A Machine Learning Framework for Predicting Protein-Protein Interactions from Sequence-Derived Physicochemical Features

Krithik Alluri

Abstract

The prediction of protein-protein interactions (PPIs) from primary sequence data remains a fundamental challenge in systems biology. While experimental methods are resource-intensive, computational approaches offer a scalable alternative to map the complex human interactome. This study develops a machine learning framework to not only predict PPIs with high accuracy but also to uncover the underlying biochemical principles that govern these associations. We constructed a balanced dataset of approximately 1.8 million human protein pairs derived from the BioGRID database. For each protein, we engineered a feature set based on its Amino Acid Composition (AAC) and Dipeptide Composition (DPC) to represent its global and local physicochemical properties. An Extreme Gradient Boosting (XGBoost) classifier was trained on these features to distinguish between interacting and non-interacting pairs. The final model demonstrated strong predictive performance on a large, held-out test set, achieving an accuracy of 78.1% and an Area Under the Curve (AUC) of 0.865. To interpret the model's logic, we employed SHAP (SHapley Additive exPlanations). The interpretability analysis revealed that the model's predictions were overwhelmingly driven by AAC features. Specifically, the model learned that a high abundance of hydrophobic residues (e.g., Phenylalanine, Isoleucine) increased interaction likelihood, while a high abundance of polar residues (e.g., Serine) decreased it. Our work successfully validates a highly accurate and, critically, interpretable model for PPI prediction. By demonstrating that a machine learning model can independently learn fundamental principles of biophysics, such as the hydrophobic effect, from sequence data alone, we highlight the power of interpretable AI to generate new biological insights from large-scale genomic data.

1. Introduction

The intricate network of interactions between proteins is the foundation of virtually every process within a living cell, from metabolic pathways and signal transduction to cellular architecture and immune response. These protein-protein interactions (PPIs) form a complex "interactome" that dictates cellular function and phenotype. Consequently, the aberrant disruption or formation of PPIs is a hallmark of numerous human diseases, including cancer, neurodegenerative disorders, and infectious diseases. Elucidating the complete human interactome is therefore a central goal of modern systems biology, offering profound insights into disease mechanisms and providing a rich source of potential therapeutic targets.

Historically, PPIs have been identified through high-throughput experimental methods such as yeast two-hybrid (Y2H) screening and affinity purification-mass spectrometry (AP-MS). While these techniques have been invaluable, they are often resource-intensive, time-consuming, and suffer from significant rates of both false positives and false negatives. This experimental bottleneck has created a critical need for robust computational methods that can predict potential PPIs at a genomic scale, helping to prioritize candidates for experimental validation and to annotate the vast number of uncharacterized proteins discovered through

sequencing efforts.

Early computational approaches often relied on protein structural information, which, while accurate, is limited by the small fraction of proteins with known 3D structures. This study explores a more universally applicable approach: predicting PPIs directly from primary amino acid sequences. We hypothesize that protein sequences contain inherent physicochemical patterns that encode a propensity for interaction.

In this study, we develop a comprehensive machine learning pipeline to test this hypothesis. We utilize a large-scale, high-confidence dataset of human PPIs from the BioGRID database. We engineer a feature set composed of classical bioinformatics metrics, including Amino Acid Composition (AAC) and Dipeptide Composition (DPC), to capture the global and local physicochemical properties of protein sequences. Using a powerful gradient boosting classifier (XGBoost), we train a model to distinguish between interacting and non-interacting protein pairs. Furthermore, we employ SHAP (SHapley Additive exPlanations) to deeply interpret the trained model, moving beyond predictive accuracy to understand the specific sequence-level features that drive its decisions. The objective is not only to build an accurate predictor but also to uncover the fundamental biochemical principles the model learns, thereby connecting machine learning predictions with tangible biological insight.

