Predicting the dynamic interaction between intrinsically disordered proteins

Yuchuan Zheng[1†], Qixiu Li[1†], Maria I. Freiberger[2], Haoyu Song[1], Guorong Hu[1], Moxin Zhang[1], Ruoxu Gu[3], Jingyuan Li[1*]

[1] School of Physics, Zhejiang University, Hangzhou 310058, PR China
[2] Protein Physiology Lab, Departamento de Quimica Biologica, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires-CONICET-IQUIBICEN, Buenos Aires, Argentina
[3] School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, PR China
[†]These authors contributed equally to this work.
[*]Corresponding author. Email: jingyuanli@zju.edu.cn (J.L.)

Abstract

Interactions between intrinsically disordered proteins (IDPs) are highly dynamic, their interaction interfaces are diverse and affected by their inherent conformation fluctuations. Comprehensive characterization of these interactions based on current techniques is challenging. Here, we present GSALIDP, a GraphSAGE-LSTM Network to capture the dynamic nature of IDP conformation and predict the behavior of IDP interaction. The training data for our method is obtained from atomistic molecular dynamics (MD) simulations. Our method models multiple conformations of IDP as a dynamic graph which can effectively describe the fluctuation of its flexible conformation. GSALIDP can effectively predict the interaction sites of IDP as well as the contact residue pairs between IDPs. Its performance to predict IDP interaction is on par with or even better than the conventional models to predict the interaction of structural proteins. To the best of our knowledge, this is the first machine-learning method to realize the prediction of interaction behavior of IDPs. Our framework can be exploited to study other IDP-mediated processes such as folding-upon-binding and fuzzy interaction, wherein the dynamic nature of IDP conformation is also essential.

Introduction

Protein-protein interactions (PPIs) play a central role in all biological processes[1–3]. Conventional PPIs are encoded in the complementarity of proteins' binding interfaces, with the premise of well-defined protein conformation[4–8]. On the other hand, there are numerous proteins that lack stable structures and are denoted as intrinsically disordered proteins (IDPs).[9–11] Their interactions are highly dynamic and coupled with the fluctuation of their flexible conformations[12–15]. Therefore, the behaviors of IDP interactions are distinct from conventional PPIs. And the studies about IDP interaction are indispensable for understanding the related biological processes such as the cellular liquid-liquid phase separation (LLPS)[16–22] and should effectively complement the paradigm of PPIs.

There are growing numbers of studies about the behavior of IDP interactions[15,23–28]. For example, the sequence regions in FUS LC and H1.0 with considerable interaction propensities were identified by the combination of NMR and FRET. Despite these progresses, comprehensive characterization of such highly dynamic and heterogeneous IDP interactions is still a challenge due to the limitation of spatial and temporal resolutions. Simulations were widely applied to study the behavior of IDP interactions[15,23,25]. Our previous molecular dynamic (MD) simulation study of the LAF-1 RGG domains identified a finite C-terminal region with high interaction propensities[15]. It should be noted that simulations are usually hampered by the overwhelming computational costs due to the high degrees of freedom of such systems. From this perspective, state-of-the-art machine-learning methods have the potential to effectively investigate the behavior of IDP interactions because of their efficiency in recognizing the patterns of high-dimensional data[29].

Machine-learning methods have been widely used in the prediction of the interaction behavior of structural proteins, i.e. the residues involved in the protein interaction[30–42]. These PPI prediction methods can be largely classified into two classes, the sequence-based methods that only require protein sequences, and the structure-based methods which also utilize protein structures. There have been attempts to use these methods to predict the behavior of folding-upon-binding processes of IDPs[43]. However, in most cases (e.g. the formation of membraneless organelles), the interactions between IDPs are seemingly irregular and the proteins remain structureless. Such interactions can't be predicted by these methods. Despite the development of the PPI prediction methods, predicting IDP interactions is still very difficult. IDPs have lower sequence conservation and higher proportions of low-complexity sequence domains (LCDs). Such sequence features provide challenges for current sequence analysis techniques as well as sequence-based models in effectively extracting information from IDP sequences[44]. In addition, due to the limitation of experimental resolution, the structures of both monomeric IDP and IDP complex are largely unavailable. Thus, the training data for structure-based methods is lacking. Besides, due to their inherent conformation fluctuations, multiple IDP conformations as well as structural flexibility need to be considered to depict the behavior of IDP interaction. This brings additional challenges to the development of prediction method. To the best

of our knowledge, there is still no machine-learning method for predicting the behavior of IDP interaction.

