Advancing Protein-DNA Binding Site Prediction: Integrating Sequence Models and Machine Learning Classifiers

Taslim Murad*, **Sarwan Ali***, **Prakash Chourasia**, **Murray Patterson**

Department of Computer Science
Georgia State University, Atlanta, GA, USA
{tmurad2, sali85, pchourasia1}@student.gsu.edu, mpatterson30@gsu.edu
* Equal Contribution

Abstract

Predicting protein-DNA binding sites is a challenging computational problem that has led to the development of advanced algorithms and techniques in the field of bioinformatics. Identifying the specific residues where proteins bind to DNA is of paramount importance, as it enables the modeling of their interactions and facilitates downstream studies. Nevertheless, the development of accurate and efficient computational methods for this task remains a persistent challenge. Accurate prediction of protein-DNA binding sites has far-reaching implications for understanding molecular mechanisms, disease processes, drug discovery, and synthetic biology applications. It helps bridge the gap between genomics and functional biology, enabling researchers to uncover the intricacies of cellular processes and advance our knowledge of the biological world. The method used to predict DNA binding residues in this study is a potent combination of conventional bioinformatics tools, protein language models, and cutting-edge machine learning and deep learning classifiers. On a dataset of protein-DNA binding sites, our model is meticulously trained, and it is then rigorously examined using several experiments. As indicated by higher predictive behavior with AUC values on two benchmark datasets, the results show superior performance when compared to existing models. The suggested model has a strong capacity for generalization and shows specificity for DNA-binding sites. We further demonstrated the adaptability of our model as a universal framework for binding site prediction by training it on a variety of protein-ligand binding site datasets. In conclusion, our innovative approach for predicting protein-DNA binding residues holds great promise in advancing our understanding of molecular interactions, thus paving the way for several groundbreaking applications in the field of molecular biology and genetics. Our approach demonstrated efficacy and versatility underscore its potential for driving transformative discoveries in biomolecular research.

1 Introduction

Protein-DNA binding site prediction is an essential area of research with significant implications in various fields, including molecular biology, genetics, drug discovery, and synthetic biology. Accurate prediction of protein-DNA binding sites has far-reaching potential, leading to groundbreaking biotechnology and drug design applications (Śledź and Caflisch 2018) along with several other applications including understanding gene regulation (Ptashne 1986), functional annotation of genomes (Pique-Regi et al. 2011), cancer research for designing targeted therapies (Xu et al. 2016; O'Connor 2015), evolutionary studies (Lichtarge, Bourne, and Cohen 1996), genetic engineering, and accelerating drug discovery (Zhao, Cao, and Zhang 2020). Protein-DNA interactions are fundamental to many biological processes, such as DNA replication (Echols 1986), transcription (Dey et al. 2012), repair, and recombination (Polo and Jackson 2011; West 2003). Accurate prediction of binding sites aids in deciphering the molecular mechanisms behind these processes, shedding light on the intricate workings of the cell. It plays a crucial role in regulating gene expression by binding to specific sites on DNA. Predicting these binding sites helps researchers understand how genes are turned on or off, which is essential for understanding normal development, disease processes, and various cellular responses. It also can help identify potential drug targets for diseases like cancer and genetic disorders. Some research applies protein-DNA binding site prediction to specific diseases, such as cancer (Zhu, Wang, and Qian 2016). By identifying altered binding sites in disease-related genes, researchers aim to uncover novel therapeutic targets. Designing drugs that interfere with these interactions could offer new therapeutic strategies. With the advent of high-throughput sequencing technologies, vast amounts of genomic data are generated. Predicting protein-DNA binding sites aids in interpreting this data by identifying regions that are likely to be functionally important. Protein-DNA binding sites in genomic analyses provide a foundational understanding of the functional elements within a genome and their roles in various biological processes. It bridges the gap between genomic sequences and biological functions, enabling researchers to unravel the complexities of gene regulation and molecular interactions.

Many recent studies leverage machine learning and deep learning techniques to predict protein-DNA binding sites (Zhang, Zhu, and shuang Huang 2019). These methods often involve training models on large datasets of known binding sites and using them to predict binding locations in genomic sequences. Researchers focus on identifying relevant features or descriptors that can capture the characteristics of protein-DNA interactions (Jones et al. 2001). These features might include sequence motifs, physicochemical properties, and structural information of DNA and protein molecules (Ivanciuc et al. 2004; Zhang and Liu 2019). Some studies integrate various types of omics data, such as genomics, transcriptomics, and proteomics, to improve the accuracy of binding site predictions. This integration allows for a more comprehensive understanding of the regulatory landscape. Evolutionarily conserved regions are often indicative of functional importance (Tatarinova et al. 2016). Research explores the use of conservation scores and comparative genomics approaches to enhance the accuracy of binding site predictions. Incorporating 3D structural information of protein-DNA complexes helps refine binding site predictions by considering the physical interactions between the molecules. This includes techniques such as molecular docking and structural modeling. Deep learning models, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), are applied to sequence data to capture complex patterns in DNA and protein sequences, improving binding site prediction accuracy. Transfer learning techniques involve pretraining models on related tasks and fine-tuning them for binding site prediction (Han, Pang, and Wu 2021). This approach benefits from the knowledge acquired in the pretraining phase. Evaluating the performance of prediction methods is crucial. Researchers design benchmark datasets, establish evaluation metrics (such as sensitivity, specificity, and AUC), and compare different algorithms to assess their effectiveness.

The information needed to understand protein structures is all contained in the protein sequence. However, extracting the structure from the sequence only is a difficult and time-consuming task. Consequently, structure-only-based models consistently outperform sequence-based models in terms of performance due to the availing of complete structures, which is not easy to infer from sequences only. Structure-based models, however, need precise protein structures as input to assure model performance. As a result, predicting DNA-binding sites from protein sequences is still a significant and urgent research issue. Because feature extraction frequently relies on manual design and does not produce a refined initial representation, the performance of current sequence-based models is still insufficient for practical application. Therefore, it is imperative to create an end-to-end model without the use of handcrafted features. The representation learning methods pre-training and contrastive learning are both frequently employed. The model is trained unsupervised during pre-training using information from a huge amount of unlabeled data, and the model parameters are then transferred to later tasks for feature extraction or fine-tuning. In contrastive learning, samples from the same class are found to be close to one another, while examples from other classes are found to be distant from one another.

In this work, we identify several limitations to the work done in (Liu and Tian 2023) and propose alternate solutions to overcome the limitations. More specifically, our contributions are the following:

1. It is a well-established fact in the literature that the neural network-based methods do not work efficiently as compared to simple Machine Learning (ML) classifiers (e.g. tree-based methods) in the case of tabular data (Grinsztajn, Oyallon, and Varoquaux 2022; Joseph and Raj 2022; Malinin, Prokhorenkova, and Ustimenko 2021). Therefore, instead of using 1DCNN for the underlying classification (and including all discussion around that) as done in (Liu and Tian 2023), we use simple ML classifiers for the underlying supervised analysis.
2. Authors in (Liu and Tian 2023) use ProtBert (Elnaggar et al. 2021) as the pre-trained model to generate the embeddings for each amino acid within protein sequences. We replace that with a more efficient SeqVec (Heinzinger et al. 2019) pre-trained model. The choice of replacing ProtBert with SeqVec is due to its demonstrated effectiveness in learning relevant features for our task (i.e. Binding site prediction).
3. Furthermore, we propose a lightweight model using the idea of Sparse Coding, which combines the power of k-mers and one-hot encoding to design efficient initial embeddings for the amino acids. The only parameter in this sparse coding-based embedding method is k (contextual window size for amino acids), which is significantly lesser compared to complex models like ProtBert and SeqVec. This *Almost Parameter Free* approach makes Sparse coding an ideal choice for *fast* binding site prediction.