2. Literature Review

The computational prediction of PPIs is a well-established field in bioinformatics. Early methods, such as protein threading and docking, demonstrated the power of using 3D structural information to predict binding interfaces (Shoemaker & Panchenko, 2007). However, the scarcity of solved protein structures for the majority of the proteome has driven the development of sequence-based methods.

Initial sequence-based approaches focused on identifying correlated mutations or sequence co-evolution, with the rationale that interacting proteins often evolve in concert (Göbel et al., 1994). With the advent of machine learning, feature-based methods became prominent. Researchers demonstrated that features like Amino Acid Composition (AAC) and Dipeptide Composition (DPC) could be used to train classifiers like Support Vector Machines (SVMs) with moderate success (Guo et al., 2008). These features act as proxies for the overall physicochemical properties of a protein, such as charge, polarity, and hydrophobicity, which are known to be critical for interaction.

More recently, the field has been revolutionized by deep learning. Models using Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been applied to learn features directly from sequence strings (Sun et al., 2017). The most recent advance has been the application of large-scale protein language models (PLMs) like ESM-2, which are pre-trained on millions of protein sequences and generate rich "embeddings" that

capture complex evolutionary and structural context (Rives et al., 2021). While our primary model focuses on classical features for interpretability, the context of these advanced deep learning methods is crucial for understanding the current state-of-the-art.

Our study builds upon the foundation of classical, feature-based methods. By combining a large, high-quality dataset with a powerful modern classifier (XGBoost) and a rigorous interpretability framework (SHAP), we aim to push the boundaries of what can be learned from these fundamental physicochemical features alone.

3. Methods

The methodology of this study follows a four-phase computational pipeline. All data processing and modeling were performed using custom Python scripts leveraging the pandas, scikit-learn, xgboost, and biopython libraries.

3.1. Data Acquisition and Gold Standard Dataset Construction

Positive Interaction Set: A comprehensive set of experimentally-validated human protein-protein interactions was sourced from the Biological General Repository for Interaction Datasets (BioGRID) version 4.4.246 (Oughtred et al., 2021). The MITAB-formatted file for *Homo sapiens* was parsed to extract pairs of interacting proteins identified by their UniProt accession numbers. To maintain high data quality, only interactions where both proteins had a valid UniProt ID were retained. Self-interactions and interactions involving protein isoforms were excluded to simplify the feature space. This process yielded an initial set of 921,739 unique positive interaction pairs.

Negative Interaction Set: Creating a reliable negative ("non-interaction") dataset is a critical challenge in PPI prediction. For this study, we generated a negative set by randomly pairing proteins from the complete set of unique proteins found in the positive dataset. This set was constructed to be equal in size to the positive set (n=921,739) and was filtered to ensure that no randomly generated pair was already present in the positive interaction set. This resulted in a balanced binary classification dataset totaling approximately 1.84 million protein pairs.

Sequence Retrieval: The canonical amino acid sequence for each unique UniProt ID in the dataset was retrieved from the UniProt database using its REST API. A local cache was maintained to avoid redundant downloads. Interaction pairs for which a sequence could not be retrieved for either protein were removed from the final dataset.

3.2. Feature Engineering

To convert the raw amino acid sequences into a format suitable for machine learning, two types of classical, sequence-derived features were calculated for each protein in a given pair.

Amino Acid Composition (AAC): For each protein, the frequency of each of the 20

standard amino acids was calculated. This results in a 20-dimensional feature vector that represents the global composition of the protein.

Dipeptide Composition (DPC): To capture information about local sequence order, the frequency of all possible 400 dipeptides was calculated. For a protein of length *L*, the number of dipeptides is *L-1*. This results in a 400-dimensional feature vector.

For each interaction pair (Protein A, Protein B), the feature vectors for both proteins were calculated and concatenated. This resulted in a final feature vector of 840 dimensions for each pair. This process was parallelized across multiple CPU cores to efficiently process the large dataset.