In this work, we develop a machine-learning method to predict the behavior of IDP interaction. The information of IDP interaction is obtained through all-atom MD simulations. A novel GraphSAGE-LSTM Network (noted as GSALIDP) is then constructed to capture the dynamic nature of IDP conformation and predict the IDP interaction. By modeling multiple IDP conformations as a dynamic graph, this framework can effectively characterize the fluctuation of its flexible conformation. GSALIDP can accomplish the prediction of the interaction sites of IDP. Its performance is evaluated by a series of metrics and is on par with or even better than the conventional models to predict the interaction of structural proteins. Besides, our model can also predict the contact residue pairs between IDPs. As revealed by our result, the highly dynamic IDP interaction can be effectively predicted by the dynamic graph neural network. This framework can be exploited to studying other IDP-mediated processes such as folding-upon-binding and fuzzy interaction, wherein the dynamic nature of IDP conformation is also essential.

Results

Obtaining the conformation of IDPs from MD simulations. The N-terminal region of LAF-1, an intrinsically disordered R/G-rich domain (LAF-1 RGG domain), is considered in this work. [45,46]. Previous studies showed that this domain is important to the LLPS of LAF-1, which drives the assembly of the membraneless organelle called P granules[45,47,48]. Besides, the isolated RGG domain can sufficiently engender LLPS *in vitro*[23]. Thus, LAF-1 RGG domain is conceived as a representative IDP system[15,23,45].

In our previous work, all-atom MD simulations were conducted to characterize the interaction between two RGG domains[15]. A finite region in LAF-1 RGG that exhibits an exceptional interaction tendency was identified[15]. The trajectories are then adopted for the current study. More specifically, 17 independent trajectories with a cumulative length of 18 μs are used. IDPs continuously undergo reversible interactions throughout the simulation. The durations of these interactions are highly varied. Apart from the transient interactions, there are stable interactions with sustained durations that dominate the IDP interaction[15]. Accordingly, we collect a series of stable interaction events from the MD trajectories using the following threshold: two proteins keeping contact for more than 30 ns and the heavy atom contact number > 200. In total, 35 stable interaction events are obtained.

To consider the fluctuation of IDP's flexible conformation, multiple IDP conformations prior to the interaction event are extracted. Specifically, 16 monomeric conformations of each IDP are obtained in 0.2 ns intervals between 0 to 3 ns prior to the contact. These conformations are used to calculate the features of the residues and construct the graphs that serve as the input of our model. The IDP complex at 3 ns after contact is used to label the interaction sites of each IDP as well as the contact residue pairs between two IDPs. The residues from two IDPs are considered to be in contact if the distance between any of their heavy atoms is less than 6 Å. The IDP complexes of the 34 interaction events are randomly divided into the training set and test set in the ratio of 6:1.