In the remaining paper, we discuss the literature review in Section 2 while the paper's main contributions are highlighted in Section 3. The experiments and dataset detail are given in Section 4 followed by the discussion of the results for proposed and baseline models in Section 5. In the end, we discuss the conclusion of the paper in Section 6.

2 Related Work

The prediction of protein-DNA binding sites is a critical task in computational biology, with applications ranging from understanding gene regulation to designing novel therapeutic agents (Collie and Parkinson 2011). Over the years, various computational methods have been developed to tackle this complex problem, each leveraging different techniques and approaches. In this section, we review the key literature in the field of protein-DNA binding site prediction, focusing on different methodologies, challenges, and advancements.

Evolutionary information has been a cornerstone of protein-DNA binding site prediction (Kuznetsov et al. 2006; Si, Zhao, and Wu 2015). Methods utilizing multiple sequence alignments (MSAs) (Ahmad and Sarai 2005; Yan et al. 2006) and phylogenetic profiles (La and Kihara 2012) have shown promising results. Techniques like DRNAPred (Yan and Kurgan 2017) and DNAPred (Zhu et al. 2019) incorporate evolutionary conservation patterns to identify potential binding sites. SVMnuc (Su et al. 2019) and NCBRPred (Zhang, Chen, and Liu 2021) also utilize evolutionary information for distinguishing binding sites.

Traditional machine-learning techniques have been extensively used in the context of binding site prediction. Methods like SVMnuc (Su et al. 2019) and DBPred (Patiyal, Dhall, and Raghava 2022) incorporate support vector machines (SVMs) to classify binding sites based on a set of engineered features derived from sequence and structure data. These methods have demonstrated reasonable predictive performance and often rely on well-curated training datasets.

Recent advancements in deep learning have led to the development of more complex models for protein-DNA binding site prediction (Zhang et al. 2022; Si, Zhao, and Wu 2015). ProtBert (Elnaggar et al. 2021), a pre-trained transformer model adapted from natural language processing, has shown its potential to capture intricate sequence patterns. The combination of ProtBert with 1D convolutional neural networks (1DCNN) has been explored to enhance performance in identifying binding sites. Transfer learning from related domains, such as language models, has become a prominent technique (Novakovsky et al. 2021; Aizenshtein-Gazit and Orenstein 2022). Pre-trained models like ProtBert and SeqVec (Heinzinger et al. 2019), inspired by NLP models, have shown success in capturing high-level features in protein sequences. These models provide a foundation for building more specialized predictors with fewer labeled samples. SeqVec introduces embeddings that capture the biophysical properties of protein sequences by training on vast unlabeled protein data. These embeddings, derived from a language model, have demonstrated their potential in improving predictions. Sparse coding techniques, which do not require labeled data, have also been explored to generate embeddings that preserve important context (Wu et al. 2021).

3 Proposed Approach

We propose a protein-ligand binding sites prediction framework to perform the binding site prediction of a given protein sequence. The overall architecture of the proposed model comprised two main modules: the sequence embedding module and the classification module.

3.1 Sequence Embedding Module

The sequence embedding module leverages two distinct techniques, namely SeqVec and Sparse Coding, to create fixed-length embeddings for individual amino acids within a protein sequence.

SeqVec (Heinzinger et al. 2019) It is a pre-trained protein language model that captures intricate sequence patterns and semantic information inherent to protein sequences. The SeqVec language model is based on Embeddings from Language Models (ELMo) (Sarzynska-Wawer et al. 2021), commonly applied in natural language processing to create continuous vector representations (embeddings) for protein sequences. These embeddings, named SeqVec (Sequence-to-Vector), capture biophysical properties from unlabeled data (UniRef50) and enable simple neural networks to excel in various tasks.

The architecture of SeqVec contains the following steps:

1. **ELMo Pre-training:** ELMo, originally designed for natural language processing, is a bi-directional language model that learns to predict the likelihood of the next word in a sentence given the surrounding words. It does so by training on massive amounts of unlabeled text data, such as Wikipedia articles. ELMo develops contextualized embeddings that capture the syntax and semantics of the language. In the context of SeqVec, ELMo is trained on large protein sequence databases, specifically UniRef50, to predict the next amino acid in a sequence based on its neighboring amino acids.
2. **Embedding Extraction:** After pre-training, ELMo produces embeddings for amino acids in a protein sequence. These embeddings capture the contextual information about each amino acid based on its surrounding amino acids in the sequence.
3. **Sequence Embeddings:** The output of ELMo for each amino acid is a continuous vector representation that captures the biophysical properties of the protein sequence. These embeddings are referred to as SeqVec embeddings and serve as the representation of the protein sequence.
4. **Embeddings for Prediction Tasks:** The SeqVec embeddings can be used as features for various protein prediction tasks, such as secondary structure prediction, intrinsic disorder prediction, subcellular localization prediction, and more. These embeddings are fed into neural networks or other machine learning models to perform these tasks.

The key innovation of SeqVec lies in its use of ELMo to capture the biophysical properties of protein sequences. ELMo's ability to learn contextualized embeddings from unlabeled protein sequences enables SeqVec to generate embeddings that encode relevant information about protein structure and function. This approach offers an alternative to the traditional use of evolutionary information and provides a scalable solution for analyzing protein sequences, particularly in scenarios involving large-scale proteomics data. Note that we justify the preference for SeqVec over other protein-based pre-trained language models, such as ProtBert (Elnaggar et al. 2021) (as used in CLAPE (Liu and Tian 2023)) due to its demonstrated effectiveness in learning relevant features for our task.

Sparse Coding Since fine tunning a language model could still be expensive, and it may not generalize better in all scenarios, we proposed a sparse coding-based alternative, which involves the power of k-mers (for neighborhood context capturing) and one-hot encoding (for generic embedding generation) to transform amino acids into numerical representations. The utilization of Sparse Coding is justifiable by its ability to capture local compositional information within amino acids of the protein sequences, enhancing our model's capability to learn meaningful patterns associated with binding sites. For this purpose, we take a k-mer (where $k = 9$, which is decided using the standard validation set approach) as a sliding window for each amino acid. Then we design a one-hot encoding-based representation for the k-mer, which acts as the local embedding for the given amino acid. In this way, we design embedding for each amino acid, which is then used as input for supervised analysis using machine learning and deep learning models.

One exception occurs in our sparse coding-based embedding when the sliding window (k-mer) reaches the end of the sequence. In that case, for the remaining $n - k$ amino acids (where n is the protein sequence length), we take the k-mers-based sliding window in reverse order and repeat the one-hot encoding step, hence preserving the neighborhood context.

