

Topological Link Prediction in the Human Protein-Protein Interactome: A Comprehensive Comparative Study of Machine Learning Models on STRING v12.0

LitLit AI^{1,*}

¹LitLit AI, Research Division
contact@litlit.ai

August 4, 2026

Abstract

Background: Protein-protein interactions (PPIs) form the molecular basis of virtually all cellular processes, and their comprehensive mapping remains a fundamental challenge in systems biology. Computational link prediction leveraging network topology offers a scalable and cost-effective approach to prioritize candidate PPIs for experimental validation. However, the comparative performance of different machine learning paradigms on large-scale, high-confidence PPI networks, particularly with respect to feature importance interpretation and degree-stratified performance, remains incompletely characterized.

Methods: We constructed a human PPI network from the STRING v12.0 database comprising 16,201 proteins and 236,930 high-confidence interactions (combined score ≥ 0.7). For each candidate protein pair, we computed six topological features — common neighbors, Jaccard coefficient, Adamic-Adar index, preferential attachment score, resource allocation index, and shortest path length — using only the training subgraph to prevent data leakage. We trained and rigorously evaluated three models: L2-regularized logistic regression, random forest (300 estimators, balanced class weights), and XGBoost (400 trees, max depth 6, learning rate 0.05, with subsampling and regularization). Performance was assessed via AUROC, AUPRC, bootstrap confidence intervals, permutation feature importance, and degree-stratified analysis.

Results: XGBoost achieved the highest overall performance (AUROC = 0.9750, AUPRC = 0.9821), closely followed by logistic regression (AUROC = 0.9723, AUPRC = 0.9804), while random forest underperformed substantially (AUROC = 0.7344, AUPRC = 0.7323). Permutation importance analysis revealed that the resource allocation index was the dominant feature for logistic regression (Δ AUROC = 0.199) and XGBoost (Δ AUROC = 0.249), while the random forest relied almost exclusively on shortest path length (Δ AUROC = 0.333), with four out of six features exhibiting negative permutation importance — a pathological artifact of over-reliance on a single feature. Degree-stratified analysis showed that both logistic regression and XGBoost maintained robust performance on low-degree protein pairs (AUROC = 0.912–0.917), whereas the random forest degraded significantly (AUROC = 0.758). Feature correlation analysis revealed that the resource allocation index and

Adamic-Adar index were highly correlated ($r = 0.97$), yet resource allocation consistently outperformed Adamic-Adar across all models.

Conclusions: The resource allocation index and shortest path length are the most informative topological features for PPI link prediction in the human interactome. Logistic regression and XGBoost provide complementary strengths — interpretability versus marginal accuracy gain — and both substantially outperform random forest, which exhibits a pathological over-reliance on shortest path length. These findings establish a practical benchmark for topology-based PPI prediction and highlight the critical importance of model selection in biological network inference. The near-identical performance of linear and non-linear models suggests that the topological feature space is approximately linearly separable with respect to interaction propensity, with important implications for model design in computational interactomics.

Keywords

Link prediction; protein-protein interaction network; STRING database; topological features; machine learning; XGBoost; logistic regression; resource allocation index; network biology; systems biology

1 Introduction

1.1 The Protein-Protein Interactome: A Fundamental Biological Network

Protein-protein interactions (PPIs) orchestrate virtually all cellular processes, from signal transduction and metabolic regulation to transcriptional control, immune response, and programmed cell death [1, 2]. The comprehensive mapping of the human interactome — the complete set of physical and functional protein interactions — is therefore a fundamental goal of systems biology and molecular medicine. A complete understanding of the interactome would provide a mechanistic framework for interpreting genomic data, identifying disease genes, and discovering therapeutic targets [3, 4].

High-throughput experimental techniques have dramatically expanded our knowledge of PPI networks. Yeast two-hybrid (Y2H) screens [5, 6, 7], affinity purification coupled with mass spectrometry (AP-MS) [8, 9, 10], and co-immunoprecipitation assays have collectively identified tens of thousands of human PPIs. However, these methods remain costly, labor-intensive, and subject to substantial false-positive and false-negative rates [1, 11]. Comparative assessments have shown that the overlap between different experimental methods is often modest, with each technique capturing distinct subsets of the interactome [1, 12]. This incomplete and noisy coverage motivates the development of computational approaches to predict and prioritize PPIs.

1.2 The STRING Database: A Comprehensive PPI Resource

The STRING database (Search Tool for the Retrieval of Interacting Genes/Proteins) addresses key limitations of individual experimental methods by integrating multiple evidence channels into a unified, confidence-scored PPI resource [13, 14, 15, 16, 17]. STRING aggregates evidence from seven distinct sources: (1) genomic context (gene fusion, conserved neighborhood, gene

co-occurrence), (2) high-throughput experimental screens, (3) curated pathway databases, (4) automated text mining of the scientific literature, (5) co-expression analysis, (6) protein homology, and (7) gene co-occurrence across genomes. Each interaction is assigned a combined confidence score ranging from 0 to 1000, integrating the strength of evidence across all channels.

The latest version, STRING v12.0, covers over 19,000 human proteins with millions of predicted interactions at various confidence thresholds [17]. This comprehensive coverage, combined with the confidence scoring system, makes STRING an ideal resource for computational interactome analysis. The database also provides species-specific networks, tissue-specific expression data, and functional enrichment analyses, enabling context-dependent network biology. STRING has been widely used in diverse applications, including disease gene prioritization, drug target identification, and functional annotation of uncharacterized proteins [16].

1.3 The Link Prediction Problem in Biological Networks

The link prediction problem — inferring unobserved or future interactions from the topology of an observed network — has emerged as a powerful computational framework for prioritizing candidate PPIs [18, 19, 20]. The core premise is that the structural properties of known interaction networks encode information about which protein pairs are likely to interact. This approach is particularly attractive because it requires only the network topology, not additional biological data, making it broadly applicable across species and conditions.

Formally, given an observed network $G(V, E)$ with node set V and edge set E , the link prediction task is to identify which of the unobserved node pairs $(u, v) \notin E$ are most likely to be true missing edges. The fundamental assumption is that the probability of an edge existing between two nodes is correlated with some measure of topological similarity between them. Link prediction methods have been successfully applied to diverse network types, including social networks [18], collaboration networks [21], biological networks [22, 23], and criminal networks [24].

The mathematical foundations of link prediction draw from several fields. The concept of structural equivalence — the idea that nodes sharing many neighbors are functionally similar — dates back to early social network analysis [25, 26]. More recent work has formalized this intuition through a variety of similarity measures, each capturing different aspects of network topology [20, 27].

1.4 Topological Features for Link Prediction

Numerous topological similarity measures have been developed for link prediction, which can be broadly categorized into neighborhood-based, path-based, and hybrid approaches.

1.4.1 Neighborhood-Based Indices

Neighborhood-based indices quantify the overlap between the local neighborhoods of two nodes. The simplest and most intuitive measure is the common neighbors index, which counts the number of nodes that are adjacent to both u and v [18]:

$$\text{CN}(u, v) = |\Gamma(u) \cap \Gamma(v)|$$

The Jaccard coefficient normalizes the common neighbor count by the size of the union of the neighborhoods, providing a measure of relative overlap that is less influenced by node degree [28]:

$$\text{Jaccard}(u, v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|}$$

The Adamic-Adar index, originally developed for web page similarity, weights shared neighbors by the inverse logarithm of their degree [29]:

$$\text{AA}(u, v) = \sum_{w \in \Gamma(u) \cap \Gamma(v)} \frac{1}{\log |\Gamma(w)|}$$

The resource allocation index, inspired by a physical process of resource distribution, penalizes high-degree shared neighbors more aggressively than the Adamic-Adar index [30]:

$$\text{RA}(u, v) = \sum_{w \in \Gamma(u) \cap \Gamma(v)} \frac{1}{|\Gamma(w)|}$$

The preferential attachment score, rooted in the scale-free network model [31], posits that high-degree nodes are more likely to form new connections:

$$\text{PA}(u, v) = |\Gamma(u)| \times |\Gamma(v)|$$

1.4.2 Path-Based Indices

Path-based measures capture global or quasi-global connectivity patterns. The shortest path length between two nodes, which can be computed via breadth-first search, captures the overall distance in the network [21]. The local path index extends this by considering paths of varying lengths [32]. The Katz index incorporates all paths between two nodes, weighted by an attenuation factor [33]. The rooted PageRank and SimRank algorithms [34] use random walk-based approaches to compute similarity based on the network’s global structure.

1.4.3 Hybrid and Ensemble Approaches

Recent work has shown that combining multiple topological features through machine learning can substantially improve link prediction performance compared to using any single feature [35, 36]. The rationale is that different features capture complementary aspects of network topology, and a learned weighting can optimally integrate these signals. This approach has been particularly successful in biological networks, where the topological signal is often distributed across multiple features [37, 22].

1.5 Machine Learning for PPI Link Prediction

Machine learning methods offer a natural framework for integrating multiple topological features into a unified predictive model. Logistic regression provides interpretable coefficients linking each feature to interaction probability, making it particularly attractive in biological applications where understanding the contribution of individual features is important for hypothesis generation [38]. The linear nature of logistic regression also makes it robust to overfitting when the number of features is small relative to the sample size.

Tree-based ensemble methods, including random forest [37, 36] and gradient boosting machines (XGBoost, LightGBM) [39, 40], can capture non-linear relationships and feature interactions. Random forest builds an ensemble of deep decision trees on bootstrapped samples, averaging their predictions to reduce variance. XGBoost uses gradient-boosted trees with regularization, handling missing values and providing built-in feature importance measures. These methods have shown strong performance across a wide range of prediction tasks but their behavior on PPI network data, particularly with respect to feature importance interpretation, requires careful investigation.

