Received: 12 December 2022 | Revised: 15 February 2023 | Accepted: 7 March 2023

DOI: 10.1002/prot.26486

RESEARCH ARTICLE

PROTEINS WILEY

Prediction of protein–protein interactions using sequences of intrinsically disordered regions

Gözde Kibar | Martin Vingron

Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Ihnestrasse 63-73, 14195 Berlin, Germany

Correspondence
Martin Vingron, Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Ihnestrasse 63-73, 14195, Berlin, Germany.
Email: vingron@molgen.mpg.de

Funding information
Bundesministerium für Bildung und Forschung; Max-Planck-Gesellschaft

Abstract

Protein–protein interactions (PPIs) play a crucial role in numerous molecular processes. Despite many efforts, mechanisms governing molecular recognition between interacting proteins remain poorly understood and it is particularly challenging to predict from sequence whether two proteins can interact. Here we present a new method to tackle this challenge using intrinsically disordered regions (IDRs). IDRs are protein segments that are functional despite lacking a single invariant three-dimensional structure. The prevalence of IDRs in eukaryotic proteins suggests that IDRs are critical for interactions. To test this hypothesis, we predicted PPIs using IDR sequences in candidate proteins in humans. Moreover, we divide the PPI prediction problem into two specific subproblems and adapt appropriate training and test strategies based on problem type. Our findings underline the importance of defining clearly the problem type and show that sequences encoding IDRs can aid in predicting specific features of the protein interaction network of intrinsically disordered proteins. Our findings further suggest that accounting for IDRs in future analyses should accelerate efforts to elucidate the eukaryotic PPI network.

KEYWORDS
intrinsic disorder, intrinsically disordered proteins, machine learning, prediction, protein–protein interactions

1 | INTRODUCTION

For a long time, interactions among proteins were seen as interactions among almost rigid bodies. However, many proteins contain regions lacking a single invariant structure and which are highly flexible. Such regions are termed intrinsically disordered regions (IDRs) and a protein containing one or more IDRs is an intrinsically disordered protein (IDP). As a consequence of their conformational flexibility, IDPs can have a multitude of binding modes by acquiring different conformations based on the shape of the target protein. For example, a single IDP can fold upon binding to targets through a mechanism called disorder to order transition.[1] Chen et al. (2021)[2] review methods for predicting the site in an IDP which is responsible for interaction with another protein. It has also been hypothesized that different modes of interaction behaviors of IDPs can be encoded by the sequence features of IDRs. For example, an IDR can select interaction partners that have similar IDR sequences while not targeting other IDRs with more distinct sequences.[3] Two IDPs with multiple binding sites can also directly interact with each other in a dynamic equilibrium.[4]

In spite of IDRs playing important roles in a wide variety of cellular functions and target selection, how IDPs interact with each other remains to be elucidated. While there already exist computational

methods predicting protein–protein interactions (PPIs) from sequence (see Casadio et al., 2022;[5] Ganapathiraju & Dunham, 2021[6] for review), few efforts have been made to capture the IDP-specific interactions. Also, interactions between IDPs are highly underrepresented in structure-based docking analysis due to their lack of well-defined structures. Many of the known structure-based PPI prediction algorithms[5] rely on interactions using the structures in the Protein Data Bank[7] which are not suitable for the interaction dynamics of IDPs. On the other hand, most of the sequence-based PPI prediction algorithms do not make a distinction between IDPs and structured proteins. Yet, IDRs are frequently characterized by their low-complexity amino acid composition, such that it appears reasonable to assume that sequence derived features will enable a machine learning classifier to predict interacting IDPs from their sequences. Thus, current approaches do not reflect the intrinsic disorder-based binding modes and IDPs still present a challenge for PPI prediction. Therefore, we want to investigate whether IDR sequences can help in predicting interactions in the human IDP-specific interaction network.

PPI prediction is a case of a pair-input prediction problem. This class of problems is particularly difficult in that both the exact formulation of the prediction task, as well as the appropriate testing procedures are tricky.[8] In terms of the exact formulation of the problem, Park and Marcotte distinguish whether, from the two proteins for which we want to predict if they interact, both, one, or neither have been part of the training set. Clearly, prediction success is likely to differ according to these classes. Furthermore, there are numerous pitfalls in the choice of negative sets, in balancing positive and negative sets, and due to the hub structure of biological networks.[9] In this situation, Ganapathiraju and Dunham recently re-implemented many PPI prediction tools from the literature and, unsurprisingly, found that their performance is overstated in the original publications.[6]

Here, we propose to break down the general question of predicting PPIs into two more specific problems. The asymmetric problem is defined as the problem of predicting new interactors for a protein that has, usually together with some known interactors, been part of the training set. The symmetric problem is the problem of predicting whether two proteins, neither of which was part of the training set, interact. We trained a random forest (RF) model for the asymmetric problem. For the symmetric problem, we still use random forests but, in a way, so as to enforce that the classification result is independent of the order in which the two proteins are provided. We focus on studying the interactions between IDPs and examine the advantage which the IDR information provides towards predicting interactions.

Our study reveals that IDR sequences aid in differentiating between interacting and non-interacting proteins compared to entire and non-IDR sequences in corresponding IDPs for both of the problem types. Our prediction method for the asymmetric problem improves over IDPpi which, to the best of our knowledge, is the only machine learning algorithm developed specifically to predict protein interactions that involve IDPs.[10] We provide evidence that the symmetric prediction problem is not a trivial problem and good accuracy is not easily achieved which again underlines the importance of distinguishing these two problem types. Our design of the algorithms allows users to predict the interactions of two given IDPs as well as listing the best candidate target proteins for a given single IDP.

2 | METHODS

2.1 | Dataset preparation

Positive interactions between IDPs have been curated from Perovic et al., 2018.[10] They developed "IDPpi" to predict binary interactions that involve IDPs. They collected IDPs from the DisProt database[11] and extracted the interactions involving at least one IDP from the HIPPIE[12] interaction database. Similar sequences have been removed using CD-Hit.[13] The dataset consisted of 19 837 interactions of 5989 human proteins. Protein sequences of IDPs in the dataset were retrieved from UniProt.[14] In this study, we focus on interactions between two IDPs. To this end, we identified IDR regions in each sequence and extracted interactions between proteins with IDRs longer than 15 amino acids from the curated IDPpi network. The final IDP-IDP interaction network we work with consists of a total of 5535 interactions between proteins with at least one IDR of at least 15 amino acids.

