Presentation: Problem F

dsmp-2024-groupt3 | Mini-Project, Bristol University

Pulkit Dhingra, Luoxi Liu, Shaivya Shankar, Shuyi Li

Overview

Introduction to Protein Binding

→ Proteins often function by forming complexes through *protein-protein interactions*

→ These interactions occur at specific binding sites involving *residue pairs* from two protein surfaces

Residue Pairs and Binding Challenge

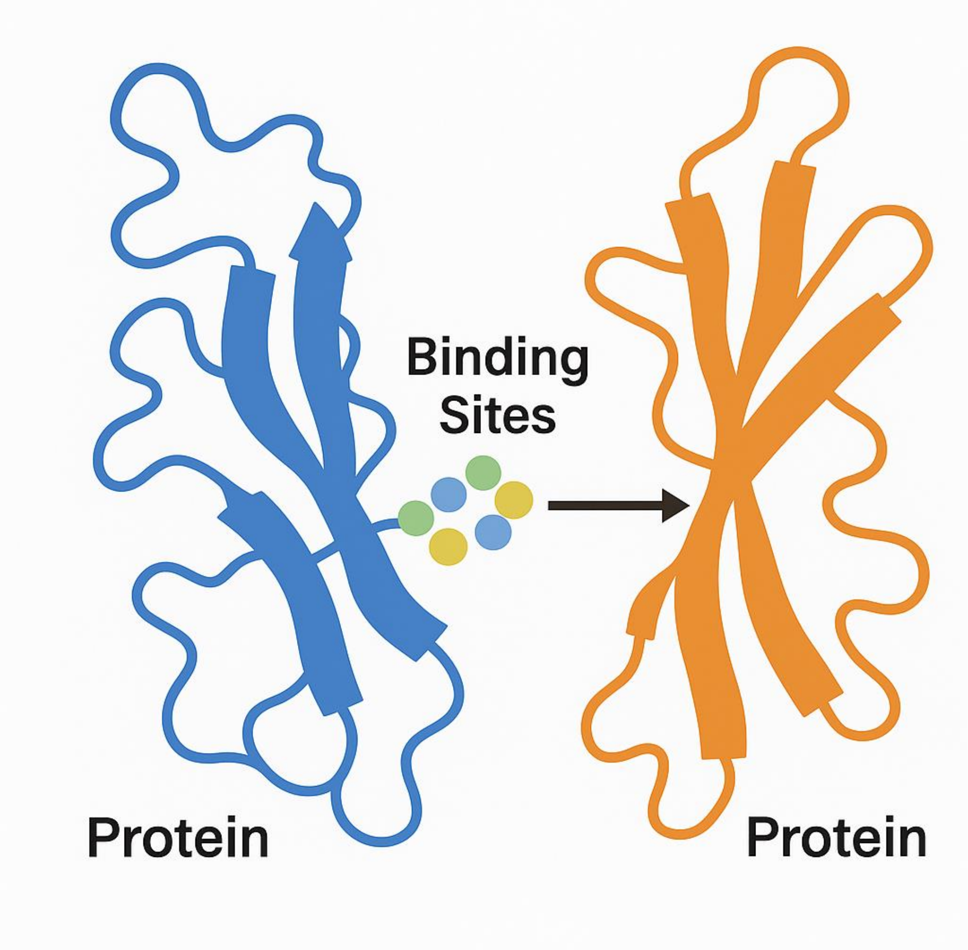
→ Not all residues participate in binding — only certain *core interacting residues* form stable contacts

→ Predicting which residue pairs bind is crucial for understanding molecular mechanisms and for drug design

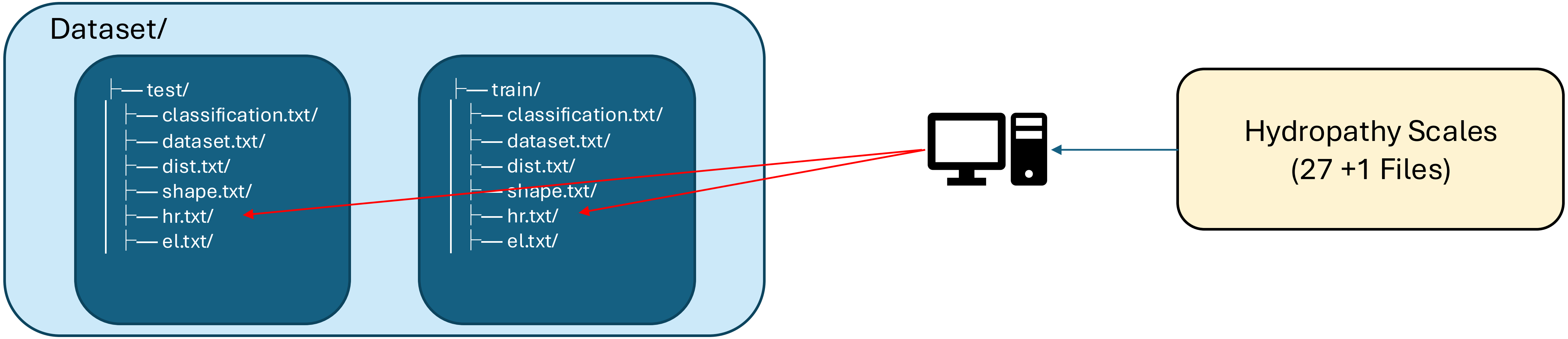
Our Task

→ **Task I:** *Analyze hydropathy impact on identifying interacting residue pairs*