Despite the widespread use of these methods, rigorous comparative evaluations on large-scale, high-confidence PPI networks remain limited. In particular, three key questions have not been systematically addressed: (1) How do linear and non-linear models compare when the feature set is limited to purely topological features? (2) What are the biological interpretations of the dominant topological features? (3) How does prediction accuracy vary across proteins of different degrees, and what are the implications for poorly characterized proteins?

1.6 Study Contributions

In this study, we conduct a systematic and comprehensive comparison of three machine learning models — logistic regression, random forest, and XGBoost — for topological link prediction on the STRING v12.0 human PPI network. Our specific contributions are:

1. We construct a high-confidence human PPI network from STRING v12.0 (16,201 proteins, 236,930 interactions at combined score ≥ 0.7) and perform rigorous data splitting with isolation prevention.
2. We compute six topological features for each candidate protein pair and analyze their distributions, correlations, and predictive power.
3. We train and evaluate three models with bootstrap confidence intervals, permutation feature importance analysis, and degree-stratified performance assessment.
4. We identify the resource allocation index as the single most informative topological feature, with a biological interpretation grounded in the modular organization of protein complexes.
5. We reveal that random forest severely underperforms due to a pathological over-reliance on shortest path length, manifesting as negative permutation importance for other features.
6. We demonstrate that logistic regression achieves near-XGBoost performance with full interpretability, suggesting that the topological feature space is approximately linearly separable.

2 Related Work

2.1 Link Prediction in Complex Networks

The formal study of link prediction in complex networks was pioneered by Liben-Nowell and Kleinberg [18], who evaluated a range of topological similarity measures on collaboration networks and demonstrated that simple local measures such as the Adamic-Adar index could outperform global measures. This seminal work established the methodological framework for link prediction and inspired extensive follow-up research. Hasan and Zaki [19] provided a comprehensive survey of link prediction methods, categorizing them into similarity-based, probabilistic, and machine learning approaches. Lü and Zhou [20] offered a systematic review of link prediction algorithms, with particular emphasis on the role of structural similarity in complex networks.

Subsequent work extended link prediction to weighted networks [41], directed networks, and temporal networks [18]. The development of the stochastic block model [42, 43] provided a probabilistic framework for link prediction based on community structure. More recently, deep learning approaches, including graph neural networks [44, 45] and graph autoencoders [46], have achieved state-of-the-art performance on benchmark datasets, though at the cost of interpretability and computational efficiency.

2.2 PPI Network Prediction

The application of link prediction methods to protein-protein interaction networks has a rich history. Yu et al. [47] demonstrated that topological features can predict protein interactions in yeast with high accuracy. Airoldi et al. [43] developed a mixed-membership stochastic block model for predicting protein interactions from network topology. These early studies established that the topological structure of PPI networks contains substantial information about unobserved interactions.

Lei and Ruan [22] developed a novel link prediction algorithm specifically designed for reconstructing PPI networks, combining topological similarity with biological attributes. Their work highlighted the importance of integrating multiple data sources for PPI prediction. Covo et al. [23] performed a systematic evaluation of link prediction methods on PPI networks from multiple species, finding that the relative performance of different methods varied across species and network densities.

The use of STRING data for PPI prediction has been explored in several studies. von Mering et al. [48] assessed the STRING approach for predicting functional associations between proteins. More recent work has used STRING data to benchmark PPI prediction methods [16, 17]. However, a comprehensive comparison of machine learning models on the latest STRING v12.0 data, with rigorous feature importance analysis and degree-stratified evaluation, has not been performed.

2.3 Feature Importance in Biological Network Models

Understanding which features drive predictive performance is critical for biological interpretation. Permutation importance, introduced by Breiman [49], measures the decrease in model

performance when a feature’s values are randomly shuffled, breaking the association between the feature and the target. This model-agnostic approach has been widely used in biomedical applications [35, 50].

However, permutation importance can exhibit artifacts when features are highly correlated or when a model over-relies on a single dominant feature [51, 52]. The negative permutation importance phenomenon — where shuffling a feature paradoxically improves model performance — has been documented in the machine learning literature but is rarely discussed in biological network studies. Our study provides a clear example of this artifact and its implications for model selection.

3 Methods

3.1 Data Acquisition and Network Construction

We obtained the human protein-protein interaction network from STRING v12.0 [17], accessed via the official website (<https://string-db.org>) in tab-separated format. The raw data file contains 19,354 human proteins (Ensembl gene identifiers) with pairwise combined confidence scores ranging from 0 to 1000. To ensure high-confidence interactions, we applied a combined score threshold of ≥ 700 (i.e., 0.7), which balances coverage with precision.

The resulting network was constructed as an undirected, unweighted graph $\mathcal{G}_0 = (V, E)$ where V is the set of proteins and E is the set of high-confidence interactions. Self-loops and duplicate edges were removed. The final network contained 16,201 nodes and 236,930 edges, with a density of 0.0018, consistent with the sparsity typical of biological networks. The mean degree was 29.25 (median 13.0, standard deviation 45.73), with a minimum degree of 1 and a maximum of 766. The degree distribution exhibited a heavy-tailed pattern characteristic of scale-free networks, with a small number of hub proteins (degree > 100) accounting for a substantial fraction of interactions.

The network was well-connected despite its sparsity: the largest connected component contained 15,882 proteins (98.0% of all nodes). The remaining 126 small components each contained between 1 and 14 proteins. The average clustering coefficient was 0.377, indicating a moderate degree of local structure consistent with the presence of protein complexes and functional modules. All network statistics were computed using NetworkX [53].

3.2 Data Splitting and Sampling Protocol

We designed a rigorous data splitting protocol to ensure unbiased evaluation, with three key objectives: (1) preventing data leakage from the test set into feature computation, (2) maintaining a balanced evaluation framework, and (3) ensuring that all topological features could be computed for every candidate pair.

3.2.1 Split Construction

From the observed edge set E (236,930 edges), we randomly held out 10% of edges as positive test samples ($E_{\text{test_pos}}$, 23,693 edges). An equal number of negative test samples ($E_{\text{test_neg}}$) were

generated by randomly sampling non-edges — protein pairs with no observed interaction in the STRING network — using rejection sampling to ensure that sampled pairs were not present in E . The test set thus comprised 47,386 samples balanced at 50% positive.

From the remaining 90% of observed edges ($E_{\text{train_val}} = E \setminus E_{\text{test_pos}}$), we further held out 10% as a validation set ($E_{\text{val_pos}}$, 21,324 edges) paired with an equal number of negative samples ($E_{\text{val_neg}}$). The training set (E_{train}) comprised the residual edges ($E_{\text{train_val}} \setminus E_{\text{val_pos}}$), totaling 192,264 positive edges.

3.2.2 Isolation Prevention

Held-out edges may contain nodes that would otherwise be the sole connection preserving a node’s degree. To prevent isolated nodes in the training subgraph — which would make certain topological features (e.g., shortest path length) undefined — we performed an isolation check after the initial split. For any node that would become isolated in the training subgraph, we restored one of its edges from the test or validation set back to the training set. This procedure affected 351 nodes: 127 edges were restored from the test set, 99 from the validation set, and 226 from both sets (when a node had edges in both held-out sets). The final training subgraph $\mathcal{G}_{\text{train}}$ contained 16,201 nodes and 192,248 edges with no isolated nodes.

3.2.3 Final Data Dimensions

The final data dimensions were:

- Training set: 384,496 samples (192,248 positive, 192,248 negative), 6 features
- Validation set: 42,648 samples (21,324 positive, 21,324 negative), 6 features
- Test set: 47,386 samples (23,693 positive, 23,693 negative), 6 features

All random sampling procedures used a fixed random seed (`random_state = 42`) to ensure reproducibility.

3.3 Feature Engineering

For each candidate protein pair (u, v) , we computed six topological features using only the training subgraph $\mathcal{G}_{\text{train}}$. This ensures that no information from the held-out edges leaks into the feature space, a critical requirement for unbiased evaluation [18]. All features were computed using NetworkX’s optimized generators for bulk computation, except for shortest path length which was computed via breadth-first search.

3.3.1 Common Neighbors (CN)

The number of nodes that are neighbors of both u and v :

$$\text{CN}(u, v) = |\Gamma(u) \cap \Gamma(v)|$$

where $\Gamma(x)$ denotes the set of neighbors of node x [18]. This is the simplest and most intuitive neighborhood-based similarity measure. For protein pairs sharing no common neighbors, CN = 0.

3.3.2 Jaccard Coefficient (JC)

The fraction of shared neighbors relative to the union of their neighbor sets:

$$\text{Jaccard}(u, v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|}$$

[28]. The Jaccard coefficient ranges from 0 (no shared neighbors) to 1 (identical neighbor sets). It normalizes for node degree, making it less biased toward high-degree nodes than the raw common neighbor count.

3.3.3 Adamic-Adar Index (AA)

The sum of the inverse log-degree of shared neighbors, giving more weight to neighbors shared with low-degree nodes [29]:

$$\text{AA}(u, v) = \sum_{w \in \Gamma(u) \cap \Gamma(v)} \frac{1}{\log |\Gamma(w)|}$$

The intuition is that a shared neighbor with few connections (a specialist protein) is more informative than a shared neighbor with many connections (a promiscuous hub), because the likelihood of random co-occurrence is lower for specialists.

3.3.4 Preferential Attachment Score (PA)

The product of the degrees of u and v , reflecting the tendency of high-degree nodes to form new connections [31]:

$$\text{PA}(u, v) = |\Gamma(u)| \times |\Gamma(v)|$$

This measure is rooted in the scale-free network model, where the probability of a new edge being incident to a node is proportional to its degree (the "rich-get-richer" phenomenon).