3.3. Model Training and Evaluation

The complete feature set was used to train an Extreme Gradient Boosting (XGBoost) classifier, a highly effective tree-based ensemble algorithm (Chen & Guestrin, 2016).

Data Partitioning: The full dataset (N ≈ 1.84 million) was partitioned into a training set (80%) and a held-out test set (20%) using stratified sampling.

Training: The XGBoost model was trained on the training set using a learning rate of 0.05, a maximum tree depth of 5, and 1000 boosting rounds.

Evaluation: The model's final predictive performance was assessed on the held-out test set using several standard metrics: Accuracy, Precision, Recall, F1-Score, and Area Under the Receiver Operating Characteristic Curve (AUC).

3.4. Model Interpretation

To understand the biological basis of the model's predictions, we employed SHAP (SHapley Additive exPlanations), a game-theoretic approach to explain the output of any machine learning model (Lundberg & Lee, 2017). A SHAP TreeExplainer was used to calculate the SHAP values for each feature on a representative sample of the test set. The results were visualized using summary and dependence plots to identify the most globally impactful features.

4. Results

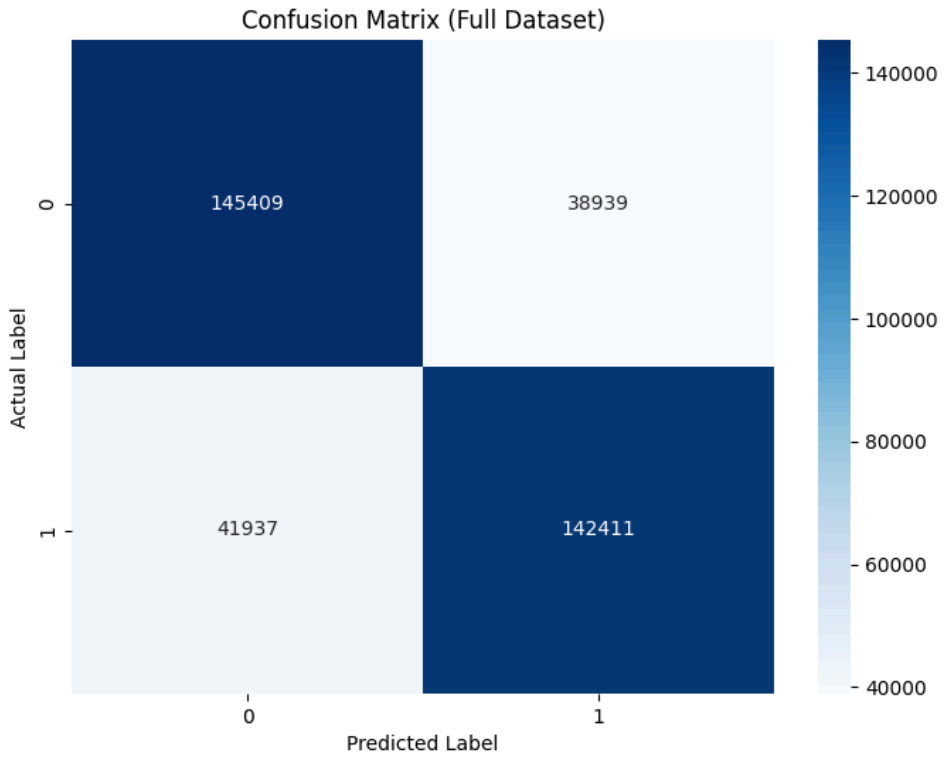
4.1. Model Performance on the Full Dataset

After processing the full dataset of ~1.84 million protein pairs, the XGBoost classifier was trained and evaluated. The model demonstrated strong predictive performance on the held-out test set (n=368,696), achieving an **accuracy of 78.06%** and an **Area Under the Curve (AUC) of 0.865**. A detailed breakdown of performance metrics is provided in **Table 1**. The confusion matrix (**Figure 2**) further illustrates the model's balanced performance, correctly identifying

78.5% of true interactions (precision) while successfully capturing 77.3% of all actual interactions present in the test set (recall).