→ **Task II:** *Compare hydropathy scales and define a new combined scale using PCA*



Data Sources



Dataset/

```
├── test/
│   ├── classification.txt/
│   ├── dataset.txt/
│   ├── dist.txt/
│   ├── shape.txt/
│   ├── hr.txt/
│   ├── el.txt/
```

```
├── train/
│   ├── classification.txt/
│   ├── dataset.txt/
│   ├── dist.txt/
│   ├── shape.txt/
│   ├── hr.txt/
│   ├── el.txt/
```

Hydropathy Scales
(27 +1 Files)

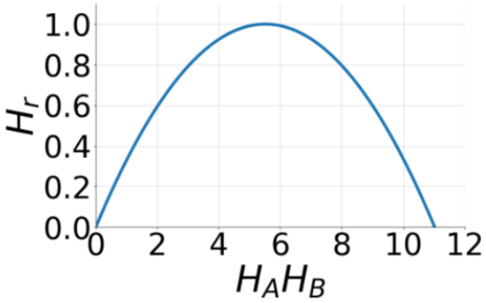
Our Understanding of the data source files and formats

- **Data.txt :-** This file represents the residueA along and residueB along with 9 different neighbours of the B.
- **Hr.txt :-** This files contains hydropathy value which has been computed using the given formula. Where Ha and Hb represents the hydrophobicity values of residue A and B. “a” and “b” value which was originally given to us 0.033 and 0.363.
- **El.txt :-** Electrostatic complementarity between A_n and B_n and between A_n and the eight neighbours of B_n.
- **Dist.txt :-**This represents the distance between residue A_n and residue B_n and between residue A_n and the nine neighbours of B_n.
- **Shape.txt :-** Shape complementarity between A_n and B_n and between A_n and the nine neighbours of B_n
- **Classification.txt :-** Contains actual labels for the training dataset

Task 1:
Analyze hydropathy impact on Identify interacting residue pairs

Formula Modifications Success and Failures

✅ Task 1 required generating hr.txt files using 27 hydropathy scales.
❌ The provided formula couldn't generate valid Hr values for all scales.



Our Work:

- Modified the original formula to adapt to new scale values.

$$
\begin{aligned}
-aX_1^2 + bX_1 + c &= 0 \quad \text{(min)}\\
-aX_2^2 + bX_2 + c &= 1 \quad \text{(mid)}\\
-aX_3^2 + bX_3 + c &= 0 \quad \text{(max)}
\end{aligned}
$$

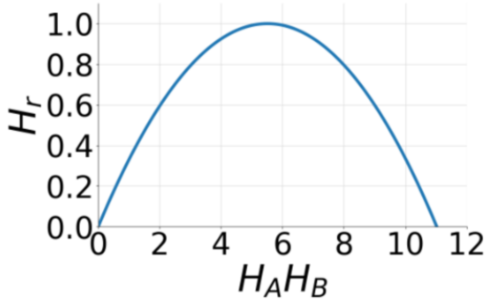
$$H_r = -a(H_A H_B)^2 + b(H_A H_B)$$

Compared with the given formula, the modified formula adds a new parameter c. Where X1 represents the minimum value of HA * HB, X2 represents the median value, and X3 represents the maximum value of HA * HB. And 0 and 1 correspond to the maximum and minimum values of Hr respectively.

It can handle negative values without shifting the scale

Skewed distribution of data made it hard to identify the median value of HA * HB accurately.

