BMC Biology

RESEARCH

Open Access

Negative sampling strategies impact the prediction of scale-free biomolecular network interactions with machine learning

Pengpai Li[1†], Bowen Shao[1†], Guoqing Zhao[1] and Zhi-Ping Liu[1,2*]

Abstract

Background Understanding protein-molecular interaction is crucial for unraveling the mechanisms underlying diverse biological processes. Machine learning (ML) techniques have been extensively employed in predicting these interactions and have garnered substantial research focus. Previous studies have predominantly centered on improving model performance through novel and efficient ML approaches, often resulting in overoptimistic predictive estimates. However, these advancements frequently neglect the inherent biases stemming from network properties, particularly in biological contexts.

Results In this study, we examined the biases inherent in ML models during the learning and prediction of protein-molecular interactions, particularly those arising from the scale-free property of biological networks—a characteristic where in a few nodes have many connections while most have very few. Our comprehensive analysis across diverse tasks, datasets, and ML methods provides compelling evidence of these biases. We discovered that the training and evaluation of ML models are profoundly influenced by network topology, potentially distorting model performance assessments. To mitigate this issue, we propose the degree distribution balanced (DDB) sampling strategy, a straightforward yet potent approach that alleviates biases stemming from network properties. This method further underscores the limitations of certain ML models in learning protein-molecular interactions solely from intrinsic molecular features.

Conclusions Our findings present a novel perspective for assessing the performance of ML models in inferring protein-molecular interactions with greater fairness. By addressing biases introduced by network properties, the DDB sampling approach provides a more balanced and precise assessment of model capabilities. These insights hold the potential to bolster the reliability of ML models in bioinformatics, fostering a more stringent evaluation framework for predicting protein-molecular interactions.

Keywords Protein-molecular interactions, Machine learning, Negative sampling, Biological networks

Background

In the realm of computational biology, machine learning (ML) methods have garnered considerable attention for predicting protein-molecular interactions, including lncRNA-protein [1, 2], protein–protein [3–6], and drug-target interactions [7]. Despite numerous studies showcasing impressive performance on testing datasets, the accuracy of these models in guiding wet-lab experiments remains a hurdle. This raises the question: Have

[†]Pengpai Li and Bowen Shao contributed equally to this work.

*Correspondence:
Zhi-Ping Liu
zpliu@sdu.edu.cn
[1] Department of Biomedical Engineering, School of Control Science and Engineering, Shandong University, Jinan 250061, Shandong, China
[2] National Center for Applied Mathematics, Shandong University, Jinan 250100, Shandong, China

ML-based computational methods been overestimated? If so, what are the reasons behind the disparity between evaluation and generalization performance? Addressing these questions is imperative.

Early studies have suggested that the remarkable performance of ML models in predicting protein–protein interactions might have been overstated [8, 9]. Yu et al. [8] revealed that model performances vary significantly with different negative sampling strategies, suggesting that sequence-based methods may not reliably predict protein–protein interactions. Similarly, Park and Marcotte [9] found that pair-input methods are significantly influenced by the paired nature of inputs and advocated for reporting predictive performances separately for each distinct test class. Despite these insights, recent studies continue to report overly optimistic model estimates when developing new approaches for predicting protein-molecular interactions. We have compiled and organized studies on protein-molecular interaction prediction using random negative sampling in recent years in Additional file 1: Note S1 [6, 10–61]. Hence, it is crucial to meticulously consider the limitations and potential biases in the development and evaluation of ML models for predicting protein-molecular interactions, particularly when interpreting and applying the results to real-world contexts.

In this paper, we thoroughly reviewed the overall development process of pair-input ML models and identified that the overestimated model performance stems from negative sampling procedures. It is well-established that the purpose of subset sampling for cross-validation differs from that of training set sampling [62]. In cross-validation, reducing bias is paramount to ensure that unbiased subsets represent the overall dataset, allowing evaluation results to generalize. Conversely, in training, the objective is to obtain subsets that facilitate effective model learning. We contend that negative sampling methods, such as random negative sampling commonly used in training, do not always yield the most suitable subsets for model learning. In fact, randomly sampled data still contain influential information, such as network topology, that significantly impacts model learning. Our findings indicate that the commonly adopted random negative sampling strategy results in a degree distribution disparity between positive and negative samples. This disparity significantly influences model learning, overshadowing the significance of the node features themselves.

Our experiments, focusing on three prevalent protein-molecular interaction prediction tasks—lncRNA-protein, protein–protein, and drug-target—provide a comprehensive investigation into both heterogeneous (lncRNA-protein, drug-target) and homogeneous (protein–protein) interactions. This ensures that our findings transcend specific molecule pairs but generalize across different molecular interactions, regardless of whether the involved nodes are of the same type. These results reveal that well-trained ML models tend to predict molecule pairs based solely on the degree of their nodes. In other words, ML models assign high interaction scores to pairs with high node degrees and low scores to those with low node degrees. Consequently, it becomes difficult for ML models to learn unique molecular representations or graph features, which are often the focal point of most novel method development publications. Our study underscores the limitations of existing ML models and highlights the need for new approaches that address these challenges to enhance the accuracy and reliability of predicting protein-molecular interactions.

Results

Overview

In this paper, we approach the task of predicting protein-molecular interactions by adopting a link prediction perspective within protein-molecular interaction graphs. Figure 1b illustrates the diverse array of ML frameworks utilized for predicting protein-molecular interactions, encompassing two pivotal stages: molecule encoding and classification. Notably, due to the absence of negative pairs (non-interactive pairs) and the presence of only experimentally verified positive pairs (interactive pairs), negative pairs are generated by sampling from the complement of the true molecular network prior to model training, as depicted in Fig. 1a.

However, conventional random negative sampling often results in a significant degree distribution disparity between positive and negative samples, stemming from the scale-free property of most biological networks, as illustrated in Fig. 1c. This disparity is evident in the boxplot on the left-hand side of Fig. 1d. Consequently, the ML model may inadvertently learn this degree-based difference and predict interaction probabilities primarily based on sample degrees, rather than capturing the intrinsic molecular features, as shown on the right-hand side of Fig. 1d. To mitigate this issue, we introduce a straightforward yet effective negative sampling strategy, termed degree distribution balanced (DDB), which neutralizes this disparity and enables the model to genuinely learn interaction relationships from the underlying molecular features, as demonstrated in Fig. 1e. Our analysis underscores the crucial role of negative sample collection and the potential ramifications of sampling biases on protein-molecular interaction prediction.

Random negative sampling induces prediction bias

In this work, we have scrutinized the influence of distribution discrepancies between positive and negative