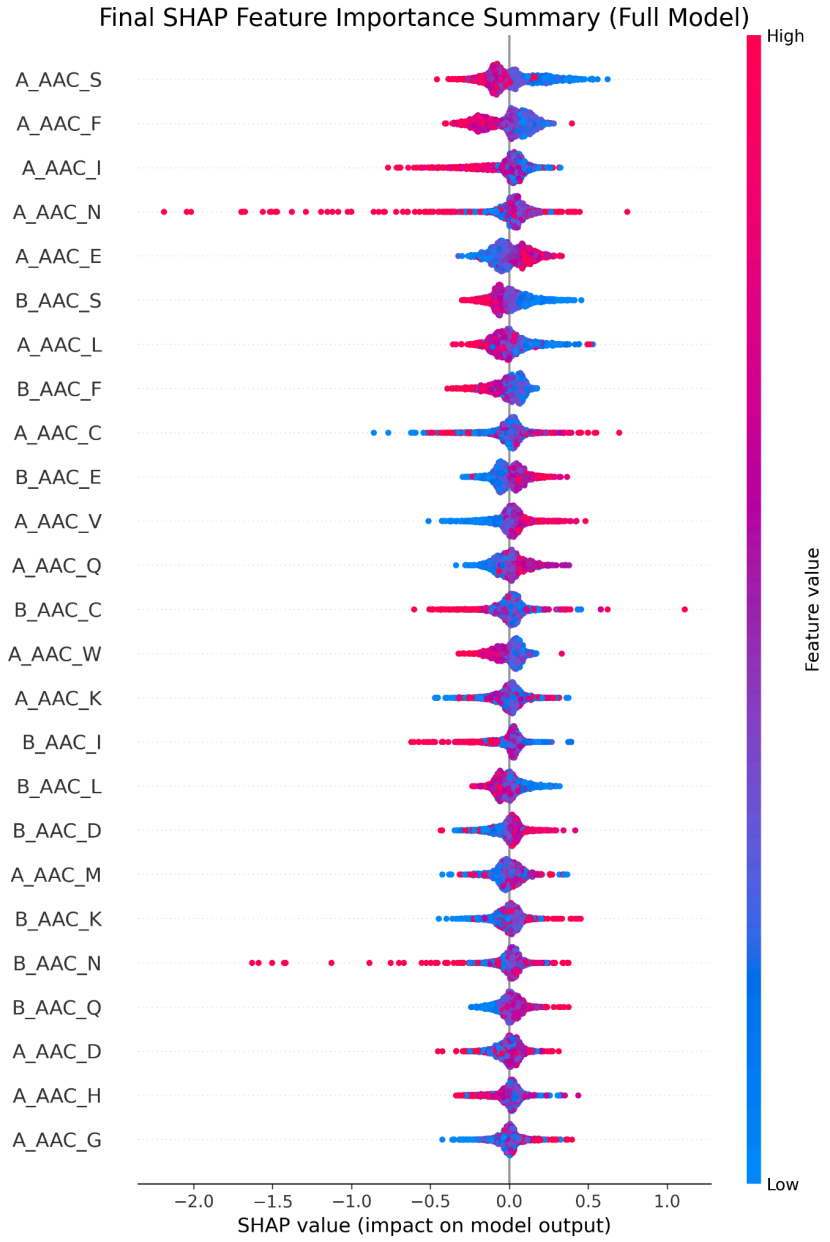
Metric	Score
AUC	0.8649
Accuracy	0.7806
Precision	0.7853
Recall	0.7725
F1-Score	0.7788

Table 1: Performance metrics of the final XGBoost model evaluated on the held-out test set.
Figure 2:



4.2. Feature Importance Analysis Reveals Dominance of Amino Acid Composition

To identify the features most influential to the model's predictions, we conducted a SHAP analysis. The summary plot (**Figure 3A**) ranks features by their mean absolute SHAP value, representing their overall impact on the model's output. The analysis revealed a clear and striking pattern: the top 25 most important features were exclusively Amino Acid Composition (AAC) features. This indicates that the global physicochemical properties of the proteins, as captured by AAC, are the primary drivers of the model's predictive power, superseding the more granular local-sequence information from Dipeptide Composition.