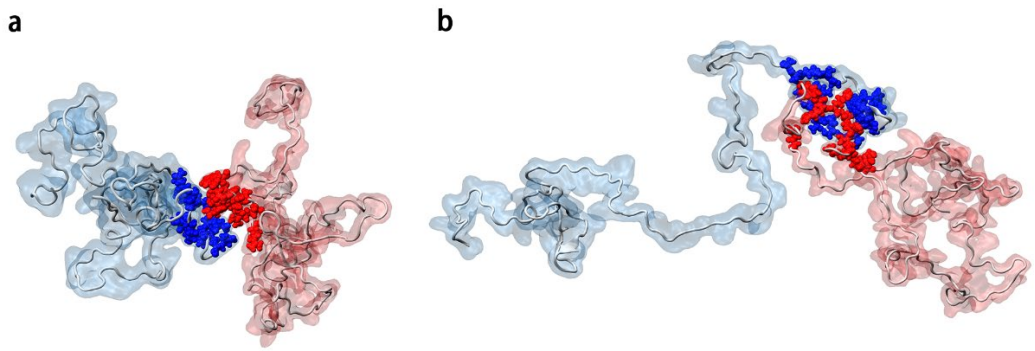

Fig. 1 | Representative complex structures of LAF-1 RGG domains obtained from

simulations. In red and blue are represented two proteins and in spheres are represented the residues involved in IDP interaction.

Representation of IDP conformations. The conformation of IDP is represented as a graph G = (V, E), where V denotes the residues of IDP (nodes) and E denotes the contacts between residues (edges). This graph is encoded by a node feature matrix X and an adjacency matrix A: X is an n×d matrix that contains the feature values of the residues, where n is the number of residues and d is the number of features; A is an n×n matrix that encodes the contacts between residues, where the element A(i, j) equals 1 if there is a contact between residues i and j, and 0 otherwise.

For an IDP complex, 16 monomeric conformations of each IDP are considered (as described in the previous section). These conformations are represented as a sequence of graphs, $\{G_1, G_2, \cdots, G_{16}\}$, where the $G_1, G_2, \ldots, G_{16}$ correspond to the conformation obtained at 3, 2.8, …, 0 ns prior to the contact, respectively. The n-th graph is denoted as $G_n = (V_n, E_n)$, and its corresponding node feature matrix and adjacency matrix are denoted as X_n and A_n. The prediction of interaction sites is based on the graph sequence of the IDP. And the prediction of the contact residue pairs is based on the graph sequences of both IDPs.

Network architecture of GSALIDP. Multiple conformations of IDP (i.e. 16 conformations) are considered in our method to describe IDP's flexibility. These conformations are represented by the sequence of graphs. In other words, IDP is characterized by a dynamic graph[49–51] instead of a single graph: the features of nodes (i.e., residues) and the connections between them (i.e., contacts) evolve over time. We develop the GraphSAGE-LSTM network that integrates the graph sample and aggregate (GraphSAGE) network[52] with the long short-term memory (LSTM) network[53] (see more discussions in Methods). LSTM serves as the main framework of the model, and GraphSAGE is embedded within LSTM to extract the structural information of IDPs. More specifically, GraphSAGE sequentially processes 16 IDP conformations and inputs the extracted structural information into the LSTM. LSTM further analyzes the hidden temporal information among these IDP conformations. This framework is similar to previous machine-learning methods of dynamic graph modeling[49–51,54–56], e.g. the integrated framework of graph convolutional network (GCN) and LSTM proposed by Seo et.al[54]. The GraphSAGE-LSTM Network we develop can effectively extract the features of IDP's conformational evolution.

The structure of our model is shown in Fig. 2. Standard LSTM comprises four core components: the input gate, forget gate, cell state, and output gate. In the GraphSAGE-LSTM network, the primary structure of LSTM is retained, but the computation of its four components is replaced by GraphSAGE. The GraphSAGE module generates node embeddings by sampling and aggregating features from a node's neighborhood. Thus, the properties of local environment of the node are effectively considered. And the architecture of LSTM can capture the long-term dependencies hidden in the sequence. The GraphSAGE-LSTM Network thus can effectively extract both the spatial and temporal information contained in the dynamic graph. It scans the entire graph sequence

and outputs the hidden layer vector of the last graph. This vector is considered as the encoding of IDP residues and is further turned into the predicted results by a multi-layer perceptron (MLP). For the prediction of interaction sites, the MLP takes in the encoding of each residue and outputs the probability of its involvement in the IDP interaction. For the prediction of contact residue pairs, the encodings of the residues of two IDPs are concatenated and the resulting vector is employed as the encoding of the residue pair between IDPs. The MLP then takes in this vector and outputs the contact probability of residue pair.