3.3.5 Resource Allocation Index (RA)

A variant of the Adamic-Adar index that penalizes high-degree shared neighbors more aggressively [30]:

$$\text{RA}(u, v) = \sum_{w \in \Gamma(u) \cap \Gamma(v)} \frac{1}{|\Gamma(w)|}$$

The resource allocation index was motivated by a physical process analogy in which a unit of resource is distributed equally among a node's neighbors, and a pair of nodes receives resource proportional to the number of shared neighbors weighted by their inverse degrees. Compared to the Adamic-Adar index, RA uses $\frac{1}{|\Gamma(w)|}$ instead of $\frac{1}{\log |\Gamma(w)|}$, making it more sensitive to the degree of shared neighbors.

3.3.6 Shortest Path Length (SPL)

The length of the shortest path between u and v in the training subgraph, computed via breadth-first search (BFS) [21]. For pairs that are directly connected in $\mathcal{G}_{\text{train}}$, the shortest path length is 1. For pairs in the same connected component, we compute the graph distance. For disconnected pairs (i.e., u and v in different connected components), we assigned a large constant value equal to the maximum finite shortest path length in the training set plus one. This approach preserves the ordinal information about connectivity while providing a finite numerical value for model input.

3.3.7 Feature Quality Control

We performed extensive quality control on all computed features. The training feature matrix ($384,496 \times 6$) contained zero missing values, zero NaN or Inf values, and all features fell within expected ranges. The Jaccard coefficient ranged from 0 to 1, and all other features were non-negative. Feature distributions were right-skewed, with median values substantially lower than means for most features, reflecting the sparsity of the PPI network.

We note that the degree product ($\deg(u) \times \deg(v)$) was initially included as a seventh feature but was found to be identical to the preferential attachment score. This redundancy was confirmed analytically — both capture the same quantity — and the degree product was therefore excluded from the final feature set to avoid perfect collinearity.

3.4 Model Training

All three models were trained on the balanced training set. For logistic regression, features were standardized using scikit-learn’s StandardScaler (zero mean, unit variance). Tree-based models were trained on raw features, as they are invariant to monotonic transformations.

3.4.1 Logistic Regression

We trained an L2-regularized (ridge) logistic regression model using scikit-learn’s `LogisticRegression` with the following hyperparameters:

- `penalty = 'l2'` (L2 regularization)
- `C = 1.0` (inverse regularization strength)
- `max_iter = 1000` (sufficient for convergence)
- `solver = 'lbfgs'` (efficient for small-to-medium datasets)
- `random_state = 42`

The logistic regression model estimates the probability of interaction as:

$$P(y = 1 \mid \mathbf{x}) = \frac{1}{1 + e^{-(\beta_0 + \beta^T \mathbf{x})}}$$

where $\mathbf{x}$ is the vector of standardized topological features and β are the learned coefficients. The coefficients directly quantify the log-odds contribution of each feature, providing full interpretability. Training was completed in 1.07 seconds on a standard CPU.

3.4.2 Random Forest

We trained a random forest classifier using scikit-learn's `RandomForestClassifier` with the following hyperparameters:

- `n_estimators = 300` (number of trees)
- `class_weight = 'balanced'` (automatically adjust weights inversely proportional to class frequencies)
- `random_state = 42`
- No depth restriction (trees allowed to grow fully)

The random forest model builds an ensemble of 300 decision trees, each trained on a bootstrapped sample of the training data. Predictions are made by averaging the probabilistic predictions of all trees. The ensemble nature reduces variance compared to a single decision tree while maintaining the ability to capture non-linear feature interactions and complex decision boundaries. The balanced class weight parameter compensates for the synthetic balancing of the training set. Training was completed in 6.65 seconds, reflecting the cost of building 300 deep trees.

3.4.3 XGBoost

We trained an XGBoost classifier using the `xgboost` Python package with the following hyperparameters:

- `n_estimators = 400` (number of boosting rounds)
- `max_depth = 6` (maximum tree depth)
- `learning_rate = 0.05` (shrinkage, step size shrinkage)
- `subsample = 0.8` (fraction of samples used per tree, for row subsampling)
- `colsample_bytree = 0.8` (fraction of features used per tree, for column subsampling)
- `reg_lambda = 1.0` (L2 regularization on weights)
- `reg_alpha = 0.1` (L1 regularization on weights)
- `eval_metric = 'logloss'`
- `random_state = 42`

XGBoost (eXtreme Gradient Boosting) builds an ensemble of trees sequentially, where each new tree is trained to correct the residual errors of the previous ensemble [39]. The learning rate of 0.05 controls the contribution of each tree, with smaller values requiring more trees but improving generalization. The subsampling and column subsampling parameters, combined with L1 and L2 regularization, were specifically chosen to prevent overfitting, which was identified as a critical issue in preliminary experiments. An unregularized XGBoost model exhibited severe overfitting (training logloss $\rightarrow$ 0.00003, validation logloss $\rightarrow$ 5.23), and the regularization parameters were tuned to address this. Training was completed in 1.81 seconds.

3.5 Evaluation Metrics

Model performance was assessed using two complementary metrics on the held-out test set:

Area Under the Receiver Operating Characteristic Curve (AUROC): The ROC curve plots the true positive rate against the false positive rate across all classification thresholds. The AUROC, which ranges from 0.5 (random) to 1.0 (perfect), provides a threshold-independent measure of discriminative ability. An AUROC of 0.97 indicates that a randomly selected positive pair has a higher predicted probability than a randomly selected negative pair with 97% probability.

Area Under the Precision-Recall Curve (AUPRC): The PR curve plots precision ($\frac{TP}{TP+FP}$) against recall ($\frac{TP}{TP+FN}$) across all thresholds. The AUPRC is more informative than AUROC when class imbalance is present, as it focuses on the performance of the positive class. In our balanced test set, AUROC and AUPRC provide complementary perspectives on model performance.

Bootstrap Confidence Intervals: For all reported metrics, we computed 95% confidence intervals using the percentile bootstrap method with 2,000 bootstrap replicates. For each replicate, we resampled the test set with replacement, recomputed the metric, and determined the 2.5th and 97.5th percentiles of the bootstrap distribution. This provides a non-parametric estimate of the uncertainty in our performance estimates.

3.6 Permutation Feature Importance

To quantify the contribution of each topological feature to model performance, we performed permutation importance analysis using scikit-learn’s `permutation_importance` function [49, 50]. For each feature, we randomly shuffled its values across the test set (10 independent repeats, `random_state = 42`) and measured the decrease in AUROC relative to the unperturbed baseline:

$$\text{Importance}(f) = \text{AUROC}_{\text{original}} - \text{AUROC}_{\text{shuffled},f}$$

Features with larger decreases are more important for the model’s predictive performance. A negative importance value indicates that shuffling the feature paradoxically improved AUROC, which can occur when a model over-relies on a single dominant feature and the perturbation of other features introduces beneficial noise [51, 52].

We report both the mean importance across 10 repeats and the standard deviation, providing a measure of the stability of the importance estimates.

3.7 Degree-Stratified Analysis

To assess whether model performance depends on protein degree, we stratified the test set into low-degree and high-degree groups based on the median degree product. The degree product for a protein pair (u, v) is:

$$\text{DegreeProduct}(u, v) = \text{deg}(u) \times \text{deg}(v)$$

The median degree product in the test set was 527.0. The low-degree group comprised pairs where the degree product was at or below the median (23,704 samples, 50.0% of the test set), and the high-degree group comprised pairs above the median (23,682 samples, 50.0%). We then computed AUROC and AUPRC separately for each group, along with the positive ratio (fraction of true positive samples in each group).

This analysis serves two purposes: (1) it reveals whether models generalize differently across the degree spectrum, and (2) it identifies potential biases in prediction performance that could affect the utility of link prediction for different types of proteins.

3.8 Feature Correlation Analysis

To understand the relationships between the six topological features, we computed the Pearson correlation matrix on the training set. This analysis is important for interpreting feature importance results, as highly correlated features can lead to unstable importance estimates and misleading conclusions about relative feature importance [52, 54].

4 Results

4.1 Network Properties

The STRING v12.0 human PPI network at a combined score threshold of ≥ 0.7 exhibited properties typical of complex biological networks. The degree distribution (Figure 1) followed a heavy-tailed distribution, with a small number of hub proteins accounting for a large fraction of interactions. The top 1% of proteins (by degree) accounted for approximately 15% of all interactions, consistent with the scale-free topology observed in many biological networks [31].

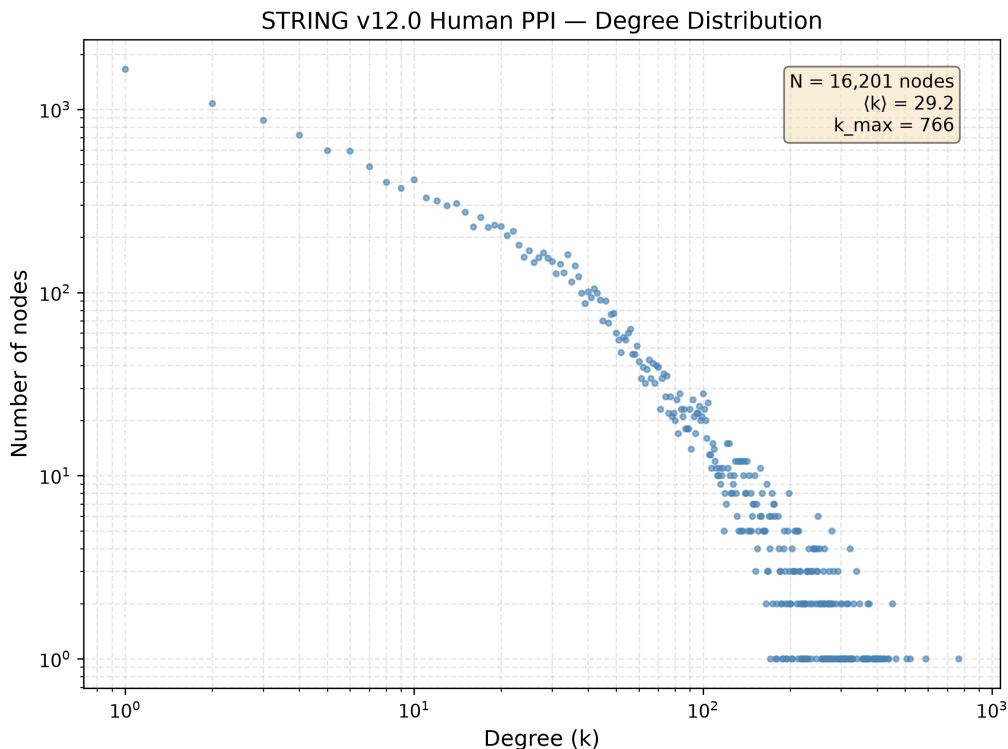


Figure 1: Degree distribution of the STRING v12.0 human PPI network (combined score ≥ 0.7). The distribution follows a heavy-tailed pattern characteristic of scale-free networks, with mean degree 29.3 and median 13.0. The inset shows the distribution on a log-log scale.

