Overview

This research focuses on predicting peptide binding using a deep learning model trained on labeled amino acid sequences. The dataset included 60,970 peptides, each 25 amino acids in length, labeled as either binding (1) or non-binding (0). The goal was to build a reliable, generalizable model capable of handling sequences of any length using padding and sliding window techniques.

Dataset & Preprocessing

The dataset was consistent in length, with each peptide consisting of exactly 25 residues. About 11.5% of sequences were labeled as binders.

- Encoding: Each amino acid was converted into an integer (A–Y → 0–19).
- Shuffling & Splitting: 80% of the dataset was used for training, 20% for validation.
- Storage: Sequences and labels were saved as NumPy arrays (`X.npy` and `y.npy`).

Scripts such as `load_and_preview.py`, `analyze_lengths.py`, and `encode_sequences.py` handled loading, validation, and formatting to prepare the data for training.

Model Architecture

The model is a 1D Convolutional Neural Network (CNN), implemented using TensorFlow/Keras. It was chosen for its ability to detect local sequence motifs predictive of binding.

- Embedding Layer: Converts amino acid integers into dense 16-dimensional vectors.
- Conv1D Layer: Applies filters to detect local patterns across the sequence.
- Global Max Pooling: Captures the most significant feature activation across the entire peptide.
- Dense Layer: Interprets pooled features to predict binding likelihood.
- Sigmoid Output: Produces a probability score between 0 and 1.

The model was trained for 5 epochs with a batch size of 64, using binary cross-entropy loss and the Adam optimizer.

Evaluation Results

Validation Performance:

