ICONIP 2024
31st International Conference on Neural Information Processing
December 2–6, 2024 · Auckland, New Zealand iconip2024.org

DeepPWM-BindingNet: Unleashing Binding Prediction with Combined Sequence and PWM Features

Sarwan Ali, Prakash Chourasia and Murray Patterson

Georgia State University
November 15, 2024

ICONIP 2024
31st International Conference on Neural Information Processing
December 2–6, 2024 · Auckland, New Zealand iconip2024.org

Sponsors: g.tec, 100% PURE NEW ZEALAND, AUT KNOWLEDGE ENGINEERING & DISCOVERY RESEARCH INNOVATION, Springer

Supported by: AUT, University of Auckland, Nottingham Trent University, APNNS

ICONIP 2024
31st International Conference on Neural Information Processing
December 2–6, 2024 · Auckland, New Zealand iconip2024.org

Table of Contents

Introduction

- Predicting binding sites and studying DNA-protein interactions is essential. They are critical for processes like gene expression, DNA repair, and signal transduction.
- Has applications in drug discovery, gene regulation, and disease prediction.
- As high-throughput sequencing advances, there is a need for computational models to predict binding interactions between DNA sequences and proteins.
- Deep learning models are becoming more popular and have proven to be effective in capturing complex relationships in biological data.
- CNNs (Convolutional Neural Networks) & RNNs (Recurrent Neural Networks) are being used for DNA-protein binding predictions.

Traditional Methods for DNA-Protein Binding Prediction

- Sequence analysis tools like MEME (Multiple Em for Motif Elicitation) [1], Gibbs Motif Sampler [2] are used to identify DNA motifs - likely binding sites for specific proteins.
- Tools like TRANSFAC [3] and JASPAR [4] provide databases of known transcription factor binding motifs.
- Position Weight Matrices (PWMs): Predict binding sites based on nucleotide frequency.
- ChIP: Experimental method to identify protein-bound DNA sequences

Challenges with existing methods?

- Struggle to capture complex sequence patterns
- Expensive and requires specialized equipment, reagents, and expertise.
- Limited sensitivity to weak/transient interactions
- Less specificity and difficulty in identifying novel binding partners
- False positives/negatives in predictions
- Existing Deep learning models are more effective in capturing complex relationships in biological data but have a few limitations
  - Overlook high-order correlations between nucleotides
  - Fixed motif length for binding site prediction
  - Miss potential interactions due to simplified models

What we propose?

- We propose DeepPWM-BindingNet, which is a novel deep-learning (DL) architecture for DNA-protein binding prediction.
- Combines DNA sequence information, protein structures, and Position Weight Matrices (PWMs). PWMs represent binding preferences at different positions in DNA sequences.
- Integration of PWM-derived features with DL enhances accuracy and interpretability.

Our Contribution:

- We integrate PWM-derived features with deep learning to improve accuracy. PWMs capture empirical data on protein-DNA binding preferences at different positions.
- Hierarchical feature extraction is made possible by utilizing CNNs and RNNs to extract local and global features from sequences.
- Attention Mechanism enhances focus on critical regions within DNA sequences to improve prediction.

Core Components

What are PWMs?

- Position Weight Matrices (PWMs) are used to encode the binding preferences of proteins at various positions in a DNA sequence. These matrices capture empirical data and provide valuable context for DNA-protein interactions.

Hierarchical Feature Extraction - CNNs + RNNs Architecture:

- CNNs (Convolutional Neural Networks): Capture local sequence patterns and motifs.
- RNNs (Recurrent Neural Networks): Capture global dependencies and long-range interactions between DNA and protein structures.

Attention weights are applied to important sequence segments, allowing the model to prioritize regions likely to interact with the protein.

DeepPWM Architecture

- Convolutional Layers: 1D convolutions capture local patterns in DNA/protein sequences with varying kernel sizes.
- Max-Pooling Layers: Down-sample feature maps to retain the important information.
- Bidirectional LSTM Layer: Captures sequential dependencies and long-range interactions, considering both past and future contexts.
- Attention Mechanism: Focuses on the most informative parts of the sequence.
- Global Average Pooling: Reduces spatial dimensions while retaining key features from the attention-weighted LSTM output.
- Dense Layers with Regularization: Extracts high-level features using ReLU activation and L2 regularization.
- Output Layer: Softmax activation for classification into binding/non-binding classes.

