in DNA binding proteins. *Nucleic Acids Res.* 41, 7606–7614. doi: 10.1093/nar/gkt544
- Pandey, A. K., and Loscalzo, J. (2023). Network medicine: an approach to complex kidney disease phenotypes. *Nat. Rev. Nephrol.* 19, 463–475. doi: 10.1038/s41581-023-00705-0
- Patiyal, S., Dhall, A., and Raghava, G. P. (2022). DBPpred: a deep learning method for the prediction of DNA interacting residues in protein sequences. *bioRxiv*. 23:bbac322. doi: 10.1093/bib/bbac322
- Patro, R., Sefer, E., Malin, J., Marçais, G., Navlakha, S., and Kingsford, C. (2012). Parsimonious reconstruction of network evolution. *Algorithms Mol. Biol.* 7:25. doi: 10.1186/1748-7188-7-25
- Peled, S., Leiderman, O., Charar, R., Efroni, G., Shav-Tal, Y., and Ofra, Y. (2016). De-novo protein function prediction using DNA binding and RNA binding proteins as a test case. *Nat. Commun.* 7:13424. doi: 10.1038/ncomms13424
- Qiu, J., Bernhofer, M., Heinzinger, M., Kemper, S., Norambuena, T., Melo, F., et al. (2020). ProNA2020 predicts protein–DNA, protein–RNA, and protein–protein binding proteins and residues from sequence. *J. Mol. Biol.* 432, 2428–2443. doi: 10.1016/j.jmb.2020.02.026
- Qiu, Y., and Wei, G.-W. (2023). Artificial intelligence-aided protein engineering: from topological data analysis to deep protein language models. *Brief. Bioinform.* 24:bbad289. doi: 10.1093/bib/bbad289
- Qu, K., Wei, L., and Zou, Q. (2019). A review of DNA-binding proteins prediction methods. *Curr. Bioinforma.* 14, 246–254. doi: 10.2174/1574893614666181212102030
- Rao, R. M., Liu, J., Verkuil, R., Meier, J., Canny, J., Abbeel, P., et al. (2021). MSA transformer. *Paper presented at the International Conference on Machine Learning*.
- Rastogi, C., Rube, H. T., Kribelbauer, J. F., Crocker, J., Loker, R. E., Martini, G. D., et al. (2018). Accurate and sensitive quantification of protein–DNA binding affinity. *Proc. Natl. Acad. Sci. USA* 115, E3692–E3701. doi: 10.1073/pnas.1714376115
- Schleif, R. (1988). DNA binding by proteins. *Science* 241, 1182–1187. doi: 10.1126/science.2842864
- Sefer, E. (2025). DRGAT: predicting drug responses via diffusion-based graph attention network. *J. Comput. Biol.* 32, 330–350. doi: 10.1089/cmb.2024.0807
- Shanahan, H. P., Garcia, M. A., Jones, S., and Thornton, J. M. (2004). Identifying DNA-binding proteins using structural motifs and the electrostatic potential. *Nucleic Acids Res.* 32, 4732–4741. doi: 10.1093/nar/gkh803
- Shiroma, Y., Takahashi, R. u., Yamamoto, Y., and Tahara, H. (2020). Targeting DNA binding proteins for cancer therapy. *Cancer Sci.* 111, 1058–1064. doi: 10.1111/cas.14355
- Si, J., Zhao, R., and Wu, R. (2015). An overview of the prediction of protein DNA-binding sites. *Int. J. Mol. Sci.* 16, 5194–5215. doi: 10.3390/ijms16035194
- Song, L., Koga, Y., and Ecker, J. R. (2016). Profiling of transcription factor binding events by chromatin immunoprecipitation sequencing (ChIP-seq). *Curr. Prot. Plant Biol.* 1, 293–306. doi: 10.1002/cppb.20014
- Steinberger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* 35, 1026–1028. doi: 10.1038/nbt.3988
- Su, H., Liu, M., Sun, S., Peng, Z., and Yang, J. (2019). Improving the prediction of protein–nucleic acids binding residues via multiple sequence profiles and the consensus of complementary methods. *Bioinformatics* 35, 930–936. doi: 10.1093/bioinformatics/bty756