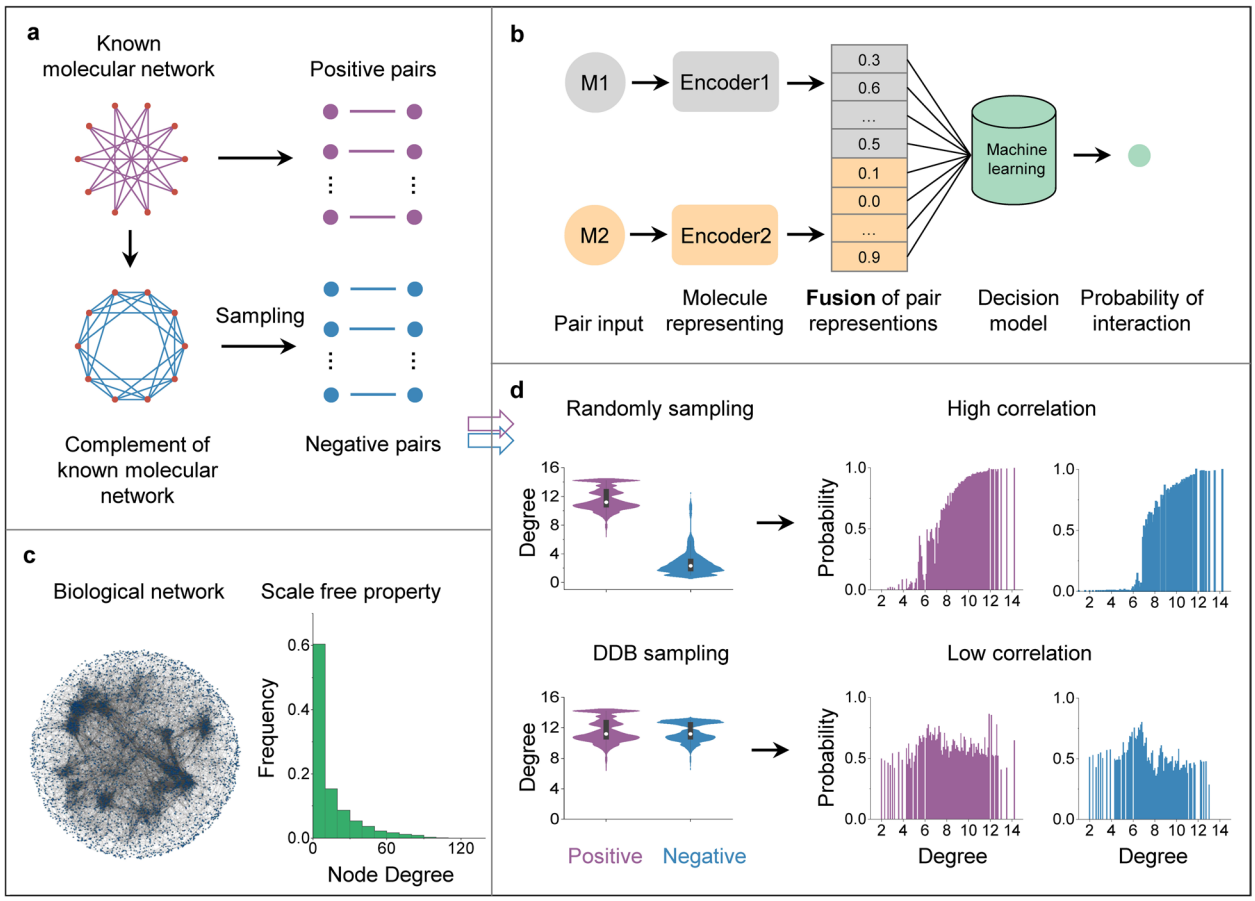

Fig. 1 This figure illustrates the influence of a scale-free network dataset on a pair-input ML model, affecting its learning bias and evaluation. **a** The compiled molecular network comprises interacting samples (positive pairs) but does not inherently include non-interacting samples (negative pairs). Traditional supervised ML methods (as shown in subfigure **b**) require both positive and negative pairs for training and evaluation. Consequently, negative pairs are typically sampled from the complement graph of the known molecular networks (i.e., node pairs not connected in the observed network). **b** A fundamental ML framework for protein-molecular interactions is demonstrated. Two molecules (M1 and M2) are encoded into feature vectors, which are then concatenated to signify their connection. These concatenated features serve as input to train an ML model predicting the interaction probability between the molecule pair. **c** An exemplar biological network, specifically the yeast gene interactive network, demonstrates the presence of a scale-free property. **d** Random sampling techniques result in an imbalanced degree distribution between positive and negative samples. Our analysis indicates that this distribution disparity correlates strongly with the predicted linkage probability and link degree in the ML model's outputs. **e** To counteract the distribution imbalance, a novel sampling strategy called "degree distribution balanced (DDB) sampling" is introduced. By utilizing this method to generate negative samples, the correlation between linkage probability and link degree is reduced

pairs, stemming from the scale-free attribute of biological networks. These differences are internalized by ML models during training, ultimately leading to biased predictions grounded on the degree of pairs within their respective scale-free networks. To elucidate this phenomenon, we benchmarked three categories of biological networks, namely lncRNA-protein interactions, protein–protein interactions, and drug-target interactions, as summarized in Table 1. We leveraged three ML methods—Noise-RF, Seq-RF, and Seq-Deep (detailed in Methods)—to learn from these networks and generate predictions. Notably, we adopted a random sampling strategy to create negative samples for both training and testing datasets in this context. To ensure clarity, the ratio of positive to negative samples was maintained at 1:1 for both datasets, which aligns with the sparsity of biological networks and does not undermine the generalizability of our test evaluation results. Instead, it addresses the extreme imbalance between positive and negative samples, enabling the model to focus more intently on the distinctions between positive and negative samples, thereby bolstering its generalization capability.

Table 1 Characteristics of datasets. LPI stands for lncRNA-protein interaction, PPI represents protein–protein interaction, and DTI denotes drug-target interaction

Dataset		Origin		Processed		Power law
		Nodes	Edges	Nodes	Edges	
LPI	NPInter v4.0 [63, 64]	LncRNA: 43,945 Pro: 3446	373,947	LncRNA: 27,257 Pro: 2440	214,957	2.12
	RAID v2.0 [65, 66]	LncRNA: 1670 Pro: 8688	30,958	LncRNA: 1093 Pro: 5523	15,384	2.38
PPI	InBioMap [67, 68]	Pro: 11,727	175,298	Pro: 5915	69,082	5.50
	STRING v11.5 [69, 70]	Pro: 14,173	178,896	Pro: 8234	79,670	6.78
	BioGRID v4.4.214 [71, 72]	Pro: 23,096	111,249	Pro: 6530	33,560	4.0
	HuRI [73, 74]	Pro: 8275	52,569	Pro: 5073	23,637	3.18
DTI	DrugBank v5.0 [75, 76]	/	/	Drug: 5994 Pro: 3502	16,598	2.595
	DrugCentral [77, 78]	/	/	Drug: 1427 Pro: 1106	9477	2.38

Evaluation in transductive prediction

During the transductive validation, overlap between validation and training dataset nodes is conceivable. This validation method entails randomly selecting a subset of experimentally verified interactive pairs to form the validation dataset. The transductive evaluation results for the eight benchmark datasets are exhaustively detailed in Table 2. Remarkably, all classifiers, including Noise-RF, exhibited commendable performance across these datasets. This favorable outcome is further elucidated by Fig. 2a.

The violin plots in the left panels of Fig. 2a depict the distribution of pair degrees, defined as the sum of the degrees of two molecules within a pair. The results unequivocally reveal a discernible difference between positive and negative sets when negative samples are randomly drawn for all three tasks. Specifically, the degrees of pairs in the positive set surpass those in the negative set.

Subsequently, we computed the average predicted scores using ML methods for pairs with identical degrees within each set and plotted histograms of predicted scores against pair degrees in Fig. 2a. A robust correlation between predicted scores and pair degrees was observed for both positive and negative datasets. Consequently, pairs with higher degrees consistently received higher interaction scores, whereas those with lower degrees garnered lower scores. Considering that pair degrees in the positive set were substantially higher than those in the negative set, the Noise-RF model's impressive performance can be attributed to this degree distribution disparity. For instance, in the NPInter 4.0 dataset, nearly 98.9% of pairs in the positive set exhibited degrees exceeding 8, whereas 96.1% of randomly sampled negative pairs had degrees below 8. Thus, the correlation between prediction scores and pair degrees established a clear boundary distinguishing the two sets. Despite Noise-RF achieving a remarkable AUC value of 0.993, it exhibited a pronounced bias.

Table 2 Transductive model evaluation based on training dataset with negative samples generated by random sampling. The three values separated by double vertical bars are AUROC, Spearman coefficient correlation (ρ) between the degree of positive samples and their predicted interaction probability, and that between the degree of negative samples and their predicted interaction probability. Results represent mean of $n = 15$ independent runs

		Noise-RF	Seq-RF	Seq-Deep
LPI	NPInter4.0	0.993|0.912|0.170	0.993|0.826|0.144	0.994|0.576|0.333
	RAIDv2.0	0.997|0.859|0.416	0.995|0.772|0.247	0.995|0.861|0.095
PPI	InBioMap	0.930|0.867|0.825	0.936|0.845|0.758	0.971|0.621|−0.003
	STRING	0.844|0.935|0.865	0.862|0.911|0.734	0.935|0.761|0.093
	BioGRID	0.815|0.738|0.538	0.857|0.819|0.662	0.874|0.739|0.411
	HuRI	0.865|0.716|0.579	0.892|0.747|0.649	0.904|0.618|0.339
DTI	DrugBank	0.808|0.742|0.599	0.888|0.765|0.497	0.894|0.674|0.346
	DrugCentral	0.864|0.845|0.667	0.920|0.865|0.454	0.934|0.738|0.223

Evaluation in inductive prediction

To decouple from the biological network's topological structure, we adopted the evaluation framework proposed by Park and Marcotte [9]. This strategy categorized pairs into three classes: fully observed in the training set (C1, where both components of the test molecular pair were present in the training data), partially observed in the training set (C2, where only one component of the pair was previously observed), and entirely unseen in the training set (C3, with no