Math's Behind Our Work



$$y = -ax^2 + bc + C$$

then $y = C$,

c can vary depending on scale

Using each scale we need to find the Hydropathy values based on residue pairs

$$H_r = -a(H_A H_B)^2 + b(H_A H_B)$$

We need to find ‘a’ and ‘b’ values for each scale

Max Point $\longrightarrow -a{X_1}^2 + bX_1 = 0$

Min Point $\longrightarrow -a{X_2}^2 + bX_2 = 0$

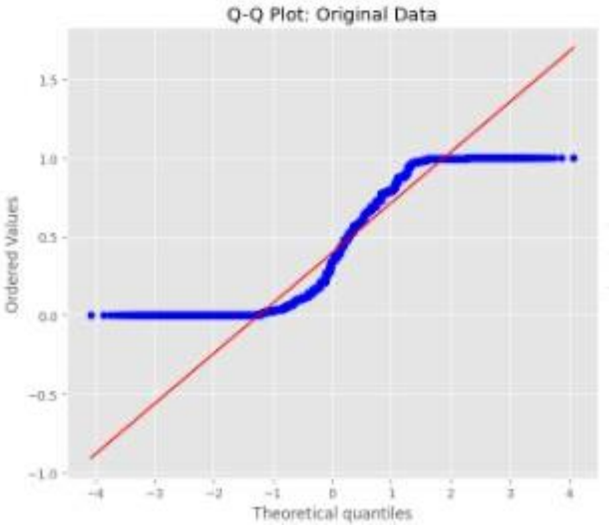
$$\begin{bmatrix} X_1 & X_1 \\ X_2 & X_2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

where, $X_1, X_2 \rightarrow$ *represents the* $H_a H_b$ *values*

Issues within the Generated data

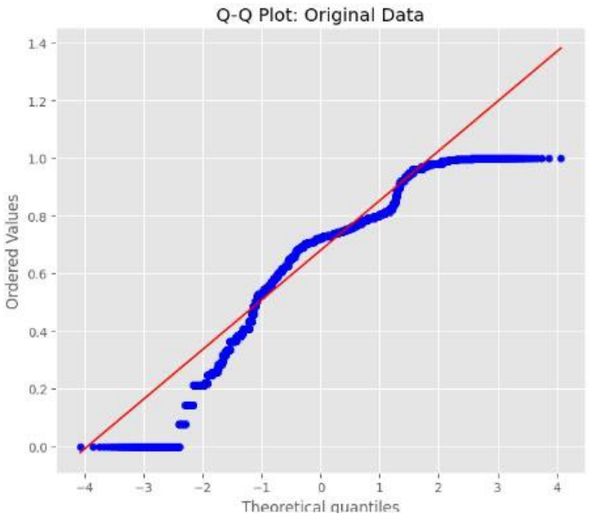
While analyzing the generated hydropathy files, we came across the following issues within the generated Hydropathy data.

Best case



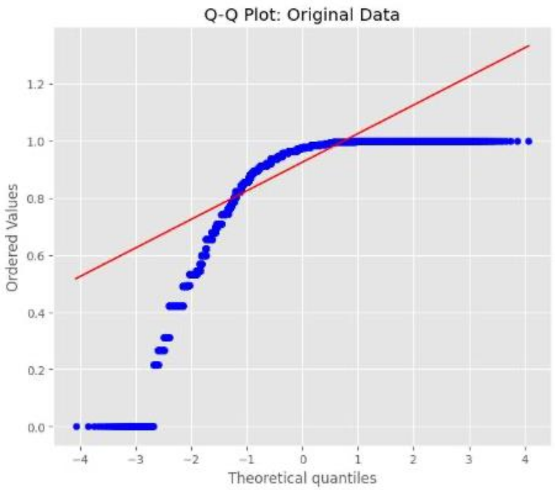
ZIMJ680101

High Variance



PRAM900101

Skewed Data



CASG920101

Almost all the data had high Kurtosis

Impact on the model training

Unstable Gradients & Slow Convergence

- Datasets with high kurtosis can lead to gradient instability during training, impairing a model’s capacity to capture complex patterns.

Batch Normalization Sensitivity

- Batch Normalization (BN) assumes that activations are approximately normally distributed. Heavy-tailed distributions can skew BN's moving mean and variance estimates, reducing its effectiveness and causing shifts in activation distributions across batches.

Overfitting to Outliers

- High-capacity CNNs may overfit to rare outlier data points, leading to poor generalization.

Gradient Explosion in Deep Networks

- In CNNs, large activations from outliers can propagate through layers, exacerbating gradient explosion or vanishing issues.

Data Engineering

Origin Shifting

Shift Ha·Hb values to ensure positivity and standardize input range.

Ensures all HaHb values are positive and compatible with the parabolic formula.

Noise Injection

Add Gaussian noise to enhance model robustness and generalization.

(Data Regularization) Improves tolerance to data irregularities and simulates real-world variation.

Power Transform

Normalize skewed data distributions for better model learning.