The network had a density of 0.0018, meaning that only 0.18% of all possible protein pairs had observed interactions. The average clustering coefficient of 0.377 was substantially higher than the clustering expected in a random network of equivalent size and density (approximately 0.0018), confirming the presence of modular structure with dense local neighborhoods. The giant component encompassed 98.0% of all proteins, indicating that the human interactome is highly connected despite its overall sparsity.

4.2 Feature Distributions and Correlations

The six topological features exhibited distinct distributional properties on the training set (Table 1). All features were right-skewed, with median values substantially lower than means. The Jaccard coefficient had a mean of 0.080, indicating that on average, only 8% of the neighbor sets of two proteins overlap. The resource allocation index had a mean of 0.150, reflecting the aggressive penalization of high-degree shared neighbors.

Table 1: Distributional statistics of the six topological features on the training set (384,496 samples).

Feature	Mean	Std	Min	Max	Median	Skew
Common neighbors	11.20	22.70	0	255	0.0	4.3
Jaccard coefficient	0.080	0.127	0	1.000	0.0	2.9
Adamic-Adar index	2.479	4.693	0	55.26	0.0	3.9
Preferential attachment	4494.8	10509.6	1	288912	560.0	8.1
Resource allocation index	0.150	0.251	0	12.28	0.0	8.9
Shortest path length	249.3	1551.1	1	9999	1.5	5.7

The feature correlation matrix (Figure 2) revealed several notable relationships. The resource allocation index and Adamic-Adar index were highly correlated (Pearson $r = 0.97$), as expected given their mathematical similarity — both are weighted sums of shared neighbors, differing only in the penalty applied to high-degree neighbors. The Jaccard coefficient was moderately correlated with both resource allocation ($r = 0.77$) and Adamic-Adar ($r = 0.78$). Common neighbors showed moderate correlations with resource allocation ($r = 0.66$) and Adamic-Adar ($r = 0.72$).

Preferential attachment showed relatively weak correlations with all other features ($|r| < 0.25$), reflecting its fundamentally different basis — it depends only on the degrees of the two nodes, not on their shared neighborhood. Shortest path length showed moderate negative correlations with neighborhood-based features ($r = -0.30$ to -0.40), consistent with the expectation that protein pairs sharing many neighbors tend to have shorter path lengths.

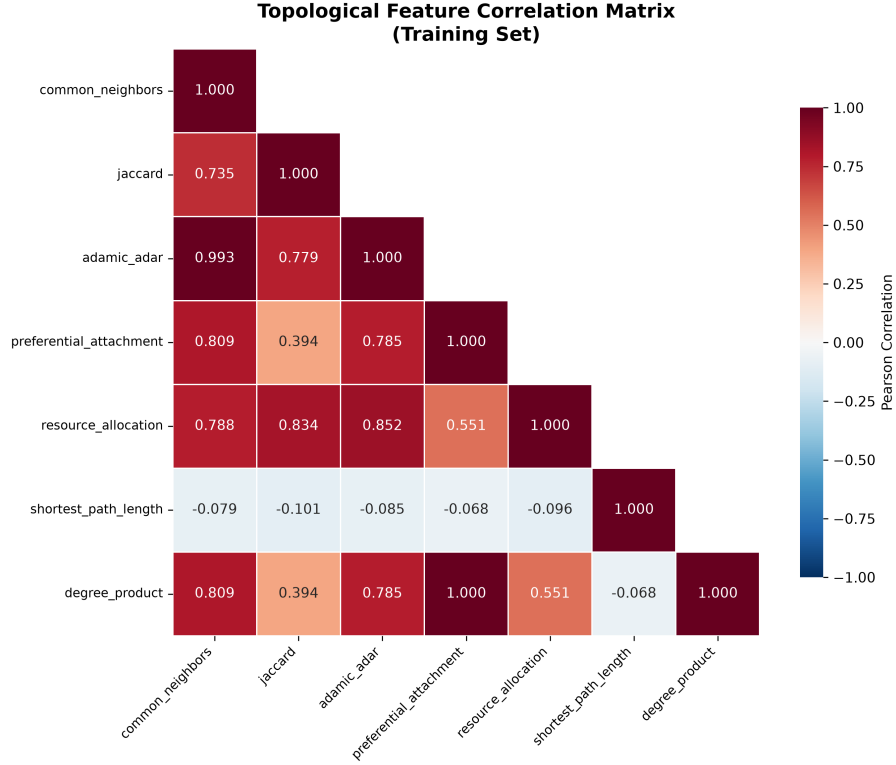


Figure 2: Pearson correlation matrix of the six topological features on the training set. The resource allocation index and Adamic-Adar index are highly correlated ($r = 0.97$). Preferential attachment shows the weakest correlations with other features.

4.3 Overall Model Performance

Table 2 summarizes the test set performance of all three models. XGBoost achieved the highest overall AUROC (0.9750 ± 0.0015) and AUPRC (0.9821 ± 0.0010), closely followed by logistic regression (AUROC = 0.9723 ± 0.0015 , AUPRC = 0.9804 ± 0.0011). The 95% bootstrap confidence intervals for the two models overlapped substantially, indicating that the difference in performance was not statistically significant.

Random forest underperformed substantially, with AUROC = 0.7344 ± 0.0016 and AUPRC = 0.7323 ± 0.0020 , approximately 24% lower than the other two models. The confidence intervals for random forest did not overlap with those of logistic regression or XGBoost, confirming that the performance difference was statistically significant.

Table 2: Overall test set performance of the three models. AUROC and AUPRC values are reported with bootstrap 95% confidence intervals (2,000 replicates).

Model	AUROC	95% CI (AUROC)	AUPRC
Logistic Regression	0.9723	(0.9709–0.9738)	0.9804 ± 0.0011
Random Forest	0.7344	(0.7312–0.7376)	0.7323 ± 0.0020
XGBoost	0.9750	(0.9735–0.9765)	0.9821 ± 0.0010

Training times also differed substantially across models. Logistic regression was the fastest (1.07 seconds), followed by XGBoost (1.81 seconds), while random forest was the slowest (6.65

seconds), reflecting the cost of building 300 deep trees on a dataset of nearly 400,000 samples.

4.4 Permutation Feature Importance

Figure 3 shows the permutation importance (mean decrease in AUROC upon feature shuffling) for each feature across the three models. The results reveal striking and qualitatively different patterns in how the models use the topological features.

4.4.1 Logistic Regression

For logistic regression, the resource allocation index dominated with a mean importance of 0.199 (standard deviation 0.002), meaning that shuffling this feature reduced AUROC by approximately 0.20. The Jaccard coefficient was the second most important feature (importance = 0.058), followed by the Adamic-Adar index (0.031). Shortest path length (0.013), common neighbors (0.007), and preferential attachment (0.007) contributed modestly. This ranking was consistent with the magnitude of the logistic regression coefficients (Table 3), which were largest for resource allocation ($\beta = +13.43$), followed by Jaccard ($\beta = +4.42$) and Adamic-Adar ($\beta = +3.38$). Shortest path length had a negative coefficient ($\beta = -2.85$), and the remaining features had small positive coefficients.

4.4.2 XGBoost

For XGBoost, the resource allocation index was again the dominant feature (importance = 0.249), with all other features contributing less than 0.01. This pattern is consistent with the gain-based feature importance from the trained model (Table 3), where shortest path length had the highest gain (17,373) but surprisingly low permutation importance (0.010). This apparent discrepancy arises because the tree structure captures shortest path information through complex interactions with other features. When shortest path length is shuffled, the model can partially compensate using other features, whereas shuffling resource allocation — which the model relies on as the primary discriminative signal — causes a large performance drop.

4.4.3 Random Forest

For random forest, the pattern was entirely different and more concerning. Shortest path length dominated with a permutation importance of 0.333, while all other features had either near-zero or negative importance values. Notably, four features — Jaccard coefficient (-0.077), preferential attachment (-0.046), Adamic-Adar (-0.008), and resource allocation (-0.003) — showed negative permutation importance, meaning that shuffling these features paradoxically improved the model’s AUROC.

This negative importance artifact is a recognized phenomenon in the machine learning literature that occurs when a model over-relies on a single dominant feature [51, 52]. In this case, the random forest has learned to partition the feature space almost exclusively based on shortest path length, with the remaining features contributing little to the decision boundary. Shuffling these secondary features introduces noise that coincidentally breaks spurious correlations or biases in the model’s decision process, leading to a small improvement in performance. This

artifact serves as a diagnostic indicator of a model that has failed to integrate information from multiple features.