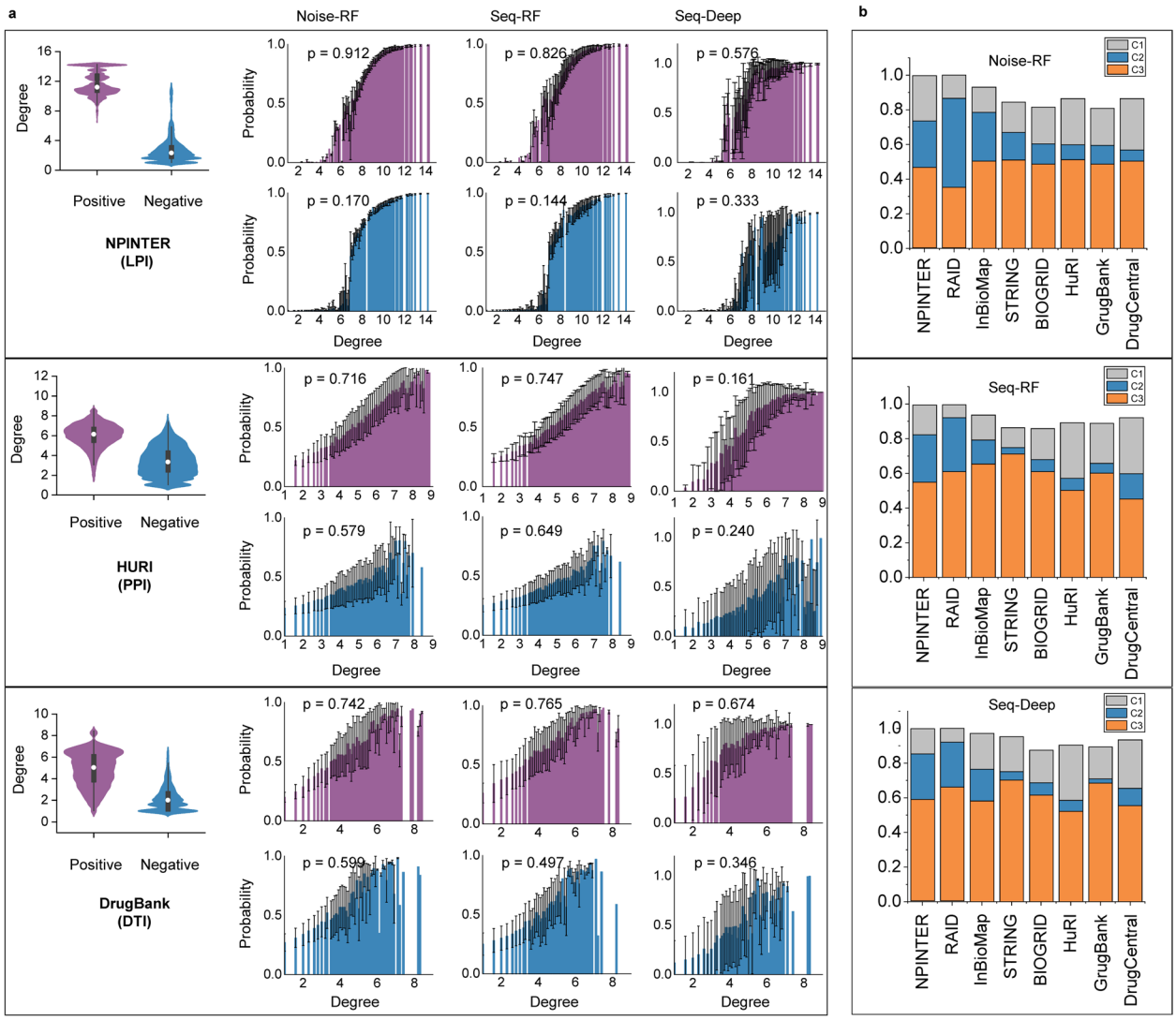

Fig. 2 The ML models capture the degree distribution characteristics of biological networks displaying the scale-free property. **a** For clarity, in this figure, the degree of a pair was computed as the logarithm of the sum of the degrees of the two nodes forming a connection. The violin plots depict the distributions of positive and negative training samples for three datasets. The histograms visually represent the relationship between the probability and the degree of the samples. These probabilities were obtained by averaging the probabilities of samples sharing the same degree (rounded to one decimal place). **b** Comparison of model performance: three types of ML models were evaluated across eight datasets with respect to testing sets C1, C2, and C3 (detailed values provided in Additional file 10: Table S1; results represent mean of $n = 15$ independent runs)

overlapping molecular components between test and training sets), collectively referred to as inductive evaluation. We comparatively assessed the ML models on C1, C2, and C3 testing set to gauge their adaptability. As shown in Fig. 2b (detailed values provided in Additional file 10: Table S1), the inductive capabilities of ML models for the three tasks declined substantially. Model performance progressively diminished from the C1 set to the C2 set and further to the C3 set. The AUC of the Noise-RF model on C3 approximated random guessing, indicating that the model was not influenced by the network structure. Additionally, performance on the C3 dataset reflected the true generalization capability of Seq-RF and Seq-Deep.

When comparing the performance of ML models on C1, C2, and C3 testing sets, we can deduce that the training of these models is primarily influenced by the implicit degree distribution of the network rather than the molecular representations.

Furthermore, we utilized multiple datasets to solidify our conclusions. Additional files present the results from five additional datasets (Additional file 5: Fig. S1, Additional file 6: Fig. S2, Additional file 7: Fig. S3, Additional file 8: Fig. S4, Additional file 9: Fig. S5). These results are consistent with our analysis. The findings across multiple datasets and methods indicate that our observations are not confined to a specific model but are generalized to all ML models.

Can constrained negative sampling alleviate bias?

In the preceding section, we highlighted that the disparity in degree distribution between positive and negative samples can substantially mislead the ML models, introducing a pronounced bias. To tackle this issue, a straightforward solution involves implementing exclusive negative sampling. This strategy strives to align the distribution of selected negative samples with that of positive samples, thereby ensuring higher consistency. This section delves into whether the distribution-constrained negative sampling strategies can mitigate the bias acquired by ML models. Specifically, for the training set, negative samples are generated using the DDB method for model parameters learning, whereas for the testing set, negative samples are randomly sampled to better generalize to the target population. Further detailed descriptions on this method are provided in Methods.

Evaluation in transductive prediction

As illustrated in the violin plots in Fig. 3a, the DDB sampling strategy ensures that the degree distribution of negative samples aligns with that of positive samples. Subsequently, the predicted results are depicted in the histograms of Fig. 3a. We observe a significant reduction in the correlation between predicted scores and sample degrees. However, as shown in Table 3, the performance of the three baseline ML models also experiences a notable decline.

Evaluation in inductive prediction

Does the DDB constraint unveil the model's genuine capability to learn interaction relationships from intrinsic molecular features? Fig. 3b (detailed values in Additional file 11: Table S2) displays the predictive performances of three models on the C1, C2, and C3 datasets. Notably, for the Noise-RF method, which does not utilize any intrinsic molecular features (like sequence information), its performance on the C1 and C2 datasets substantially decreases, approaching the predictive level of the C3 dataset. This implies that much of the network topology information previously embedded in the training data has been eliminated by the DDB constraint. Figure 2b presents the experimental results for negative samples generated by random sampling, comparing the performances of Noise-RF, Seq-RF, and Seq-Deep on the C1, C2, and C3 datasets. It is evident that the gap between Seq-RF, Seq-Deep, and Noise-RF widens, further demonstrating that the DDB constraint uncovers the models' ability to genuinely learn interaction relationships from intrinsic molecular features, such as sequence information.

In Fig. 4, we observe a general downward trend in the predictive performances of the two sequence-based ML models across the C1, C2, and C3 datasets after applying the DDB constraint. This phenomenon can be attributed to the scale-free nature of biological networks, as previously discussed. Specifically, due to the scale-free property of most biological networks, interactions tend to have higher degrees, causing a significant degree distribution difference between positive and negative samples across the entire dataset. In other words, although network topology is not an intrinsic molecular feature, ML models still tend to learn this information as it helps them achieve better results to some extent. Whether this is beneficial depends on whether network topology is viewed as a feature to be learned or as a confounding factor to be excluded. From our perspective, when predicting protein-molecular interactions using strictly sequence-based methods, it is more objective to exclude the influence of network topology information.

DDB method interacts with network topology and functional attributes

In predicting protein-molecular interactions, in addition to intrinsic molecular features such as sequence and structural information, there exists a wealth of additional information that can influence negative sample selection or model training, either directly or indirectly. This additional information can be broadly categorized into two types: those derived from the topological structure of the biomolecular network and those obtained from probing the functional characteristics of biomolecules. Notably, the degree distribution of samples is information sourced from the network's topology. Similarly, the shortest path distance between two nodes is another topological feature. In contrast, subcellular localization and gene ontology (GO) similarity of molecules reflect information gained from exploring biomolecular functions. For detailed calculation methods of subcellular localization and GO similarity between sample node pairs, please refer to Additional file 4: Note S4. These methods, namely shortest path distance, subcellular localization, and GO similarity, are commonly employed to sample negative samples in the ML task of molecular interaction prediction. For instance, when the shortest distance between a pair of nodes exceeds a certain threshold, they