3.2 Classification Module

After designing the embeddings for each amino acid within the protein sequence for the purpose of binding site prediction, the next step is to select efficient classification models to perform the actual site prediction. For this purpose, the features generated by the sequence embedding module (i.e. SeqVec and Sparse Coding) are fed into the classification module, which is composed of multiple machine-learning classifiers. For the same binding site prediction problem, authors in (Liu and Tian 2023) propose the use of a four "one-dimensional convolutional neural network" (1DCNN) model as the backbone network. The raw dimension of the input is 1024, and the output dimensions of the four layers are 1024, 128, 64, and 2, respectively. Each layer has a stride of 1 and is followed by a batch normalization layer (except for the last one). The layers are also accompanied by varying sizes of kernel filters and paddings. The kernel sizes are 7, 5, 3, 5, and the paddings are 3, 2, 1, 2, respectively. The 1DCNN is designed to capture neighboring information within protein sequences and employs operations like max pooling for down-sampling. Padding is applied for different convolutional kernel sizes to maintain the same sequence length for input and output features, ensuring a unified token-level classification outcome. The activation function ReLU (rectified linear unit) introduces nonlinearity to the model, and techniques such as dropout and batch normalization are utilized to enhance model robustness and generalization. The classification head is an integral part of the model, employing a Softmax function. This function scales the output values between 0 and 1, representing mutually exclusive prediction scores. These scores reflect the probability of a given residue being a DNA-binding site. The classification head is then used to predict DNA-binding sites within protein sequences.

While deep learning model, such as 1DCNN, exhibits remarkable capacity in various tasks, the dataset size, task complexity, and interpretability considerations have guided our choice towards machine-learning classifiers (i.e. Naive Bayes, Multi-Layer Perceptron, K-Nearest Neighbors, Random Forest, Logistic Regression, and Decision Tree). These classifiers collectively analyze the encoded features and make predictions about the presence of ligand binding sites in the protein sequence. Moreover, it is well known in the literature that the neural network-based methods do not perform optimally as compared to simple Machine Learning (ML) classifiers (e.g. tree-based methods) in the case of tabular data (Grinsztajn, Oyallon, and Varoquaux 2022; Joseph and Raj 2022; Malinin, Prokhorenkova, and Ustimenko 2021). Therefore, we decided to use simple ML models for the downstream supervise analysis (i.e. binding site prediction).

By integrating these modules, our proposed framework strives to provide accurate predictions of protein-ligand binding sites, leveraging the strengths of SeqVec and Sparse Coding for feature representation and harnessing machine-learning classifiers for classification tasks. This design rationale ensures a well-rounded approach to predicting binding probabilities while considering the intricacies of the protein-ligand interaction problem.

To demonstrate the power of simple ML models over the deep learning models, we fine-tuned the existing ProtBert (Elnaggar et al. 2021) model (as used in CLAPE (Liu and Tian 2023)) to generate embeddings for the amino acids and performed binding site predictions as well. The fine-tuning hyper-parameters are ADAM optimizer, 25 batch size, and 10 training epochs. A loss function is formed by combining the focal loss (Lin et al. 2017) and triplet center loss (TCL) (He et al. 2018) to handle the data imbalance issue effectively, and it's defined as,

$$Loss = L_{focal} + \lambda L_{tcl} \tag{1}$$

where λ is a hyperparameter with 0.1 value.

4 Experimental Setup

This section discusses the details of the datasets used for conducting the experiments along with the employed evaluation metrics and baseline methods. All experiments are conducted using a server having Intel(R) Xeon(R) CPU E7-4850 v4 @ $2.40GHz$ with Ubuntu 64 bit OS (16.04.7 LTS Xenial Xerus) having 3023 GB memory.

4.1 Dataset Statistics

We perform the binding site classification task using DNA-based datasets i.e. Dataset 1 and Dataset 2 (Liu and Tian 2023). These datasets were preprocessed to improve the model's robustness and avoid data imbalance bias. In both datasets, the binding sites were defined as residues with a distance of < 0.5 (threshold value) $+R$, where R represents the sum of the Van der Waals radius of the two nearest atoms between the residue and the nucleic acid molecule. The details of each of the datasets are as follows,

Dataset 1 It comprises 646 protein sequences as the training set and 46 as the test set. This data was introduced by (Patiyal, Dhall, and Raghava 2022) after extracting it from (Qiu et al. 2020) and (Zhang, Ma, and Kurgan 2019). The statistical detail of this data is given in Table 1.

Dataset 2 It has 573 protein sequences as a train set and 129 as a test. It was introduced in (Xia et al. 2021) after collecting from the BioLiP database (Yang, Roy, and Zhang 2012). This dataset consisted of protein-DNA complex structural data, and its statistical detail is given in Table 1.

4.2 Evaluation Metrics

To evaluate the performance of the binding site prediction task, we used various evaluation metrics. The metrics are specificity, precision, recall, F1-score, ROC AUC, and Matthews correlation coefficient (MCC). As this is a binary classification problem, the formulas used to compute some of the popular metrics are as follows,

		Binding	Non-Binding
Dataset 1	Train	15636	298503
	Test	965	9911
Dataset 2	Train	14479	145404
	Test	2240	35275

Table 1: The number of binding and non-binding sites present in the test and train sets of each DNA-based dataset respectively (Liu and Tian 2023).

$$specificity = \frac{TN}{TN + FP} \tag{2}$$

$$precision = \frac{TP}{TP + FP} \tag{3}$$

$$recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1 - score = 2 * \frac{precision * recall}{precision + recall} \tag{5}$$

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

where TN, TP, FN, and FP represent true negative, true positive, false negative, and false positive, respectively. Moreover, the ROC AUC from SKLearn is used by us to get the values of ROC AUC. We have reported the average score for each of the metrics after 5 runs. In terms of interpretability, the description of different evaluation metrics along with their corresponding interpretability detail, is given in Table 2.

Metric	Interpretability
TP	Indicates the number of residues that are correctly classified as DNA-binding sites
FP	Indicates the number of residues that are incorrectly classified as DNA-binding sites
TN	Indicates the number of residues that correctly classified as non-binding sites
FN	Indicates the number of residues that were incorrectly classified as non-binding sites
Specificity	Indicates the portion of correctly predicted non-binding sites
Precision	Indicates the accuracy of residues predicted as DNA-binding sites
Recall	Indicates the portion of DNA-binding residues that were successfully discovered by the model
F1-score	Indicates the harmonic mean of precision and recall
MCC	Indicates the prediction ability of both positive and negative classes
ROC-AUC	Indicates the overall performance of the model

Table 2: Interpretability of different evaluation metrics in terms of binding site prediction

4.3 Baseline Models

We have compared the performance of our proposed methods with various baselines, and the baseline models are as follows,

DRNAPred (Yan and Kurgan 2017) A sequence-based method for predicting and differentiating between DNA- and RNA-binding residues, called DRNApred, is proposed in (Yan and Kurgan 2017). Protein-DNA and protein-RNA interactions are fundamental in cellular functions, yet many remain uncharacterized. Existing methods often misclassify binding residues, and their runtime limitations hinder large-scale applications. DRNApred addresses these challenges by using a new dataset of both DNA- and RNA-binding proteins, employing regression to penalize cross-predictions, and employing a unique two-layered architecture. The method outperforms state-of-the-art predictors by significantly reducing cross-predictions, providing high-quality false positives near-native binding residues, and improving accuracy in predicting binding proteins. Application to the human proteome validates its ability to reduce cross-predictions and identifies novel DNA/RNA-binding proteins with similar characteristics to known ones. This method showcases efficiency and accuracy in sequence-based binding prediction for nucleic acids.