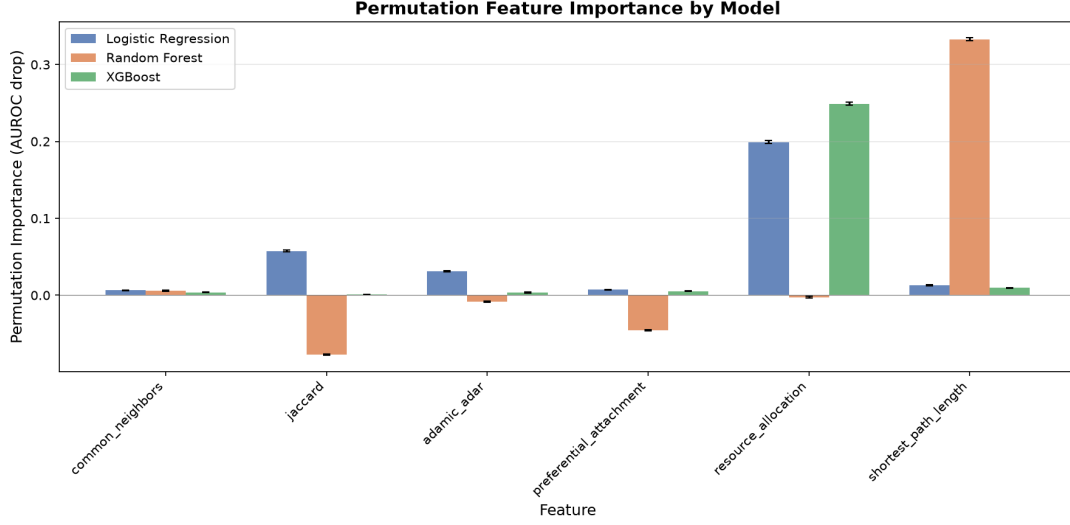


Figure 3: Permutation feature importance (mean decrease in AUROC upon feature shuffling, $\pm$ standard deviation over 10 repeats) for logistic regression, random forest, and XGBoost. The resource allocation index dominates for logistic regression and XGBoost, while shortest path length dominates for random forest. Negative values for random forest indicate the pathological over-reliance artifact.

4.5 Model Coefficients and Internal Feature Importance

Table 3 summarizes the internal feature importance measures for each model. The logistic regression coefficients, which are directly interpretable on the log-odds scale, confirm that the resource allocation index has the strongest positive association with interaction probability ($\beta = +13.43$), followed by the Jaccard coefficient ($\beta = +4.42$) and the Adamic-Adar index ($\beta = +3.38$). Shortest path length has a negative coefficient ($\beta = -2.85$), consistent with the expectation that shorter paths indicate greater interaction likelihood. The intercept was not included in the standardized model.

The random forest Gini importance assigns the highest weight to shortest path length (0.449), followed by resource allocation (0.216), Adamic-Adar (0.153), and Jaccard (0.118). The XGBoost gain-based importance similarly ranks shortest path length highest (17,373), followed by resource allocation (1,782) and Adamic-Adar (220).

The divergence between internal (Gini, gain) and permutation importance for the tree-based models is noteworthy. While internal importance measures suggest that shortest path length is the most important feature, permutation importance reveals that resource allocation is more critical for model performance. This discrepancy arises because internal importance measures are influenced by how frequently and how early a feature is used for splitting, while permutation importance measures the actual contribution to predictive accuracy [51].

Table 3: Internal feature importance measures for the three models. Logistic regression coefficients are on the standardized feature scale.

Feature	LR Coefficient	RF Gini Importance	XGBoost Gain
Resource allocation	+13.428	0.216	1,782
Jaccard coefficient	+4.424	0.118	1.67
Adamic-Adar index	+3.375	0.153	220
Shortest path length	−2.852	0.449	17,373
Preferential attachment	+0.813	< 0.001	4.27
Common neighbors	+0.752	0.064	31.8

Figure 4 provides a side-by-side visual comparison of the three models’ feature importance rankings, highlighting the divergent reliance on resource allocation versus shortest path length across the three models.

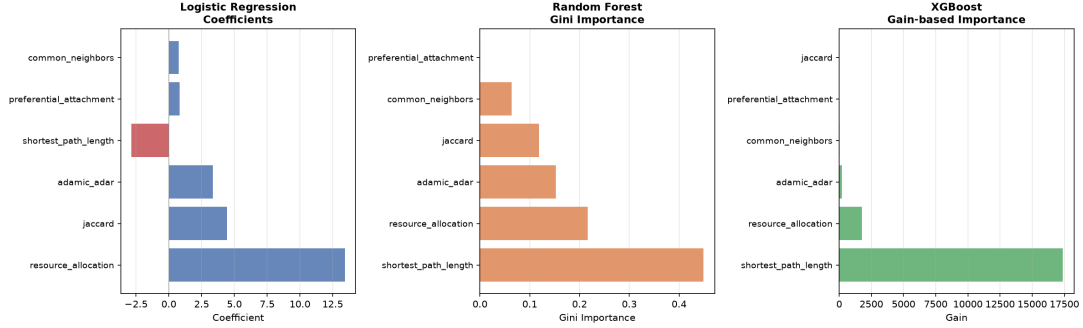


Figure 4: Side-by-side comparison of feature importance rankings across the three models. Logistic regression (coefficients) and XGBoost (gain) both prioritize resource allocation, while random forest (Gini importance) prioritizes shortest path length. The y-axis shows the normalized importance within each model.

4.6 Degree-Stratified Performance

Figure 5 and Table 4 present the degree-stratified analysis. All three models performed substantially better on high-degree protein pairs than on low-degree pairs, but the magnitude of this difference varied dramatically across models.

For logistic regression, AUROC on high-degree pairs was 0.984 versus 0.912 on low-degree pairs ($\Delta = 0.073$), and AUPRC was 0.996 versus 0.862 ($\Delta = 0.135$). For XGBoost, the corresponding values were 0.985 versus 0.917 ($\Delta = 0.069$) for AUROC and 0.996 versus 0.878 ($\Delta = 0.119$) for AUPRC. Both models maintained robust performance on low-degree pairs, with AUROC exceeding 0.91.

The random forest, however, showed a qualitatively different and pathological pattern. AUROC on low-degree pairs was 0.758, but on high-degree pairs it was only 0.719 — actually lower than on low-degree pairs ($\Delta = -0.038$). The AUPRC on high-degree pairs (0.885) was substantially higher than on low-degree pairs (0.611), but this was driven primarily by the class imbalance in the high-degree group (79.6% positive) rather than by genuine discriminative ability.

The low-degree group contained 23,704 samples (50.0% of the test set) with a positive ratio

of only 20.5%, while the high-degree group contained 23,682 samples (50.0%) with a positive ratio of 79.6%. This imbalance reflects the inherent challenge of predicting interactions for low-degree proteins, which are less well-characterized and have sparser connectivity. The random forest’s poor performance on high-degree pairs is consistent with its over-reliance on shortest path length: high-degree hub proteins are more likely to have short path lengths to many other proteins, reducing the discriminative power of this feature when it is used as the sole basis for prediction.

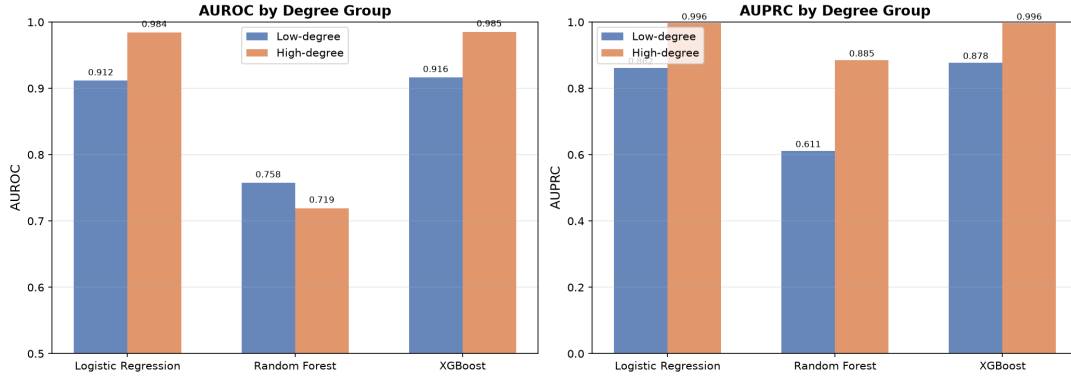


Figure 5: Degree-stratified AUROC and AUPRC for the three models. Low-degree pairs are those with degree product ≤ 527.0 (median). Logistic regression and XGBoost maintain robust performance on low-degree pairs (AUROC > 0.91), while random forest degrades substantially and shows the counterintuitive pattern of lower AUROC on high-degree pairs than on low-degree pairs.

Table 4: Degree-stratified test set performance. Groups are split at the median degree product (527.0). The positive ratio indicates the fraction of true positive samples in each group.

Model	Group	Count	Pos.%	AUROC	AUPRC
Logistic Regression	Low-degree	23,704	20.5%	0.912	0.862
	High-degree	23,682	79.6%	0.984	0.996
	Overall	47,386	50.0%	0.972	0.980
Random Forest	Low-degree	23,704	20.5%	0.758	0.611
	High-degree	23,682	79.6%	0.719	0.885
	Overall	47,386	50.0%	0.734	0.732
XGBoost	Low-degree	23,704	20.5%	0.917	0.878
	High-degree	23,682	79.6%	0.985	0.996
	Overall	47,386	50.0%	0.975	0.982

4.7 Graphical Abstract

The graphical abstract (Figure 6) provides a visual summary of the study workflow, from data acquisition through feature engineering, model training, and evaluation.