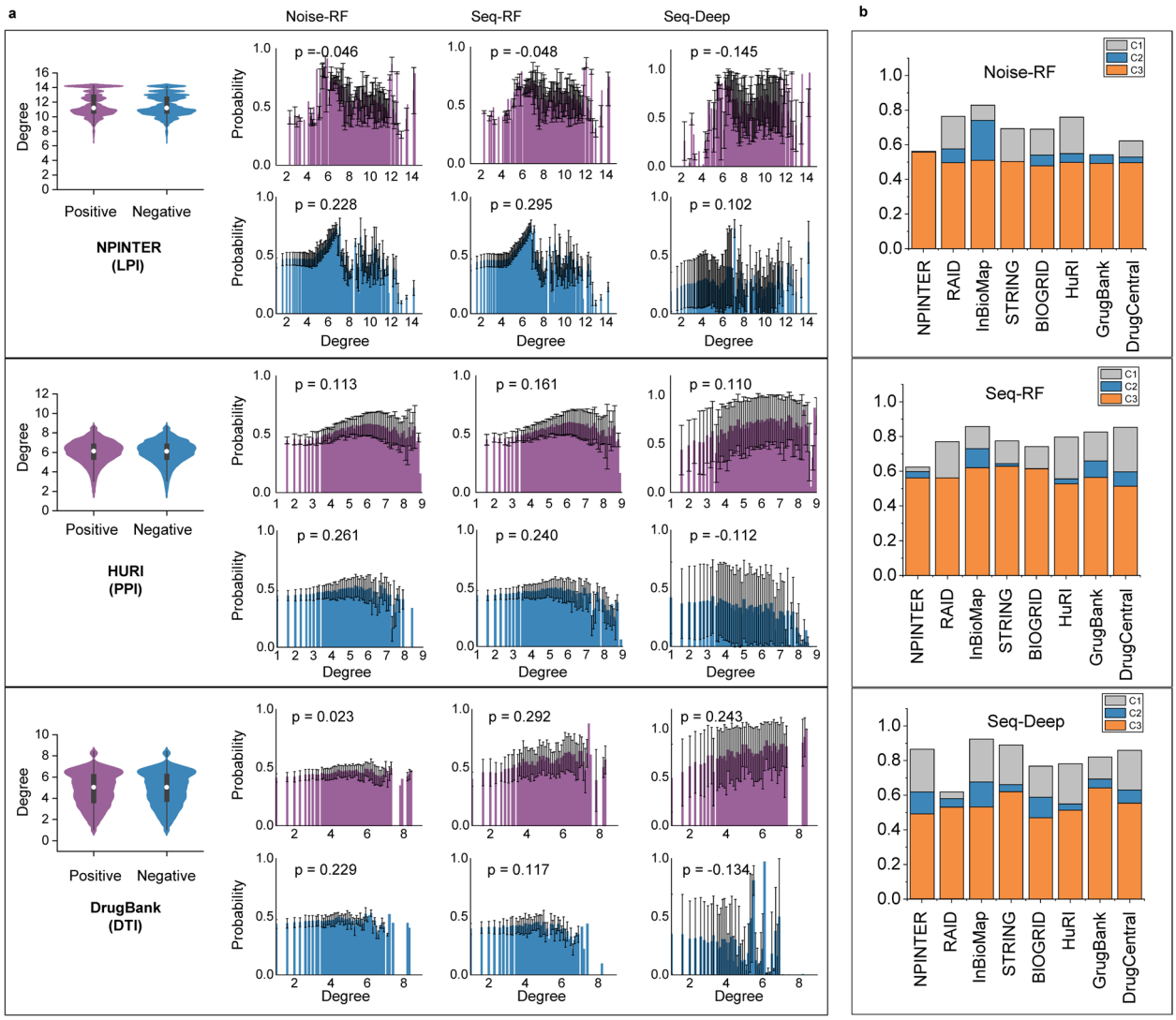

Fig. 3 The DDB constraint assists ML models in focusing on molecular interactive patterns rather than network topology. **a** For clarity in presentation, in this figure, the degree of a pair is computed by taking the logarithm of the sum of the degrees of the two nodes forming a linkage. The violin plots depict the distributions of positive and negative training samples across three datasets. The histograms visualize the relationship between the probability and degree of the samples, where the probabilities are calculated by averaging the probabilities of samples with the same degree (rounded to one decimal place). **b** Comparison of model performance: three types of ML models are evaluated across eight datasets with respect to testing sets C1, C2, and C3 (detailed values provided in Additional file 11: Table S2; results represent mean of $n = 15$ independent runs)

are considered as a negative sample. This also underscores the topology property based on the six degrees of separation hypothesis in complex network theory. In contrast, functional implications derived from GO annotations and subcellular localization are often leveraged to generate negative samples.

Investigating the interplay between the DDB sampling method and the sample degree distribution, along with other pertinent attributes, offers a further vivid illustration of the proposed DDB methodology. As depicted in Fig. 5a, a robust correlation emerges between the degree of a sample and the shortest path distance separating two nodes within that sample, whereas neither subcellular localization nor GO similarity among node pairs exhibits a clear link with degree distribution. Figure 5b elucidates how the DDB method influences the topological information within the training data via sample degree, disrupting the strong correlation between the model's predictions and degree distribution in the test set when random negative sampling is employed. This

Table 3 Transductive model evaluation based on training dataset with negative samples generated by DDB. The three values divided by the two vertical bars relatively are AUROC, Spearman coefficient correlation (ρ) between the positive sample degree and sample predicted interactive probability, and ρ between the negative sample degree and sample predicted interactive probability. Results represent mean of $n = 15$ independent runs

		Noise-RF	Seq-RF	Seq-Deep
LPI	NPInterv4.0	0.548|−0.046|0.228	0.624|−0.048|0.295	0.862|−0.145|0.102
	RAID v2.0	0.764|−0.284|0.634	0.770|−0.286|0.603	0.617|−0.072|−0.116
PPI	InBioMap	0.828|0.077|0.420	0.858|0.144|0.265	0.924|−0.026|−0.285
	STRING	0.693|0.141|0.542	0.774|0.306|0.259	0.889|0.265|−0.227
	BioGRID	0.689|0.171|0.329	0.743|0.250|0.246	0.767|0.122|−0.108
	HuRI	0.759|0.113|0.261	0.797|0.161|0.240	0.782|0.110|−0.112
DTI	DrugBank	0.494|0.023|0.229	0.825|0.292|0.117	0.821|0.243|−0.134
	DrugCentral	0.622|0.300|0.491	0.852|0.354|0.309	0.859|0.206|−0.139

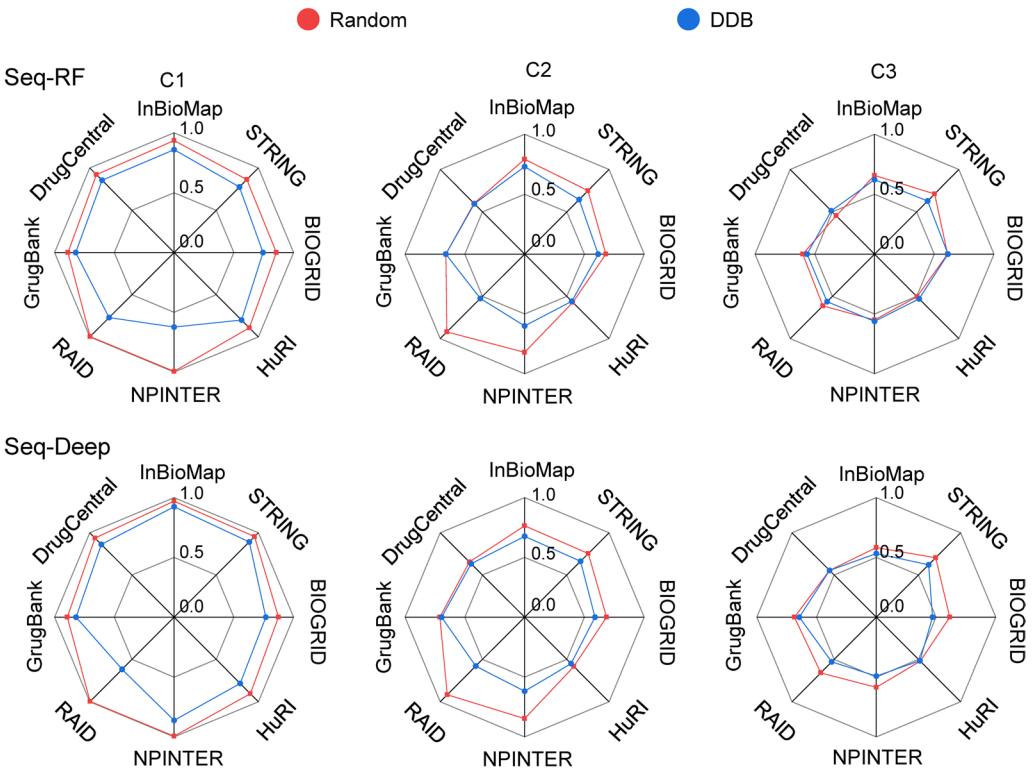

Fig. 4 Comparison of predictive performance between ML models trained on DDB and those trained on randomly sampled datasets

phenomenon is also reflected in the shortest path distance between sample nodes.

Discussion

Our study brings to light a pivotal challenge in utilizing ML models for predicting protein-molecular interactions: the impact of network topology on model predictions. By examining the degree distribution disparity between positive and negative samples, we illustrated that ML models, particularly those employing random negative sampling, frequently learn to predict interactions based on the network structure rather than intrinsic molecular features. This introduces a substantial bias that distorts the evaluation of these models, especially in transductive learning scenarios where test data may overlap with training data.