DNAPred (Zhu et al. 2019) The method presented in the paper introduces a novel two-stage imbalanced learning algorithm called Ensembled Hyperplane-Distance-Based Support Vector Machines for the prediction of protein-DNA binding sites. The data imbalance problem, where the number of negative-class samples (nonbinding residues) significantly outweighs the positive-class samples (binding residues), often limits the performance of machine learning predictors. This paper addresses this issue by first using a hyperplane-distance-based under-sampling (HD-US) algorithm to generate multiple training subsets and training individual Support Vector Machines (SVMs) on them. In the second stage, an enhanced AdaBoost (EAdaBoost) algorithm is employed to ensemble the trained SVMs. The approach outperforms several other imbalanced learning algorithms and achieves a significant improvement in identifying protein-DNA binding sites.

SVMnuc (Su et al. 2019) The paper presents a new method called SVMnuc, which is an ab-initio method devised to predict nucleic acids-binding residues, addressing the challenge of accurately identifying these biologically significant sites. Leveraging the fact that binding residues are evolutionarily conserved, SVMnuc employs three distinct sequence profiles to extract a comprehensive set of features that capture residue characteristics. These profiles encompass information from PSI-BLAST, generating a position-specific scoring matrix (PSSM) through sequence-profile alignment; PSIPRED, offering probabilities for secondary structure states; and HHblits, producing a hidden Markov model (HMM) profile based on database searches.

Each of these profiles is processed to ensure its applicability, such as transforming PSSM values to a logistic scale between 0 and 1. The essence of binding residues' mutual influence within a binding pocket is embraced through the application of a sliding window approach, accounting for neighboring residues.

NCBRPred (Zhang, Chen, and Liu 2021) Authors in (Zhang, Chen, and Liu 2021) propose a method, called NCBRPred, which is designed to predict nucleic acid binding residues within proteins. NCBRPred adopts a multilabel sequence labeling model (MSLM). By employing bidirectional Gated Recurrent Units (BiGRUs), NCBRPred effectively captures intricate interactions among residues. This approach treats the prediction of both DNA-binding and RNA-binding residues as a unified multilabel learning task, integrating data from both DNA- and RNA-binding proteins during training to alleviate cross-prediction problems.

DBPred (Patiyal, Dhall, and Raghava 2022) Authors in (Patiyal, Dhall, and Raghava 2022) proposed a server-based tool, called DBPred, for predicting DNA-binding residues in proteins. A range of traditional machine learning and deep learning (1D-CNN) techniques-based models are developed within the tool, incorporating binary, physicochemical properties, and Position-Specific Scoring Matrix (PSSM)/evolutionary profiles. The study's rigorous methodology, encompassing thorough training and unbiased evaluation, contributes a powerful tool for uncovering the intricacies of DNA-protein interactions, amplifying the potential to unravel genetic regulatory mechanisms.

CLAPE (Liu and Tian 2023) The CLAPE (Contrastive Learning And Pre-trained Encoder) approach, introduced in (Liu and Tian 2023), offers a novel solution for predicting DNA binding residues in proteins. This approach effectively combines the power of a pre-trained protein language model (i.e. ProBert) with contrastive learning techniques (i.e. 1DCNN model). CLAPE leverages a dataset of protein-DNA binding sites for training and computes classification results using the CNN model.

5 Results And Discussion

The classification results for the proposed method and its comparison with the baselines are shown in Table 3 for Dataset 1. Compared to the baselines such as DRNAPred, DNAPred, SVMnuc, NCBRPred, DBPred, and ProtBert + 1DCNN, we can observe that ProtBert + ML classifiers (our pre-trained model) i.e. Naive Bayes, Multi-layer Perceptron, K-Nearest Neighbors, Random Forest, Logistic Regression, and Decision Tree, show near-perfect specificity and precision scores. This eventually means that the number of correctly predicted non-binding sites is higher. Moreover, the accuracy of the residues predicted as DNA-binding sites is also higher. However, for Recall and ROC-AUC, the baseline NCBRPred shows higher performance, while ProtBert + 1DCNN shows superior performance in the case of F1 and MCC scores. More complex models like ProtBert + 1DCNN may have a higher capacity to capture intricate patterns in the data, which could lead to better F1 and MCC scores.

For our SeqVec + ML classifiers and Sparse coding + ML/DL classifiers, we can again observe a near-perfect specificity score. One interesting insight to note here is that since the Sparse coding-based embedding method is completely unsupervised and does not involve any expensive model training, it is still able to achieve a higher specificity score. This is due to the fact that it preserves the neighborhood context efficiently within the generated embeddings. The reason for the simpler models to excel in specific metrics, such as ProtBert + ML classifiers achieving high specificity and precision due to their focused decision boundaries.

The classification results for the proposed method and its comparison with the baselines are shown in Table 4 for Dataset 2. We can observe that for the specificity, precision, and MCC, both pre-trained models (i.e. ProBert and SeqVec) and the Sparse coding-based method show higher scores using simple ML classifiers rather than using comparatively more complex 1DCNN model. For F1 and ROC-AUC, we can observe that DNAPred performs the best.

5.1 Statistical Significance

Since we report results for 5 experimental runs (as discussed in Section 4.2), we analyzed the standard deviation values for the 5 runs and computed P-values using the student t-test. We observed that the P-values were < 0.05 due to lower standard deviation values. This validation confirmed the statistical significance of the results.

6 Conclusion

In this study, we have addressed the challenging problem of predicting protein-DNA binding sites using an innovative and comprehensive approach. Protein-DNA interactions play a crucial role in various biological processes, and accurate prediction of binding sites has broad implications in molecular biology, genetics, drug discovery, and beyond. Our work capitalizes on the synergy between conventional bioinformatics techniques, state-of-the-art protein language models, and advanced machine learning classifiers. Through meticulous experimentation and rigorous evaluation, we have demonstrated the superiority of our proposed approach over existing models. Our model, trained on a dataset of protein-DNA binding sites, exhibits robust predictive behavior as evidenced by higher predictive values on benchmark datasets. The flexibility and generalization capacity of our models is highlighted by their adaptability as a universal framework for binding site prediction across diverse protein-ligand binding scenarios. Our approach introduces several contributions to enhance the accuracy and efficiency of protein-DNA binding site prediction. By leveraging SeqVec, a powerful pre-trained model, we capture intricate sequence features effectively. Additionally, we propose a lightweight model based on Sparse Coding, which combines k-mers and one-hot encoding to generate efficient initial embeddings. This approach's parameter efficiency positions it as a promising candidate for rapid binding site prediction. As we continue to refine and expand our approach, we envision its potential to drive breakthroughs across various domains of biology and genetics. Exploring the integration of epigenetic information and investigating ensemble