Model Training

We train the deep learning model using the prepared dataset with the following configurations:

- **Loss Function**: Binary cross-entropy loss [5] is used for the classification (see Equation 1).
- **Optimizer**: We use the Adam optimizer to update model weights during training.
- **Callbacks**: Callbacks such as learning rate reduction and early stopping are employed to optimize training and prevent overfitting.
- **Batch Size and Epochs**: Training is performed in mini-batches with a specified batch size, and the process is repeated for a predefined number of epochs.

$$\text{Loss} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)\right] \tag{1}$$

where y is the class label and p is the probability for prediction.

Baselines

- MLapSVM [6] :
  - This method combines features from protein sequences—pseudo-position specific scoring matrix (PsePSSM), global encoding (GE), and normalized Moreau–Broto autocorrelation (NMBAC)—and uses a novel edge weight calculation.
  - The use of multiple Laplacian regularizations creates a robust multigraph model that is less sensitive to neighborhood size.
- LapSVM [7] :
  - A semi-supervised learning method for classification that applies manifold regularization to traditional SVM.
  - They use the same features as MLapSVM—PsePSSM, global encoding (GE), NMBAC, and their concatenation—to create embeddings for LapSVM input.

Baselines

- SeqVec [8] :
  - An ELMo-based method for processing input sequences.
  - It begins by padding sequences and using character convolutions to map amino acids to a fixed-length latent space.
  - A bidirectional LSTM layer adds context, while another LSTM predicts the next word. Both passes are independently optimized during training.
- PDBP-Fusion [9]: The model combines CNNs for local feature extraction and Bi-LSTMs for capturing long-term dependencies in DNA sequences
  - Local Feature Learning: A CNN layer detects functional domains in the protein sequences.
  - Long-Term Context Learning: A Bi-LSTM layer captures long-term sequence dependencies.

Dataset

- We use the following 4 DNA-binding and non-binding protein sequences datasets.

Datasets	Total Samples			Length Statistics		
	Negative	Positive	Total	Min	Max	Mean
PDB14189	7060	7129	14819	51	4911	425.313
PDB2272	1119	1153	2272	51	5183	459.907
PDB1075	550	525	1075	51	1323	240.213
PDB186	93	93	186	64	1323	264.693

