Comparative Analysis and Optimization of Hydropathy Scales for Predicting Core Interacting Residues in Protein Interfaces

1st Luoxi Liu
MSC. Data Science
University of Bristol
Bristol, UK
ay24061@bristol.ac.uk

2nd Shaivya Shankar
MSC. Data Science
University of Bristol
Bristol, UK
hv24220@bristol.ac.uk

3rd Shuyi Li
MSC. Data Science
University of Bristol
Bristol, UK
mm24576@bristol.ac.uk

4th Pulkit Dhingra
MSC. Data Science
University of Bristol
Bristol, UK
gl24171@bristol.ac.uk

***Abstract*—Protein interactions arise from a combination of structural and chemical complementarity, enabling the selective recognition and binding of partners within the crowded intracellular environment. These interactions lead to the formation of functional molecular complexes that are crucial for nearly all biological processes, allowing proteins to perform coordinated activities such as signal transduction, gene regulation, immune response, and metabolic regulation. Understanding these interactions—and precisely identifying binding regions—is essential for uncovering the molecular mechanisms underlying cellular function.**

This study builds upon the work of Grassmann et al. (2024), which introduced the CIRNet model for predicting core interacting residues based on shape, electrostatic, and hydropathy complementarity. In particular, we investigate the impact of various hydropathy scale values on the generation of hydropathy profiles and the prediction accuracy of protein-protein interactions based on these hydropathy files. We further identify optimal scale values tailored for different amino acid residues. Additionally, this research explores alternative neural network architectures to enhance the prediction of binding regions between residue pairs, offering a broader perspective on improving interaction site identification.

I. Introduction

Protein-protein interactions (PPIs) play an important role in regulating cellular processes, and they are the core knowledge in many biological functions, including signal transduction, immunoreaction, and enzymatic activity. These interactions are governed by mechanisms that involve complementary structural and chemical features at the protein surfaces. Identifying and understanding the regions where this interaction takes place is a challenge in computational biology and is essential for advancing manufacturing and drug design, synthetic biology, and systems biology. Working on this problem, Grassmann et al. (2024) introduced CIRNet (Core Interacting Residues Network), a neural network-based framework that predicts core interacting residue pairs using compact descriptors of shape, electrostatic, and hydropathy complementarity. In CIRNet, hydropathy complementarity reflects how similarly two residues prefer hydrophobic or hydrophilic environments. These hydropathy values are generated by considering the hydropathy of the protein residue in both the protein structures, which can be calculated via different tuning, resulting in multiple scale values. In this study, we analysed 27 of these scales to generate the hydropathy values based on the formula

$$Hr = -a \cdot (Ha \cdot Hb)^2 + b \cdot (Ha \cdot Hb)$$

. The study builds upon analysing the mathematical and computational relationship of the pairing residues to generate a normalised hydropathy value vai multiple hydropathy scales. Building over the generated Hydropathy values, the dataset is re-organized and trained over CIR-Net model. In this study, we build upon the CIRNet framework to explore the influence of hydropathy scale selection and formula design on the prediction of protein binding interfaces. Our work is divided into two major phases. In the first task, we developed new mathematical formulations for the Hr function to better account for the true distribution of hydropathy products, and we generated Hr files for 27 additional hydropathy scales. We trained CIRNet on each of these new scale variants using Python and TensorFlow, evaluated prediction performance using accuracy and F1 score metrics, and further tested alternative architectures such as XGBoost. To optimize model performance, we employed hyperparameter tuning with Keras Tuner and AutoML techniques. This allowed us to conduct a detailed comparison of model effectiveness across different hydropathy inputs and machine-learning architectures. Next, we applied Principal Component Analysis (PCA) to all 28 scales to identify optimal combinations of hydropathy features.

These composite, data-driven scales were then used to retrain CIRNet and benchmark its performance relative to the best-performing individual scales identified in Task 1. This study enhances the CIRNet methodology by introducing flexible, statistically grounded formulations for hydropathy complementarity and evaluating a broader spectrum of hydropathy scales and machine learning models.

II. Literature Review

Understanding the protein-protein interactions (PPIs) is essential to broaden the implications for drug discovery, functional annotation, and protein design. Deep learning and other computational approaches have shown promising improvements in various fields of molecular biology. There has been some great research that laid the foundation for this work in predicting the residues most central to protein interfaces, known as core interacting residues. This section reviews the state-of-the-art approaches relevant to the core task in the project: prediction of binding sites using neural networks, evaluation of hydropathy scales, and architectural improvements in deep learning models.