Method	Model	Spec.	Prec.	Recall	F1 (Bina.)	ROC-AUC	MCC
DRNAPred	-	0.692	0.185	0.677	0.291	0.755	0.226
DNAPred	-	0.655	0.157	0.671	0.254	0.730	0.194
SVMnuc	-	0.666	0.154	0.668	0.250	0.715	0.192
NCBRPred	-	0.674	0.165	0.677	0.265	0.713	0.207
DBPred	-	0.784	0.243	**0.708**	0.362	**0.794**	0.320
ProBert	1DCNN	0.834	0.307	0.658	**0.380**	0.746	**0.339**
	NB	0.775	0.167	0.464	0.246	0.619	0.093
	MLP	0.971	0.466	0.254	0.329	0.613	0.294
	KNN	0.981	0.500	0.191	0.277	0.586	0.271
	RF	**0.999**	**0.999**	0.002	0.004	0.501	0.043
	LR	0.994	0.672	0.117	0.199	0.555	0.265
	DT	0.942	0.211	0.159	0.182	0.550	0.096
SeqVec	1DCNN	0.972	0.418	0.280	0.293	0.626	0.276
	NB	0.807	0.178	0.431	0.252	0.619	0.103
	MLP	0.963	0.379	0.231	0.287	0.597	0.228
	KNN	0.991	0.692	0.191	0.300	0.591	0.356
	RF	**0.999**	0.851	0.023	0.046	0.511	0.135
	LR	0.997	0.715	0.075	0.136	0.536	0.219
	DT	0.933	0.197	0.167	0.181	0.551	0.088
Sparse Coding (kmers+OHE)	1DCNN	**0.999**	0.000	0.000	0.000	0.500	0.000
	NB	0.938	0.096	0.067	0.079	0.503	0.007
	MLP	**0.999**	0.000	0.000	0.000	0.500	0.000
	KNN	0.997	0.115	0.003	0.006	0.500	0.009
	RF	0.997	0.289	0.011	0.021	0.504	0.041
	LR	**0.999**	0.000	0.000	0.000	0.500	0.000
	DT	0.944	0.102	0.065	0.079	0.507	0.011

Table 3: Binding site prediction (classification) results for different evaluation metrics using the proposed and baseline methods for **Dataset 1**. The best values are shown in bold. Dashes "-" in the model column mean they were end-to-end models and used as described in respective original studies.

Method	Model	Spec.	Prec.	Recall	F1 (Bina.)	ROC-AUC	MCC
DRNAPred	-	0.937	0.190	0.233	0.210	0.693	0.155
DNAPred	-	0.954	0.353	0.396	**0.373**	**0.845**	0.332
SVMnuc	-	0.966	0.371	0.316	0.341	0.812	0.304
NCBRPred	-	0.969	0.312	0.392	0.347	0.823	0.313
ProBert	1DCNN	0.830	0.242	**0.619**	0.317	0.725	0.221
	NB	0.761	0.141	0.618	0.230	0.690	0.100
	MLP	0.954	0.305	0.318	0.311	0.636	0.225
	KNN	0.974	0.310	0.179	0.227	0.577	0.181
	RF	**0.999**	0.545	0.002	0.005	0.501	0.035
	LR	0.979	0.489	0.305	0.376	0.642	**0.354**
	DT	0.910	0.123	0.198	0.157	0.554	0.059
SeqVec	1DCNN	0.960	0.328	0.287	0.283	0.623	0.292
	NB	0.753	0.089	0.382	0.145	0.568	0.035
	MLP	0.954	0.262	0.253	0.257	0.604	0.175
	KNN	0.986	0.503	0.215	0.301	0.601	0.303
	RF	**0.999**	**0.782**	0.051	0.096	0.525	0.194
	LR	0.991	0.512	0.146	0.227	0.568	0.252
	DT	0.882	0.107	0.222	0.144	0.552	0.046
Sparse Coding (kmers+OHE)	1DCNN	**0.999**	0.000	0.000	0.000	0.500	0.000
	NB	0.986	0.058	0.012	0.021	0.499	0.000
	MLP	**0.999**	0.103	0.001	0.002	0.500	0.005
	KNN	0.993	0.066	0.007	0.012	0.500	0.002
	RF	0.997	0.109	0.004	0.007	0.500	0.009
	LR	**0.999**	0.000	0.000	0.000	0.500	0.000
	DT	0.900	0.062	0.104	0.078	0.502	0.002

Table 4: Binding site prediction (classification) results for different evaluation metrics using the proposed and baseline methods for **Dataset 2**. The best values are shown in bold. Dashes "-" in the model column mean they were end-to-end models and used as described in respective original studies.

methods to combine predictions from diverse models could enhance the performance of binding site prediction, thus advancing our understanding of intricate cellular processes.

References

Ahmad, S.; and Sarai, A. 2005. PSSM-based prediction of DNA binding sites in proteins. *BMC bioinformatics*, 6: 1–6.

Aizenshtein-Gazit, S.; and Orenstein, Y. 2022. DeepZF: improved DNA-binding prediction of C2H2-zinc-finger proteins by deep transfer learning. *Bioinformatics*, 38(Supplement_2): ii62–ii67.

Brandes, N.; Ofer, D.; Peleg, Y.; Rappoport, N.; and Linial, M. 2022. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8): 2102–2110.

Collie, G. W.; and Parkinson, G. N. 2011. The application of DNA and RNA G-quadruplexes to therapeutic medicines. *Chemical Society Reviews*, 40(12): 5867–5892.

Dey, B.; Thukral, S.; Krishnan, S. A.; Chakrobarty, M.; Gupta, S.; Manghani, C.; and Rani, V. 2012. DNA–protein interactions: methods for detection and analysis. *Molecular and Cellular Biochemistry*, 365: 279–299.

Echols, H. 1986. Multiple DNA-protein interactions governing high-precision DNA transactions. *Science*, 233 4768: 1050–6.

Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M.; Bhowmik, D.; and Rost, B. 2021. ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Learning. *bioRxiv*.

Grinsztajn, L.; Oyallon, E.; and Varoquaux, G. 2022. Why do tree-based models still outperform deep learning on tabular data? *arXiv preprint arXiv:2207.08815*.

Han, W.; Pang, B.; and Wu, Y. N. 2021. Robust Transfer Learning with Pretrained Language Models through Adapters. *ArXiv*, abs/2108.02340.

He, X.; Zhou, Y.; Zhou, Z.; Bai, S.; and Bai, X. 2018. Triplet-center loss for multi-view 3d object retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1945–1954.

Heinzinger, M.; Elnaggar, A.; Wang, Y.; Dallago, C.; Nechaev, D.; Matthes, F.; and Rost, B. 2019. Modeling aspects of the language of life through transfer-learning protein sequences. *BMC bioinformatics*, 20(1): 1–17.

Ivanciuc, O.; Oezguen, N.; Mathura, V. S.; Schein, C. H.; Xu, Y.; and Braun, W. 2004. Using property based sequence motifs and 3D modeling to determine structure and functional regions of proteins. *Current medicinal chemistry*, 11 5: 583–93.

Jones, S.; van Heyningen, P.; Berman, H. M.; and Thornton, J. M. 2001. Protein-RNA interactions: a structural analysis. *Nucleic acids research*, 29 4: 943–54.

Joseph, M.; and Raj, H. 2022. GATE: Gated Additive Tree Ensemble for Tabular Classification and Regression. *arXiv preprint arXiv:2207.08548*.

Kuznetsov, I. B.; Gou, Z.; Li, R.; and Hwang, S. 2006. Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins. *PROTEINS: Structure, Function, and Bioinformatics*, 64(1): 19–27.

La, D.; and Kihara, D. 2012. A novel method for protein–protein interaction site prediction using phylogenetic substitution models. *Proteins: Structure, Function, and Bioinformatics*, 80(1): 126–141.