The results of our transductive evaluation confirm that models, including simple classifiers like Noise-RF, can achieve high predictive performance, but this success is predominantly attributed to their reliance on node

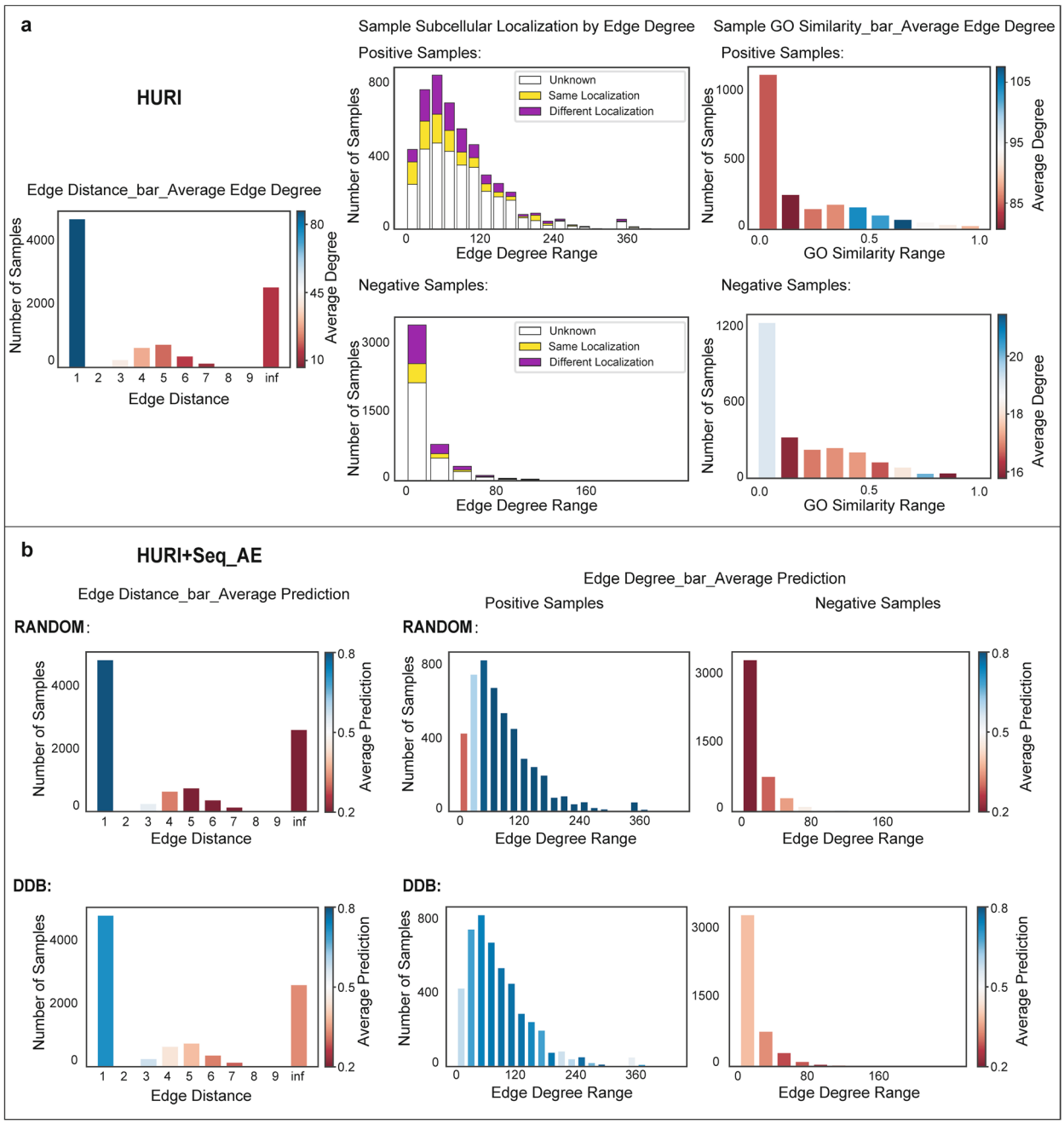

Fig. 5 Analysis of the degree distribution and its correlated features within a test set, generated via random negative sampling for the HURI dataset, under a transductive evaluation paradigm. The figure explores the interplay between sample degree and three key pivotal factors: the shortest path distance between nodes, subcellular localization of node pairs, and gene ontology (GO) similarity of node pairs. Additionally, it evaluates the relationship between predictions made by Seq_AE model on the test set and both degree distribution and shortest path distance, considering scenarios where the training set employs either random or DDB-based negative sampling. **a** *Edge Distance_bar_Average Edge Degree* illustrates the average degree of samples characterized by varying shortest path distances between node pairs within the test set. *Sample Subcellular Localization by Edge Degree* depicts the subcellular localization among node pairs across different degree ranges for both positive and negative samples. *Sample GO Similarity_bar_Average Edge Degree* represents the average degree of samples falling within various GO similarity ranges for both positive and negative samples. **b** *Edge Distance_bar_Average Prediction* showcases the average Seq_AE prediction values for test samples with distinct shortest path distances between node pairs, comparing the impact of random and DDB negative sampling in the training set. *Edge Degree_bar_Average Prediction* displays the average Seq_AE prediction values for positive and negative samples spanning various degree ranges in the test set, under both random and DDB sampling strategies employed during training

degree distribution. The violin plots and histograms in Fig. 2a distinctly show that the models exploit the degree difference between positive and negative pairs, resulting in inflated AUROC values. However, when we adopted an inductive evaluation framework (C1, C2, and C3 sets), which eliminates the influence of shared nodes between the training and test sets, the models' performance dropped significantly—especially on the C3 set, where no nodes were seen during training. This decline in performance, particularly for the Noise-RF model, indicates that the model's predictive capabilities were largely driven by the degree of nodes rather than by meaningful protein-molecular interactions. Consequently, this underscores the overestimation of model generalization ability in previous studies that relied on random negative sampling.

The introduction of the DDB sampling method provides a solution to mitigate this bias. By aligning the degree distribution of negative samples with that of positive samples, we observed a considerable reduction in the correlation between predicted scores and node degree. This suggests that the DDB sampling method effectively removes the confounding influence of network topology, revealing the true capability of ML models to learn protein-molecular interaction patterns from intrinsic features such as sequence information. The experimental results, particularly in the inductive evaluation (C1, C2, C3 sets), show that the performance gap between sequence-based models (Seq-RF, Seq-Deep) and the topology-based Noise-RF model widens, emphasizing that sequence information plays a more significant role when degree bias is eliminated.

However, this improvement comes at the expense of overall predictive performance. As shown in Table 3, the performance of all models, including Seq-RF and Seq-Deep, declines after applying the DDB constraint, especially in transductive prediction settings. This observation implies that, while DDB method effectively reduces bias, it also strips away useful topological information that models may leverage to enhance performance. Whether this is desirable hinges on the application's goals. For tasks where network topology is deemed a valid feature, excluding it may impede performance. However, for applications aiming to predict protein-molecular interactions purely based on molecular properties, removing the influence of network topology is crucial for obtaining more objective and generalizable results.

In light of our findings, several promising avenues for future research have emerged. Firstly, integrating higher-quality and more comprehensive datasets could more effectively capture the intrinsic properties of protein-molecular interactions, thereby reducing noise and enhancing model reliability. Secondly, incorporating additional molecular features—such as post-translational modifications, binding affinities, or other biochemical properties—may provide a more nuanced understanding of interaction dynamics. Third, considering the temporal dynamics of interactions, including time-resolved data and network evolution, could further refine predictions and offer insights into the transient nature of these interactions. These future directions may not only have the potential to enhance the predictive performance of ML models but also to extend the applicability of our proposed DDB sampling strategy.

Conclusions

In this study, we systematically analyzed various random negative sampling strategies within the framework of inferring biomolecular interactions. Across numerous datasets and ML models, we validated and visualized the predictive bias on network topology. Through inductive evaluation, we uncovered a significant reduction in the generative performance of ML models when network influences were eliminated.

A notable limitation of this research is the absence of a proposed effective method to mitigate the bias issue in pair-input ML models, which we intend to explore further in our future studies. We anticipate that our finding will garner attention within the bioinformatics community and contribute positively to the field's progression. When assessing models for linkage prediction, it is imperative to meticulously select the evaluation pipeline to ensure reliable validation performance. Random negative sampling can introduce bias into ML models, thereby necessitating an independent test set for accurate assessment of model performance.

In conclusion, while ML models have exhibited promising potential in predicting protein-molecular interactions, their dependence on network topology often leads to inflated performance estimates. The DDB sampling method offers a straightforward yet effective approach to minimize this bias, albeit while also revealing the limitations of current models in learning from intrinsic molecular features. To enhance the accuracy and applicability of these models in real-world biological research, the development of advanced ML methods that can better capture interaction relationships from inherent molecular features, while circumventing the influence of network topology, will be paramount.

Methods

Datasets

The negative sampling approach was evaluated across three distinct types of protein-centric molecular interaction predictions: lncRNA-protein, protein–protein, and

drug-target interactions. These categories encompass both heterogeneous and homogeneous interaction types, ensuring the broad applicability of the approach and demonstrating its effectiveness across various protein-molecular interaction scenarios, regardless of the nature or type of interacting molecules. To underscore the generalizability of our conclusions, multiple datasets were utilized for each prediction task.

For lncRNA-protein interactions and protein–protein interactions, the datasets were refined through several steps: initially, only both interacting molecules from human were retained. Subsequently, proteins with unavailable sequences were excluded. To ensure dataset non-redundancy, the CD-HIT tool [79] was employed to reduce sequence identity, with a cutoff value of 0.3 for protein–protein interactions and 0.8 for lncRNA-protein interactions. For drug-target interactions, two datasets sourced from DrugBank and DrugCentral were utilized, which were further processed by KG-MTL [42]. A detailed description of all eight datasets is provided in Additional file 2: Note S2 [63–82].