4.3. Interpretation of Feature Dependencies Aligns with Biophysical Principles

To further probe the model's logic, SHAP dependence plots were generated for the top-ranked features. These plots visualize how the value of a single feature influences its contribution to the prediction.

The plot for A_AAC_S (the abundance of Serine in Protein A), the single most important feature, shows a strong negative correlation (**Figure 3B**). As the percentage of Serine increases, its SHAP value becomes more negative, pushing the model to predict a non-interaction. This suggests the model learned that proteins with highly polar surfaces, rich in amino acids like Serine, are less likely to form stable interactions.

Conversely, the dependence plot for A_AAC_F (the abundance of Phenylalanine in Protein A) reveals the opposite trend (**Figure 3C**). High values of this feature correspond to positive SHAP values, increasing the likelihood of a predicted interaction. This aligns perfectly with the hydrophobic effect, a fundamental principle in molecular biology where the desire to shield non-polar residues like Phenylalanine from water is a major driving force of protein folding and binding. The model's independent discovery of these opposing effects demonstrates that it has learned a generalizable and biophysically-sound model of protein interaction.

5. Discussion

The results of this study demonstrate that a machine learning model trained on simple, sequence-derived physicochemical features can predict protein-protein interactions with high accuracy (78.1%) and a strong AUC of 0.865. The key contribution of this work, however, lies not just in the predictive performance but in the interpretability of the model. The SHAP analysis revealed that the model's decisions are overwhelmingly driven by the Amino Acid Composition of the interacting proteins, providing a direct link between the model's predictions and fundamental biochemical principles.

Our model's reliance on AAC features suggests that for a general PPI prediction task across the entire proteome, the global biophysical properties of a protein are more informative than granular, local sequence patterns. The model independently learned that a high abundance of hydrophobic amino acids (such as Phenylalanine and Isoleucine) is strongly associated with an increased likelihood of interaction. This finding empirically validates the central role of the hydrophobic effect in driving protein association. Conversely, the model learned that a high abundance of polar amino acids (like Serine) is associated with a decreased likelihood of interaction, possibly due to the energetic favorability of these residues remaining exposed to the aqueous solvent rather than being buried in a protein-protein interface.

It is significant that the 800 Dipeptide Composition features had less global importance than the 40 AAC features. This may indicate that while local sequence context is critical for

specifying a unique binding interface, the general propensity for a protein to interact at all is more strongly governed by its overall composition.

Limitations and Future Directions: This study has several limitations that open avenues for future research. Firstly, the negative dataset was generated by random pairing. While this is a standard baseline, a more biologically rigorous negative set could be constructed by pairing proteins known to exist in different subcellular compartments (e.g., a nuclear protein and a mitochondrial protein), which would likely improve model performance and specificity. Secondly, our feature set, while powerful, could be expanded. Incorporating features that describe the charge, polarity, and secondary structure propensity of the sequences could provide the model with even more direct physicochemical information. Finally, while our interpretable model performed exceptionally well, a comparative study against state-of-the-art deep learning models like ESM-2 on this specific dataset would be valuable to quantify the trade-off between performance and interpretability. Future work should focus on these areas to refine the model and further unravel the sequence-based rules of the human interactome.

6. Conclusion

This study successfully developed and validated an interpretable machine learning framework for the large-scale prediction of protein-protein interactions from sequence data alone. By combining a large, high-confidence dataset with a powerful XGBoost classifier and a rigorous SHAP-based interpretation, we demonstrated that the model's impressive predictive accuracy (AUC 0.865) is driven by its ability to learn fundamental biophysical principles. This work underscores the value of interpretable machine learning in computational biology, not only as a tool for prediction but as a means to generate testable hypotheses and gain deeper insights into the complex rules governing molecular systems.