2.2 | Training and test data, negative sampling

For the selection of test and training interactions, nodes of the final PPI network (proteins) are partitioned randomly into training (90%) and test nodes (10%) (Figure 1). After the nodes were partitioned, we went through the neighbors of each training node in the network. If the neighbor of a training node is a training node as well, the edge between two training nodes is considered a training positive interaction.

It has become common practice to assume that a pair of proteins for which we do not know of an interaction actually does not interact,[15] that is, such a pair constitutes a negative example. Of course, the number of negative pairs will by far outnumber the positive pairs such that one needs to subsample them in order to create a balance between positive and negative pairs. Exactly how one subsamples a negative training set has a large impact on performance evaluation of the machine learning models. In particular, hub proteins in the training positive dataset can dominate the network and create a bias towards hub proteins in the training process. To avoid this, we follow the recommendation of Westhead et al., 2010[9] and perform a balanced sampling. This approach aims at creating a negative set where the nodes have equal degrees as in the positive set. The procedure has been implemented in the BRS-nonint program and it randomly subsamples a negative training dataset from the noninteracting pairs, such that the degree of each protein in the positive training dataset equals the one in the negative training dataset. Note that this method is recommended for subsampling the negative training set but not for the negative test set.[15]

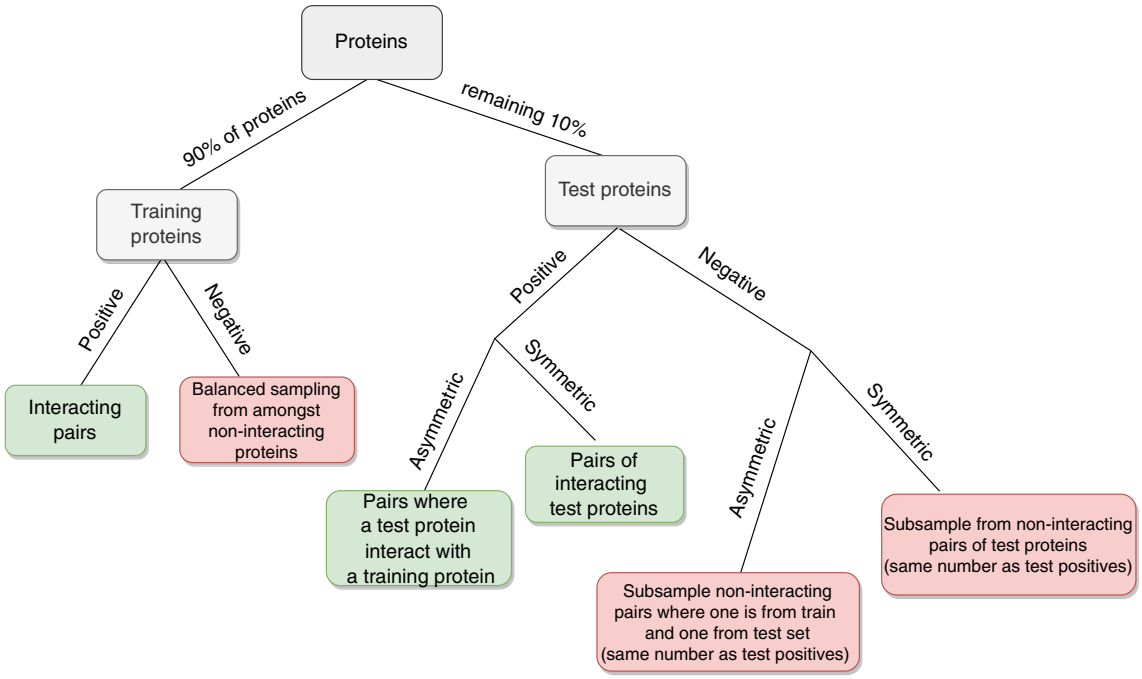

FIGURE 1 Workflow of the sampling process. First, proteins in the dataset are randomly split into test and training proteins. Interactions between training proteins are taken as a positive training set, whereas balanced sampling is applied to generate a negative training set. For the test sets, first we extracted all the interactions and noninteractions involving at least one test protein, then we separated the test pairs from each other depending on the problem type. Symmetric test pairs are test pairs involving one test and one training pair, whereas asymmetric test pairs are interactions between two test proteins.

For the selection of a positive test set, edges involving the test nodes are taken as an initial positive test set. For the selection of the negative test set, first we removed negative training interactions from the set of all negative PPIs. Then we distinguish two cases reflecting the two different questions asked in the asymmetric and the symmetric problem. For the asymmetric problem, the aim is to have a test set consisting of interactions between one known and one unknown protein. Consequently, we extracted the interactions that share only one node with the training nodes from the initial positive test set.

For the symmetric problem, the aim is to test on interactions between two unseen proteins. Hence, we extracted interactions such that none of the components in the pair are present in the training set, that is, both come from the test set. With this distinction in the design of the test sets according to problem type, we aim at delineating the most appropriate validation scheme for the respective classifier. We consider this a particularly challenging test scenario since we do not test on queries which the classifier has seen before. Finally, we randomly subsampled a negative test set such that the positive to negative test set ratio is 1:1.

2.3 | Definition of features

Our classifier is based on sequence derived features with different versions tested for different subsequences of a protein sequence. For each protein, IDR sequences are annotated using PONDR®-VSL2.[16] After filtering out the proteins with all disordered subregions shorter than 15 amino acids, identified IDR sequences were concatenated into one sequence for each protein. In order to assess the predictive performance of IDR sequences, we also compare using (i) the entire protein sequence, or (ii) the non-IDR part of the sequences. For the non-IDR sequences that are shorter than 15 amino acids, we randomly selected a start position in the entire protein and chose a stretch of subsequence of corresponding IDR length.

An amino acid sequence can be represented as a feature vector consisting of numerical features. In this study, we used the same features utilized by Perovic et al., 2018.[12] They used Pseudo-amino acid composition (PAAC)[17] and dipeptide composition (DC) to extract features from the sequences. Additional features like Moreau-Broto autocorrelation (AC),[18] Quasi-sequence-order (QSO)[19] and composition[20] were calculated using protr package.[21]

Autocorrelation features are helpful in representing protein sequences without completely losing their sequence-order information. The AC is another autocorrelation feature that was originally proposed to predict membrane protein types. It calculates the product of feature values encoding each peptide separated by a distance and averages these values across the peptide sequence. For the calculation of product of feature values, we used disorder related feature measures as proposed by Perovic et al., 2018[10] and additional feature values such as hydrophobicity, average flexibility indices, polarizability and the amino acid free energy of solution in water.