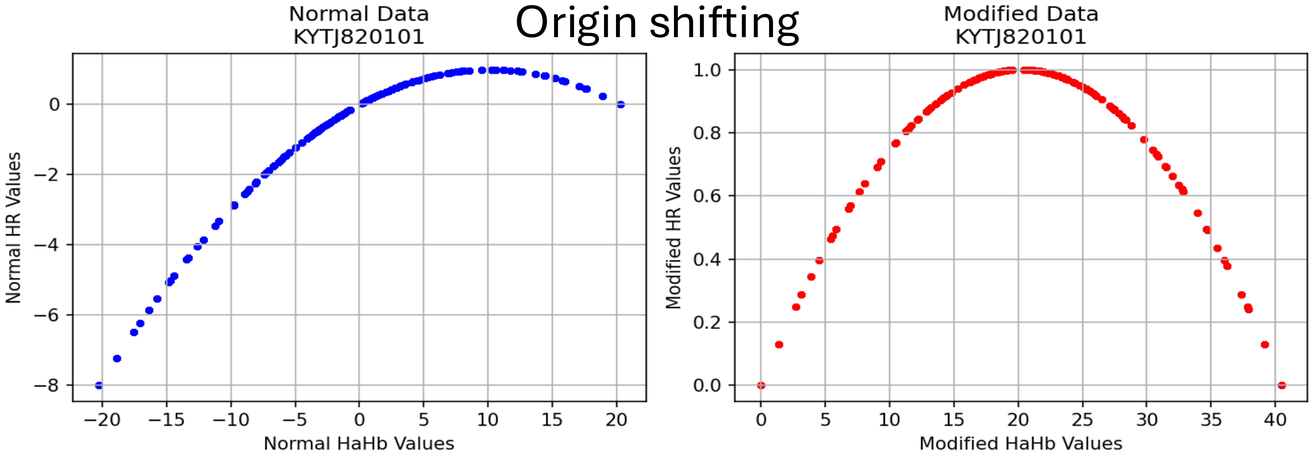
Reduces skewness, making feature distributions more Gaussian for stability.

Min-Max Scala

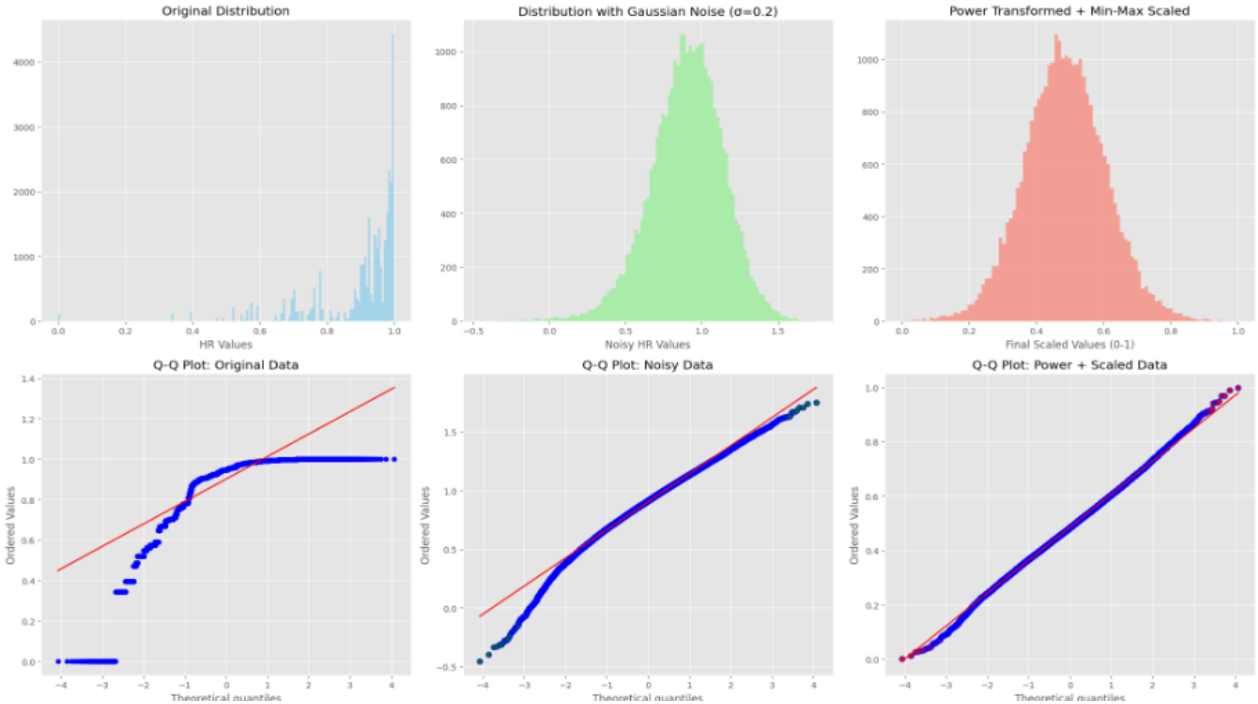
Scale features to [0, 1] for consistent neural network input.

Facilitates faster convergence and balanced learning in neural networks.