Negative sampling strategy

The positive dataset is constructed by utilizing all interactive pairs within each dataset. Equally important for training and validating ML models is the inclusion of high-quality negative data, which consists of non-interactive pairs. However, acquiring such data is challenging due to its scarcity. In the context of ML for predicting protein-molecular interactions, selecting a designated number of negative samples from the complementary network of the real network is a crucial procedure.

Random sampling

This method involves randomly selecting two nodes from the candidate node set. If the selected pair of nodes does not constitute a validated pair, it is considered as a potential candidate. The number of randomly sampled negative examples is matched to the count of positive examples.

Degree distribution balanced (DDB) sampling

We proposed the DDB sampling strategy to select a negative set with a degree distribution similar to that of the positive set. In this strategy, the pair degree is defined as the sum of the degrees of the two nodes in the pair. The core idea is to generate negative samples by selecting pairs of nodes whose combined degree closely matches the combined degree of the positive (real) pairs, thereby preserving the overall degree distribution in the negative set.

During the sampling process, we utilize a two-step approach to identify suitable negative samples:

Initial matching: For each positive pair, we first search for a non-existing edge in the set of negative candidates with the same degree sum as the positive pair. If a suitable negative pair is found, it is added to the negative set.

Adaptive search: If no suitable negative pair is found with the exact same degree sum, the search continues in a "neighboring degree" approach. Specifically, the algorithm gradually expands the search by checking degree sums that are slightly higher or lower than the current pair's degree sum. This adaptive search continues until a suitable negative pair is found or the search range is exhausted.

The negative sample selection process is designed to be efficient. The precomputed degree distribution is randomly shuffled to ensure diversity in the sampling process. If an exact match is not available, the search for a matching negative pair is adaptive, exploring both the left (lower degree sum) and right (higher degree sum) neighbors in the degree distribution. The number of negative pairs sampled is matched to the number of positive examples to maintain a 1:1 ratio between positive and negative samples.

Machine learning models

Figure 1b depicts the core structure of the pair-input ML model. Regardless of the molecule encoding utilized in the preceding stages, it is crucial to concatenate the features of the pair-wise molecules prior to inputting them into a decision-making classifier. To substantiate our conclusion, we evaluated three ML models: (i) Noise-RF, where molecules are represented by random Gaussian noise vector, and these noise vectors, alongside network interactions, are learned by a random forest classifier. (ii) Seq-RF, combining handcrafted molecular feature extraction with traditional classifier-based models, akin to those utilized in previous studies [12, 19, 83, 84]; (iii) Seq-Deep (deep learning), integrating auto-encoder feature extraction with neural network-based decision models, as described in earlier works [22, 26, 28, 30, 31, 33, 35, 41, 44, 85]. The Noise-RF model serves to assess the influence of distribution discrepancies on ML models.

Noise-RF

The feature vectors of molecules are initialized with Gaussian noise. The length of each type of molecule's vector is consistent with that of the subsequent Seq-RF model. Noise-RF utilizes the random forest as the decision model.

Seq-RF

Initially, we extracted handcrafted molecular features for three types of molecules. For proteins, protein sequences are represented by reduced amino acid sets [5], and the protein's feature vector is created by normalizing the counts of each possible conjoint triad, resulting in a vector length of 343. For LncRNA, normalized 4-mer frequencies are computed from the RNA sequences, yielding a vector length of 256. Drugs are encoded using the MinHashed Atom Pair fingerprint [86] with a radius of 2 and 128 output dimensions.

For a pair of molecules, the two encoded vectors are concatenated into one vector for representing the pair. Based on the molecular representation described, the concatenated vector dimensions for the three tasks are 686 for protein–protein interactions, 599 for lncRNA-protein interactions, and 471 for drug-target interactions. Subsequently, the link prediction task is transformed into a conventional classification task, and we applied the random forest classifier with 500 trees for binary classification.

Seq-Deep

The Seq-Deep approach comprises two channels, each utilizing a neural network for molecular auto-encoding. The auto-encoded features from both channels are concatenated and fed into a multi-layer perceptron for the final decision. To capture the sequence information of proteins and lncRNAs, we employed a 1-dimensional convolutional neural network (CNN). This CNN module learns representations of each residue or base in the sequence, followed by a max-pooling readout function to obtain the overall representation of the protein or lncRNA sequence. For drugs, we utilized a graph convolutional network (GCN) to encode the structural information represented by the drug's SMILES notation. After the GCN, a max-pooling readout function was applied to derive the representation of the drug's molecular graph. A more extensive and detailed explanation of the Seq-Deep method can be found in Additional file 3: Note S3.

Prediction validations

Transductive validation

In transductive validation, the validation set may potentially contain common nodes with the training dataset. This method involves randomly selecting a subset of the experimental interactive pairs to form the validation dataset. To demonstrate the impact of the degree distribution disparity between positive and negative samplings on ML models, we adopted two strategies for generating negative samples for the training dataset. The first strategy randomly pairs nodes from the training set to select negative samples, while the second strategy employs the DDB strategy for sampling negative samples.

For the validation dataset, negative samples were chosen by randomly pairing nodes from the validation set, regardless of the strategies used for the training dataset. This approach facilitates a comparison of predictive outcomes from models trained on both degree-biased and non-biased training datasets.

Inductive validation

To comparatively assess the generative capabilities of ML models trained on datasets using negative sampling and DDB negative sampling strategies, we evaluated their predictive performance on distinct classes of test pairs. Following the approach proposed by Park and Marcotte [9], we utilized Kernighan–Lin algorithm [87] to divide the molecular network into two separate sub-networks with no shared components. The intra-links within each sub-network were categorized as C1 and C3, respectively, while the links connecting the two sub-networks were categorized as C2. We trained the ML models using C1 and subsequently reported their predictive performance on the C2 and C3 datasets. Consequently, the C2 dataset comprises test pairs that share only one protein with the training set, whereas the C3 dataset consists of test pairs that share no proteins with the training set.

Abbreviations

ML	Machine learning
DDB	Degree distribution balanced
AUROC	Area under the receiver operating characteristic
CNN	Convolutional neural network
GCN	Graph convolutional network
LPI	LncRNA-protein interaction
PPI	Protein-protein interaction
DTI	Drug-target interaction
GO	Gene ontology

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12915-025-02231-w.

Additional file 1: Note S1. A survey of studies employing random negative sampling in their predictions of protein-molecular interactions

Additional file 2: Note S2. A detailed description of the eight datasets employed for analyzing three types of protein-molecular interactions

Additional file 3: Note S3. An expanded and thorough explanation of the Seq-Deep method

Additional file 4: Note S4. Detailed information on the calculation methods for subcellular localization and GO similarity between sample node pairs

Additional file 5: Figure S1. The figure depicts the degree distribution of positive and negative samples in the RAID dataset under two different negative sampling strategies: random negative sampling and DDB negative sampling. It also showcases the correlation between sample prediction probability and node degree for three methods, Noise_RF, Seq_RF, and Seq_Deep, under both sampling strategies. The violin plots display the distribution of positive and negative training samples, while

the histograms illustrate the relationship between sample probability and node degree. The probability is derived by averaging the probabilities of samples sharing the same degree (rounded to one decimal place)

Additional file 6: Figure S2. Similar to Additional file 5: Fig.ure S1, it presents the relevant details for the InBioMap dataset

Additional file 7: Figure S3. Similar to Additional file 5: Fig.ure S1, it displays the relevant information for the STRING dataset

Additional file 8: Figure S4. Similar to Additional file 5: Fig.ure S1, it shows the relevant details for the BioGRID dataset

Additional file 9: Figure S5. Similar to Additional file 5: Fig.ure S1, it exhibits the relevant details for the DrugCentral dataset

Additional file 10: Table S1. Inductive model evaluation based on a training dataset with negative samples generated by random sampling. The results are presented based on the model's performance on C1 | C2 | C3 set

Additional file 11: Table S2. Inductive model evaluation based on a training dataset with negative samples generated by the DDB negative sampling method. The results are shown based on the model's performance on C1 | C2 | C3 set

Acknowledgements
Thanks are due to our lab members for their insightful comments during the long project.

Authors' contributions
Z.L. conceived the study, Z.L. and P.L. designed the experiments, P.L., B.S. and G.Z. conducted the experiments, P.L., B.S. and G.Z. analyzed the results. P.L. B.S. and Z.L. wrote and revised the manuscript. All authors read and approved the final manuscript.

Funding
This work was partially supported by the National Natural Science Foundation of China (Nos. 62373216, 92374107); National Key Research and Development Program of China (No. 2020YFA0712402); and the Fundamental Research Funds for the Central Universities (No. 2022JC008).