Similar to autocorrelation features, QSO also consider the effects of sequence order calculates physicochemical distances between residues based on two physicochemical 20×20 distance matrices: Schneider-Wrede[22] and Grantham physicochemical distance matrix.[23] These distance matrices are computed based on the residue properties of hydrophobicity, hydrophilicity, polarity, and side-chain volumes.

Finally, we computed composition which is a mapping based method which maps every amino acid to a particular property and calculates the percentage of amino acids with a particular property in a sequence. These properties include hydrophobicity, normalized Van der Waals Volume, polarity, polarizability and charge. In this way, a total 541-dimensional feature vector is used to represent a protein sequence (Table 1).

TABLE 1 Features considered in the feature extraction step

Feature	Feature type	Dimension
Pseudo amino acid composition	Distance-based	25
Normalized Moreau-Broto autocorrelation	Distance-based	45
Quasi-sequence-order	Distance-based	50
Composition	Mapping-based	21
Dipeptide composition	Sequence-based	400

2.4 | Model, training, and performance evaluation

A particular difficulty in a pair-input problem is how to fuse the feature vectors of the two proteins into a single feature vector. For the asymmetric problem we subtract the feature vector of the unknown protein from the feature vector of the known protein. The hope is that this will reflect potential interactions between two proteins. After subtracting the feature vectors, we train a RF model for each input sequence type (Figure 2).

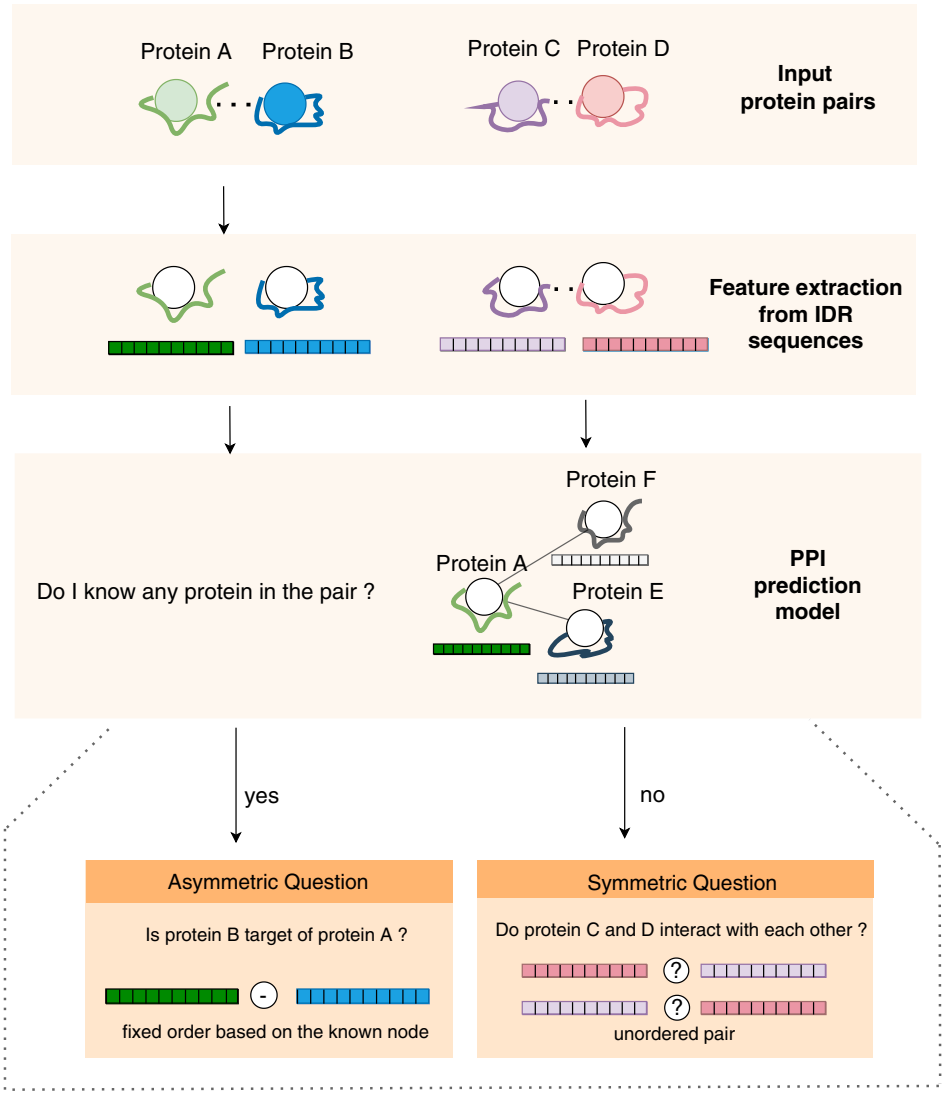

FIGURE 2 Workflow of our method. Given a set of protein sequences, we first extract features from the corresponding IDR sequences. Following this, we identify the problem type and use appropriate training models to predict the interaction between proteins.

In the symmetric problem, we predict the interactions between two unknown proteins and therefore it is not possible to fix the order of the input pairs. This led us to enforce symmetry in the design of the predictor such that the prediction outcome of both orderings of the two input proteins would be the same. To this end, we combine two machine learning models. First the two prediction outcomes of both orderings were computed using the asymmetric RF model that uses subtraction as feature fusion technique. Then, we additionally trained a second RF model using absolute subtraction as an operator. Together, this provides us with three votes, one each for the two input orders to the asymmetric classifier and a third one from the classifier with absolute difference as input. Finally, we combine the outcomes of the three predictors by a majority vote. Together, we call this the ensemble model.

Two parameters of the asymmetric RF model namely min_samples_split (the minimum number of samples required to split an internal node) and max_depth (the maximum depth of the tree) were optimized. Optimization was done based on one random 90:10 split by obtaining the minimal gap between training and test accuracy through a grid search in a parameter space. After optimizing these two parameters we again randomly partitioned the nodes of the curated PPI network to obtain 10 different training and test sets.

We evaluated the performance of both the asymmetric and symmetric model on test sets specific to problem types (see Figure 1). The area under the ROC curve (AUC) scores of all tests were averaged to obtain the overall performance of the predictors. To measure the performance of the model, we used the AUC, accuracy, F score, precision and recall scores.