Results for PDB14189 Dataset

Method	Model	Acc. ↑	Prec. ↑	NPV ↑	Sensitivity ↑	Specificity ↑	MCC ↑	F1 ↑	ROC-AUC ↑	ROC-Pr ↑
Local	GE	85.52 ± 0.20	83.83 ± 0.24	87.42 ± 0.75	88.19 ± 0.87	82.82 ± 0.47	71.13 ± 0.46	85.95 ± 0.29	92.85 ± 0.20	90.80 ± 0.19
Behavior	NMBAC	89.70 ± 0.01	85.29 ± 0.03	95.45 ± 0.03	96.07 ± 0.03	83.27 ± 0.04	80.04 ± 0.01	90.36 ± 0.01	95.93 ± 0.05	94.90 ± 0.04
Similarity	MCD	74.84 ± 0.29	73.48 ± 0.63	76.49 ± 1.49	78.16 ± 2.34	71.49 ± 1.77	49.81 ± 0.71	75.72 ± 0.76	82.45 ± 0.13	80.67 ± 0.02
(LapSVM) [7]	PSSM	76.88 ± 0.83	71.91 ± 1.52	85.19 ± 1.13	88.76 ± 1.56	64.89 ± 3.24	55.34 ± 1.06	79.42 ± 0.30	86.42 ± 0.09	84.50 ± 0.50
	Combined	74.00 ± 0.08	71.38 ± 0.05	77.43 ± 0.12	80.54 ± 0.13	67.39 ± 0.03	48.37 ± 0.17	75.69 ± 0.09	82.11 ± 0.17	81.98 ± 0.07
Local	GE	74.64 ± 0.40	72.57 ± 0.76	77.19 ± 0.28	79.63 ± 0.66	69.59 ± 1.38	49.49 ± 0.73	75.93 ± 0.20	82.39 ± 0.40	82.26 ± 0.66
Behavior	NMBAC	74.07 ± 0.71	67.12 ± 0.61	91.14 ± 1.36	94.88 ± 0.88	53.06 ± 1.31	52.85 ± 1.47	78.62 ± 0.55	87.08 ± 0.40	85.20 ± 0.42
Similarity	MCD	76.99 ± 0.83	77.22 ± 0.36	76.80 ± 1.48	76.88 ± 2.13	77.10 ± 0.73	54.00 ± 1.62	77.04 ± 1.14	84.40 ± 0.75	82.72 ± 0.94
(MLapSVM) [6]	PSSM	91.27 ± 0.33	88.11 ± 0.65	95.07 ± 0.58	95.53 ± 0.58	86.97 ± 0.85	82.83 ± 0.62	91.66 ± 0.29	96.50 ± 0.40	95.57 ± 0.72
	Combined	87.39 ± 0.50	85.71 ± 0.84	89.29 ± 0.70	89.91 ± 0.79	84.84 ± 1.08	74.88 ± 0.99	87.75 ± 0.46	93.93 ± 0.53	92.05 ± 1.06
	SVM	71.45 ± 0.01	71.51 ± 0.01	71.74 ± 0.01	72.31 ± 0.01	69.12 ± 0.02	42.42 ± 0.01	71.17 ± 0.01	71.81 ± 0.01	67.62 ± 0.01