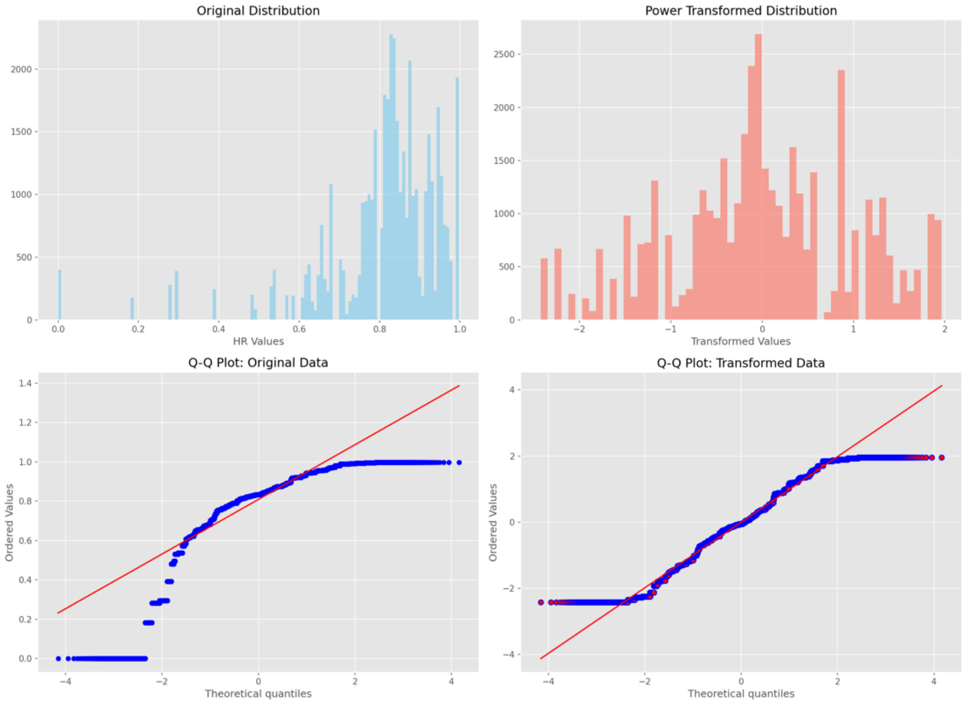
Origin shifting



Noise Injection



Power Transform



Model Design Phase

Phases of Model Training

Phase 1

Took the existing CIR-Net architecture and trained the model across all 28 scales.

Phase 2

Extended the architecture and applied hyperparameter tuning to optimize performance.

Phase 3

Adopted a statistical-based approach and implemented machine learning models for training.

Result Analysis of all the phases

L hydrophobicity scale
Accuracy = 76.93
Precision = 0.7660
Recall = 0.7619
F1- Score = 0.7640

	Predicted Positives	Predicted Negatives
Actual Positives	480 (TP)	150 (FN)
Actual Negatives	146 (FP)	506 (TN)

ENGD860101
Accuracy = 80.52
Precision = 0.8077
Recall = 0.7937
F1- Score = 0.8006

	Predicted Positives	Predicted Negatives
Actual Positives	500 (TP)	130 (FN)
Actual Negatives	119 (FP)	533 (TN)

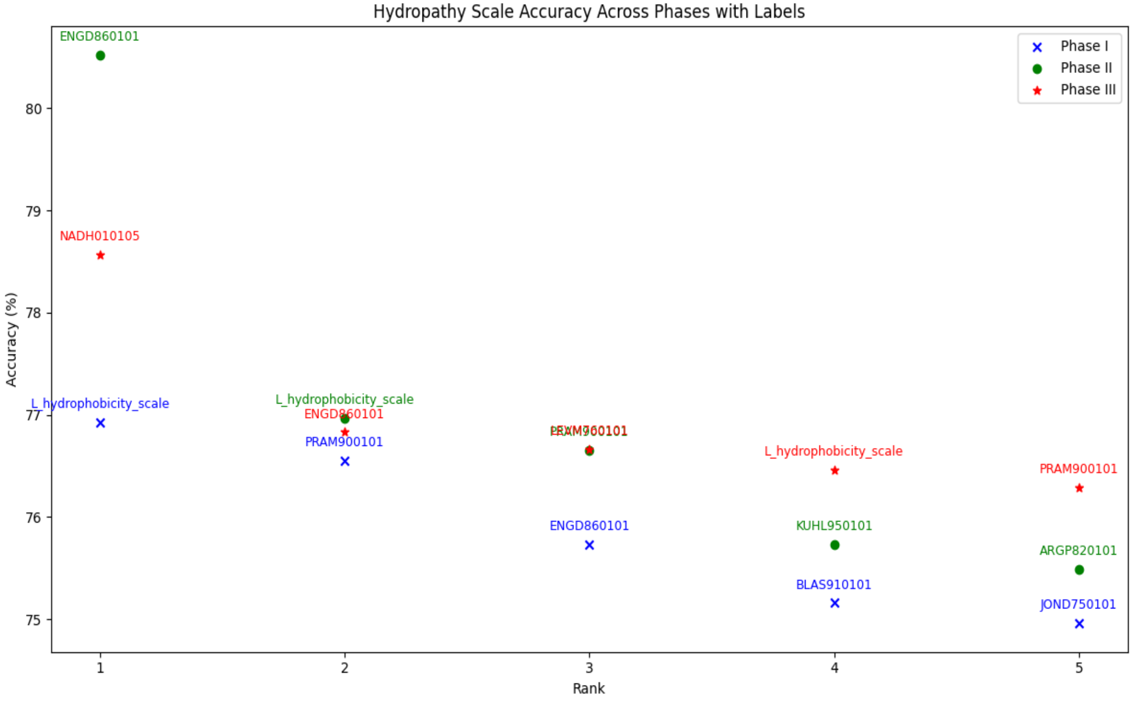
(XG-Boost)