Data availability
All data generated or analysed during this study are included in this published article, its supplementary information files and the following publicly available repositories:
NPInter v4.0 [article ref [63]; dataset ref [64]]
RAID v2.0 [article ref [65]; dataset ref [66]]
InWeb_IM human PPI network [article ref [67]; dataset ref [68]
STRING v11 [article ref [69]; dataset ref [70]]
BioGRID v3.5.187 [article ref [71]; dataset ref [72]]
HuRI human binary interactome [article ref [73]; dataset ref [74]]
DrugBank v5.0 [article ref [75]; dataset ref [76]]
DrugCentral 2021 [article ref [77]; dataset ref [78]]
The source code implementing the DDBSampling algorithm is permanently archived in Zenodo under https://doi.org/10.5281/zenodo.15303227 and is also accessible at the GitHub repository: https://github.com/zpliulab/DDBSampling.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 3 December 2024 Accepted: 2 May 2025
Published online: 09 May 2025

References

1. Ferre F, Colantoni A, Helmer-Citterich M. Revealing protein–lncRNA interaction. Brief Bioinform. 2016;17(1):106–16.
2. Suresh V, Liu L, Adjeroh D, Zhou X. RPI-Pred: predicting ncRNA-protein interaction using sequence and structural information. Nucleic Acids Res. 2015;43(3):1370–9.
3. Zhang QC, Petrey D, Deng L, Qiang L, Shi Y, Thu CA, Bisikirska B, Lefebvre C, Accili D, Hunter T, Maniatis T. Structure-based prediction of protein–protein interactions on a genome-wide scale. Nature. 2012;490(7421):556–60.
4. Kovács IA, Luck K, Spirohn K, Wang Y, Pollis C, Schlabach S, Bian W, Kim DK, Kishore N, Hao T, Calderwood MA. Network-based prediction of protein interactions. Nat Commun. 2019;10(1):1240.
5. Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, Li Y, Jiang H. Predicting protein–protein interactions based only on sequences information. Proc Natl Acad Sci U S A. 2007;104(11):4337–41.
6. Lei Y, Li S, Liu Z, Wan F, Tian T, Li S, Zhao D, Zeng J. A deep-learning framework for multi-level peptide–protein interaction prediction. Nat Commun. 2021;12(1):5465.
7. Ye Q, Hsieh CY, Yang Z, Kang Y, Chen J, Cao D, He S, Hou T. A unified drug–target interaction prediction framework based on knowledge graph and recommendation system. Nat Commun. 2021;12(1):6775.
8. Yu J, Guo M, Needham CJ, Huang Y, Cai L, Westhead DR. Simple sequence-based kernels do not predict protein–protein interactions. Bioinformatics. 2010;26(20):2610–4.
9. Park Y, Marcotte EM. Flaws in evaluation schemes for pair-input computational predictions. Nat Methods. 2012;9:1134–6.
10. Zhao G, Li P, Qiao X, Han X, Liu ZP. Predicting lncRNA–protein interactions by heterogenous network embedding. Front Genet. 2022;12: 814073.
11. Wekesa JS, Meng J, Luan Y. Multi-feature fusion for deep learning to predict plant lncRNA-protein interaction. Genomics. 2020;112(5):2928–36.
12. Zhou L, Wang Z, Tian X, Peng L. LPI-deepGBDT: a multiple-layer deep framework based on gradient boosting decision trees for lncRNA–protein interaction identification. BMC Bioinformatics. 2021;22:1–24.
13. Zhang SW, Zhang XX, Fan XN, Li WN. LPI-CNNCP: prediction of lncRNA-protein interactions by using convolutional neural network with the copy-padding trick. Anal Biochem. 2020;601: 113767.
14. Han S, Yang X, Sun H, Yang H, Zhang Q, Peng C, Fang W, Li Y. LION: an integrated R package for effective prediction of ncRNA–protein interaction. Brief Bioinform. 2022;23(6):bbac420.
15. Peng L, Wang C, Tian X, Zhou L, Li K. Finding lncRNA-protein interactions based on deep learning with dual-net neural architecture. IEEE/ACM Trans Comput Biol Bioinform. 2021;19(6):3456–68.
16. Peng L, Tan J, Tian X, Zhou L. EnANNDeep: an ensemble-based lncRNA–protein interaction prediction framework with adaptive k-nearest neighbor classifier and deep models. Interdiscip Sci Comput Life Sci. 2022;14(1):209–32.
17. Cheng S, Zhang L, Tan J, Gong W, Li C, Zhang X. DM-RPIs: predicting ncRNA-protein interactions using stacked ensembling strategy. Comput Biol Chem. 2019;83: 107088.
18. Li Y, Sun H, Feng S, Zhang Q, Han S, Du W. Capsule-LPI: a LncRNA–protein interaction predicting tool based on a capsule network. BMC Bioinformatics. 2021;22(1):246.
19. Tian X, Shen L, Wang Z, Zhou L, Peng L. A novel lncRNA–protein interaction prediction method based on deep forest with cascade forest structure. Sci Rep. 2021;11(1):18881.
20. Wekesa JS, Luan Y, Chen M, Meng J. A hybrid prediction method for plant lncRNA-protein interaction. Cells. 2019;8(6):521.
21. Wekesa JS, Meng J, Luan Y. A deep learning model for plant lncRNA-protein interaction prediction with graph attention. Mol Genet Genomics. 2020;295:1091–102.
22. Li X, Han P, Wang G, Chen W, Wang S, Song T. SDNN-PPI: self-attention with deep neural network effect on protein-protein interaction prediction. BMC Genomics. 2022;23(1):474.
23. Zhang L, Yu G, Xia D, Wang J. Protein–protein interactions prediction based on ensemble deep neural networks. Neurocomputing. 2019;324:10–9.
24. Yu B, Chen C, Wang X, Yu Z, Ma A, Liu B. Prediction of protein–protein interactions based on elastic net and deep forest. Expert Syst Appl. 2021;176: 114876.