	NB	55.11 ± 0.01	65.41 ± 0.02	78.45 ± 0.03	96.21 ± 0.01	13.90 ± 0.01	17.87 ± 0.02	45.14 ± 0.01	55.46 ± 0.00	52.85 ± 0.00
	MLP	74.43 ± 0.01	74.56 ± 0.01	74.38 ± 0.02	74.12 ± 0.03	75.66 ± 0.01	48.97 ± 0.01	74.32 ± 0.01	74.66 ± 0.01	69.32 ± 0.01
SeqVec [8]	KNN	72.57 ± 0.00	72.62 ± 0.00	72.88 ± 0.01	71.92 ± 0.01	73.77 ± 0.01	44.16 ± 0.01	72.73 ± 0.00	72.54 ± 0.00	69.51 ± 0.01
	RF	76.75 ± 0.00	76.74 ± 0.00	77.94 ± 0.00	78.17 ± 0.01	73.65 ± 0.01	52.31 ± 0.01	76.65 ± 0.00	76.11 ± 0.00	68.32 ± 0.01
	LR	72.54 ± 0.00	72.76 ± 0.00	73.99 ± 0.01	74.14 ± 0.01	70.11 ± 0.01	44.84 ± 0.01	72.95 ± 0.00	72.81 ± 0.00	67.56 ± 0.01
	DT	65.54 ± 0.00	65.11 ± 0.00	65.84 ± 0.01	64.33 ± 0.01	66.74 ± 0.01	31.69 ± 0.00	65.57 ± 0.00	65.30 ± 0.00	67.56 ± 0.00
	2 Layer CNN (OH)	80.34 ± 0.89	81.64 ± 1.89	79.46 ± 2.93	78.75 ± 4.55	81.95 ± 3.15	60.90 ± 1.64	80.04 ± 1.62	88.74 ± 0.50	86.69 ± 0.85
PDBP-	3 Layer CNN (OH)	81.64 ± 1.38	80.98 ± 3.38	83.41 ± 4.68	83.61 ± 7.01	79.66 ± 5.80	63.82 ± 2.28	81.95 ± 2.35	90.01 ± 0.63	88.57 ± 0.88
Fusion [9]	Fusion (Embed)	53.35 ± 3.07	64.24 ± 8.87	41.52 ± 21.24	38.89 ± 37.31	67.95 ± 39.57	9.72 ± 7.38	36.27 ± 22.55	56.74 ± 7.05	56.20 ± 5.41
	Fusion (OH)	80.49 ± 1.61	77.66 ± 4.03	85.35 ± 4.15	86.79 ± 5.37	74.13 ± 7.67	61.94 ± 2.25	81.72 ± 1.10	89.25 ± 0.72	87.66 ± 0.98
DeepPWM-BindingNet	-	79.80 ± 0.94	77.53 ± 1.65	82.69 ± 1.75	84.32 ± 2.29	75.24 ± 2.71	59.89 ± 1.87	80.74 ± 0.90	87.69 ± 0.88	86.44 ± 0.94

- Results for different embedding methods on **SARS-CoV-2 Variant Dataset**.
- Although it is not better the advantage of our proposed method is that it provides interpretability due to the inclusion of the attention mechanism.