2.5 | Comparison with other methods

Firstly, we compared our method to IDPpi[10] which is a sequence based RF model that predicts interactions of IDPs. Like many other methods, IDPpi uses features derived from the amino acid sequence, and it concatenates feature vectors generated from the entire sequences of the two proteins. IDPpi follows the discipline suggested by Park and Marcotte (2011)[15] in using balanced sampling for training and random sampling from negatives for testing.

In the original paper, IDPpi achieved AUC score of 0.745. The authors also reported that IDPpi performs better than the prediction algorithms reported in (Martin et al., 2005),[24] (Guo et al., 2008),[25] and (Shen et al., 2007)[26] such that we did not ourselves evaluate those methods again.

We also wanted to compare our method to the recently published state-of-the-art approach D-SCRIPT[27] which is a deep learning method developed for predicting PPIs. It can be trained using only protein sequences and uses 150 convolutional layers in total to predict the interaction between the input proteins. D-SCRIPT was trained originally on 38 345 human proteins and reached an AUC score of 0.833 on the human PPI network. It allows users to choose a dataset to train the model from scratch or provides a pre-trained human model. We evaluated the performance of the pre-trained model as well as training the model from scratch. We had to decrease the number of convolutional layers in the model from 150 to 15 in order to stay within our limits on the available memory which is 1 terabyte.

Many of the sequence-based approaches suffer from the unclear prediction questions. For example, many of the PPI prediction tools including IDPpi[10] and DeepPPI[28] produce results which depend on the order in which two input proteins were provided. This is due to concatenation feature vectors of protein pairs in the training process which is keeping the order information and therefore not a symmetric operation. This creates asymmetrical results and therefore leads to misinterpretation of the results, thus making it difficult for the users to choose the prediction algorithm that is suited to the particular problem they have. Given that there is no known protein in the test pairs for the symmetrical problem, in our design of a classifier for the symmetric problem we will try to avoid the dependence of the input order in which the proteins are supplied to the classifier.

In addition, methods that use the random sampling approach for the selection of negative training set including D-SCRIPT and DeepPPI, can lead to a bias in their reported accuracies.[9]

3 | RESULTS

Instead of predicting whether two IDPs, irrespective of whether we have seen either or both in the training set, interact, we introduce two different machine learning problems. The asymmetric problem asks whether a protein that was in the training set interacts with a query protein. The symmetric problem asks whether two proteins, neither of which was in the training set, interact. We have designed two machine learning procedures for these tasks (Figure 2). Our methods extract IDR sequences from the given proteins and capture the physicochemical properties of the corresponding IDR sequences via a large set of sequence derived features. Since we are particularly interested in the contribution of IDRs towards PPI, we will also study the effect of focusing the physicochemical properties on the IDR regions. Corresponding to the two problems we also designed two setups for testing.

3.1 | The asymmetric model: Interaction partners for a known protein

After testing a number of machine learning algorithms, we decided on a RF classifier. The RF method gave good results, is simple, and the parameters can be chosen so as to avoid overtraining. From the IDPs, the IDR regions of minimally 15 amino acids were extracted and, for each of the two proteins, those regions were concatenated. Then a vector of 541 statistical and physicochemical characteristics was computed for these concatenated IDR sequences as described in Methods, and the vector for the unknown protein was subtracted from the vector for the known protein. The RF classifier was trained on those vectors representing the difference between features.

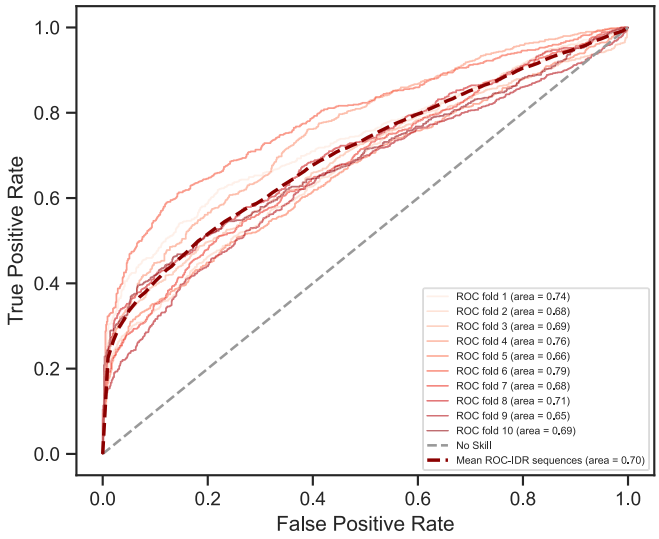

FIGURE 3 ROC AUC scores of the 10-fold cross validation of IDR sequences in the asymmetric model.

In order to create independent test sets that reflect our asymmetric problem, we split our network randomly (90%:10%) into training and test sets. We repeated this process 10 times and, in each split, making sure that test pairs share only one protein with the training set. Sampling from the negative set was done as described in Methods. The performance of the RF classifier was evaluated on these 10 independent test sets. The proposed RF model reached an average 10-fold cross validation AUC score of 0.70 (Figure 3). Thus, using IDR sequences alone it is indeed possible to predict interactions between a known IDP and an unknown IDP with about the accuracy that the best alternative methods would achieve.

3.2 | IDRs are better predictors of interaction than entire or non-IDR sequences in the asymmetric problem

The program IDPpi as the other machine learning algorithm based on disordered proteins does not distinguish between disordered and ordered segments of the protein sequence. Instead, IDPpi uses the entire protein sequence to compute the feature vector for a protein. Above, we showed the predictive performance of our algorithm based on IDR sequences. Thus, the question arises whether the focus on IDRs aids the prediction of interactions. To answer this, we compare the performance of IDRs to two alternatives: (1) Computing the feature vectors from the entire sequence; and (2) computing the feature vector from the non-IDR part of the sequence as described in Methods. For both options we trained a RF model. One would expect that an entire protein sequence would contain more information than an IDR sequence, which in turn would be more informative than its parts, be it the IDRs or the non-IDR part. Contrary to this intuition, Figure 4A shows that the average ROC curve for the IDR-based classifier dominates the other two ROC curves. Since this is based on our 10 independent test sets, Figure 4B shows boxplots over the 10 individual AUC-values for each option, with the IDR based one significantly better than the other two, which are not significantly different from each other in terms of their AUC. We conclude that IDRs are particularly informative with respect to protein–protein interaction among IDPs.