- Takeda, Y., Ohlendorf, D., Anderson, W., and Matthews, B. (1983). DNA-binding proteins. *Science* 221, 1020–1026. doi: 10.1126/science.6308768

- Tjong, H., and Zhou, H.-X. (2007). DISPLAR: an accurate method for predicting DNA-binding sites on protein surfaces. *Nucleic Acids Res.* 35, 1465–1477. doi: 10.1093/nar/gkm008
- UniProt Consortium (2015). UniProt: a hub for protein information. *Nucleic Acids Res.* 43, D204–D212. doi: 10.1093/nar/gku989
- UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47, D506–D515. doi: 10.1093/nar/gky1049
- Uryu, K., Nakashima-Yasuda, H., Forman, M. S., Kwong, L. K., Clark, C. M., Grossman, M., et al. (2008). Concomitant TAR-DNA-binding protein 43 pathology is present in Alzheimer disease and corticobasal degeneration but not in other tauopathies. *J. Neuropathol. Exp. Neurol.* 67, 555–564. doi: 10.1097/NEN.0b013e31817713b5
- Wang, L., Huang, C., Yang, M. Q., and Yang, J. Y. (2010). BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. *BMC Syst. Biol.* 4, 1–9. doi: 10.1186/1752-0509-4-S1-S3
- Wang, S., Li, W., Fei, Y., Cao, Z., Xu, D., and Guo, H. (2019). An improved process for generating uniform PSSMs and its application in protein subcellular localization via various global dimension reduction techniques. *IEEE Access* 7, 42384–42395. doi: 10.1109/access.2019.2907642
- Xia, Y., Xia, C.-Q., Pan, X., and 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, e51–e51. doi: 10.1093/nar/gkab044
- Xiong, Y., Liu, J., and Wei, D. Q. (2011). An accurate feature-based method for identifying DNA-binding residues on protein surfaces. *Proteins* 79, 509–517. doi: 10.1002/prot.22898
- Yan, J., and Kurgan, L. (2017). DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues. *Nucleic Acids Res.* 45:e84–e84. doi: 10.1093/nar/gkx059
- Zaru, R., and Orchard, S. UniProt Consortium (2023). UniProt tools: BLAST, align, peptide search, and ID mapping. *Curr. Protoc.* 3:e697. doi: 10.1002/cpz1.697
- Zhang, J., Chen, Q., and Liu, B. (2021). NCBRPred: predicting nucleic acid binding residues in proteins based on multilabel learning. *Brief. Bioinform.* 22:bbaa397. doi: 10.1093/bib/bbaa397
- Zhang, L., and Liu, T. (2024). PDNAPred: interpretable prediction of protein-DNA binding sites based on pre-trained protein language models. *Int. J. Biol. Macromol.* 281:136147. doi: 10.1016/j.ijbiomac.2024.136147
- Zhou, J., Xu, R., He, Y., Lu, Q., Wang, H., and Kong, B. (2016). PDNAsite: identification of DNA-binding site from protein sequence by incorporating spatial and sequence context. *Sci. Rep.* 6:27653. doi: 10.1038/srep27653
- Zhu, Y.-H., Hu, J., Song, X.-N., and Yu, D.-J. (2019). DNAPred: accurate identification of DNA-binding sites from protein sequence by ensemble hyperplane-distance-based support vector machines. *J. Chem. Inf. Model.* 59, 3057–3071. doi: 10.1021/acs.jcim.8b00749
- Zhu, Y.-H., Liu, Z., Liu, Y., Ji, Z., and Yu, D.-J. (2024). ULDNA: integrating unsupervised multi-source language models with LSTM-attention network for high-accuracy protein-DNA binding site prediction. *Brief. Bioinform.* 25:bbae040. doi: 10.1093/bib/bbae040