Results for PDB2272 Dataset

Method	Model	Acc. ↑	Prec. ↑	NPV ↑	Sensitivity ↑	Specificity ↑	MCC ↑	F1 ↑	ROC-AUC ↑	ROC-Pr ↑
Local	GE	55.24 ± 1.47	53.15 ± 0.82	96.44 ± 4.36	99.74 ± 0.35	9.38 ± 2.75	21.08 ± 4.56	69.35 ± 0.74	74.11 ± 2.38	70.59 ± 1.94
Behavior	NMBAC	51.36 ± 0.33	51.07 ± 0.18	80.95 ± 18.87	99.57 ± 0.39	1.70 ± 0.52	6.29 ± 3.33	67.51 ± 0.20	74.20 ± 2.18	71.91 ± 1.85
Similarity	MCD	60.91 ± 2.30	58.30 ± 1.89	67.36 ± 2.97	81.35 ± 2.30	39.85 ± 5.32	23.31 ± 4.66	67.89 ± 1.39	65.04 ± 2.68	63.52 ± 3.62
(LapSVM) [7]	PSSM	67.25 ± 2.25	61.60 ± 1.64	87.09 ± 3.98	94.36 ± 1.64	39.32 ± 3.83	40.48 ± 4.79	74.53 ± 1.50	83.82 ± 2.87	84.61 ± 2.88
	Combined	61.97 ± 2.50	57.59 ± 1.72	85.41 ± 3.86	95.58 ± 0.92	27.35 ± 4.73	31.34 ± 5.26	71.86 ± 1.44	80.64 ± 2.80	81.19 ± 2.28
Local	GE	55.24 ± 1.48	53.16 ± 0.82	95.57 ± 3.98	99.65 ± 0.32	9.47 ± 2.79	20.89 ± 4.55	69.33 ± 0.74	74.10 ± 2.38	70.58 ± 1.94
Behavior	NMBAC	51.32 ± 0.32	51.05 ± 0.18	80.95 ± 18.87	99.57 ± 0.39	1.61 ± 0.54	6.03 ± 3.11	67.49 ± 0.19	74.21 ± 2.18	71.91 ± 1.86
Similarity	MCD	60.96 ± 2.26	58.32 ± 1.86	67.47 ± 2.90	81.44 ± 2.32	39.85 ± 5.32	23.42 ± 4.57	67.93 ± 1.35	65.04 ± 2.68	63.52 ± 3.62
(MLapSVM) [6]	PSSM	67.30 ± 2.36	61.67 ± 1.75	86.80 ± 3.78	94.19 ± 1.56	39.59 ± 4.13	40.44 ± 4.89	74.53 ± 1.55	83.83 ± 2.86	84.61 ± 2.88
	Combined	62.06 ± 2.56	57.65 ± 1.77	85.48 ± 3.84	95.58 ± 0.92	27.52 ± 4.88	31.51 ± 5.34	71.91 ± 1.47	80.64 ± 2.80	81.18 ± 2.28
	SVM	51.42 ± 0.03	51.45 ± 0.03	51.65 ± 0.03	63.78 ± 0.12	40.12 ± 0.17	03.34 ± 0.06	50.44 ± 0.05	51.12 ± 0.03	60.87 ± 0.05
	NB	56.56 ± 0.01	59.43 ± 0.01	53.67 ± 0.01	26.63 ± 0.02	85.67 ± 0.01	14.45 ± 0.02	51.56 ± 0.01	56.11 ± 0.01	75.56 ± 0.01
	MLP	57.45 ± 0.01	57.41 ± 0.01	56.86 ± 0.01	53.22 ± 0.04	61.56 ± 0.03	14.33 ± 0.02	57.56 ± 0.01	57.12 ± 0.01	66.87 ± 0.01
SeqVec [8]	KNN	57.86 ± 0.01	57.54 ± 0.01	56.32 ± 0.02	49.75 ± 0.03	64.77 ± 0.03	13.43 ± 0.02	56.43 ± 0.01	57.36 ± 0.01	67.57 ± 0.01
	RF	61.24 ± 0.01	62.52 ± 0.01	63.41 ± 0.03	68.86 ± 0.03	55.44 ± 0.03	23.57 ± 0.03	61.17 ± 0.01	61.47 ± 0.01	63.13 ± 0.02
	LR	58.77 ± 0.01	59.42 ± 0.01	56.26 ± 0.01	44.37 ± 0.02	72.22 ± 0.02	17.90 ± 0.03	57.45 ± 0.01	58.36 ± 0.01	70.27 ± 0.02
	DT	56.22 ± 0.03	56.74 ± 0.03	56.31 ± 0.03	58.67 ± 0.03	55.78 ± 0.05	12.67 ± 0.06	56.44 ± 0.03	56.68 ± 0.03	64.24 ± 0.02
	2 Layer CNN (OH)	60.78 ± 6.07	61.33 ± 8.35	75.99 ± 22.07	79.86 ± 20.72	41.12 ± 31.55	26.27 ± 9.57	66.65 ± 4.96	74.29 ± 2.45	72.06 ± 3.00
PDBP-	3 Layer CNN (OH)	54.60 ± 4.75	53.56 ± 5.27	91.49 ± 21.00	97.80 ± 8.80	10.09 ± 17.33	16.95 ± 9.96	68.61 ± 1.94	77.78 ± 2.48	75.48 ± 3.70
Fusion [9]	Fusion (Embed)	51.02 ± 2.81	45.20 ± 28.67	35.69 ± 22.11	37.18 ± 44.41	65.29 ± 45.91	3.65 ± 8.11	28.89 ± 29.44	52.11 ± 6.75	54.46 ± 5.53
	Fusion (OH)	69.44 ± 3.32	70.15 ± 3.94	70.74 ± 6.79	70.70 ± 13.00	68.12 ± 10.03	39.82 ± 6.31	69.54 ± 6.30	78.11 ± 2.90	76.11 ± 3.61
DeepPWM-BindingNet	-	69.40 ± 4.50	66.86 ± 4.05	71.50 ± 15.06	80.50 ± 5.88	57.96 ± 12.75	39.38 ± 9.43	72.81 ± 2.57	75.88 ± 3.37	73.62 ± 3.35