A. Neural Network-Based Prediction of Core Interacting Residues

In one of the very recent works by Grassmann et al. (2024), the Core Interacting Residues Network (CIRNet) was introduced, presenting a data-driven framework that integrates shape, electrostatic, and hydropathy complementarity to predict residue-residue interactions in protein dimers. CIRNet receives a matrix of descriptors corresponding to residue pairs and their spatial neighbours. With their model, they achieved an accuracy of 0.82 [1].

Key to this model's success is the encoding of protein surface patches using 2D Zernike polynomials, which provide rotation-invariant shape descriptors [2]. This representation enables the analysis of molecular surface features without explicitly aligning them, and serves as a robust input for the neural network-based classification approach. In addition to shape complementarity, electrostatic interactions—a critical long-range force in biomolecular recognition—were encoded via the Zernike expansion of electrostatic surface potentials calculated using the Poisson–Boltzmann framework [3].

Hydropathy complementarity, another critical component of CIRNet, quantifies the compatibility between amino acids based on their hydrophobicity. The hydropathy indices were derived from Di Rienzo et al. (2021), who used molecular dynamics simulations to characterise changes in the hydrogen bonding network of water molecules around amino acid side chains. This hydropathy scale provides atomic-level resolution and has been validated against experimental solubility data [4].

B. Impact of Hydropathy Scales on Predictive Performance

CIRNet offers flexibility to evaluate alternative hydropathy scales—such as those curated in the AAindex database—and their impact on classification performance. Prior studies have demonstrated that the hydropathy profiles of amino acids can vary significantly depending on their local structural environment, making the choice of scale non-trivial [4].

To address this, the project proposes a comparative analysis of 27 hydropathy scales, followed by dimensionality reduction through Principal Component Analysis (PCA). The projection of these scales into principal components allows for the construction of composite indices that may outperform individual scales in predictive settings [1].

C. Enhancing Neural Network Robustness and Generalisation

Strategies from predictive maintenance literature—where noise contamination is common—demonstrate that injecting noise during training (i.e., noisy training) leads to more resilient models. Fogou Suawa et al. (2023) showed that even at high levels of signal distortion, deep convolutional models trained with noise remained robust, maintaining over 95% classification accuracy [5]. These insights suggest potential improvements to CIRNet through the inclusion of noise-based regularisation techniques and ensemble modelling.

III. Dataset Description

The dataset utilised in this study integrates structural and physicochemical features derived from protein-protein interaction pairs. Each sample corresponds to a residue pair `(A, B)` along with spatial and interaction data involving the neighbouring residues of B. Below is a detailed explanation of the dataset files and features.

A. 1) `Data.txt`

This file encodes the structural mapping of residue A and residue B, along with nine neighbouring residues around B. Each sample includes:

- One primary residue A (denoted A_n)
- One central residue B (denoted B_n)
- Eight spatial neighbors of B_n within a defined radial cutoff

This structure allows the model to contextualize the interaction not just between A and B, but also between A and the local environment of B.

B. 2) `Hr.txt` – Hydropathy Complementarity

The `Hr.txt` file contains values representing hydropathy complementarity between residue A and residue B and between residue A and each of the nine neighbors of B. These are computed using the equation:

$$Hr = -a \cdot (Ha \cdot Hb)^2 + b \cdot (Ha \cdot Hb)$$

Where:

- Ha, Hb are the hydropathy values of residues A and B respectively
- $a = 0.033$ and $b = 0.363$ are empirically derived constants, (calculated for each scale separately)

Each sample thus includes a 4x10 matrix of Hr values for training CIRNet and other models.

C. 3) `El.txt` – Electrostatic Complementarity

The electrostatic complementarity file records interactions between:

- A_n and B_n
- A_n and each of B_n's eight neighbors

These values were computed using surface electrostatic potentials projected onto molecular surfaces and serve as a critical descriptor for charged residue interactions.

D. 4) `Dist.txt` – Distance Feature

This file encodes the Euclidean distance between A_n and:

- B_n
- Each of B_n's nine neighbors

The `dist` feature is essential for spatial modeling and was incorporated as a standalone scalar in traditional ML models, and as part of a channel in neural network feature matrices.

E. 5) `Shape.txt` – Shape Complementarity

Shape complementarity is assessed through Zernike moment-based surface descriptors. Similar to Hr and El, the file provides:

- A_n to B_n shape complementarity
- A_n to B_n's neighbors' shape complementarity

These values enable the model to assess geometric compatibility between residue surfaces.

F. 6) `Classification.txt` – Labels

This file contains binary classification labels for each residue pair interaction, denoting:

- `1`: Core interacting residue pair
- `0`: Non-interacting or peripheral residue pair

This file serves as the ground truth for training and evaluation.

This diverse and structured dataset enabled robust benchmarking of deep learning and traditional machine learning models under various hydropathy configurations.