25. Zhang D, Kabuka M. Multimodal deep representation learning for protein interaction identification and protein family classification. BMC Bioinformatics. 2019;20:1–4.
26. Chen M, Ju CJ, Zhou G, Chen X, Zhang T, Chang KW, Zaniolo C, Wang W. Multifaceted protein–protein interaction prediction based on Siamese residual RCNN. Bioinformatics. 2019;35(14):i305–14.
27. Chen C, Zhang Q, Ma Q, Yu B. LightGBM-PPI: predicting protein-protein interactions through LightGBM with multi-information fusion. Chemometr Intell Lab Syst. 2019;191:54–64.
28. Chen C, Zhang Q, Yu B, Yu Z, Lawrence PJ, Ma Q, Zhang Y. Improving protein-protein interactions prediction accuracy using XGBoost feature selection and stacked ensemble classifier. Comput Biol Med. 2020;123: 103899.
29. Mahapatra S, Sahu SS. Improved prediction of protein–protein interaction using a hybrid of functional-link Siamese neural network and gradient boosting machines. Brief Bioinform. 2021;22(6):bbab255.
30. Mahapatra S, Gupta VR, Sahu SS, Panda G. Deep neural network and extreme gradient boosting based hybrid classifier for improved prediction of protein-protein interaction. IEEE/ACM Trans Comput Biol Bioinform. 2021;19(1):155–65.
31. Zhao L, Wang J, Hu Y, Cheng L. Conjoint feature representation of GO and protein sequence for PPI prediction based on an inception RNN attention network. Mol Ther Nucleic Acids. 2020;22:198–208.
32. Tsubaki M, Tomii K, Sese J. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. Bioinformatics. 2019;35(2):309–18.
33. Nasiri E, Berahmand K, Rostami M, Dabiri M. A novel link prediction algorithm for protein-protein interaction networks by attributed graph embedding. Comput Biol Med. 2021;137: 104772.
34. Wang Y, You ZH, Yang S, Li X, Jiang TH, Zhou X. A high efficient biological language model for predicting protein–protein interactions. Cells. 2019;8(2):122.
35. Ji BY, You ZH, Jiang HJ, Guo ZH, Zheng K. Prediction of drug-target interactions from multi-molecular network based on LINE network representation method. J Transl Med. 2020;18:1–11.
36. Zheng S, Li Y, Chen S, Xu J, Yang Y. Predicting drug–protein interaction using quasi-visual question answering system. Nat Mach Intell. 2020;2:134–40.
37. Wang J, Wang H, Wang X, Chang H. Predicting drug-target interactions via FM-DNN learning. Curr Bioinform. 2020;15(1):68–76.
38. Hu S, Zhang C, Chen P, Gu P, Zhang J, Wang B. Predicting drug-target interactions from drug structure and protein sequence using novel convolutional neural networks. BMC Bioinformatics. 2019;20:1–2.
39. Wang W, Wang Y, Zhang Y, Liu D, Zhang H, Wang X. PPDTS: predicting potential drug–target interactions based on network similarity. IET Syst Biol. 2022;16(1):18–27.
40. Huang K, Xiao C, Glass LM, Sun J. MolTrans: molecular interaction transformer for drug–target interaction prediction. Bioinformatics. 2021;37(6):830–6.
41. Xuan P, Hu K, Cui H, Zhang T, Nakaguchi T. Learning multi-scale heterogeneous representations and global topology for drug-target interaction prediction. IEEE J Biomed Health Inform. 2021;26(4):1891–902.
42. Ma T, Lin X, Song B, Yu PS, Zeng X. Kg-mtl: knowledge graph enhanced multi-task learning for molecular interaction. IEEE Trans Knowl Data Eng. 2022;35(7):7068–81.
43. Yu Z, Lu J, Jin Y, Yang Y. KenDTI: an ensemble model for predicting drug-target interaction by integrating multi-source information. IEEE/ACM Trans Comput Biol Bioinform. 2021;18(4):1305–14.
44. Xu X, Xuan P, Zhang T, Chen B, Sheng N. Inferring drug-target interactions based on random walk and convolutional neural network. IEEE/ACM Trans Comput Biol Bioinform. 2021;19(4):2294–304.
45. Zhao Q, Zhao H, Zheng K, Wang J. HyperAttentionDTI: improving drug–protein interaction prediction by sequence-based deep learning with attention mechanism. Bioinformatics. 2022;38(3):655–62.
46. Zhao Q, Duan G, Zhao H, Zheng K, Li Y, Wang J. Gifdti: prediction of drug-target interactions based on global molecular and intermolecular interaction representation learning. IEEE/ACM Trans Comput Biol Bioinform. 2022;20(3):1943–52.
47. Liu Z, Chen Q, Lan W, Pan H, Hao X, Pan S. GADTI: graph autoencoder approach for DTI prediction from heterogeneous network. Front Genet. 2021;12: 650821.
48. El-Behery H, Attia AF, El-Fishawy N, Torkey H. Efficient machine learning model for predicting drug-target interactions with case study for COVID-19. Comput Biol Chem. 2021;93: 107536.
49. Chu Y, Kaushik AC, Wang X, Wang W, Zhang Y, Shan X, Salahub DR, Xiong Y, Wei DQ. DTI-CDF: a cascade deep forest model towards the prediction of drug-target interactions based on hybrid features. Brief Bioinform. 2021;22(1):451–62.
50. Ye Q, Zhang X, Lin X. Drug-target interaction prediction via graph autoencoder and multi-subspace deep neural networks. IEEE/ACM Trans Comput Biol Bioinform. 2022;20(5):2647–58.
51. Eslami Manoochehri H, Nourani M. Drug-target interaction prediction using semi-bipartite graph model and deep learning. BMC Bioinformatics. 2020;21:1–6.
52. Ren ZH, You ZH, Zou Q, Yu CQ, Ma YF, Guan YJ, You HR, Wang XF, Pan J. DeepMPF: deep learning framework for predicting drug–target interactions based on multi-modal representation with meta-path semantic analysis. J Transl Med. 2023;21:48.
53. Zhang P, Wei Z, Che C, Jin B. DeepMGT-DTI: transformer network incorporating multilayer graph information for drug–target interaction prediction. Comput Biol Med. 2022;142: 105214.
54. Öztürk H, Özgür A, Ozkirimli E. DeepDTA: deep drug–target binding affinity prediction. Bioinformatics. 2018;34(17):i821–9.
55. Lee I, Keum J, Nam H. DeepConv-DTI: prediction of drug-target interactions via deep learning with convolution on protein sequences. PLoS Comput Biol. 2019;15(6): e1007129.
56. Wu Y, Gao M, Zeng M, Zhang J, Li M. BridgeDPI: a novel graph neural network for predicting drug–protein interactions. Bioinformatics. 2022;38(9):2571–8.
57. Kim Q, Ko JH, Kim S, Park N, Jhe W. Bayesian neural network with pretrained protein embedding enhances prediction accuracy of drug-protein interaction. Bioinformatics. 2021;37(20):3428–35.
58. Wan X, Wu X, Wang D, Tan X, Liu X, Fu Z, Jiang H, Zheng M, Li X. An inductive graph neural network model for compound–protein interaction prediction based on a homogeneous graph. Brief Bioinform. 2022;23(3):bbac073.
59. Peng J, Li J, Shang X. A learning-based method for drug-target interaction prediction based on feature representation learning and deep neural network. BMC Bioinformatics. 2020;21(Suppl 13):394.
60. Shen Y, Zhang Y, Yuan K, Li D, Zheng H. A knowledge-enhanced multi-view framework for drug-target interaction prediction. IEEE Trans Big Data. 2021;8(5):1387–98.
61. Hu S, Xia D, Su B, Chen P, Wang B, Li J. A convolutional neural network system to discriminate drug-target interactions. IEEE/ACM Trans Comput Biol Bioinform. 2019;18(4):1315–24.
62. Park Y, Marcotte EM. Revisiting the negative example sampling problem for predicting protein–protein interactions. Bioinformatics. 2011;27(21):3024–8.
63. Teng X, Chen X, Xue H, Tang Y, Zhang P, Kang Q, Hao Y, Chen R, Zhao Y, He S. NPInter v4.0: an integrated database of ncRNA interactions. Nucleic Acids Res. 2020;48(D1):D160–5.
64. Teng X, Chen X, Xue H, Tang Y, Zhang P, Kang Q, Hao Y, Chen R, Zhao Y, He S. NPInter v4.0 dataset . Institute of Biophysics, Chinese Academy of Sciences; 2020. http://bigdata.ibp.ac.cn/npinter4/download/
65. Yi Y, Zhao Y, Li C, Zhang L, Huang H, Li Y, Liu L, Hou P, Cui T, Tan P, Hu Y. RAID v2.0: an updated resource of RNA-associated interactions across organisms. Nucleic Acids Res. 2017;45(D1):D115–8.
66. Yi Y, Zhao Y, Li C, Zhang L, Huang H, Li Y, Liu L, Hou P, Cui T, Tan P, Hu Y. RAID v2.0 dataset . RNA Society; 2017. http://www.rna-society.org/raid/
67. Li T, Wernersson R, Hansen RB, Horn H, Mercer J, Slodkowicz G, Workman CT, Rigina O, Rapacki K, Stærfeldt HH, Brunak S. A scored human protein–protein interaction network to catalyze genomic interpretation. Nat Methods. 2017;14(1):61–4.
68. Rapacki K, Brunak S, Workman CT, et al. InWeb_IM human PPI network dataset . Intomics/Inbiomap; 2017. https://www.intomics.com/inbio/map
69. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, Jensen LJ. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019;47(D1):D607–13.
70. STRING Consortium. STRING v11 protein–protein association network data . STRING; 2019. https://string-db.org/cgi/download.pl

71. Oughtred R, Stark C, Breitkreutz BJ, Rust J, Boucher L, Chang C, Kolas N, O'Donnell L, Leung G, McAdam R, Zhang F. The BioGRID interaction database: 2019 update. Nucleic Acids Res. 2019;47(D1):D529–41.
72. The BioGRID Consortium. BioGRID 3.5.187 interaction dataset . BioGRID; 2019. https://downloads.thebiogrid.org/BioGRID/Release-Archive/BIOGRID-3.5.187/
73. Luck K, Kim DK, Lambourne L, Spirohn K, Begg BE, Bian W, Brignall R, Cafarelli T, Campos-Laborie FJ, Charloteaux B, Choi D. A reference map of the human binary protein interactome. Nature. 2020;580(7803):402–8.
74. Luck K, Kim DK, Lambourne L, et al. HuRI human binary protein interactome data . Interactome Atlas; 2020. http://interactome-atlas.org/download
75. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2018;46(D1):D1074–82.
76. Wishart DS, Feunang YD, Guo AC, et al. DrugBank 5.0 dataset . University of Alberta/Bioinformatics Group; 2018. https://go.drugbank.com/releases/5-0-0
77. Avram S, Bologa CG, Holmes J, Bocci G, Wilson TB, Nguyen DT, Curpan R, Halip L, Bora A, Yang JJ, Knockel J. DrugCentral 2021 supports drug discovery and repositioning. Nucleic Acids Res. 2021;49(D1):D1160–9.
78. Avram S, Bologa CG, Holmes J, et al. DrugCentral 2021 dataset . DrugCentral Consortium; 2021. https://drugcentral.org/download
79. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28(23):3150–2.
80. Zhao L, Wang J, Li Y, Song T, Wu Y, Fang S, Bu D, Li H, Sun L, Pei D, Zheng Y. NONCODEV6: an updated database dedicated to long non-coding RNA annotation in both animals and plants. Nucleic Acids Res. 2021;49(D1):D165–71.
81. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019;47(D1):D506–15.
82. Eyre TA, Ducluzeau F, Sneddon TP, Povey S, Bruford EA, Lush MJ. The HUGO gene nomenclature database, 2006 updates. Nucleic Acids Res. 2006;34(suppl_1):D319–21.
83. Das S, Chakrabarti S. Classification and prediction of protein–protein interaction interface using machine learning algorithm. Sci Rep. 2021;11(1):1761.
84. Muppirala UK, Honavar VG, Dobbs D. Predicting RNA-protein interactions using only sequence information. BMC Bioinformatics. 2011;12:1–11.
85. Gao KY, Fokoue A, Luo H, Iyengar A, Dey S, Zhang P. Interpretable drug target prediction using deep neural representation. In: Proc IJCAI. 2018;3371–3377.
86. Capecchi A, Probst D, Reymond JL. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. J Cheminform. 2020;12:1–5.
87. Kernighan BW, Lin S. An efficient heuristic procedure for partitioning graphs. Bell Syst Tech J. 1970;49(2):291–307.

Publisher's Note