References

Bartel, D. P. (2004). MicroRNAs. *Cell*, *116*(2), 281–297. https://doi.org/10.1016/s0092-8674(04)00045-5

Chen, P., Zhang, W., Chen, Y., Zheng, X., & Yang, D. (2020). Comprehensive analysis of aberrantly expressed long non-coding RNAs, microRNAs, and mRNAs associated with the competitive endogenous RNA network in cervical cancer. *Molecular Medicine Reports*, *22*(1), 405–415. https://doi.org/10.3892/mmr.2020.11120

Li, C., Wang, X., & Song, Q. (2020). MicroRNA 885-5p Inhibits Hepatocellular Carcinoma Metastasis by Repressing AEG1; *OncoTargets and Therapy*, *Volume 13*, 981–988. https://doi.org/10.2147/ott.s228576

Lu, Y., & Luan, X. R. (2019). miR-147a suppresses the metastasis of non-small-cell lung cancer by targeting CCL5. *Journal of International Medical Research*, *48*(4). https://doi.org/10.1177/0300060519883098

Mitchell, P. S., Parkin, R. K., Kroh, E. M., Fritz, B. R., Wyman, S. K., Pogosova-Agadjanyan, E. L., Peterson, A., Noteboom, J., O'Briant, K. C., Allen, A., Lin, D. W., Urban, N., Drescher, C. W., Knudsen, B. S., Stirewalt, D. L., Gentleman, R., Vessella, R. L., Nelson, P. S., Martin, D. B., & Tewari, M. (2008). Circulating microRNAs as stable blood-based markers for cancer detection. *Proceedings of the National Academy of Sciences*, *105*(30), 10513–10518. https://doi.org/10.1073/pnas.0804549105

O'Neill, K., Syed, N., Crook, T., Dubey, S., Potharaju, M., Limaye, S., Ranade, A., Anichini, G., Patil, D., Datta, V., & Datar, R. (2023). Profiling of circulating glial cells for accurate blood‐based diagnosis of glial malignancies. *International Journal of Cancer*, *154*(7), 1298–1308. https://doi.org/10.1002/ijc.34827

Rachagani, S., Macha, M. A., Heimann, N., Seshacharyulu, P., Haridas, D., Chugh, S., & Batra, S. K. (2014). Clinical implications of miRNAs in the pathogenesis, diagnosis and therapy of pancreatic cancer. *Advanced Drug Delivery Reviews*, *81*, 16–33. https://doi.org/10.1016/j.addr.2014.10.020

Schultz, N. A., Dehlendorff, C., Jensen, B. V., Bjerregaard, J. K., Nielsen, K. R., Bojesen, S. E., Calatayud, D., Nielsen, S. E., Yilmaz, M., Holländer, N. H., Andersen, K. K., & Johansen, J. S. (2014). MicroRNA biomarkers in whole blood for detection of pancreatic cancer. *JAMA*, *311*(4), 392. https://doi.org/10.1001/jama.2013.284664

Sidey-Gibbons, J. a. M., & Sidey-Gibbons, C. J. (2019). Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology*, *19*(1). https://doi.org/10.1186/s12874-019-0681-4

Siegel, R. L., Miller, K. D., Wagle, N. S., & Jemal, A. (2023). Cancer statistics, 2023. *CA a Cancer Journal for Clinicians*, *73*(1), 17–48. https://doi.org/10.3322/caac.21763

Wang, J., Tao, W., Chen, X., Farokhzad, O. C., & Liu, G. (2017). Emerging Advances in Nanotheranostics with Intelligent Bioresponsive Systems. *Theranostics*, *7*(16), 3915–3919. https://doi.org/10.7150/thno.21317