- Results for different embedding methods on **SARS-CoV-2 Variant Dataset**.
- We can observe that in terms of average accuracy, the proposed method shows value in the top 5% accuracy.

Results for PDB1075 Dataset

Method	Model	Acc. ↑	Prec. ↑	NPV ↑	Sensitivity ↑	Specificity ↑	MCC ↑	F1 ↑	ROC-AUC ↑	ROC-Pr ↑
Local Behavior Similarity (LapSVM) [7]	GE	51.16 ± 0.00	0.00 ± 0.00	51.16 ± 0.00	0.00 ± 0.00	100.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	77.50 ± 2.46	74.97 ± 1.55
	NMBAC	51.16 ± 0.00	0.00 ± 0.00	51.16 ± 0.00	0.00 ± 0.00	100.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	77.11 ± 4.24	74.41 ± 4.08
	MCD	61.21 ± 1.37	81.29 ± 5.55	57.41 ± 0.89	27.05 ± 3.16	93.82 ± 2.47	28.35 ± 3.58	40.41 ± 3.51	76.20 ± 1.78	73.95 ± 2.87
	PSSM	75.07 ± 4.55	80.06 ± 5.50	71.84 ± 4.08	65.14 ± 5.83	84.55 ± 4.11	50.78 ± 9.23	71.79 ± 5.49	83.06 ± 3.58	79.57 ± 4.49
	Combined	70.70 ± 3.87	78.79 ± 5.30	66.70 ± 3.43	54.86 ± 6.64	85.82 ± 4.13	43.00 ± 7.84	64.48 ± 5.62	80.55 ± 3.34	76.41 ± 4.46
Local Behavior Similarity (MLapSVM) [6]	GE	51.16 ± 0.00	0.00 ± 0.00	51.16 ± 0.00	0.00 ± 0.00	100.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	77.49 ± 2.45	74.98 ± 1.53
	NMBAC	51.07 ± 0.19	0.00 ± 0.00	51.12 ± 0.09	0.00 ± 0.00	99.82 ± 0.36	-1.34 ± 2.67	0.00 ± 0.00	77.11 ± 4.24	74.41 ± 4.08
	MCD	61.12 ± 1.23	81.21 ± 5.48	57.34 ± 0.80	26.86 ± 3.04	93.82 ± 2.47	28.17 ± 3.34	40.18 ± 3.30	76.21 ± 1.77	73.94 ± 2.87
	PSSM	74.88 ± 4.44	79.98 ± 5.45	71.61 ± 3.97	64.76 ± 5.68	84.55 ± 4.11	50.44 ± 9.02	71.52 ± 5.35	83.07 ± 3.60	79.58 ± 4.50
	Combined	70.42 ± 3.49	78.98 ± 5.18	66.29 ± 2.99	53.90 ± 5.83	86.18 ± 4.04	42.58 ± 7.17	63.90 ± 5.00	80.56 ± 3.33	76.42 ± 4.45
SeqVec [8]	SVM	48.56 ± 0.05	48.43 ± 0.05	50.26 ± 0.05	37.23 ± 0.11	59.98 ± 0.05	-4.11 ± 0.10	47.12 ± 0.05	48.45 ± 0.05	63.32 ± 0.02
	NB	57.45 ± 0.04	60.67 ± 0.04	66.43 ± 0.04	80.22 ± 0.03	36.56 ± 0.06	18.88 ± 0.07	55.26 ± 0.05	58.68 ± 0.03	56.89 ± 0.03
	MLP	52.87 ± 0.02	52.89 ± 0.03	54.55 ± 0.04	55.93 ± 0.05	50.72 ± 0.05	4.45 ± 0.05	52.57 ± 0.02	52.32 ± 0.03	61.45 ± 0.02
	KNN	58.56 ± 0.02	58.67 ± 0.03	60.66 ± 0.03	62.35 ± 0.06	54.88 ± 0.03	16.43 ± 0.05	58.32 ± 0.02	58.56 ± 0.02	62.22 ± 0.02
	RF	61.78 ± 0.02	61.43 ± 0.02	61.22 ± 0.03	55.56 ± 0.04	67.67 ± 0.03	22.78 ± 0.04	61.32 ± 0.02	61.45 ± 0.02	66.57 ± 0.02
	LR	53.32 ± 0.04	53.45 ± 0.04	56.38 ± 0.05	62.92 ± 0.06	44.58 ± 0.02	6.59 ± 0.08	52.26 ± 0.03	53.43 ± 0.04	59.44 ± 0.01
	DT	57.67 ± 0.03	57.55 ± 0.03	59.43 ± 0.02	57.81 ± 0.04	57.99 ± 0.06	15.64 ± 0.06	57.32 ± 0.03	57.67 ± 0.03	63.28 ± 0.03
PDBP-Fusion [9]	2 Layer CNN (OH)	68.65 ± 2.86	66.72 ± 5.99	75.58 ± 8.05	75.73 ± 15.14	61.96 ± 15.27	39.85 ± 5.06	69.45 ± 5.47	78.08 ± 2.41	74.29 ± 2.97
	3 Layer CNN (OH)	68.33 ± 3.69	62.96 ± 4.57	82.47 ± 5.29	87.64 ± 7.20	50.15 ± 12.33	41.27 ± 5.33	72.87 ± 1.94	79.25 ± 1.78	76.38 ± 2.72
	Fusion (Embed)	51.10 ± 3.27	28.06 ± 28.71	40.51 ± 23.96	45.46 ± 47.69	56.40 ± 46.73	0.63 ± 8.38	31.31 ± 32.45	68.03 ± 5.07	66.48 ± 5.64
	Fusion (OH)	65.04 ± 5.67	59.57 ± 4.87	86.05 ± 6.42	92.01 ± 6.47	39.64 ± 14.98	37.15 ± 9.32	71.96 ± 2.82	80.09 ± 1.67	76.92 ± 3.31
DeepPWM-BindingNet	-	72.05 ± 2.56	69.37 ± 3.48	75.66 ± 3.54	76.37 ± 5.43	68.00 ± 6.19	44.70 ± 5.03	72.52 ± 2.57	78.45 ± 2.70	74.13 ± 5.23

- The accuracy is comparable and the F1 score is almost the same as the best.