NADH010105
Accuracy = 78.57
Precision = 0.7795
Recall = 0.7857
F1- Score = 0.7826

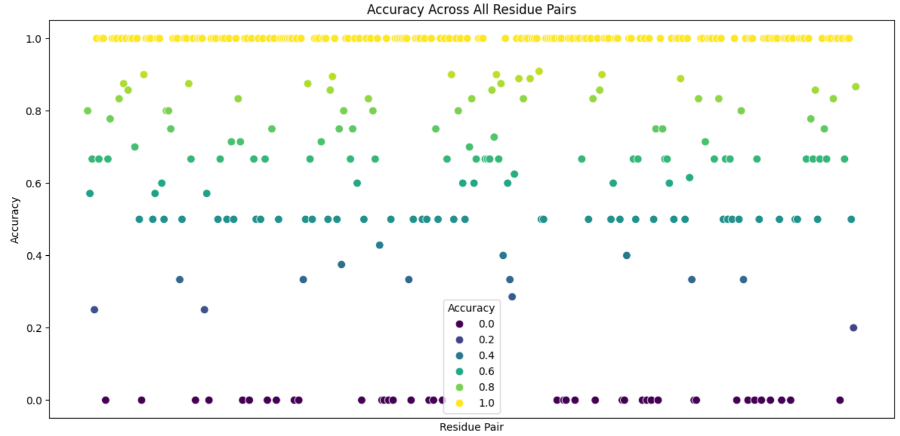
	Predicted Positives	Predicted Negatives
Actual Positives	495 (TP)	135 (FN)
Actual Negatives	140 (FP)	512 (TN)

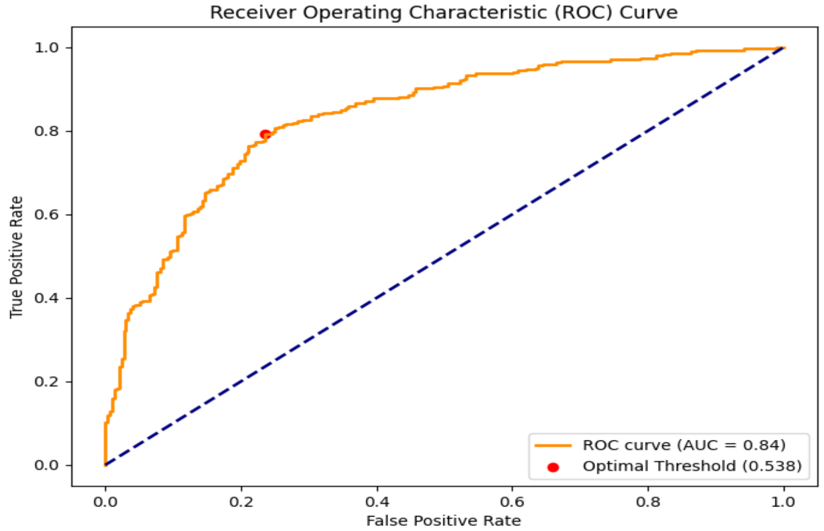
Threshold Value for Best scale:- 0.513

Visualize the Results



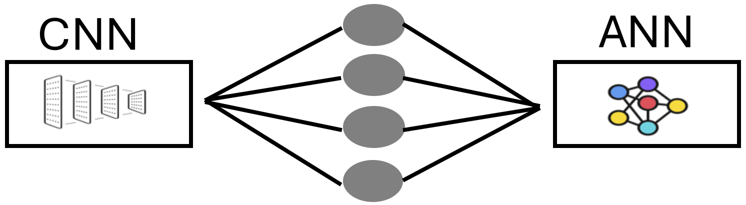
Visualizing accuracy over various phases of model training