Lichtarge, O.; Bourne, H. R.; and Cohen, F. E. 1996. An evolutionary trace method defines binding surfaces common to protein families. *Journal of molecular biology*, 257 2: 342–58.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Liu, Y.; and Tian, B. 2023. Protein-DNA binding sites prediction based on pre-trained protein language model and contrastive learning. *arXiv preprint arXiv:2306.15912*.

Malinin, A.; Prokhorenkova, L.; and Ustimenko, A. 2021. Uncertainty in gradient boosting via ensembles. In *International Conference on Learning Representations (ICLR)*.

Novakovsky, G.; Saraswat, M.; Fornes, O.; Mostafavi, S.; and Wasserman, W. W. 2021. Biologically relevant transfer learning improves transcription factor binding prediction. *Genome biology*, 22(1): 1–25.

O'Connor, M. J. 2015. Targeting the DNA Damage Response in Cancer. *Molecular cell*, 60 4: 547–60.

Patiyal, S.; Dhall, A.; and Raghava, G. P. 2022. A deep learning-based method for the prediction of DNA interacting residues in a protein. *Briefings in Bioinformatics*, 23(5): bbac322.

Pique-Regi, R.; Degner, J. F.; Pai, A. A.; Gaffney, D. J.; Gilad, Y.; and Pritchard, J. K. 2011. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome research*, 21 3: 447–55.

Polo, S. E.; and Jackson, S. P. 2011. Dynamics of DNA damage response proteins at DNA breaks: a focus on protein modifications. *Genes & development*, 25 5: 409–33.

Ptashne, M. 1986. Gene regulation by proteins acting nearby and at a distance. *Nature*, 322: 697–701.

Qiu, J.; Bernhofer, M.; Heinzinger, M.; Kemper, S.; Norambuena, T.; Melo, F.; and Rost, B. 2020. ProNA2020 predicts protein–DNA, protein–RNA, and protein–protein binding proteins and residues from sequence. *Journal of molecular biology*, 432(7): 2428–2443.

Sarzynska-Wawer, J.; Wawer, A.; Pawlak, A.; Szymanowska, J.; Stefaniak, I.; Jarkiewicz, M.; and Okruszek, L. 2021. Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, 304: 114135.

Si, J.; Zhao, R.; and Wu, R. 2015. An overview of the prediction of protein DNA-binding sites. *International journal of molecular sciences*, 16(3): 5194–5215.

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(6): 930–936.

Tatarinova, T. V.; Chekalin, E.; Nikolsky, Y.; Bruskin, S. A.; Chebotarov, D.; McNally, K. L.; and Alexandrov, N. N. 2016. Nucleotide diversity analysis highlights functionally important genomic regions. *Scientific Reports*, 6.

Van der M., L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research (JMLR)*, 9(11).

West, S. C. 2003. Molecular views of recombination proteins and their control. *Nature Reviews Molecular Cell Biology*, 4: 435–445.

Wu, J.; Dong, Q.; Gui, J.; Zhang, J.; Su, Y.; Chen, K.; Thompson, P. M.; Caselli, R. J.; Reiman, E. M.; Ye, J.; et al. 2021. Predicting brain amyloid using multivariate morphometry statistics, sparse coding, and Correntropy: Validation in 1,101 individuals from the ADNI and OASIS databases. *Frontiers in Neuroscience*, 15: 669595.

Xia, Y.; Xia, C.-Q.; Pan, X.; and Shen, H.-B. 2021. Graph-Bind: protein structural context embedded rules learned by hierarchical graph neural networks for recognizing nucleic-acid-binding residues. *Nucleic acids research*, 49(9): e51–e51.

Xu, D.; Jalal, S. I.; Sledge, G. W.; and Meroueh, S. O. 2016. Small-molecule binding sites to explore protein-protein interactions in the cancer proteome. *Molecular bioSystems*, 12 10: 3067–87.

Yan, C.; Terribilini, M.; Wu, F.; Jernigan, R. L.; Dobbs, D.; and Honavar, V. 2006. Predicting DNA-binding sites of proteins from amino acid sequence. *BMC bioinformatics*, 7: 1–10.

Yan, J.; and Kurgan, L. 2017. DRNApred, fast sequence-based method that accurately predicts and discriminates DNA-and RNA-binding residues. *Nucleic acids research*, 45(10): e84–e84.

Yang, J.; Roy, A.; and Zhang, Y. 2012. BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic acids research*, 41(D1): D1096–D1103.

Zhang, J.; Chen, Q.; and Liu, B. 2021. NCBRPred: predicting nucleic acid binding residues in proteins based on multilabel learning. *Briefings in bioinformatics*, 22(5): bbaa397.

Zhang, J.; and Liu, B. 2019. A Review on the Recent Developments of Sequence-based Protein Feature Extraction Methods. *Current Bioinformatics*.

Zhang, J.; Ma, Z.; and Kurgan, L. 2019. Comprehensive review and empirical analysis of hallmarks of DNA-, RNA- and protein-binding residues in protein chains. *Briefings in bioinformatics*, 20(4): 1250–1268.

Zhang, Q.; Zhu, L.; and shuang Huang, D. 2019. High-Order Convolutional Neural Network Architecture for Predicting DNA-Protein Binding Sites. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 16: 1184–1192.

Zhang, Y.; Bao, W.; Cao, Y.; Cong, H.; Chen, B.; and Chen, Y. 2022. A survey on protein–DNA-binding sites in computational biology. *Briefings in Functional Genomics*, 21(5): 357–375.

Zhao, J.; Cao, Y.; and Zhang, L. 2020. Exploring the computational methods for protein-ligand binding site prediction. *Computational and Structural Biotechnology Journal*, 18: 417 – 426.

Zhu, H.; Wang, G.; and Qian, J. 2016. Transcription factors as readers and effectors of DNA methylation. *Nature Reviews Genetics*, 17(9): 551–565.

Zhu, Y.-H.; Hu, J.; Song, X.-N.; and Yu, D.-J. 2019. DNAPred: accurate identification of DNA-binding sites from protein sequence by ensembled hyperplane-distance-based support vector machines. *Journal of chemical information and modeling*, 59(6): 3057–3071.

Śledź, P.; and Caflisch, A. 2018. Protein structure-based drug design: from docking to molecular dynamics. *Current opinion in structural biology*, 48: 93–102.

Appendix

A Social Impact

DNA-binding proteins perform various functions in living cells like replication, packaging, rearrangement, transcription, gene regulation, recombination, and DNA repair. Therefore the study of this type of interaction is essential to understand the underlying processes. This understanding can potentially drive novel applications in biotechnology, agriculture, and drug design. For example, it can help improve agricultural yields and quality, reduce the loss caused by biotic and abiotic stresses, increase breeding efficiency, etc. It may be possible to develop a web application using the proposed model, where users can input protein sequences, and the model predicts whether the protein is a DNA-binding protein. This could be useful for researchers, students, or anyone interested in molecular biology. AI-driven classification of DNA-binding proteins has the potential to revolutionize biological research, drive innovation across industries, and empower individuals to engage with and contribute to advancements in molecular biology, public health monitoring, citizen science projects, and genetics.