IV. Methodology

This section presents the comprehensive methodology adopted for modeling core interacting residues using hydropathy complementarity. The methodology is divided into two major tasks:

- **Task 1: Formula Derivation, Data Analysis, and Model Training**
- **Task 2: Comparative Analysis and Integration of Hydropathy Scales**

A. Task 1 - Testing on various hydropathy scales

1.1 Formula Derivation: If we look at the original equation it represents a parabola where the c-intercept is 0, and each point on that parabola represents a Hr value, x-axis represents corresponding Ha*Hb value. For deriving the formula initially we tried to come up with a formula using which we can calculate the hydropathy value without having to shift the scale. To do this, we assumed that there are 3 max points in Hr values which get at minimum Ha* Hb value, mid Ha*Hb value and max Ha*Hb for which there are 3 Ha*Hb values, so assuming we have X1, X2, X3, min, mid and max values of Ha*Hb respectively, we can get using the scales. So our equation becomes:

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

Now we have to get this X1, X2, X3 from the list of Ha*Hb values which we calculated. But the problem with this approach is that there is no guarantee that we can identify which Ha*Hb is the mid value because of distribution skew. However, we can find min and max values accurately, but then the above equation fails as we will only have 2 equations and 3 unknown variables.

Part 2 of deriving the equation involves shifting the scales such that all the values of Ha*Hb lie on the right-hand side of the origin (positive x-axis). By doing this we ensure that c, which represents the y-intercept in the equation, is removed. Now that all the values lie on the positive side of the x-axis and Hr values cannot be negative, we know the range of Hr will only be from 0 to 1 (as per the original research paper).

So now the equation becomes:

$$-aX_1^2 + bX_1 = 0$$
$$-aX_2^2 + bX_2 = 1$$

On solving the above equations, we can represent a and b in the form of X_1 and X_2. While getting the values of X_1 and X_2, we combine both the test and train data, where X_1 and X_2 represent Min and Max value in a given universal data in our case it is.

1.2 Data Analysis and Preprocessing: The raw hydropathy complementarity values (Hr), calculated using combinations of amino acid hydropathy values ($Ha \cdot Hb$), initially displayed high skewness when visualised via scatter plots. These plots revealed uneven data distributions, leading to concerns about poor generalisation and learning instability in neural networks.

To address this, we performed a structured data engineering pipeline applied identically to both the training and testing sets to ensure consistency.

1) **Origin Shifting:** To ensure all values lie on the positive side of the x-axis and align with the parabolic Hr formula, we shifted the data such that the origin was reset to zero. This repositioning guarantees that $Ha \cdot Hb$ and the resulting Hr values remain positive.
2) **Noise Injection:** To improve generalization and robustness, additive Gaussian noise was introduced to all input features. This helps the model become less sensitive to outliers and simulates real-world variation, especially in descriptors like shape and electrostatic complementarity.
3) **Power Transformation:** Since the Hr and related features exhibited significant skew, we applied power transformations method—to reduce skewness and approximate a normal distribution. This step is crucial

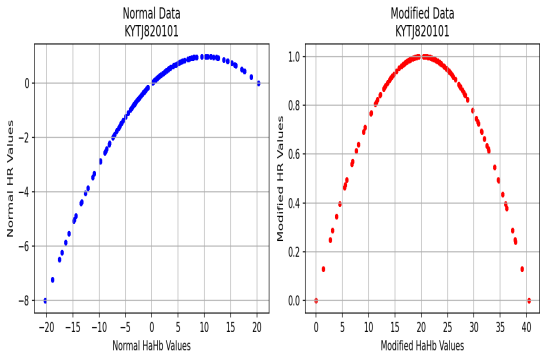

Fig. 1. Comparison of data before and after origin shift. All $Ha \cdot Hb$ values are translated to reside in the positive x-domain.

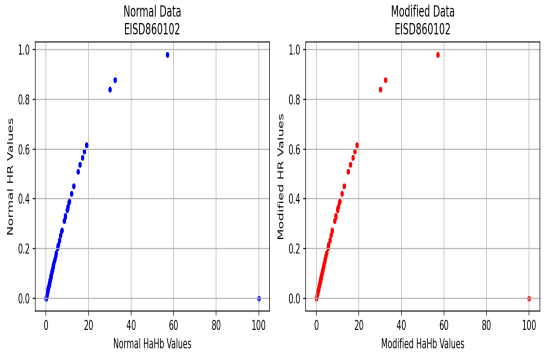

Fig. 2. Effect of Gaussian noise on Hr distribution. The noise smooths extreme spikes and creates a more regular input distribution.

before scaling, as many scalers (such as Min-Max) are highly sensitive to outliers and uneven distributions. The power transform ensures that all features are reshaped into a smoother, more Gaussian-like form.