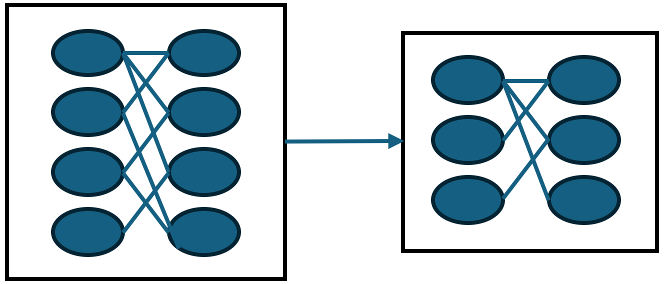
ROC curve showing threshold for the best scale

Future Enhancement Suggestions



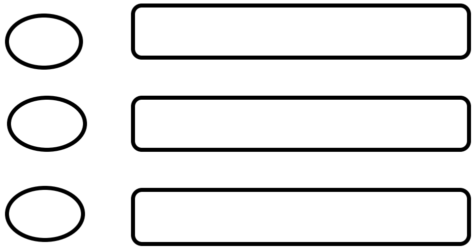
A. Explore Alternative Architectures

Compare CNNs, ANNs and RNNs to better model residue pairs from the given features and hydropathy scales



B. Incorporate Transfer Learning

Pre-train on one hydropathy scale and fine tune on others to improve the results.



C. Integrate more features

Indicates how exposed each residue is to the surrounding solvent.
➔ **Surface-exposed residues** are more likely to participate in interactions.

Measures how much a residue stays unchanged across species.
➔ **Highly conserved residues** often play critical functional roles, including binding.

Task 2:
New Hydropathy Scale Construction

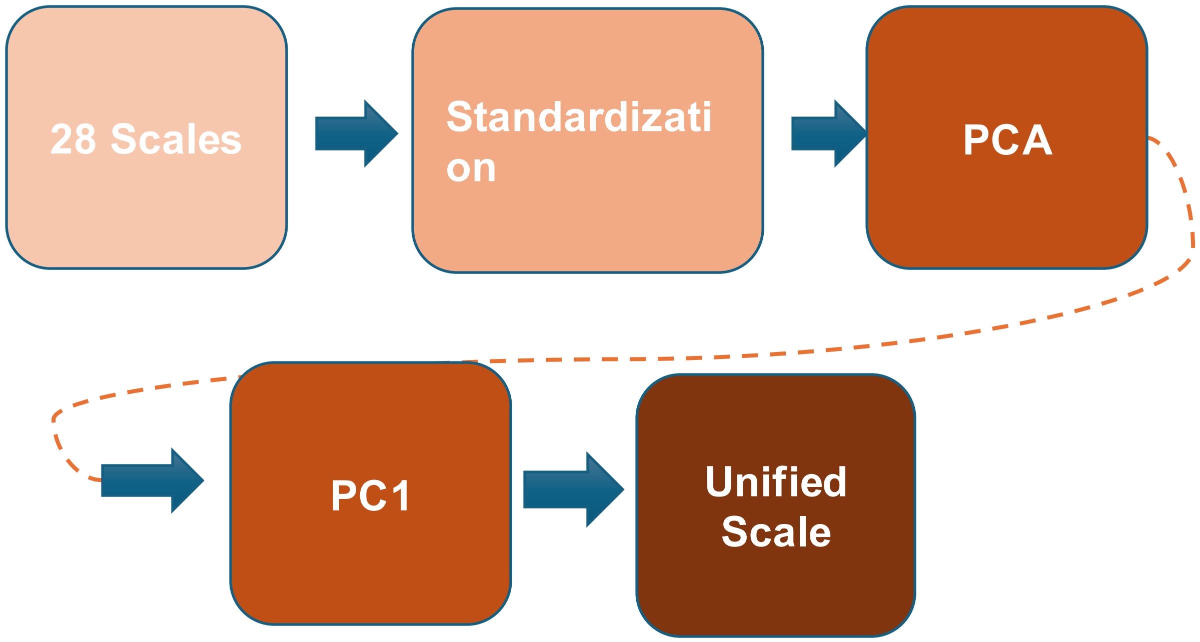
Method Selection

Derive an optimal unified hydropathy scale from individual 28 scales

Why PCA?

- More robust than simple averaging or clustering.
- Captures maximum shared variance
- Objective, avoids subjective weighting

Workflow



28 Scales → Standardization → PCA → PC1 → Unified Scale

PCA Method for Unified Hydropathy Scale

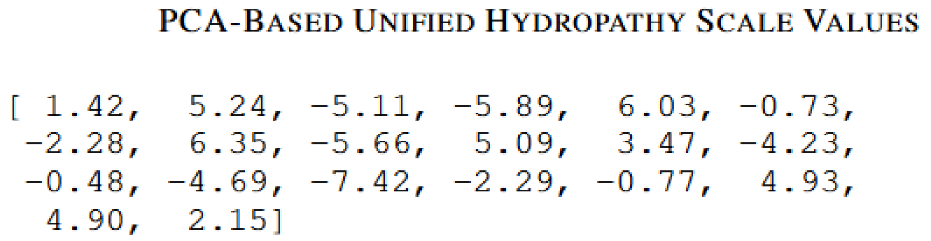
Preprocessing

- Align all scales to a fixed amino acid order.
- Standardize using Z-score normalization.