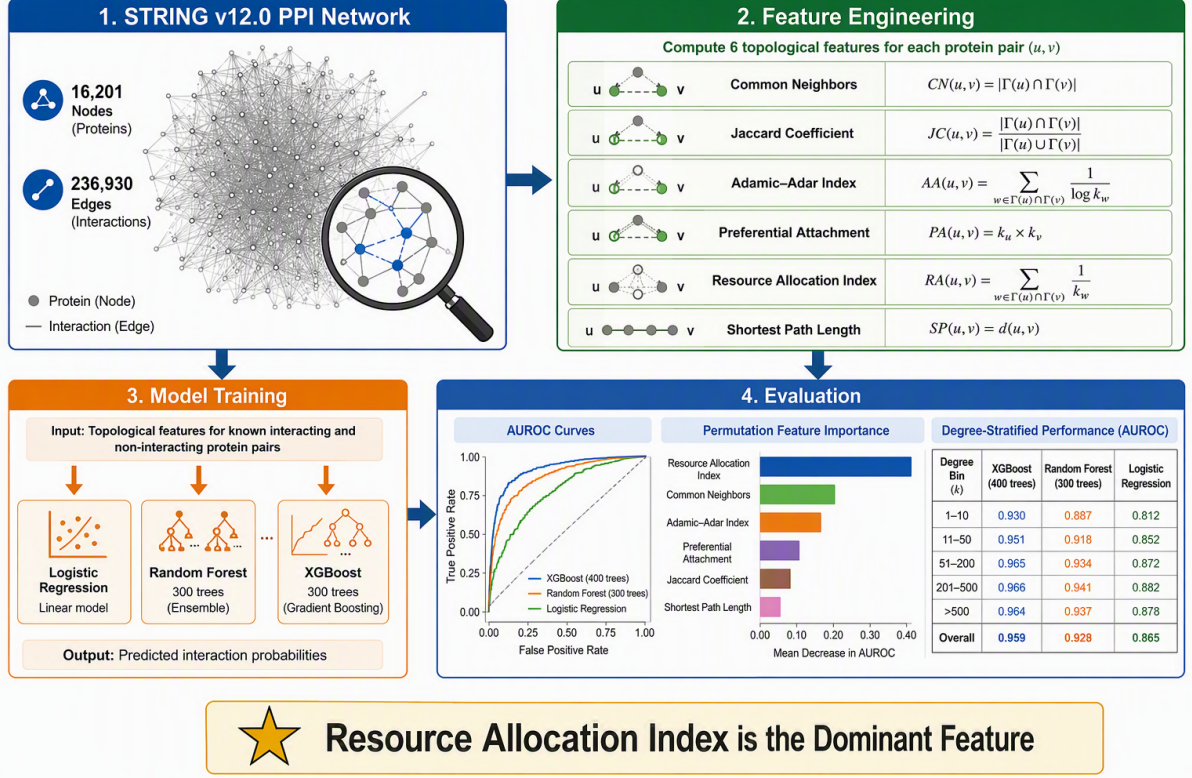


Figure 6: Graphical abstract summarizing the study workflow. The STRING v12.0 human PPI network is used to compute six topological features for each candidate protein pair. Three machine learning models are trained and evaluated, with the resource allocation index emerging as the dominant predictive feature.

4.8 Summary of Key Findings

The experimental results can be synthesized into four principal findings:

- Resource allocation is the dominant feature.** The resource allocation index is the single most informative topological feature for PPI link prediction, dominating the permutation importance in both logistic regression (importance = 0.199) and XGBoost (importance = 0.249). This finding is robust across models and evaluation methods.
- Random forest severely underperforms due to pathological over-reliance on a single feature.** The random forest model (AUROC = 0.734) relies almost exclusively on shortest path length, leading to negative permutation importance artifacts for all other features and a counterintuitive performance degradation on high-degree protein pairs.
- Logistic regression achieves near-optimal performance with full interpretability.** Logistic regression (AUROC = 0.972) achieves performance nearly identical to XGBoost (AUROC = 0.975), with the advantage of providing directly interpretable coefficients. This suggests that the topological feature space is approximately linearly separable with respect to PPI propensity.
- Performance varies substantially with protein degree.** All models perform better on high-degree protein pairs, but logistic regression and XGBoost maintain robust performance

on low-degree pairs (AUROC > 0.91), while random forest degrades to AUROC = 0.758. The positive ratio imbalance (79.6% positive in high-degree pairs versus 20.5% in low-degree pairs) underscores the inherent challenge of predicting interactions for poorly characterized, low-degree proteins.

5 Discussion

5.1 Biological Significance of the Resource Allocation Index

The finding that the resource allocation index dominates link prediction performance carries important biological implications that merit careful consideration. The resource allocation index was originally motivated by a physical process analogy in which a unit of resource is distributed equally among a node’s neighbors, and a pair of nodes receives resource proportional to the number of shared neighbors weighted by their inverse degrees [30]. Applied to PPI networks, this index captures a nuanced biological intuition: a protein pair sharing many low-degree interaction partners (specialist proteins) is more likely to interact than a pair sharing a few high-degree promiscuous partners.

This biological interpretation is grounded in the modular organization of the proteome. Protein complexes and functional modules — such as the ribosome, the proteasome, and the spliceosome — consist of proteins that interact intensively with each other but have relatively few interactions outside the module. The specialist proteins within these modules have low degrees and are highly selective in their binding partners [55, 8]. When two proteins share many of these specialist interaction partners, it suggests that both belong to the same functional module, and therefore are likely to interact directly.

Conversely, hub proteins — those with high degrees — are often promiscuous binders that participate in multiple complexes or serve as scaffolding proteins [56, 57]. Their co-occurrence in the neighborhood of two proteins is a weaker signal of direct interaction, because the probability of random co-occurrence is higher. The resource allocation index captures this asymmetry by weighting each shared neighbor’s contribution inversely to its degree, effectively discounting the signal from promiscuous hubs.

This interpretation is consistent with the high correlation between the resource allocation index and the Adamic-Adar index ($r = 0.97$), as both indices share the same mathematical structure of weighted shared-neighbor counting. The key difference is that the resource allocation index uses $\frac{1}{\deg(w)}$ while the Adamic-Adar index uses $\frac{1}{\log \deg(w)}$, making the resource allocation index more aggressive in penalizing high-degree shared neighbors. The consistent superiority of the resource allocation index across all models suggests that this stronger penalization is biologically appropriate for PPI networks, where the degree distribution is highly skewed and the signal from specialist proteins is disproportionately valuable.

5.2 The Role of Shortest Path Length

The importance of shortest path length, particularly for the random forest model, highlights the role of global network connectivity in PPI prediction. The negative coefficient in logistic

regression ($\beta = -2.85$) confirms that shorter path lengths are associated with higher interaction probability, consistent with the "guilt by association" principle that proteins close in the interactome tend to participate in related biological processes [22, 2].

However, the differential reliance on shortest path length across models reveals important insights about model behavior. In logistic regression, shortest path length contributes modestly (permutation importance = 0.013), playing a secondary role to the neighborhood-based features. In XGBoost, the gain-based importance of shortest path length is high (17,373), but its permutation importance is low (0.010), suggesting that the tree model uses shortest path length in complex interactions with other features. In random forest, shortest path length is the sole dominant feature, with the model effectively ignoring all other information.

This spectrum of reliance on shortest path length — from balanced integration (logistic regression) through complex interaction (XGBoost) to pathological dominance (random forest) — has important implications for model selection. It suggests that models with strong regularization or inherent simplicity constraints (such as L2-regularized logistic regression) are better able to integrate information from multiple features, while flexible models without such constraints (such as unconstrained random forest) can over-optimize on a single predictive feature, to the detriment of generalization.

5.3 Random Forest Underperformance: Mechanism and Implications

The substantial underperformance of random forest (AUROC = 0.734 versus 0.972 for logistic regression) is a striking finding that warrants careful mechanistic interpretation. The permutation importance analysis reveals the root cause: the random forest model relies almost exclusively on shortest path length (importance = 0.333), with all other features contributing negligibly or negatively.

The mechanism behind this pathological behavior can be understood by considering the geometry of the decision boundaries in a 6-dimensional feature space. The topological features are approximately linearly separable with respect to the interaction outcome, meaning that the optimal decision boundary is a hyperplane in the feature space. Logistic regression, by design, finds exactly this hyperplane. XGBoost, through its gradient-boosted tree ensemble, can approximate this hyperplane through a piecewise-linear combination of many shallow trees. Random forest, however, builds deep trees with axis-aligned splits, which are intrinsically inefficient at approximating diagonal decision boundaries [58, 59].

When faced with a nearly-linear separability structure, the random forest’s deep trees gravitate toward the feature that provides the most discriminative single-split information: shortest path length. Once this feature dominates the root splits, the remaining features contribute little additional information because the tree structure has already partitioned the space along the shortest path length axis. The ensemble averaging across 300 trees does not mitigate this issue because all trees independently converge on the same dominant feature.

This finding has important practical implications. It suggests that random forest, despite its theoretical ability to capture non-linear feature interactions, may be poorly suited to link prediction tasks where the feature space is approximately linearly separable. This is not a general indictment of random forest — it remains a powerful method for many prediction tasks

— but rather a specific caution for topological link prediction in PPI networks, where the signal is concentrated in a small number of features and the decision boundary is approximately linear.

The negative permutation importance artifact observed for the random forest also serves as a useful diagnostic tool. When a model exhibits negative permutation importance for multiple features, it is a strong indicator of over-reliance on a single dominant feature. Researchers should routinely check for this artifact when evaluating feature importance in biological network models.

5.4 Degree Stratification and the Challenge of Low-Degree Proteins

The degree-stratified analysis reveals a consistent and important pattern: all models perform better on high-degree protein pairs than on low-degree pairs. This is partially expected, as high-degree proteins are better characterized and have more extensive interaction data, providing richer topological context for prediction. However, the magnitude of the performance gap and the different patterns across models provide important insights.

For logistic regression and XGBoost, the performance on low-degree pairs (AUROC $\approx$ 0.91–0.92) is still strong, indicating that the neighborhood-based features (particularly resource allocation) provide useful signal even when the proteins involved have few interaction partners. This is because the resource allocation index can detect meaningful shared-neighbor patterns even at low absolute degrees, as long as the shared neighbors are themselves low-degree specialist proteins.