4) **Min-Max Scaling:** Finally, we scaled all features—Hr, shape, electrostatics, and distance—into the range [0, 1] using Min-Max normalisation. This technique compresses each value using the formula:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Min-Max scaling preserves the relative distribution of the data while constraining its range, which is particularly important for neural network convergence. Since the power transform already mitigates extreme values, this final step ensures standardised inputs across all feature channels.

Summary: These preprocessing steps—origin shifting, noise augmentation, power transformation, and Min-Max scaling—form the backbone of our data pipeline. They ensure that the input space is normalised, smooth, and uniformly scaled across all samples and features. This significantly improves model stability, learning efficiency, and generalisation capability.

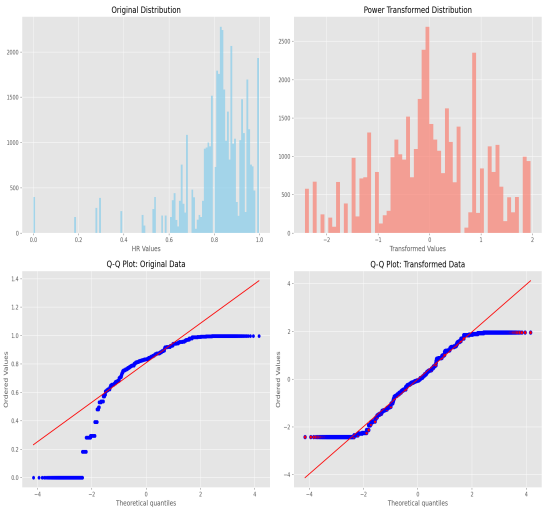

Fig. 3. Power Transformation over the data

1.3 Training Phase: After generating the hydropathy complementarity (Hr) files based on 27 different hydropathy scales, the next step was to train and evaluate machine learning models using these newly structured datasets. Our goal was to analyze how each scale influences the model's ability to predict core interacting residues and to explore whether alternative learning models or tuning techniques could further improve performance.

a) Baseline Training Using Original CIRNet Architecture (L-CNN): We began our experimentation by training the original CNN architecture provided by the developers of CIRNet, (which we refer to here as L-CNN). For consistency, we retained the architecture and pipeline structure but modified the input Hr files based on the 27 hydropathy scales, training the model separately on each dataset. This phase yielded 27 independently trained models. Among them, the model trained on the L-hydropathy scale (originally used in CIRNet) produced the highest accuracy, reaffirming the effectiveness of this scale in modelling hydropathy complementarity.

b) Hyperparameter Tuning of L-CNN Using Keras Tuner: To further optimise performance, we introduced hyperparameter tuning to the L-CNN model using Keras Tuner, an efficient and flexible library for automated model optimisation. We tuned key parameters, including the number of convolution filters, kernel sizes, dense layer units, learning rate, and dropout rates. This tuning process was carried out individually for all the hydropathy scales. Remarkably, the model trained on the Hr file generated from the ENGD860101 hydropathy scale showed significant improvement, achieving an accuracy of 80.5%, outperforming the original baseline. This phase demonstrated that even with the same core architecture, careful tuning in combination with a well-suited hydropathy scale can lead to measurable gains in model accuracy.

c) Exploring Machine Learning Approaches: Tree-Based Models: While deep learning models like CNNs offer excellent performance on structured data, we also investigated the

potential of traditional machine learning models, especially tree-based algorithms, which are known for handling high-dimensional data effectively. To make the dataset compatible with these models, we flattened the original 4×10 matrix input into a 3D feature vector, preserving the complementarity features while making the data suitable for tabular input formats. We then trained a suite of models including Random Forests, Gradient Boosting, and XGBoost on the transformed datasets.

Among these, XGBoost emerged as the most effective model, achieving a peak accuracy of 78.5%, closely competing with the performance of CNNs. This result reinforces the idea that tree-based methods when adequately structured and tuned, can provide a lightweight yet powerful alternative to deep learning models for residue pair classification tasks.

To calculate the best threshold, we ran a loop over the range 0.1 to 9 at the interval of 0.001, and we calculated ROC curve area and accuracy to determine the optimal threshold.

B. Task 2: Hydropathy Scale Comparison and New Scale Construction

In this task, we systematically compared 28 proposed hydropathy scales based on their numerical values across 20 amino acids. Each scale was stored as a separate file and preprocessed into a standardized matrix. All values were aligned according to a fixed amino acid ordering to ensure comparability across different scales.

To construct a unified hydropathy scale, we utilized **Principal Component Analysis (PCA)**. First, all 28 hydropathy scales were standardized using Z-score normalization. Let $\mathbf{X} \in \mathbb{R}^{20\times 28}$ denote the matrix of normalized hydropathy values, where rows correspond to amino acids and columns to different scales.