B Sparse Coding Architecture

The overall workflow of the Sparse Coding embedding generation technique is illustrated in Figure 1. For a given protein sequence and k-mer length k, it works by computing the k-mers of the sequence. Note that here we are using $k = 3$ to depict the workflow, but in our experiments, we have used $k = 9$. In step (a), a sequence and $k = 3$ are given as inputs. Then as long as the condition $(n - k + 1)$ is satisfied, the forward k-mers are extracted in step (b). Otherwise, the reverse k-mers are obtained in step (c). Here n represents the length of the sequence. Once the k-mers are generated, they are combined in a list in step (d). Then each k-mer is passed on to the OHE method to get its respective binary vector in (e). These OHE-based vectors are concatenated to design the final numerical embedding of the corresponding sequence.

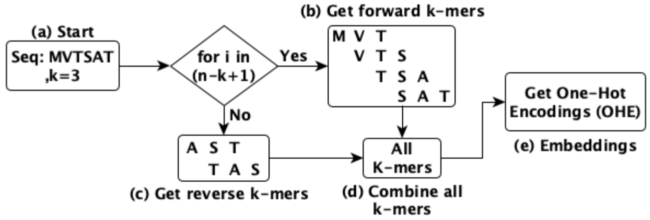


Figure 1: Workflow of Sparse Coding embedding generation method for a given sequence.

C Dataset Statistics

As our datasets contain varying sizes of sequences, the details of the maximum, minimum, and average lengths of sequences present in the test and train sets in each dataset are shown in Table 5. In each dataset, protein sequences are present along with their respective binding site indication as labels. For instance, a sample sequence from the train set of Dataset1 is "ARRIGHPYQNRTPPKRKK", where the alphabets represent the amino acids of the respective protein sequence and the labels are "001111100110111100". These labels indicate if the corresponding amino acid has DNA binding site capacity or not i.e. 1 is for the binding site & 0 for the non-binding site.

		Length Statistics		
		Max	Min	Avg
Dataset 1	Train	1937	36	279.02
	Test	743	62	236.43
Dataset 2	Train	3969	45	486.28
	Test	968	55	290.81

Table 5: The details of maximum, minimum & average lengths of sequences in the datasets 1 & 2 respectively.

D Classification Results

We utilized a deep language model, ProteinBert (Brandes et al. 2022), to get the feature embeddings of the protein sequences from our datasets and employed those embeddings to perform DNA binding site classification using our classification models (given in Section 3.2). ProteinBert is designed explicitly for protein sequences, and it consists of both local and global representations. We obtain the global representation in our experiments.

The DNA-binding site classification results of Dataset 1 is given in Table 6. The results illustrate that the NB classifier depicts maximum performance in terms of recall, f1 score, and ROC-AUC metrics, while MCC & precision is optimal for the RF classifier and specificity for MLP. However, overall ProteinBert is not showing optimal performance as compared to the other methods (mentioned in Table 3).

Moreover, the classification results obtained from Dataset 2 are reported in Table 7. We can observe that the NB model is outperforming others in terms of recall, f1 score, and ROC-AUC metrics. Precision and MCC have the highest values against the DT model, while specificity is optimal for the MLP classifier. However, yet again, the ProteinBert is unable to achieve optimal performance as compared to the other methods (mentioned in Table 4).

Method	Model	Spec.	Prec.	Recall	F1 (Bina.)	ROC-AUC	MCC
Protein Bert	1DCNN	0.998	0.000	0.044	0.000	0.521	0.060
	NB	0.529	0.106	**0.574**	**0.179**	**0.551**	0.025
	MLP	**0.999**	0.000	0.000	0.000	0.500	0.000
	KNN	0.936	0.154	0.119	0.134	0.527	0.050
	RF	0.972	**0.215**	0.078	0.115	0.525	**0.074**
	LR	0.998	0.000	0.000	0.000	0.500	0.000
	DT	0.967	0.181	0.073	0.104	0.520	0.055

Table 6: Binding site prediction (classification) results for different evaluation metrics using the ProteinBert embedding generation method for **Dataset 1**. The best values are shown in bold.

Method	Model	Spec.	Prec.	Recall	F1 (Bina.)	ROC-AUC	MCC
Protein Bert	1DCNN	0.998	0.000	0.002	0.000	0.510	0.000
	NB	0.724	0.119	**0.536**	**0.194**	**0.630**	0.067
	MLP	**0.999**	0.000	0.004	0.000	0.500	0.000
	KNN	0.979	0.068	0.021	0.032	0.500	0.001
	RF	0.996	0.230	0.016	0.029	0.506	0.044
	LR	0.998	0.002	0.003	0.000	0.500	0.010
	DT	0.991	**0.241**	0.040	0.068	0.515	**0.071**

Table 7: Binding site prediction (classification) results for different evaluation metrics using the ProteinBert embedding generation method for **Dataset 2**. The best values are shown in bold.

E t-SNE Visualization

A popular visualization technique, named t-SNE (Van der M. and Hinton 2008), is employed by us to visualize the feature vectors generated by various embedding generation methods for both datasets. We have selected the top 3 longest sequences (S1, S2, S3) from our datasets respectively, to compute the t-SNE plots. The details are discussed below.

E.1 t-SNE Dataset1

The t-SNEs against Dataset1 for the top 3 longest sequences are depicted in Figure 2. We can observe that in all the plots, the clusters are overlapping and non-definite. The binding instances are less visible and scattered throughout the plots in each figure, and a reason for it could be the data imbalance issue in the dataset i.e. number of binding instances is much less than the non-binding ones. Moreover, the Sparse Coding technique yields very similar cluster structures for all three sequences, while ProtBert and SeqVec show some variation in the structures. Overall, the patterns illustrate that none of the embedding methods for any sequence can generate very clear clusters for both the binding and non-binding classes in a 2-dimensional space.

E.2 t-SNE Dataset2

The t-SNE visualization of the top 3 longest sequences from Dataset2 are shown in Figure 3. We can observe that for any sequences against the ProtBert technique, the binding class instances are almost invisible. This indicates that this method can not preserve a good structure for less frequent classes from the dataset in a 2-dimensional space. Furthermore, the SeqVec and Sparse Coding mechanisms illustrate binding clusters being scattered across the plots for all the sequences. Overall, yet again, no definite and clear cluster structures can be viewed for Dataset2 too, and it can also be due to the class imbalance challenge.