Since the IDPs appear to provide crucial information towards interaction prediction, the question arises whether the most predictive features in the RF classifier trained on the IDR sequences are related to disorder in proteins. We therefore ranked the features according to their importance for the RF classifier under the Python scikit-learn package[29] (Figure 4C). At the top of the list there is the frequency of positively charged residues (Lys, Arg) which has been observed to play a key role in disorder.[30,31] Position 4 is occupied by the frequency of the dipeptide Ser-Pro and the frequency of Pro is on position 6, in line with the general description of IDRs.[32]

3.3 | Prediction performance on the symmetric problem is low

The symmetric problem tries to predict whether two IDPs unknown to the classifier interact. Recall (see Methods) that for the symmetric classifier we implemented a majority vote among the two asymmetric classifiers and a third classifier, where we use the absolute difference of the feature vectors, and which thus is symmetric in the input. We wanted to assess the performance of the IDR, entire and non-IDR sequences using a more stringent criterion. We again generated 10 new independent test sets without overlap with the corresponding training sets. Input order plays no role for this type of problem. We evaluated the performances by feeding entire, non-IDR or IDR sequences to train symmetric models and testing the models using 10-fold CV.

Based on the averaged 10-fold CV AUC scores, the best performing model was the symmetric model trained on IDR sequences which reached a mean AUC score of 0.54 (Figure 5). Although this finding is consistent with the performance of the IDRs for the asymmetric problem, this is still only marginally better than random classification performance. Although the algorithm we propose for the symmetric problem is the result of many tests and experiments we did, we cannot exclude that our algorithm is the cause of the failure. Nevertheless, this failure indicates that the symmetric problem is indeed very hard, which in turn underlines the importance of distinguishing the two problem types.

3.4 | State of the art methods for comparison

We proceed to compare the asymmetric and symmetric model to two existing approaches for sequence-based PPI prediction. IDPpi concatenates the feature vectors of the two input IDPs, which is not a symmetric operation. Indeed, the program returns different results depending on the order in which the proteins are provided. Therefore, IDPpi can serve as a competitor to our asymmetric predictor, but we

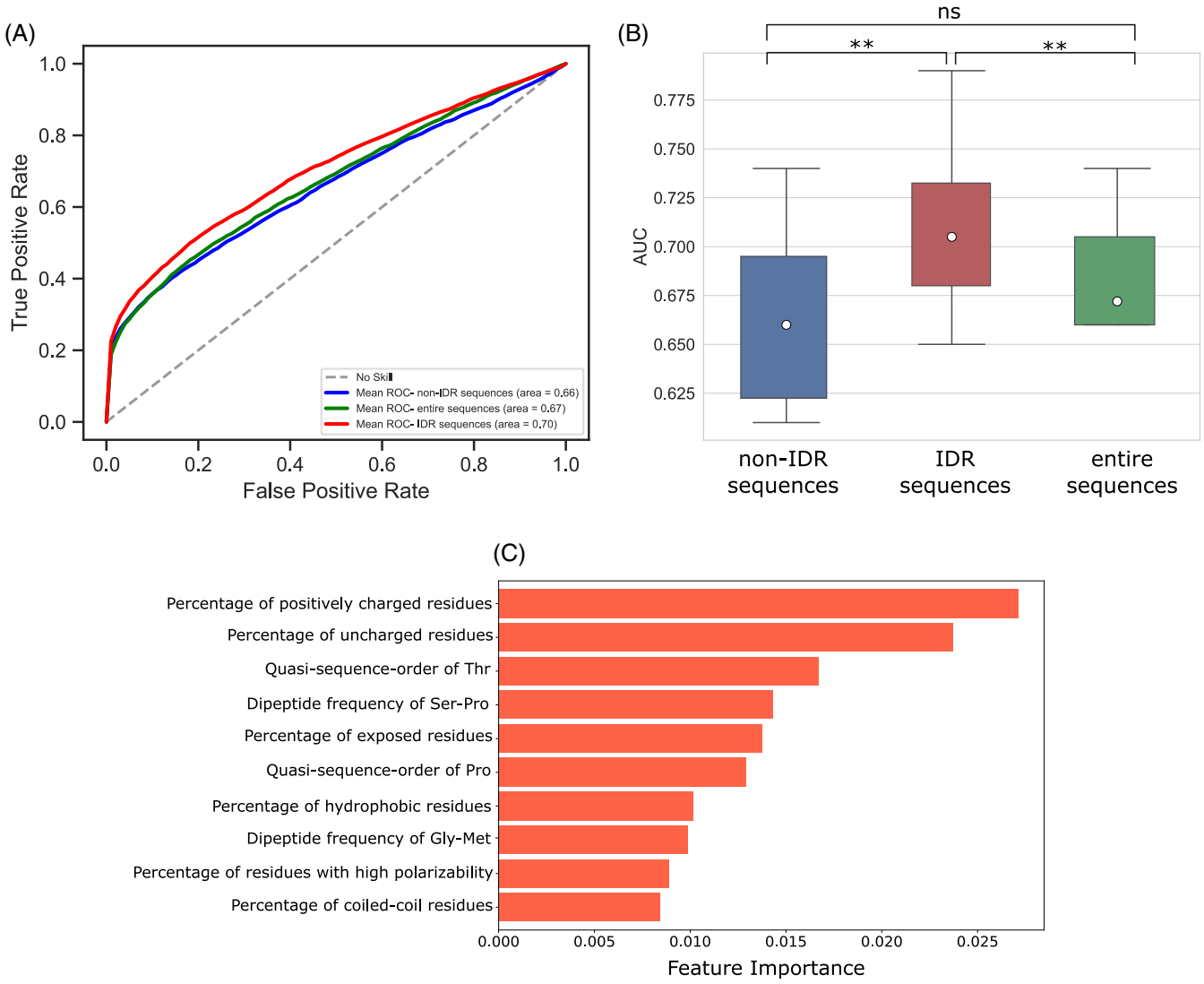

FIGURE 4 Performance of asymmetric model on three type of input sequences: IDR sequences (red), entire sequences (green) and non-IDR sequences (blue) (A) Mean ROC AUC scores of the asymmetric models on IDR sequences, entire sequences and non-IDR sequences based on 10-fold cross-validation. (B) Boxplots over the AUC-values reached by each model on 10 folds in the asymmetric problem. The white circle in the boxplot represents the mean AUC scores. Asterisks above figure bars indicate statistical significance. One asterisk (*) indicates p value smaller than 0.1 ($p < .1$). Two asterisks (**) indicate p value smaller than .05 ($p < .05$). "ns" indicates not significant ($p > .05$) difference. (C) Feature importance of asymmetric model on the best performing model based on IDR sequences.

cannot apply it on the symmetric problem and we only assessed the performance of IDPpi on the test cases for the asymmetric problem. The other alternative method we tested is D-SCRIPT, which we needed to adapt as described in Methods. D-SCRIPT is applicable for both problem types. We trained both IDPpi and D-SCRIPT using the entire protein sequences as expected by these programs. Finally, we evaluated the models using our test sets for both of the problem types.