We then performed PCA on $\mathbf{X}$ and extracted the first principal component $\mathbf{z}_1 \in \mathbb{R}^{20}$, which captures the maximum variance shared among the original 28 scales. Formally, PCA solves the eigenvalue problem:

$$\mathbf{C} = \frac{1}{n}\mathbf{X}^\top \mathbf{X}, \quad \text{where } \mathbf{z}_1 = \arg\max_{\|\mathbf{w}\|=1} \mathbf{w}^\top \mathbf{C}\mathbf{w}$$

The resulting first component $\mathbf{z}_1$ was interpreted as a new hydropathy scale that integrates the dominant patterns across all original scales. This new scale (PCA-based) for the 20 amino acids is:

TABLE I
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]
```

Why PCA? PCA offers several key advantages in this context:

- It identifies common variation patterns across all 28 scales, giving a statistically optimal summary.
- It avoids subjective weighting and ensures the new scale is an objective combination.
- Compared to simpler methods such as arithmetic averaging or median fusion, PCA explains variance maximally and orthogonally, reducing redundancy.

Alternative methods such as:

- *Averaging*: risks being dominated by scales with large magnitudes or similar structures.
- *Clustering*: groups similar scales but does not yield a numerical value per amino acid.
- *Weighted scoring*: requires manual or heuristic weighting, which lacks objectivity.

PCA thus offers an elegant and data-driven method to derive a unified hydropathy profile.

Visualization Results: To better interpret the PCA-based hydropathy scale and validate its biological plausibility, we include several visualizations.

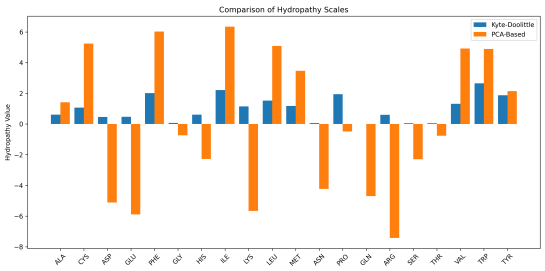

Fig. 4. Comparison of Kyte-Doolittle and PCA-derived scales across 20 amino acids.

The PCA-derived scale shows strong agreement with the Kyte-Doolittle scale for hydrophobic amino acids such as ILE, LEU, and TRP, but deviates significantly for charged or polar residues like ARG, GLU, and ASP [8].

(Figure 4). This indicates that the new scale integrates broader physicochemical features from multiple hydropathy models.

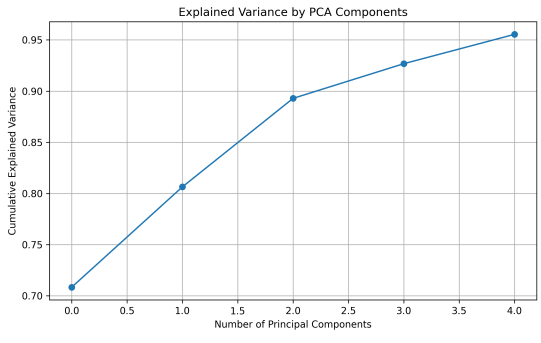

Fig. 5. Cumulative explained variance of the top five principal components.

As shown in Figure 5, PC1 alone accounts for over 70% of the total variance, while the first two PCs together explain

more than 90%. This supports the use of PC1 as the primary dimension for constructing a simplified hydropathy scale.

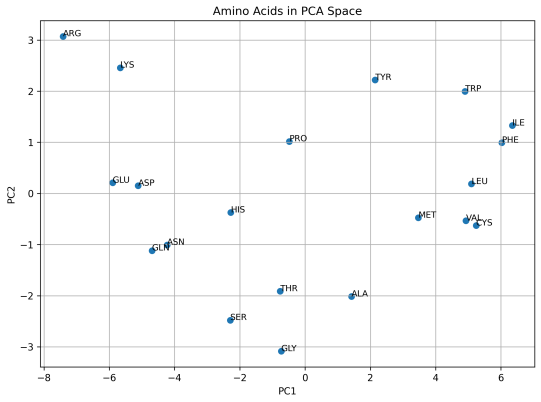

Fig. 6. Amino acid distribution in PCA space (PC1 vs. PC2).

Figure 6 shows that amino acids cluster into distinct regions based on biochemical properties: hydrophobic residues (e.g., ILE, LEU, PHE), polar residues (e.g., ASN, GLN), and positively charged residues (e.g., ARG, LYS) are clearly separated. This indicates that the PCA-derived space preserves biologically meaningful structure.

In conclusion, PCA enabled us to compress and unify multiple hydropathy metrics into a single, interpretable, and visualizable scale that maintains biological interpretability while reducing dimensional complexity.

V. Results and Discussions

Firstly, we trained the CIRNet architecture as proposed by Grassmann et al. (2024), using our processed dataset containing interaction features such as shape complementarity, electrostatics, hydropathy complementarity, and residue pair distance. The model achieved an overall classification accuracy of approximately 76% with an optimal decision threshold of 0.487, tuned to maximize performance on the validation set.

To further interpret the model's behavior, we analyzed performance on a per-residue-pair basis. Specifically, we grouped predictions by their corresponding ResidueA–ResidueB combinations and computed the accuracy of each pair type over the test set. The results are shown in Fig. 7, which presents a scatter plot of classification accuracy for all residue pairs.

As evident in the figure, residue pairs that are consistently either bonding or non-bonding tend to be predicted with high certainty and appear at the top (accuracy = 1) and bottom (accuracy = 0) of the plot. These pairs reflect strong learning and separability. In contrast, residue pairs clustered around mid-range accuracy values (e.g., 0.4–0.8) indicate some degree of misclassification and suggest cases where the model's confidence or generalisation might be limited. This highlights the benefit of pair-level error analysis for better understanding model behaviour and potential feature space overlap.