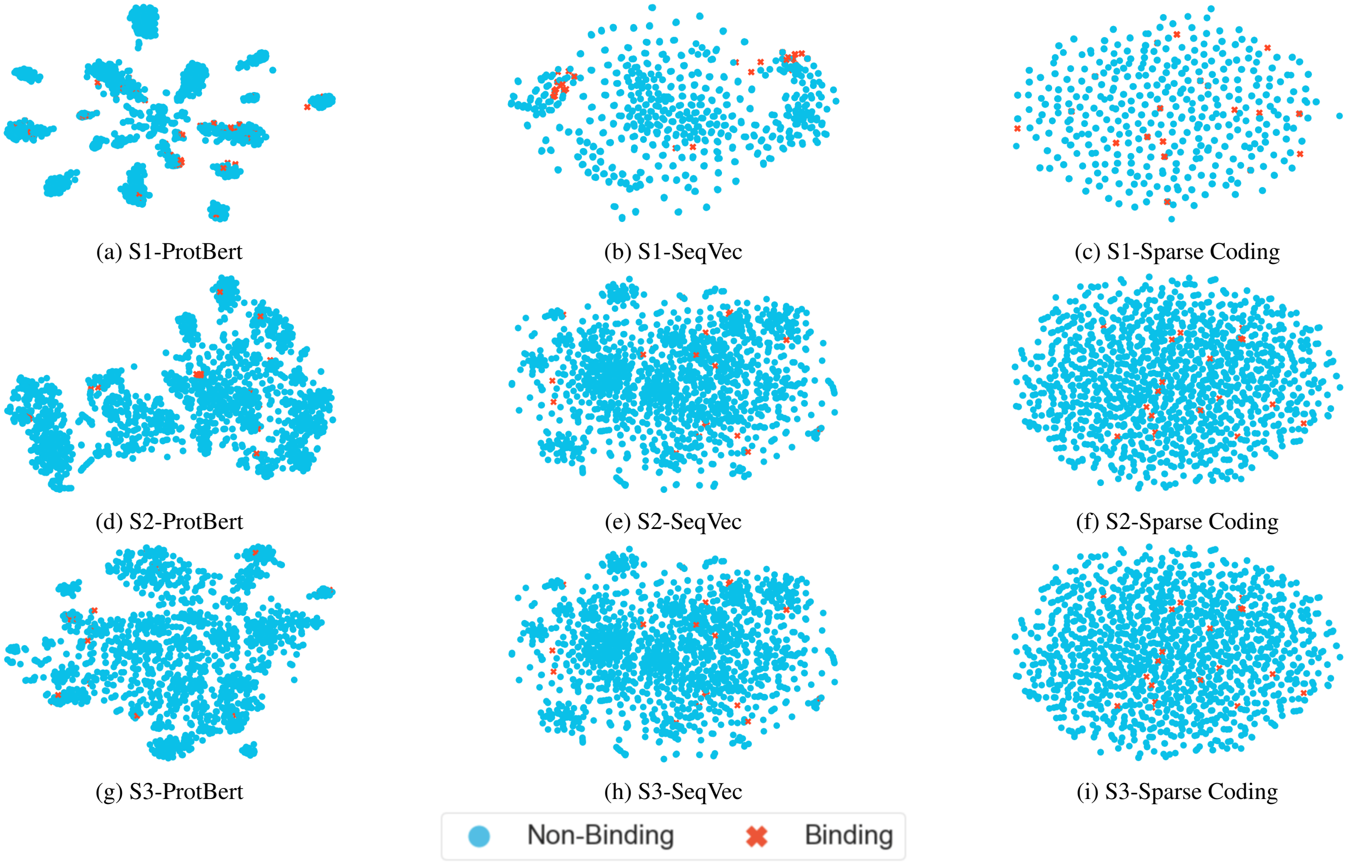
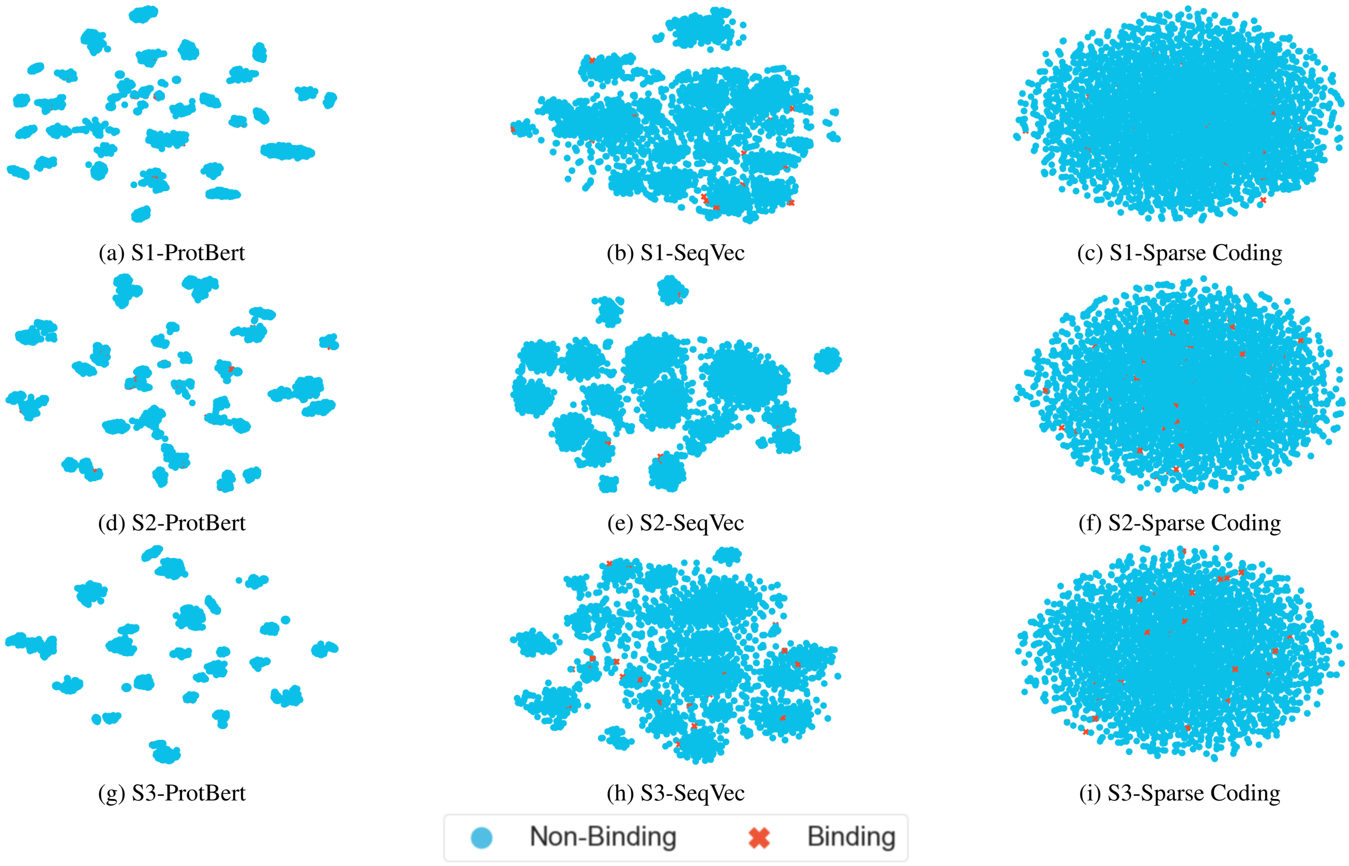

(a) S1-ProtBert
(b) S1-SeqVec
(c) S1-Sparse Coding
(d) S2-ProtBert
(e) S2-SeqVec
(f) S2-Sparse Coding
(g) S3-ProtBert
(h) S3-SeqVec
(i) S3-Sparse Coding

Figure 2: t-SNE visualization of embeddings generated by different embedding generation methods (ProtBert, SeqVec, & Sparse Coding) using top 3 longest sequences (S1, S2, S3) from **Dataset 1**. The Figure is best seen in color.



(a) S1-ProtBert

(b) S1-SeqVec

(c) S1-Sparse Coding

(d) S2-ProtBert

(e) S2-SeqVec

(f) S2-Sparse Coding

(g) S3-ProtBert

(h) S3-SeqVec

(i) S3-Sparse Coding

Figure 3: t-SNE visualization of embeddings generated by different embedding generation methods (ProtBert, SeqVec, & Sparse Coding) using top 3 longest sequences (S1, S2, S3) from **Dataset 2**. The Figure is best seen in color.