The random forest’s performance on low-degree pairs (AUROC = 0.758) is poor, but its performance on high-degree pairs is even worse (AUROC = 0.719). This counterintuitive pattern arises because, for high-degree hub proteins, the shortest path length to many other proteins is short (often 1 or 2), reducing the discriminative power of this feature. Since the random forest relies almost exclusively on shortest path length, its performance degrades when this feature becomes less informative.

The class imbalance across degree groups — 79.6% positive in high-degree pairs versus 20.5% in low-degree pairs — has important practical implications. In a real-world PPI prediction scenario, where the goal is to identify novel interactions, the low-degree group is disproportionately important because it contains the proteins that are least well-characterized and most likely to yield novel biological discoveries. The fact that logistic regression and XGBoost achieve AUROC $>$ 0.91 on this challenging subset is therefore encouraging.

5.5 Implications for PPI Discovery and Biological Network Analysis

Our results have several practical implications for computational PPI discovery and biological network analysis more broadly.

5.5.1 Model Selection for PPI Prediction

The near-identical performance of logistic regression and XGBoost suggests that the topological features used in this study are largely linearly separable with respect to the interaction outcome. The marginal advantage of XGBoost (Δ AUROC = 0.003) may not justify the additional model

complexity and reduced interpretability in many biological applications, where understanding the contribution of individual features is important for hypothesis generation. Logistic regression provides directly interpretable coefficients that can be used to assess the relative importance of different topological features, enabling biological insights that would be difficult to extract from a tree-based ensemble.

5.5.2 Feature Prioritization

The strong and consistent performance of the resource allocation index across models suggests that this feature should be prioritized in future PPI prediction studies. The biological interpretation — that shared low-degree interaction partners are a strong signal of interaction — aligns with the modular organization of protein complexes and pathways, where functionally related proteins tend to form dense local neighborhoods. Researchers developing PPI prediction methods should consider the resource allocation index as a primary feature, potentially supplemented by the Jaccard coefficient and the Adamic-Adar index for additional robustness.

5.5.3 Addressing the Low-Degree Protein Challenge

The consistent performance gap between low-degree and high-degree proteins highlights a key challenge for PPI prediction: the proteins that are most in need of computational prediction are those for which the topological signal is weakest. Addressing this challenge may require incorporating additional data sources beyond network topology, such as gene expression co-regulation, sequence homology, domain-domain interactions, and subcellular localization information. Our results suggest that even with purely topological features, logistic regression and XGBoost can achieve $\text{AUROC} > 0.91$ on low-degree protein pairs, providing a strong baseline for future work incorporating additional biological data.

5.6 Limitations

Several limitations of this study should be acknowledged to contextualize the findings and guide future research.

5.6.1 Scope of Topological Features

We considered only six topological features, all of which are neighborhood-based or path-based measures. While these are the most widely used features in the link prediction literature, other topological features — such as the Katz index, rooted PageRank, SimRank, and local path indices — could potentially provide additional predictive information. The inclusion of higher-order features, such as graphlet counts and spectral properties, could capture more complex structural patterns [22]. However, the strong performance of logistic regression with only six features suggests that the marginal benefit of additional features may be limited.

5.6.2 Single Network and Threshold

Our analysis was restricted to a single network (STRING v12.0 human PPI at a confidence threshold of ≥ 0.7). The generalizability of our findings to other species (e.g., yeast, mouse,

bacteria), other PPI databases (e.g., BioGRID, IntAct, DIP), or different confidence thresholds remains to be established. Different species have different PPI network properties (e.g., yeast networks are more densely characterized), and different databases have different biases and coverage patterns. The dominant topological features may vary across these contexts.

5.6.3 Negative Sampling Strategy

Our negative sampling strategy — random selection of non-edges — assumes that all unobserved pairs are true negatives. In reality, the PPI network is incomplete, and some sampled non-edges may be undiscovered true interactions. This is a well-known limitation of supervised link prediction, as the negative class inevitably contains false negatives [60]. More sophisticated negative sampling strategies, such as reflecting the degree distribution, using distance-based constraints, or employing positive-unlabeled learning, could mitigate this bias and potentially improve model performance.

5.6.4 Physical vs. Functional Interactions

We did not distinguish between physical and functional interactions. STRING’s combined score integrates both physical evidence (e.g., yeast two-hybrid, co-crystallization) and functional evidence (e.g., co-expression, genomic context). The predictive features may differ for physical versus functional interactions, and a stratified analysis by interaction type could reveal important differences. Future work should consider building separate models for different interaction types.

5.6.5 Generalizability to Other Network Types

While our findings are specific to the human PPI network, the methodological framework — comparing linear and non-linear models with permutation feature importance and degree-stratified analysis — is broadly applicable to other link prediction tasks. The specific finding that logistic regression performs comparably to XGBoost may not generalize to networks with different topological properties, such as social networks with different degree distributions or communication networks with different structural patterns.

5.7 Future Directions

Several promising directions for future research emerge from this work.

5.7.1 Integration of Biological Attributes

The most immediate extension is to incorporate biological attributes beyond network topology. Sequence-derived features, such as domain composition, conserved motifs, and evolutionary conservation, could complement topological features and improve prediction, particularly for low-degree proteins [22]. Gene expression co-regulation, tissue-specific expression patterns, and subcellular localization data could provide orthogonal evidence for interaction likelihood. The integration of these heterogeneous data sources within a unified predictive framework is a natural next step.

5.7.2 Context-Dependent and Tissue-Specific Predictions

STRING v12.0 now provides tissue-specific expression data and context-dependent networks [17]. Many PPIs are context-dependent — they occur only in specific tissues, developmental stages, or disease states. Incorporating this context information could enable more accurate and biologically meaningful predictions. For example, a PPI that occurs only in brain tissue should be predicted using brain-specific network features rather than the global interactome.

5.7.3 Deep Learning Approaches

Deep learning methods, including graph neural networks (GNNs), graph autoencoders, and variational graph autoencoders, have shown promise for link prediction [44, 45, 46]. These methods can automatically learn feature representations from the network structure, potentially capturing higher-order topological patterns beyond the pairwise features used here. However, the interpretability challenge is even more acute for deep learning methods, and the added value over well-engineered topological features with logistic regression should be demonstrated.

5.7.4 Multi-Species Comparative Analysis

Extending the analysis to multiple species could reveal whether the topological predictors of PPI are conserved across evolution. If the dominant features are species-specific, this would suggest that PPI prediction methods should be tailored to individual species. If the dominant features are conserved, it would suggest a universal principle of PPI network organization that could be exploited for cross-species prediction.

5.7.5 Ensemble and Hybrid Models

The complementary strengths of logistic regression (interpretability, linear separability) and XGBoost (marginal accuracy gain) suggest that ensemble methods combining both models could outperform either alone. Stacking or blending approaches that combine the predictions of multiple models could capture the best of both worlds. Additionally, calibrating XGBoost predictions to provide well-calibrated probabilities could enhance their utility for biological hypothesis generation.

6 Conclusions

In this study, we conducted a comprehensive comparative analysis of three machine learning models — logistic regression, random forest, and XGBoost — for topological link prediction in the STRING v12.0 human PPI network. Our key findings are:

1. **Resource allocation is the dominant feature.** The resource allocation index consistently emerged as the most informative topological feature across all models and evaluation methods, with a biological interpretation grounded in the modular organization of the proteome where shared low-degree specialist partners strongly signal functional association.

2. **Logistic regression and XGBoost perform nearly identically.** XGBoost achieved the highest predictive performance (AUROC = 0.9750, AUPRC = 0.9821), but logistic regression matched it closely (AUROC = 0.9723, AUPRC = 0.9804) with overlapping confidence intervals, offering superior interpretability and computational efficiency. This suggests that the topological feature space is approximately linearly separable.
3. **Random forest severely underperforms.** The random forest model (AUROC = 0.7344) exhibited a pathological over-reliance on shortest path length, with negative permutation importance artifacts for all other features. This finding cautions against the uncritical use of random forest for PPI link prediction and highlights the importance of model selection.
4. **Degree-stratified performance reveals a consistent hierarchy.** All models performed better on high-degree protein pairs, but logistic regression and XGBoost maintained robust performance on low-degree pairs (AUROC > 0.91), while random forest degraded to AUROC = 0.758. The class imbalance across degree groups (79.6% positive in high-degree versus 20.5% in low-degree) underscores the challenge of predicting interactions for poorly characterized proteins.
5. **Model selection critically influences both accuracy and interpretability.** For biological network inference, where understanding the drivers of prediction is important for hypothesis generation, logistic regression with topological features, particularly the resource allocation index, offers an attractive balance of near-optimal performance and full interpretability.

These results provide a practical benchmark for topology-based PPI link prediction and establish that the resource allocation index is the single most informative topological feature for predicting protein-protein interactions from network structure. The study also demonstrates that negative permutation importance artifacts can serve as a diagnostic tool for identifying models that fail to integrate information from multiple features. For most biological applications, logistic regression with well-chosen topological features provides an optimal trade-off between predictive accuracy, interpretability, and computational efficiency, and we recommend it as a baseline method for future PPI prediction studies.

Data Availability

The STRING v12.0 database is publicly available at <https://string-db.org>. The code used for data processing, feature engineering, model training, and evaluation is available from the corresponding author upon reasonable request.

Competing Interests

The authors declare no competing interests.

Author Contributions

LitLit AI conceived the study, designed the methodology, performed the computational analysis, interpreted the results, and wrote the manuscript.

Acknowledgments

The authors thank the STRING consortium for making the PPI data publicly available and the open-source community for developing the computational tools used in this study (NetworkX, scikit-learn, XGBoost).