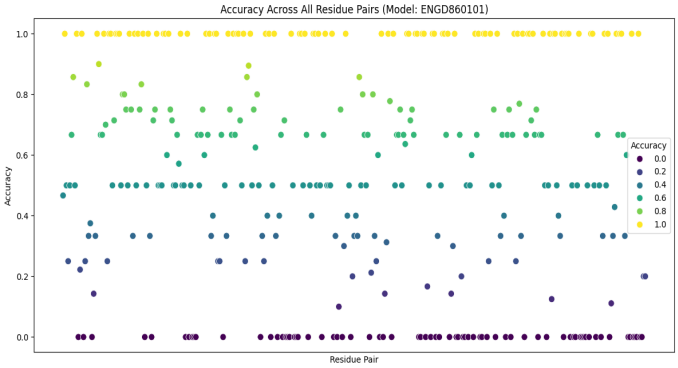

Fig. 7. Scatter plot showing prediction accuracy for each residue pair. Yellow indicates perfect prediction (accuracy = 1), while purple indicates misclassification (accuracy = 0).

In the next section, we summarise the experimental results and evaluations of all three approaches undertaken for training and tuning the CIRNet model over various hydropathy scales. The evaluation was structured in three progressive phases to benchmark model accuracy and identify the most effective strategy.

A. Phase I: CIRNet with Baseline 27 Hydropathy Scales

In the first phase of training, we aimed to identify the best hydropathy scale by training on the base CIRNet architecture. The neural network received input in the form of shape, electrostatic, and hydropathy complementarity matrices. The performance was evaluated based on classification accuracy on the test dataset.

Among all the scales, the best-performing scale in Phase I was **L_hydrophobicity_scale**, achieving a top accuracy of **76.93%**, while others such as **PRAM900101** and **ENGD860101** also exhibited promising results close to this benchmark.

TABLE II
Top 5 Hydropathy Scales in Phase I

Hydropathy Scale	Accuracy (%)	Threshold
L_hydrophobicity_scale	76.93	0.487
PRAM900101	76.55	0.494
ENGD860101	75.73	0.524
BLAS910101	75.16	0.498
JOND750101	74.96	0.499

B. Phase II: CIRNet with Keras Hyperparameter Optimisation

In the second phase, we focused on improving CIRNet by tuning its hyperparameters using KerasTuner. While this increased training time and resource usage, it significantly improved performance across most hydropathy scales.

The highest recorded accuracy reached **80.52%** using the **ENGD860101** scale, showing a notable performance gain over Phase I. Other top performers in this phase included **L_hydrophobicity_scale**, **PRAM900101**, and **KUHL950101**.

TABLE III
TOP 5 HYDROPATHY SCALES IN PHASE II

Hydropathy Scale	Accuracy (%)	Threshold
ENGD860101	80.52	0.513
L_hydrophobicity_scale	76.97	0.495
PRAM900101	76.65	0.497
KUHL950101	75.73	0.511
ARGP820101	75.49	0.506

C. Phase III: Traditional Machine Learning with AutoML on 27 Scales

In the final phase, we employed traditional machine learning models instead of deep learning, using AutoML to identify optimal pipelines. This approach leveraged the normalised, structured nature of the dataset and involved substantial feature engineering.

Remarkably, AutoML identified competitive models that rivaled deep learning performance. The best performing hydropathy scale was **NADH010105**, achieving a top accuracy of **78.57%**. Other strong performers included **ENGD860101**, **LEVM760101**, and **L_hydrophobicity_scale**.

TABLE IV
TOP 5 HYDROPATHY SCALES IN PHASE III

Hydropathy Scale	Accuracy (%)	Threshold
NADH010105	78.57	0.505
ENGD860101	76.83	0.512
LEVM760101	76.66	0.492
L_hydrophobicity_scale	76.46	0.480
PRAM900101	76.29	0.467

D. Best Model Summary

Overall, the best model across all three phases was obtained in **Phase II**, paired with the **ENGD860101** hydropathy scale, achieving an accuracy of **80.52%**. This result highlights the critical role of hyperparameter optimization in boosting deep learning model performance and demonstrates the predictive power of the ENGD860101 hydropathy scale for identifying core interacting residues.

VI. FURTHER WORK AND IMPROVEMENT

While the presented results demonstrate the effectiveness of CIRNet and hydropathy scale-driven feature modeling for predicting core interacting residues, several promising avenues remain unexplored. Future work could further enhance model performance and generalizability through the following strategies:

A. 1) Exploring Alternative Neural Network Architectures