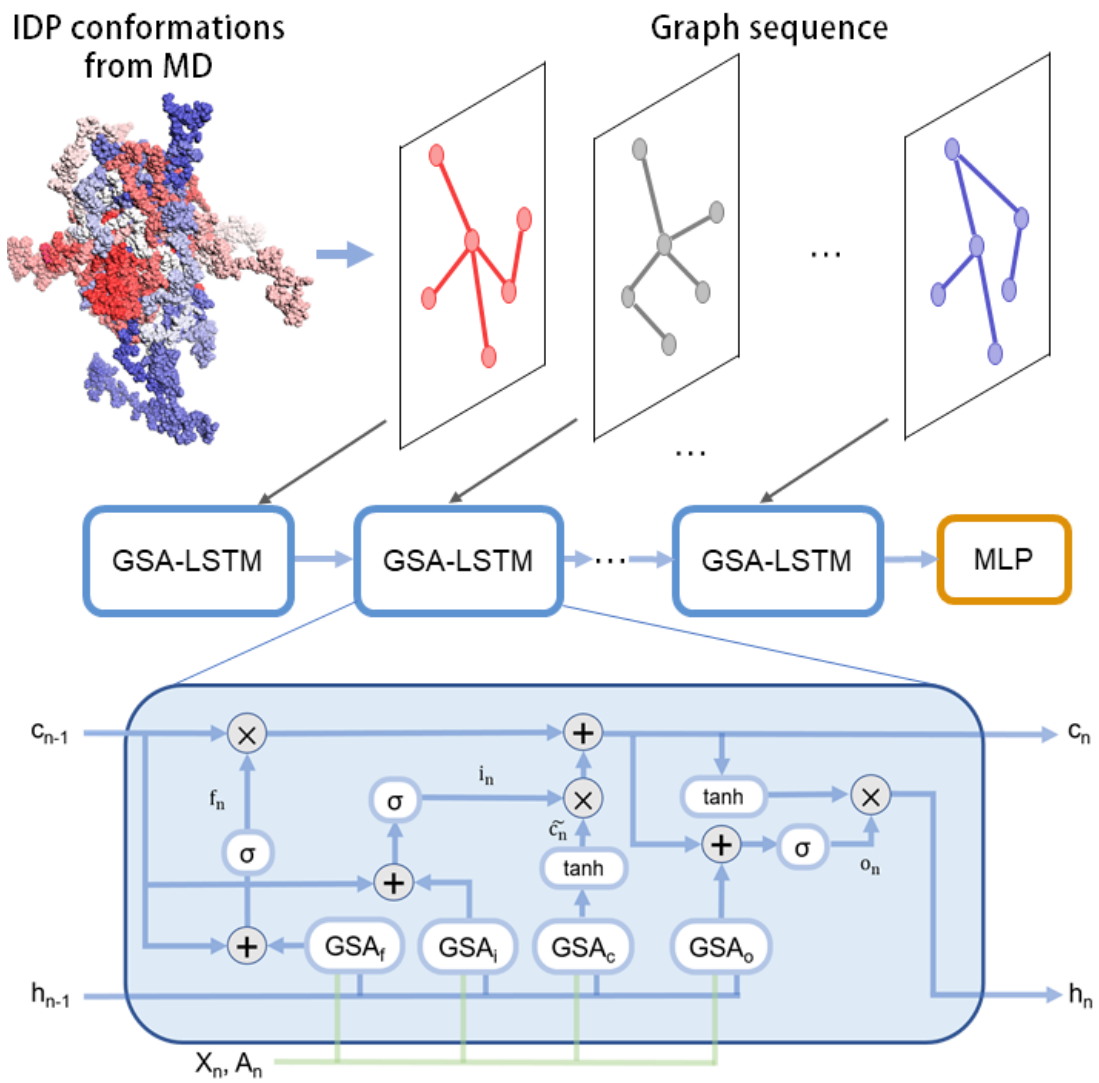

Fig. 2 | The overall framework of GSALIDP. The input of GSALIDP is a sequence of graphs $\{G_1, \ldots, G_N\}$ where each graph represents an IDP conformation and is encoded by the node feature matrix X_n and the residue contact matrix A_n. The GraphSAGE-LSTM Network (GSA-LSTM) scans the entire graph sequence and outputs the hidden layer vector of the last graph. A multi-layer perceptron (MLP) then converts this vector into the predicted probabilities. Here, h_n and c_n respectively represent the hidden layer vector and cell state. GSA_f, GSA_i, GSA_c, and GSA_o represent the GraphSAGE-embedded gates of the LSTM framework.

Performance of GSALIDP in predicting the interaction sites of IDP. The receiver operating characteristic (ROC) curve shows the pairs of true positive rate and false positive rate calculated at all possible thresholds. The ROC curve for the model with a random performance is a straight diagonal line from (0, 0) to (1, 1) and the corresponding area under the curve (AUROC) is 0.5[57]. It serves as the baseline of the ROC curve. The ROC curve of GSALIDP's prediction on the interaction sites is

represented by the yellow curve in Fig. 3. This curve apparently exceeds the baseline and the corresponding AUROC is up to 0.81. Thus, our model can effectively predict the interaction site between IDPs. And the performance is even better than the mainstream PPI prediction methods of structural proteins[30–38], whose AUROCs are in the range of 0.58 to 0.79[38].