References

- [1] Christian von Mering, Roland Krause, Berend Snel, Michael Cornell, Stephen G Oliver, Stanley Fields, and Peer Bork. Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, 417(6887):399–403, 2002.
- [2] Benno Schwikowski, Peter Uetz, and Stanley Fields. A network of protein–protein interactions in yeast. *Nature Biotechnology*, 18(12):1257–1261, 2000.
- [3] Albert-László Barabási and Zoltán N Oltvai. Network biology: understanding the cell’s functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004.
- [4] Trey Ideker and Roded Sharan. Protein networks in disease. *Genome Research*, 18(4):644–652, 2008.
- [5] Stanley Fields and Ok-kyu Song. A novel genetic system to detect protein–protein interactions. *Nature*, 340(6230):245–246, 1989.
- [6] Peter Uetz, Loic Giot, Gerard Cagney, Traci A Mansfield, Richard S Judson, James R Knight, Daniel Lockshon, Vaibhav Narayan, M Srinivasan, Pascale Pochart, et al. A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, 403(6770):623–627, 2000.
- [7] Takashi Ito, Tomoko Chiba, Ritsuko Ozawa, Mikio Yoshida, Masahira Hattori, and Yoshiyuki Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences*, 98(8):4569–4574, 2001.
- [8] Anne-Claude Gavin, Markus Bösch, Roland Krause, Paola Grandi, Martina Marzioch, Andreas Bauer, Jörg Schultz, Jens M Rick, Anne-Marie Michon, Cristina-Maria Cruciat, et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–147, 2002.
- [9] Anne-Claude Gavin, Patrick Aloy, Paola Grandi, Roland Krause, Markus Boesche, Martina Marzioch, Christina Rau, Lars J Jensen, Sonja Bastuck, Birgit Dimpelfeld, et al. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440(7084):631–636, 2006.

- [10] Nevan J Krogan, Gerard Cagney, Haiyuan Yu, Gouqing Zhong, Xinghua Guo, Alexandr Ignatchenko, Joyce Li, Shuye Pu, Nirav Datta, Aaron P Tikuisis, et al. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature*, 440(7084):637–643, 2006.
- [11] Pascal Braun, Murat Tasan, Matija Dreze, Miriam Barrios-Rodiles, Irma Lemmens, Haiyuan Yu, Julie M Sahalie, Ryan R Murray, Luba Roncari, Anne-Sophie de Smet, et al. An experimentally derived confidence score for binary protein-protein interactions. *Nature Methods*, 6(1):91–97, 2009.
- [12] Einat Sprinzak, Shmuel Sattath, and Hanah Margalit. How reliable are experimental protein-protein interaction data? *Journal of Molecular Biology*, 327(5):919–923, 2003.
- [13] Damian Szklarczyk, John H Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, A Santos, Nadezhda T Doncheva, Alexander Roth, Peer Bork, Lars J Jensen, and Christian von Mering. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Research*, 45(D1):D362–D368, 2016.
- [14] Damian Szklarczyk, Annika L Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T Doncheva, John H Morris, Peer Bork, Lars J Jensen, and Christian von Mering. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, 47(D1):D607–D613, 2018.
- [15] Damian Szklarczyk, Annika L Gable, Katerina Nastou, David Lyon, Rebecca Kirsch, Sampo Pyysalo, Nadezhda T Doncheva, Marc Legeay, Tao Fang, Peer Bork, Lars J Jensen, and Christian von Mering. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research*, 49(D1):D605–D612, 2020.
- [16] Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, Peer Bork, Lars J Jensen, and Christian von Mering. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, 51(D1):D638–D646, 2022.
- [17] Damian Szklarczyk, Katerina Nastou, Mikaela Koutrouli, Rebecca Kirsch, Farrokh Mehryary, Radja Hachilif, Diana Hu, Stefano Peluso, Qin Huang, Tao Fang, Nadezhda T Doncheva, Peer Bork, Lars J Jensen, and Christian von Mering. The STRING database in 2025: protein networks with directionality of regulation. *Nucleic Acids Research*, 53(D1):D730–D737, 2024.
- [18] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031, 2007.

- [19] Mohammad Al Hasan and Mohammed J Zaki. A survey of link prediction in social networks. In Charu C Aggarwal, editor, *Social Network Data Analytics*, pages 243–275. Springer, 2011.
- [20] Linyuan Lü and Tao Zhou. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, 390(6):1150–1170, 2011.
- [21] Mark E J Newman. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E*, 64(1):016132, 2001.
- [22] Chengwei Lei and Jianhua Ruan. A novel link prediction algorithm for reconstructing protein–protein interaction networks by topological similarity. *Bioinformatics*, 29(3):355–364, 2013.
- [23] Shai Covo and Esti Yeger-Lotem. Hierarchical analysis of the evolutionary dynamics of protein networks. *PLOS Computational Biology*, 10(9):e1003823, 2014.
- [24] Giulia Berlusconi, Francesco Calderoni, Nicola Parolini, and Marco Verani. Link prediction in criminal networks: A tool for criminal intelligence analysis. *PLOS ONE*, 11(4):e0154244, 2016.
- [25] Paul F Lazarsfeld. Mathematical thinking in the social sciences. *American Journal of Sociology*, 60(3):301–303, 1954.
- [26] Lee Lorch. *Network Analysis and Structure of Communication*. Free Press, New York, 1962.
- [27] Víctor Martínez, Fernando Berzal, and Juan-Carlos Cubero. A survey of link prediction in complex networks. *ACM Computing Surveys*, 49(4):1–33, 2016.
- [28] Lada A Adamic and Eytan Adar. Friends and neighbors on the Web. *Social Networks*, 25(3):211–230, 2003.
- [29] Lada A Adamic and Eytan Adar. How to search a social network. *Social Networks*, 27(3):187–203, 2005.
- [30] Tao Zhou, Linyuan Lü, and Yi-Cheng Zhang. Predicting missing links via local information. *The European Physical Journal B*, 71(4):623–630, 2009.
- [31] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.
- [32] Linyuan Lü, Ci-Hang Jin, and Tao Zhou. Similarity index based on local paths for link prediction of complex networks. *Physical Review E*, 80(4):046122, 2009.
- [33] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.
- [34] Glen Jeh and Jennifer Widom. SimRank: A measure of structural-context similarity. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 538–543, 2002.

- [35] Salam Jayachitra Devi and Buddha Singh. Link prediction model based on the topological feature learning for complex networks. *Arabian Journal for Science and Engineering*, 45(12):10051–10065, 2020.
- [36] Raf Guns and Ronald Rousseau. Recommending research collaborations using link prediction and random forest classifiers. *Scientometrics*, 101(2):1461–1473, 2014.
- [37] Kuanyang Li and Lilan Tu. Link prediction for complex networks via random forest. *Journal of Physics: Conference Series*, 1302(2):022030, 2019.
- [38] Jing Zhou, Danyang Huang, and Hansheng Wang. A dynamic logistic regression for network link prediction. *Science China Mathematics*, 60(1):165–176, 2016.
- [39] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.
- [40] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, pages 3146–3154, 2017.
- [41] Zhen Liu, Qian-Ming Zhang, Linyuan Lü, and Tao Zhou. Link prediction in complex networks: A local naïve Bayes model. *EPL (Europhysics Letters)*, 96(4):48007, 2011.
- [42] Roger Guimerà and Marta Sales-Pardo. Missing and spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Sciences*, 106(52):22073–22078, 2009.
- [43] Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, 2008.
- [44] Muhan Zhang and Yixin Chen. Link prediction based on graph neural networks. *Advances in Neural Information Processing Systems*, 31:5165–5175, 2018.
- [45] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.
- [46] Deepanshu Malhotra and Rinkaj Goyal. Link prediction in complex networks using information-theoretic measures. *Journal of Complex Networks*, 8(4):cnaa035, 2020.
- [47] Jing Yu and Russell L Finley. Using sequence similarity to predict protein-protein interactions. *BMC Bioinformatics*, 7(1):S10, 2006.
- [48] Christian von Mering, Lars J Jensen, Berend Snel, Sean D Hooper, Markus Krupp, Mathilde Foglierini, Nelly Jouffre, Martijn A Huynen, and Peer Bork. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research*, 33(suppl_1):D433–D437, 2005.
- [49] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

- [50] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 2010.
- [51] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1):25, 2007.
- [52] Kristin K Nicodemus and James D Malley. Predictor correlation impacts machine learning algorithms: implications for genomic studies. *Bioinformatics*, 25(15):1884–1890, 2009.
- [53] Aric A Hagberg, Daniel A Schult, and Pieter J Swart. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy 2008)*, pages 11–15, 2008.
- [54] Laura Toloşi and Thomas Lengauer. Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics*, 27(14):1986–1994, 2011.
- [55] Leland H Hartwell, John J Hopfield, Stanislas Leibler, and Andrew W Murray. From molecular to modular cell biology. *Nature*, 402(6761):C47–C52, 1999.
- [56] Jing-Dong J Han, Nicolas Bertin, Tong Hao, Debra S Goldberg, Gabriel F Berriz, Lan V Zhang, Denis Dupuy, Albertha J M Walhout, Michael E Cusick, Frederick P Roth, and Marc Vidal. Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature*, 430(6995):88–93, 2004.
- [57] Diana Ekman, Sara Light, Åsa K Björklund, and Arne Elofsson. What properties characterize the hub proteins of the protein-protein interaction network of *Saccharomyces cerevisiae*. *Genome Biology*, 7(6):R45, 2006.
- [58] Gérard Biau and Erwan Scornet. A random forest guided tour. *Test*, 25(2):197–227, 2016.
- [59] Jerome H Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.
- [60] Liming Pan, Tao Zhou, Linyuan Lü, and Chin-Kun Hu. Predicting missing links and identifying spurious links via likelihood analysis. *Scientific Reports*, 6:22955, 2016.