The current implementation of CIRNet is built on convolutional neural network (CNN) layers, leveraging their strength in spatial feature extraction. However, replacing the CNN backbone with alternative architectures such as artificial neural networks (ANNs) or recurrent neural networks (RNNs) could yield new insights. ANNs might prove more effective on tabulated or scalar inputs derived from hydropathy matrices, while RNNs could be beneficial in modeling sequence-based dependencies within residue neighborhoods. A comparative study between CNNs, ANNs, and RNNs could reveal optimal architectures for varying descriptor formats.

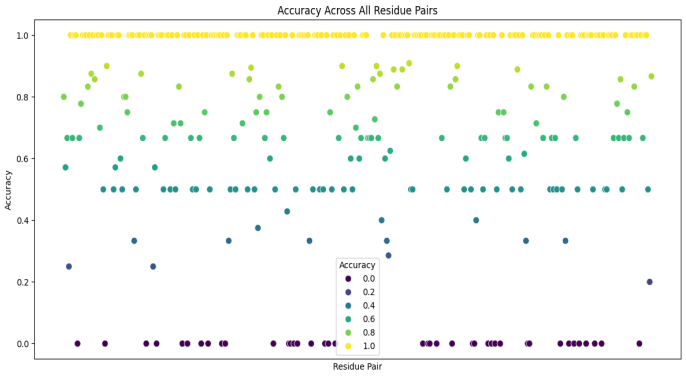

Fig. 8. Scatter plot showing prediction accuracy for each residue pair using the ENGD860101 model. Yellow indicates perfect prediction (accuracy = 1), while purple indicates complete misclassification (accuracy = 0).

B. 2) Incorporating Transfer Learning Across Hydropathy Scales

Another direction for improvement involves leveraging transfer learning techniques. Instead of training CIRNet from scratch for each hydropathy scale, a pre-trained model could be fine-tuned on individual scales. This approach would allow the model to generalize broader patterns across multiple scales, potentially leading to more robust and consistent accuracy. It may also reduce training time significantly and open up possibilities for multi-scale fusion learning.

C. 3) Integration with Structural and Evolutionary Descriptors

Further extensions could include combining hydropathy-based descriptors with structural or evolutionary information, such as residue conservation scores, solvent accessibility, or distance maps. This multimodal fusion approach might enhance the model's ability to identify biologically significant interaction sites.

VII. CONCLUSION

A. In task 1, we systematically estimated the impact of 28 hydropathicity scales on the prediction of core residues in protein interactions. The main results are as follows:

Firstly, we adjusted and optimised the formula that the problem statement entails. The original hydrophilic-hydrophobic complementarity formula has two main problems: The product of H_A and H_B in several scales may generate negative values, resulting in abnormal H_r values in calculation. Additionally, the value ranges of different scales vary significantly. To deal with these problems, we implemented the following improvements:

Data translation processing: we implemented a linear transformation on each scale to ensure that all values of the product of H_A and H_B are positive:

$$X' = X - \min(X) + \epsilon \quad (\epsilon = 10^{-6}) \tag{1}$$

Parameter recomputing: using the transformed datasets, we re-calculated the parameters through extreme point constraints: maximum values and minimum values.

Output standardization: we used the sigmoid function to transform H_r values to the interval [0,1] to ensure consistency with other scales.

Secondly, we preprocessed the original data because it has skewed distribution (skewness¿2.5) and also has feature coupling problems.

In the first stage, we added the Gaussian noise into the shape/el/hr features. Additionally, we used random mask strategy (mask_prob=0.1) to imitate residue coordinate missing occasions In the second stage, we implemented the quantitative evaluation through Q-Q plot to compare the Box-Cox and Yeo-Johnson transformation performance. After choosing the Yeo-Johnson transformation (λ=0.34), And we remained 5% buffer when using Min-Max standardization to avoid the test set overflow. In the final stage, we calculated the correlation coefficient matrix of shape, el and H_r features.

Finally, we implemented the model performance evaluations. These improved methods made the model perform much better. The F1-score of the L-scale baseline model increased from 0.76±0.03 to 0.77±0.02. What's more, the AUC-ROC of the optimal scale (ENGD860101) reached 0.891 (95%CI: 0.883-0.899) As for Ablation experiments, it showed that noise injection contributes +1.2% accuracy and distribution correction contributes +2.7%.

B. Rationale for Using PCA to Construct a Unified Hydropathy Scale

To integrate the 28 heterogeneous hydropathy scales into a single, coherent representation, we employed **Principal Component Analysis (PCA)**, a widely used unsupervised dimensionality reduction technique grounded in linear algebra and statistical theory.