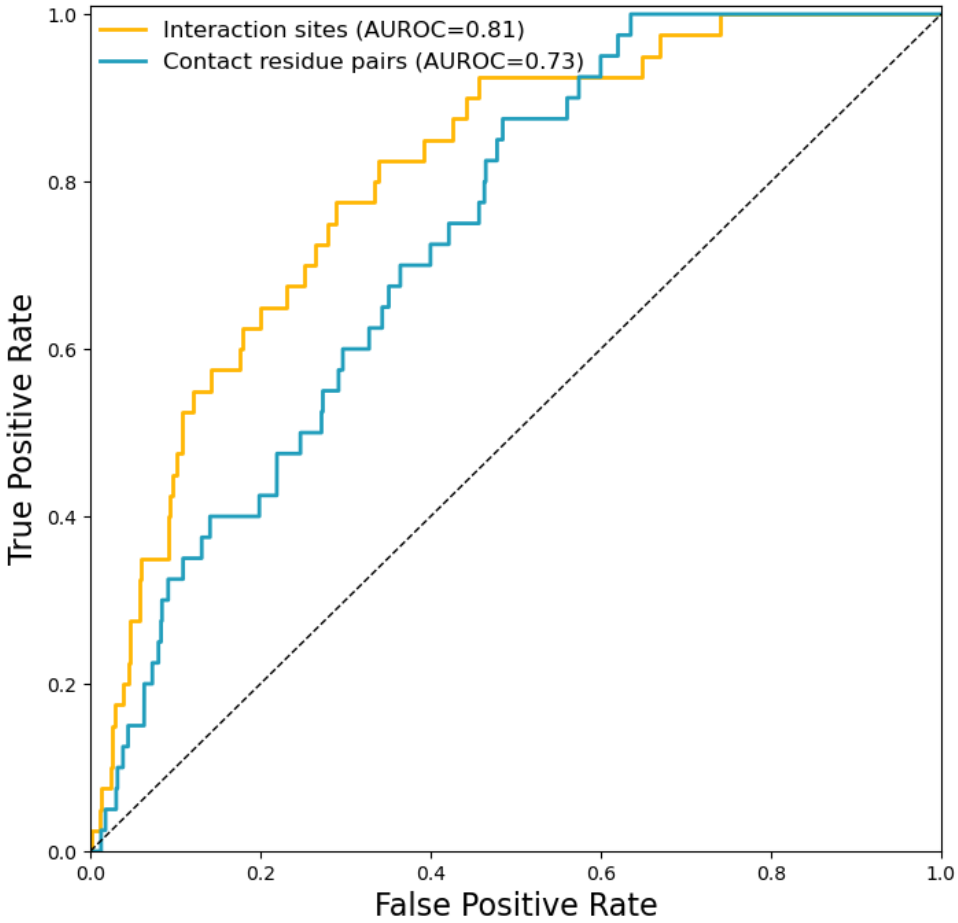

Fig. 3 | ROC Curves of GSALIDP. The curves for interaction sites, contact residue pairs, as well as random predictions are shown by the yellow solid line, blue solid line, and black dotted line, respectively.

In addition to AUROC, we also consider the area under the precision-recall curve (AUPRC), F1-score (F1), and Matthews correlation coefficient (MCC). These metrics can also effectively measure the performance of PPI prediction models[30–32,38]. Other metrics including accuracy (ACC), precision, and recall are also used. For all metrics we considered (Table 1), our model achieves comparable performance with the conventional PPI prediction models of structural protein[30–38]. Our model outperforms some mainstream models of structural protein with respect to metrics like AUROC, MCC, etc. These results further demonstrate our model's ability to distinguish between

the interaction and non-interaction sites of IDPs. It should be noted that the available data for IDP interactions is highly scarce as compared to the structural protein counterpart. Moreover, the dynamic nature of IDP interactions provides additional challenges to the prediction. Such impressive performance of our model clearly demonstrates the seemingly random interaction between IDPs is actually predictable.

Table 1 | Performance of interaction site prediction.

Method	**AUROC**	**AUPRC**	**MCC**	**F1**	Precision	Recall	ACC
Conventional PPI models	0.58 ~ 0.79	0.19 ~ 0.44	0.08 ~ 0.34	0.28 ~ 0.46	0.19 ~ 0.37	0.41 ~ 0.59	0.57 ~ 0.78
GSALIDP	0.81	0.16	0.26	0.28	0.19	0.53	0.87

Note: Performance of the prediction models of structural proteins are obtained from Yuan et al[38].