Results are shown in Table 2. Generally, under the stringent, problem-type specific test scenarios that we use, IDPpi and D-SCRIPT yield inferior performance compared to our two classifiers. For the asymmetric model, the difference in performance is quite remarkable. This is also shown in Figure 6, where the average ROC curves of the 10-fold cross-validation for the asymmetric problem are displayed. In the symmetric case, neither our method nor D-SCRIPT perform well, with a particularly low recall for D-SCRIPT. We need to qualify this comparison, though, since the original D-SCRIPT uses a large PPI network for training and has more hidden layers, which might put the version that we were able to run at a disadvantage. However, for the symmetric problem, the pre-trained D-SCRIPT model is marginally better than our model but still fails to solve this problem type.

3.5 | Case study for asymmetric model: Predicting the interactors of RB1

To illustrate the performance of our asymmetric model, we wanted to predict interactors of the retinoblastoma protein RB1 in an

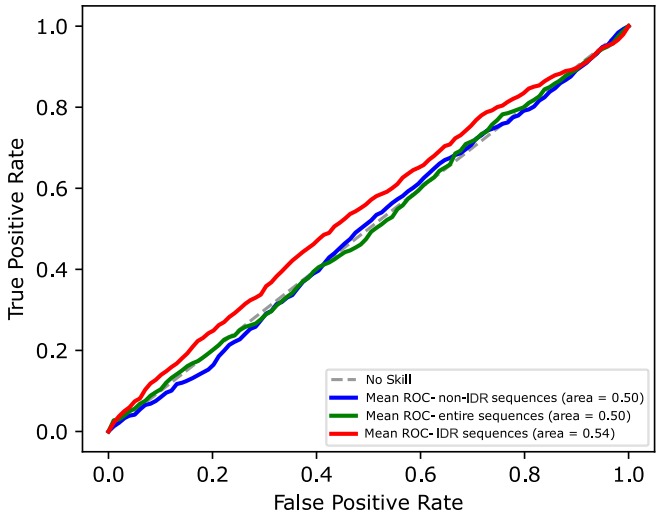

FIGURE 5 Mean ROC AUC scores of the symmetric models on the three type of input sequences based on 10-fold cross-validation.

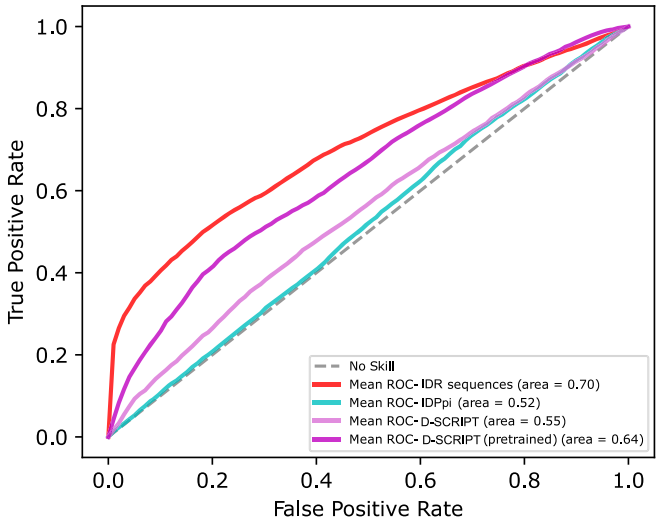

FIGURE 6 Prediction performance comparison of different classifiers using ROC curves in predicting IDP specific protein–protein interactions for the asymmetric problem. Shown in the plot are the ROC curves for IDPpi, D-SCRIPT, D-SCRIPT (pre-trained) and IDR sequences.

independent sequence set that has not been used in developing our model at all. RB1 is an important tumor suppressor protein and plays a critical role in the cell cycle. Based on disorder prediction by PONDR® VSL2, RB1 has five distinct IDRs.[16] It was reported that these disordered regions contain conserved phosphorylation sites, which prevent interaction with its target protein.[33,34] It is known that RB1 interacts with proteins with diverse IDR content, for example, its target protein E2F3 is largely disordered whereas CDK4 consists of only one disordered region based on PONDR® VSL2 predictions.[16,35]

We wanted to test a number of likely interacting proteins for RB1 to see how many of them would be predicted correctly by our algorithm. To this end, we retrieved a further five interactors of RB1 from the STRING database,[35] The STRING database had not been used for our model and these proteins did not enter the training procedure. We also selected five more interactors of RB1 from the training set used in our model as "positive controls". Together, this resulted in 10 interactors of RB1 which were input into our asymmetric model. Seven of the ten interactors of RB1 were correctly predicted by the asymmetric model (Figure 7). Of the five positive controls, four were identified correctly as interactors. This suggests that the model is not overtrained and has not simply memorized the sequences it had seen. Three out of the five sequences that had not been part of the training were correctly predicted as interactors.

4 | DISCUSSION AND CONCLUSION

Disordered regions are assumed to play a key role in mediating interactions among proteins. In our endeavor to design a predictor for PPI for IDPs we were faced with the intricacies of this problem as a special case of a pair-input prediction problem. We therefore developed two different machine learning algorithms to address two different PPI prediction problems: the asymmetric problem and the symmetric problem. For the asymmetric problem, where one of the proteins had been seen before by the classifier, we developed a method to predict disordered protein partners of known proteins present in our dataset. For the symmetric problem, we developed a method to predict novel PPIs. The methods work differently, firstly to account for the available information, and secondly because one expects a classifier for the

TABLE 2 Comparison between our proposed method based on IDR sequences and other state-of-the-art-methods according to accuracy, precision, recall, F1-score and area under the ROC curve (AUC)

Problem type	Asymmetric problem				Symmetric problem		
Methods	IDR sequences	D-SCRIPT	D-SCRIPT pretrained	IDPpi	IDR sequences	D-SCRIPT	D-SCRIPT pretrained
AUC	**0.70**	0.55	0.64	0.52	0.54	0.50	**0.57**
Accuracy	0.61	0.52	0.54	0.51	0.54	0.50	0.51
Recall	0.74	0.11	0.06	0.48	0.52	0.10	0.03
Precision	0.57	0.24	0.78	0.49	0.54	0.08	0.53
F1-Score	0.65	0.08	0.10	0.48	0.53	0.07	0.05