Results for PDB186 Dataset

Method	Model	Acc. ↑	Prec. ↑	NPV ↑	Sensitivity ↑	Specificity ↑	MCC ↑	F1 ↑	ROC-AUC ↑	ROC-Pr ↑
Local Behavior Similarity (LapSVM) [7]	GE	52.70 ± 5.92	40.71 ± 21.36	56.03 ± 8.71	54.80 ± 36.51	50.99 ± 30.86	6.36 ± 12.92	45.22 ± 26.86	61.03 ± 8.05	59.99 ± 6.81
	NMBAC	49.47 ± 1.60	0.00 ± 0.00	49.73 ± 1.32	0.00 ± 0.00	98.95 ± 2.11	-3.29 ± 6.58	0.00 ± 0.00	68.62 ± 6.32	65.23 ± 6.79
	MCD	53.23 ± 3.15	55.25 ± 7.51	52.24 ± 1.82	34.15 ± 10.88	71.99 ± 11.22	6.77 ± 8.10	41.21 ± 9.53	57.49 ± 4.36	60.72 ± 5.85
	PSSM	66.15 ± 7.08	66.08 ± 9.23	67.61 ± 6.23	69.71 ± 9.35	62.40 ± 14.50	32.87 ± 14.11	67.31 ± 6.61	70.15 ± 8.44	71.77 ± 6.48
	Combined	65.59 ± 11.00	63.71 ± 11.10	69.53 ± 12.18	73.98 ± 13.07	57.08 ± 14.29	32.09 ± 22.01	68.16 ± 10.85	67.32 ± 8.77	67.99 ± 5.46
Local Behavior Similarity (MLapSVM) [6]	GE	52.70 ± 5.92	40.71 ± 21.36	56.03 ± 8.71	54.80 ± 36.51	50.99 ± 30.86	6.36 ± 12.92	45.22 ± 26.86	61.09 ± 8.03	60.10 ± 6.72
	NMBAC	49.47 ± 1.60	0.00 ± 0.00	49.73 ± 1.32	0.00 ± 0.00	98.95 ± 2.11	-3.29 ± 6.58	0.00 ± 0.00	68.68 ± 6.29	65.26 ± 6.78
	MCD	53.23 ± 3.15	55.25 ± 7.51	52.24 ± 1.82	34.15 ± 10.88	71.99 ± 11.22	6.77 ± 8.10	41.21 ± 9.53	57.43 ± 4.33	60.60 ± 5.74
	PSSM	65.60 ± 6.40	65.80 ± 8.81	66.98 ± 5.83	68.60 ± 9.96	62.40 ± 14.50	31.85 ± 12.83	66.51 ± 6.13	70.10 ± 8.36	71.72 ± 6.41
	Combined	65.59 ± 11.00	63.71 ± 11.10	69.53 ± 12.18	73.98 ± 13.07	57.08 ± 14.29	32.09 ± 22.01	68.16 ± 10.85	67.31 ± 8.78	67.98 ± 5.44
SeqVec [8]	SVM	50.11 ± 0.04	49.25 ± 0.05	48.32 ± 0.11	53.91 ± 0.12	46.45 ± 0.20	-1.54 ± 0.11	49.57 ± 0.05	50.43 ± 0.05	61.56 ± 0.04
	NB	60.43 ± 0.05	61.61 ± 0.04	62.67 ± 0.04	65.17 ± 0.08	56.57 ± 0.11	21.52 ± 0.09	60.12 ± 0.05	60.76 ± 0.05	63.43 ± 0.04
	MLP	50.32 ± 0.06	51.64 ± 0.06	52.87 ± 0.09	54.43 ± 0.16	48.32 ± 0.13	2.43 ± 0.12	50.41 ± 0.06	51.89 ± 0.06	61.65 ± 0.05
	KNN	56.24 ± 0.06	58.46 ± 0.06	57.67 ± 0.09	52.55 ± 0.15	63.57 ± 0.17	15.13 ± 0.12	56.92 ± 0.07	57.59 ± 0.06	66.26 ± 0.08
	RF	53.35 ± 0.04	54.15 ± 0.03	54.23 ± 0.06	51.15 ± 0.13	55.73 ± 0.12	7.91 ± 0.07	52.35 ± 0.04	53.91 ± 0.03	63.55 ± 0.05
	LR	51.46 ± 0.08	52.32 ± 0.08	53.57 ± 0.11	52.68 ± 0.12	51.43 ± 0.07	3.57 ± 0.16	51.79 ± 0.08	52.32 ± 0.08	62.46 ± 0.02
	DT	49.45 ± 0.02	50.43 ± 0.03	50.42 ± 0.04	48.56 ± 0.14	51.78 ± 0.14	-1.48 ± 0.06	48.36 ± 0.02	50.78 ± 0.03	62.47 ± 0.06
PDBP-Fusion [9]	2 Layer CNN (OH)	56.20 ± 6.36	55.71 ± 5.94	54.67 ± 30.92	79.31 ± 22.77	33.54 ± 27.62	15.38 ± 13.17	62.89 ± 10.17	65.30 ± 5.84	67.28 ± 5.55
	3 Layer CNN (OH)	51.83 ± 4.26	51.20 ± 3.14	22.07 ± 30.17	95.18 ± 8.03	8.34 ± 14.15	4.20 ± 9.43	66.35 ± 2.67	64.12 ± 5.21	64.52 ± 5.35
	Fusion (Embed)	50.86 ± 1.83	40.33 ± 28.42	22.27 ± 28.85	64.43 ± 47.44	36.63 ± 47.62	2.50 ± 7.05	43.78 ± 31.24	65.37 ± 7.42	67.24 ± 5.53
	Fusion (OH)	53.55 ± 6.67	52.70 ± 5.40	24.46 ± 33.68	93.31 ± 10.97	13.81 ± 20.83	8.01 ± 14.64	66.75 ± 3.58	67.31 ± 5.57	69.13 ± 4.90
DeepPWM-BindingNet	-	49.89 ± 2.81	50.02 ± 2.22	6.55 ± 17.86	97.61 ± 5.05	2.20 ± 7.83	-1.58 ± 7.77	66.05 ± 1.93	60.00 ± 11.82	64.34 ± 8.68