- Accuracy: ~90%
- Area Under ROC Curve (AUC): 0.84

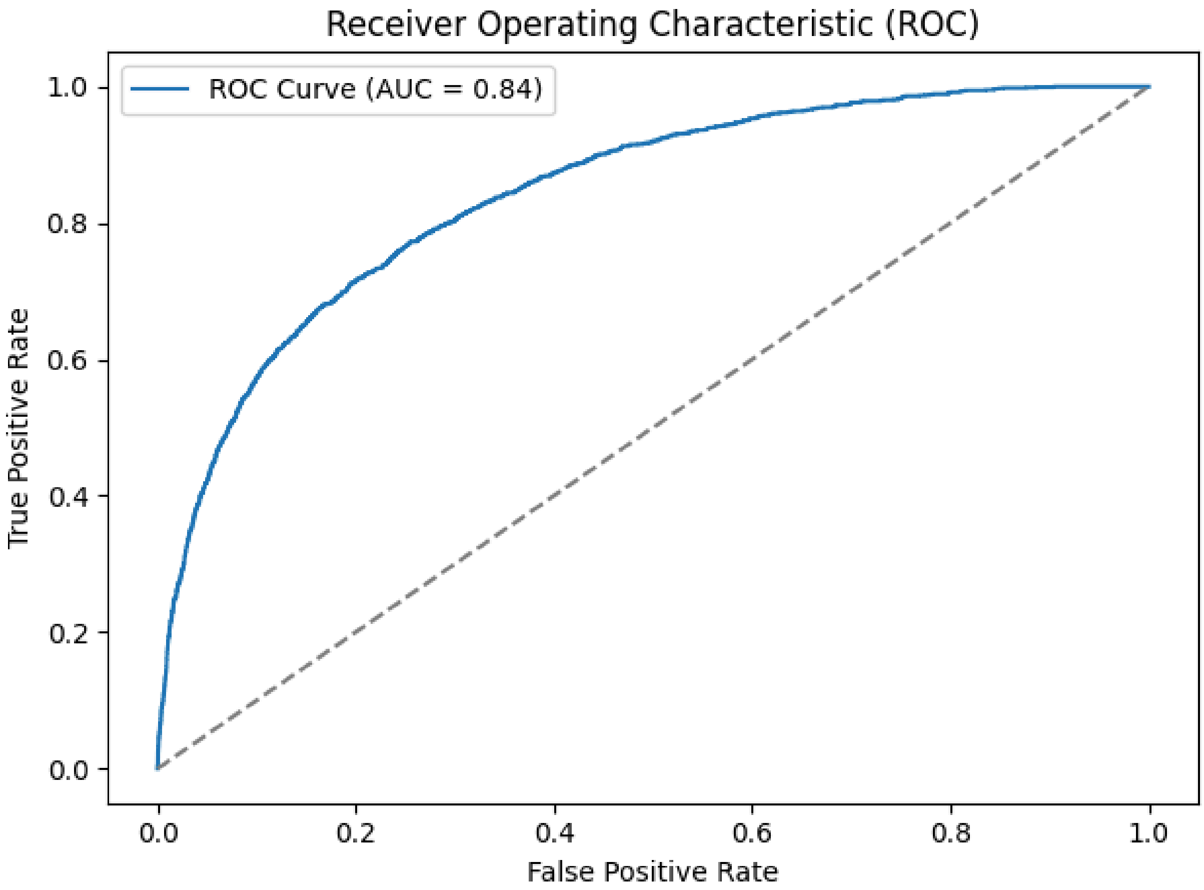

Figure 1: *ROC curve (roc_curve.png)*
This shows strong class separation and reliable ranking of binding probabilities. The AUC of 0.84 indicates good predictive power.

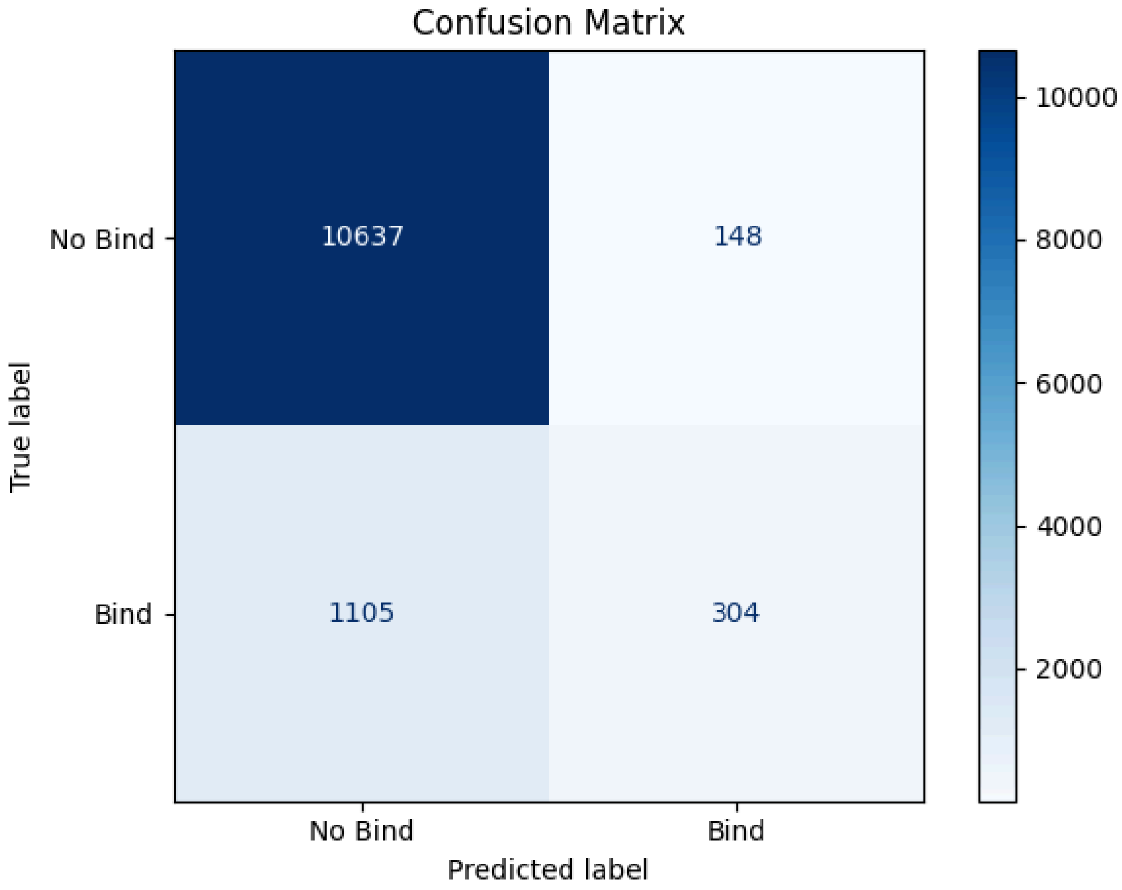

Figure 2: *Confusion matrix (confusion_matrix.png)*
The model correctly predicts most binders and non-binders. A small number of false negatives suggest that some true binders are being missed.