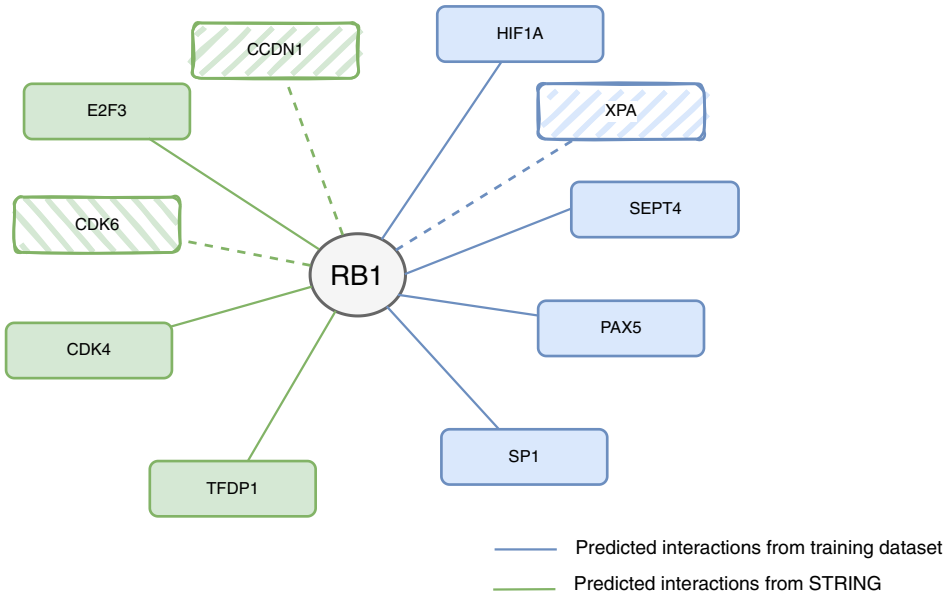

FIGURE 7 Case study for true positive predictions of RB1. Blue nodes show predicted positive interactions from training data. Green nodes show the correctly predicted positive interactions from the new interaction set. Dashed nodes are the interactors not identified by our method.

symmetric problem to be symmetric in terms of the two input proteins.

We also sought to determine if amino acid sequences of IDRs provide an advantage over other sequence regions for predicting the interactions between IDPs. To address this question, we utilized IDR sequences to predict human IDP-IDP interactions. Furthermore, we extracted entire and non-IDR subsequences to develop controls and compare the performance of IDR sequences to control models. We conclude that the disordered regions are particularly informative for prediction of interaction among IDPs.

An essential part of our approach is the design of suitable training and, in particular, testing procedures for the respective problem type. For the asymmetric problem, our designed model based on IDR sequences achieved the best performance (AUC 0.70) on the evaluations on 10 independent test sets. We compared our asymmetric model to two state-of-the-art PPI prediction models and demonstrated that our IDR-based asymmetric model performs better than the already existing models. Note, however, that this was measured with our testing procedure, which we think of as appropriate for the question and rigorous, but other testing procedures yield other quality measurements. This likely explains some of the differences in reported performances among published methods.

For the symmetric problem, our measurements of method performance indicate that there is no method that would be particularly successful on this problem. We believe that the rigorous testing according to the C3 group (defined as "Neither protein in the test pair is found in the training set") as introduced by Park and Marcotte (2012) makes the inherent difficulty of the symmetric problem apparent. Again, we do concede that other methods report better performance, albeit with other testing scenarios. This is also in line with the results by Dunham and colleagues[6] who reported that most of the PPI tools perform much lower than originally reported. In their work they even show that randomly chosen feature values may, under common testing scenarios, appear to produce more accurate results than proper prediction results.

The problem we are dealing with here is related to the search for molecular recognition features (MoRFs) which can enhance PPIs.[36] Importantly, MoRFs undergo transitions from disordered to ordered states upon binding to their interaction partners, thus they have been thought to promote binding. Currently, there are several MoRF predictors that can identify these regions using different predictive models.[37] Our method differs from protein-binding site predictors, which are mainly developed to find binding sites within a single disordered protein sequence. We would expect that identification of MoRF regions could in the future aid also in predicting binary interaction between disordered proteins.

To the best of our knowledge, our study presents the first effort that not only respects the problem class in design and sampling of positive and negative testing and training sets, but that also puts forward two different classifiers for the asymmetric and the symmetric problem. This has advanced our ability to predict new interactors for known proteins, and has shed some light on the inherent difficulty in predicting entirely new interactions. Our results also support the paradigm that IDR sequences are particularly informative when it comes to predicting interactions among proteins.

Given our experience that the asymmetric problem is easier to handle than the symmetric one, for the future one has the choice whether to improve the asymmetric model, or to make an effort towards solving the symmetric problem type.

AUTHOR CONTRIBUTIONS

Gözde Kibar: Writing – original draft; conceptualization; investigation; software; methodology; visualization; formal analysis; data curation. **Martin Vingron:** Writing – review and editing; conceptualization; funding acquisition; methodology; resources; supervision; writing – original draft.

ACKNOWLEDGMENTS

This work was supported by the IMPRS-CBSC doctoral programme. GK was partly funded by grant Bildung und Forschung (BMBF) - iGenVar FKZ 031L0169A. Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The authors declare no competing interests.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1002/prot.26486.

DATA AVAILABILITY STATEMENT

Our asymmetric model is available at Github at https://github.com/gozdekibar/IDR_PPI_prediction. User instructions can be found in the README document. Test and training data used by this study for both of the models are also available in the repository.