- The proposed method shows a near-perfect score for the sensitivity metric.

Results Summary

- While our method may not always surpass baselines in raw metrics, its unique strengths offer significant value.
- Advantages:
  - Resource Efficiency, Interpretability, and Adaptability make it a practical addition to the field.
  - Complements existing techniques (e.g., PWM), enhancing the overall toolkit for researchers
- Provide strong potential for improved real-world applicability and ethical considerations.
- Opens new avenues for exploration and positions itself as a solid foundation for future research.

Conclusion and Future Work

Conclusion

- DeepPWM-BindingNet combines deep learning with PWM-derived features for improved DNA-protein binding predictions.
- The use of hierarchical feature extraction and an attention mechanism enhances both predictive performance and model interpretability.

Future Work

- Explore Transfer Learning: Investigating novel deep learning techniques to improve efficiency.
- Broader Applications: Testing the model on other biological tasks to assess its generalizability.

Thank You

ICONIP 2024
31st International Conference on Neural Information Processing
December 2–6, 2024 · Auckland, New Zealand iconip2024.org

Questions !!

- T. L. Bailey, N. Williams, C. Misleh, and W. W. Li, "Meme: discovering and analyzing dna and protein sequence motifs," *Nucleic acids research*, vol. 34, 2006.
- A. F. Neuwald, J. S. Liu, and C. E. Lawrence, "Gibbs motif sampling: detection of bacterial outer membrane protein repeats," *Protein science*, vol. 4, no. 8, pp. 1618–1632, 1995.
- E. Wingender, P. Dietze, H. Karas, and R. Knüppel, "Transfac: a database on transcription factors and their dna binding sites," *Nucleic acids research*, vol. 24, no. 1, pp. 238–241, 1996.
- A. Sandelin, W. Alkema, P. Engström, W. W. Wasserman, and B. Lenhard, "Jaspar: an open-access database for eukaryotic transcription factor binding profiles," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D91–D94, 2004.
- Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," *Advances in neural information processing systems*, vol. 31, 2018.

- M. Sun, P. Tiwari, Y. Qian, Y. Ding, and Q. Zou, "Mlapsvm-lbs: Predicting dna-binding proteins via a multiple laplacian regularized support vector machine with local behavior similarity," *Knowledge-Based Systems*, vol. 250, p. 109174, 2022.
- M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples." *JMLR*, vol. 7, no. 11, 2006.
- M. Heinzinger, A. Elnaggar, Y. Wang, C. Dallago, D. Nechaev, F. Matthes, and B. Rost, "Modeling aspects of the language of life through transfer-learning protein sequences," *BMC bioinformatics*, vol. 20, no. 1, pp. 1–17, 2019.
- G. Li, X. Du, X. Li, L. Zou, G. Zhang, and Z. Wu, "Prediction of dna binding proteins using local features and long-term dependencies with primary sequences based on deep learning," *PeerJ*, vol. 9, p. e11262, 2021.