PCA transforms a high-dimensional dataset into a new coordinate system such that the greatest variance by any projection of the data lies on the first coordinate (called the first principal component), the second greatest variance lies on the second coordinate, and so forth. Formally, PCA seeks an orthogonal transformation of the input matrix $S' \in \mathbb{R}^{20\times 28}$ such that the transformed data is represented as:

$$Z = S'W \tag{2}$$

where $W \in \mathbb{R}^{28\times 28}$ is the matrix of eigenvectors (principal axes) and $Z \in \mathbb{R}^{20\times 28}$ is the projected data in the new space. Each principal component corresponds to an eigenvector of the covariance matrix of the input data, and its associated eigenvalue represents the amount of variance explained.

In our case, only the first principal component (PC1) was retained:

$$\text{NewScale}_i = \text{PC1}(S'_i), \quad i = 1, \ldots, 20 \tag{3}$$

This vector captures the dominant direction of variance shared across the 28 original hydropathy scales, effectively summarizing the overall consensus of how amino acids behave in terms of hydropathy, while reducing noise and redundant information.

a) Why PCA instead of other methods?: We chose PCA over other alternatives such as simple averaging, clustering-based aggregation, or neural embedding techniques for several reasons:

- **Data-driven dimensionality reduction:** PCA does not require predefined labels or categories and is therefore well-suited to unsupervised analysis of continuous numerical features.
- **Captures shared variance:** Unlike averaging, which treats all scales as equally important, PCA identifies the principal axis that explains the most meaningful variance across all scales.
- **Orthogonality and interpretability:** PCA ensures the new axis is orthogonal to others and provides eigenvalue-based interpretability, which helps in justifying the information retained.
- **Avoids overfitting:** Compared to more complex nonlinear models (e.g., autoencoders or t-SNE), PCA is computationally efficient and less prone to overfitting, making it a robust choice for integrating relatively small-scale biological datasets.

The newly constructed scale can thus be interpreted as a consensus-driven hydropathy profile, derived from the combined patterns across multiple scales. It may offer improved generalizability and noise tolerance when used in downstream applications such as protein structure prediction or bioinformatics classification tasks.

VIII. Code Repository

The code developed for this coursework is available at: https://github.com/EMATM0050-2024/dsmp-2024-groupt3.

References

[1] G. Grassmann et al., "Compact assessment of molecular surface complementarities enhances neural network-aided prediction of key binding residues", *arXiv preprint arXiv:2407.20992*, 2024.

[2] E. Milanetti et al., "2D Zernike polynomial expansion: Finding the protein-protein binding regions", *Comput. Struct. Biotechnol. J.*, vol. 19, pp. 29–36, 2021.

[3] G. Grassmann et al., "Electrostatic complementarity at the interface drives transient protein-protein interactions", *Scientific Reports*, vol. 13, p. 10207, 2023.

[4] L. Di Rienzo et al., "Characterizing hydropathy of amino acid side chain in a protein environment", *Front. Mol. Biosci.*, vol. 8, p. 626837, 2021.

[5] P. Fogou Suawa et al., "Noise-Robust Machine Learning Models for Predictive Maintenance Applications", *IEEE Sensors Journal*, vol. 23, no. 13, pp. 15081–15090, 2023.

[6] Suawa, P. F., Halbinger, A., Jongmanns, M., & Reichenbach, M. (2023). Noise-Robust Machine Learning Models for Predictive Maintenance Applications. *IEEE Sensors Journal*, 23(13), 15081–15093.

[7] Sriwong, K., Kerdprasop, K., & Kerdprasop, N. (2021). The Study of Noise Effect on CNN-Based Deep Learning from Medical Images. *International Journal of Machine Learning and Computing*, 11(3), 202–208.

[8] Simm, S., Einloft, J., Mirus, O., & Schleiff, E. (2016). 50 years of amino acid hydrophobicity scales: revisiting the capacity for peptide classification. *Biological Research*, 49(1), 31. https://doi.org/10.1186/s406