REFERENCES

1. Fuxreiter M. Classifying the binding modes of disordered proteins. *Int J Mol Sci*. 2020;21(22):8615. doi:10.3390/ijms21228615
2. Chen R, Li X, Yang Y, Song X, Wang C, Qiao D. Prediction of protein-protein interaction sites in intrinsically disordered proteins. *Front Mol Biosci*. 2022;9:9. doi:10.3389/fmolb.2022.985022
3. Chong S, Mir M. Towards decoding the sequence-based grammar governing the functions of intrinsically disordered protein regions. *J Mol Biol*. 2021;433(12):166724. doi:10.1016/j.jmb.2020.11.023
4. Wang W, Wang D. Extreme fuzziness: direct interactions between two IDPs. *Biomolecules*. 2019;9(3):81. doi:10.3390/biom9030081
5. Casadio R, Martelli PL, Savojardo C. Machine learning solutions for predicting protein–protein interactions. *WIREs computational molecular. Science*. 2022;12(6):e1618. doi:10.1002/wcms.1618
6. Dunham B, Ganapathiraju MK. Benchmark evaluation of protein–protein interaction prediction algorithms. *Molecules*. 2021;27(1):41. doi:10.3390/molecules27010041
7. Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. *Nucleic Acids Res*. 2000;28(1):235-242. doi:10.1093/nar/28.1.235
8. Park Y, Marcotte EM. Flaws in evaluation schemes for pair-input computational predictions. *Nat Methods*. 2012;9(12):1134-1136. doi:10.1038/nmeth.2259
9. 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):2610-2614. doi:10.1093/bioinformatics/btq483
10. Perovic V, Sumonja N, Marsh LA, et al. IDPpi: protein-protein interaction analyses of human intrinsically disordered proteins. *Scientific Reports*. 2018;8(1):1-10. doi:10.1038/s41598-018-28815-x
11. Piovesan D, Tabaro F, Mičetić I, et al. DisProt 7.0: a major update of the database of disordered proteins. *Nucleic Acids Res*. 2017;45(D1):D227. doi:10.1093/nar/gkw1056
12. Alanis-Lobato G, Andrade-Navarro MA, Schaefer MH. Hippie v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks. *Nucleic Acids Res*. 2016;45(D1):D414. doi:10.1093/nar/gkw985, 45.
13. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. 2006;22(13):1658-1659. doi:10.1093/bioinformatics/btl158
14. Bateman A, Martin MJ, Orchard S, et al. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res*. 2021;49(D1):D489. doi:10.1093/nar/gkaa1100
15. Park Y, Marcotte EM. Revisiting the negative example sampling problem for predicting protein-protein interactions. *Bioinformatics*. 2011;27(21):3024-3028. doi:10.1093/bioinformatics/btr514
16. Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics*. 2006;7(1):208-212. doi:10.1186/1471-2105-7-208
17. Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins*. 2001;43(3):246-255. doi:10.1002/prot.1035
18. Broto P, Moreau G, Vandycke C. Molecular structures: perception, autocorrelation descriptor and sar studies. Autocorrelation descriptor. *European J Med Chem*. 1984;19:61-65.
19. Chou KC. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. *Biochem Biophys Res Commun*. 2000;278(2):477-483. doi:10.1006/bbrc.2000.3815
20. Dubchak I, Muchnik I, Holbrook SR, Kim SH. Prediction of protein folding class using global description of amino acid sequence. *Proc Natl Acad Sci U S A*. 1995;92(19):8700-8704. doi:10.1073/pnas.92.19.8700
21. Xiao N, Cao DS, Zhu MF, Xu QS. Protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics*. 2015;31(11):1857-1859. doi:10.1093/bioinformatics/btv042
22. Schneider G, Wrede P. The rational design of amino acid sequences by artificial neural networks and simulated molecular evolution: de novo design of an idealized leader peptidase cleavage site. *Biophys J*. 1994;66(2 Pt 1):335-344. doi:10.1016/S0006-3495(94)80782-9
23. Grantham R. Amino acid difference formula to help explain protein evolution. *Science*. 1974;185(4154):862-864. doi:10.1126/science.185.4154.862
24. Martin S, Roe D, Faulon JL. Predicting protein-protein interactions using signature products. *Bioinformatics*. 2005;21(2):218-226. doi:10.1093/bioinformatics/bth483
25. Guo Y, Yu L, Wen Z, Li M. Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. *Nucleic Acids Res*. 2008;36(9):3025-3030. doi:10.1093/nar/gkn159
26. Shen J, Zhang J, Luo X, et al. Predicting protein-protein interactions based only on sequences information. *Proc Natl Acad Sci U S A*. 2007;104(11):4337-4341. doi:10.1073/pnas.0607879104
27. Sledzieski S, Singh R, Cowen L, Berger B. D-SCRIPT translates genome to phenome with sequence-based, structure-aware, genome-scale predictions of protein-protein interactions. *Cell Syst*. 2021;12(10):969-982.e6. doi:10.1016/j.cels.2021.08.010
28. Du X, Sun S, Hu C, Yao Y, Yan Y, Zhang Y. DeepPPI: boosting prediction of protein–protein interactions with deep neural networks. *J Chem Inf Model*. 2017;57(6):1499-1510. doi:10.1021/acs.jcim.7b00028
29. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Mach Learn Res*. 2011;12:2825-2830.
30. Müller-Späth S, Soranno A, Hirschfeld V, et al. Charge interactions can dominate the dimensions of intrinsically disordered proteins. *Proc Natl Acad Sci*. 2010;107(33):14609-14614. doi:10.1073/pnas.1001743107
31. Jo Y, Jang J, Song D, Park H, Jung Y. Determinants for intrinsically disordered protein recruitment into phase-separated protein condensates. *Chem Sci*. 2022;13(2):522-530. doi:10.1039/D1SC05672G
32. Uversky VN. The alphabet of intrinsic disorder. *Intrinsic Disorder Protein*. 2013;1(1):e24684. doi:10.4161/idp.24684
33. Burke JR, Deshong AJ, Pelton JG, Rubin SM. Phosphorylation-induced conformational changes in the retinoblastoma protein inhibit E2F transactivation domain binding. *J Biol Chem*. 2010;285(21):16286-16293. doi:10.1074/JBC.M110.108167

34. Desvoyes B, Gutierrez C. Roles of plant retinoblastoma protein: cell cycle and beyond. *EMBO J*. 2020;39(19):e105802. doi:10.15252/embj.2020105802
35. Szklarczyk D, Franceschini A, Wyder S, et al. String V10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res*. 2014;43(D1):D452. doi:10.1093/nar/gku1003
36. van der Lee R, Buljan M, Lang B, et al. Classification of intrinsically disordered regions and proteins. *Chem Rev*. 2014;114(13):6589-6631. doi:10.1021/cr400525m
37. Katuwawala A, Peng Z, Yang J, Kurgan L. Computational prediction of MoRFs, short disorder-to-order transitioning protein binding regions. *Comput Struct Biotechnol J*. 2019;17:454-462. doi:10.1016/j.csbj.2019.03.013

How to cite this article: Kibar G, Vingron M. Prediction of protein–protein interactions using sequences of intrinsically disordered regions. *Proteins*. 2023;91(7):980-990. doi:10.1002/prot.26486