Generalization to Variable-Length Sequences

To support real-world use cases, the model was extended to handle peptides of arbitrary length.

- Shorter Sequences: Padded with neutral tokens (encoded as 0) to reach 25 residues.
- Longer Sequences: Processed using a sliding window approach. Each 25-residue segment is scored independently, and overall predictions are aggregated using max and mean scores.

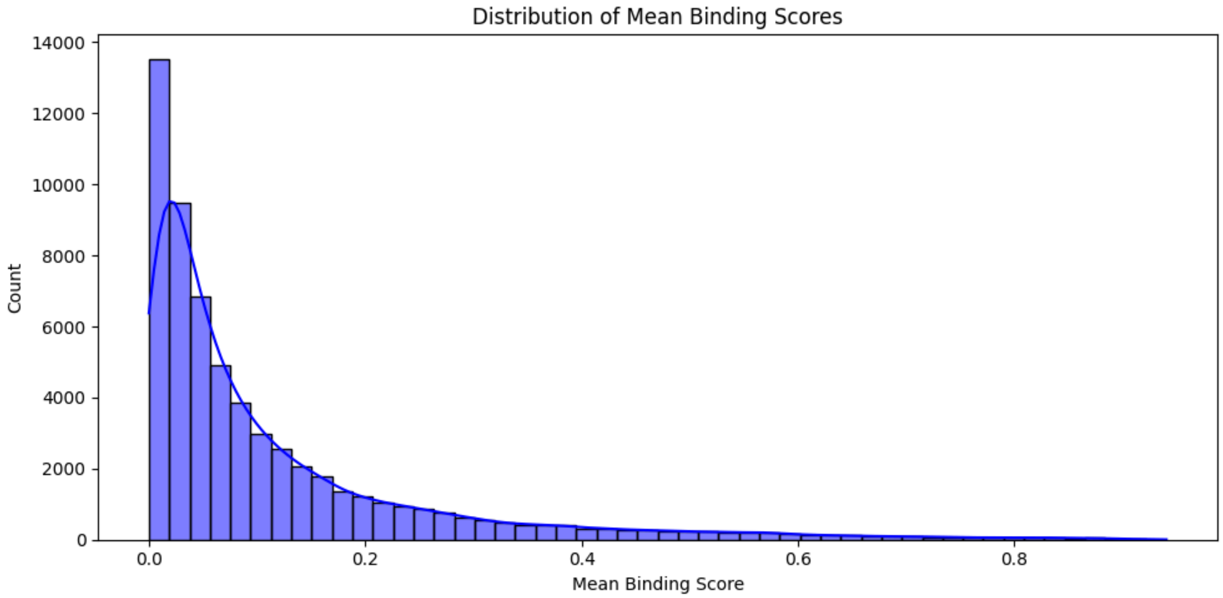

Figure 3: *Mean binding score distribution (mean_score_distribution.png)*
This histogram shows that most sequences have low average binding scores, with a smaller group of high-scoring candidates that are likely binders.

Key Scripts

- `predict_sequence.py`: Runs predictions for any single peptide input.
- `scan_windows.py`: Applies sliding window prediction to longer sequences.
- `plot_predictions.py`: Visualizes binding score distributions and ranks top-scoring peptides.

Strengths & Limitations

Strengths:

- High accuracy and AUC with minimal training time.
- Sliding window and padding make the model usable for any sequence length.
- Modular code and clear visual outputs enhance usability and reproducibility.

Limitations:

- The model only captures short-range patterns within 25 residues.
- It does not use biochemical or structural features explicitly.
- It may miss true binders that depend on long-range or 3D structural context.

Future Work

- Incorporate structure-derived features (e.g., predicted secondary structure, solvent exposure).
- Explore transformer-based or recurrent architectures for capturing long-range dependencies.
- Adjust prediction thresholds to optimize for recall in detecting binders.
- Validate results using independent datasets or experimental feedback.

Conclusion

This project demonstrates that a simple convolutional model can accurately predict peptide binding based on sequence alone. With techniques for generalization and scoring full-length proteins, the model offers a flexible foundation for further development. Future work will focus on integrating structural context and improving sensitivity to detect additional true binders.