PCA Approach

- Input: X ∈ ℝ^{20×28} (20 amino acids × 28 scales).
- Extract First Principal Component (PC1).

PCA Based Scale

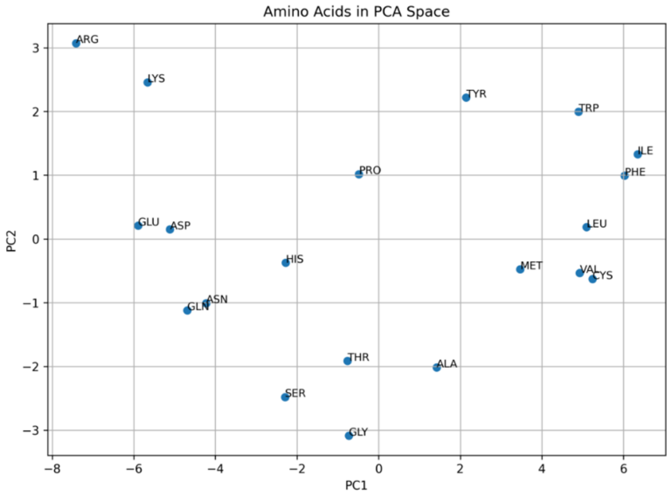


PCA-Based Unified Hydropathy Scale Values

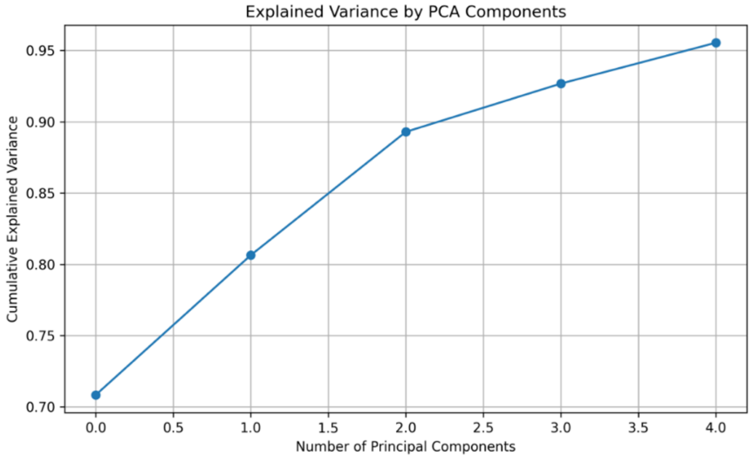
```
[ 1.42,  5.24, -5.11, -5.89,  6.03, -0.73,
 -2.28,  6.35, -5.66,  5.09,  3.47, -4.23,
 -0.48, -4.69, -7.42, -2.29, -0.77,  4.93,
  4.90,  2.15]
```

PCA Results

Amino Acid Clustering

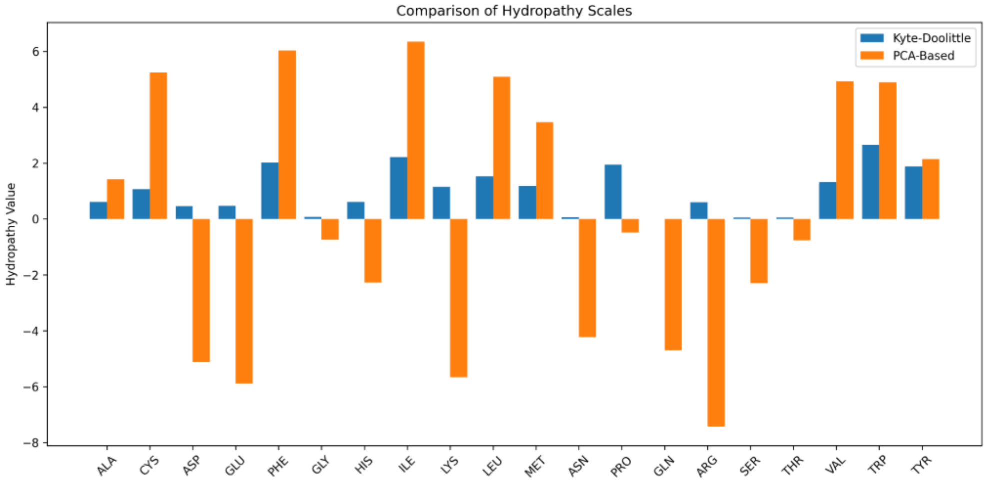


Explained Variance



- PC1 explains >70% variance.
- PC1 + PC2 explain >90%.

Scale Comparison with Kyte-Doolittle



- Indicates broader physicochemical features from multiple hydropathy models.