Performance of GSALIDP in predicting the contact residue pairs between IDPs. It should be noted that the stability of IDP interaction is attributed to the formation of contact residue pairs[15]. Under such consideration, the contact residue pairs between IDPs are also predicted. The ROC curve of GSALIDP on predicting the contact residue pairs is represented in Fig.3. The AUROC is up to 0.73. Such performance is also comparable with previous models for predicting the contact residue pairs between structural proteins[39–42], whose AUROCs are in the range of 0.69 to 0.91[41,42].

Table 2 | Performance of interaction site prediction using different feature sets

Features	**AUROC**	**AUPRC**	**F1**	**MCC**	Precision	Recall	ACC
All features	0.81	0.16	0.26	0.28	0.19	0.53	0.87
Exclude RHFD	0.66	0.09	0.18	0.14	0.12	0.38	0.84
Exclude ICA	0.69	0.11	0.17	0.17	0.10	0.70	0.68
Using only ICA and RHFD	0.68	0.10	0.14	0.13	0.08	0.73	0.59

Interpretations of our model. The impressive performance of GSALIDP indicates it can effectively predict the behavior of IDP interaction. Previous studies highlighted the importance of interpretation analysis of machine-learning models, which can facilitate obtaining new insights and extracting the knowledge that was originally hidden in the data[58–60]. Thus, we discuss the possible mechanism underlying IDP interaction by evaluating the relative importance of the utilized features.

The residues of IDP are characterized by a set of features (see Methods). Specifically, both the conventional geometric/physicochemical-based features and the statistics-based features related to the frustrations of residue interactions are considered. To demonstrate the relative importance of these features, a strategy for interpreting machine-learning models called sensitivity analysis is conducted[60–64]. Sensitivity analysis involves excluding one feature from the feature set and measuring the impact on the model's performance. Supplementary Table 2 summarizes the sensitivity analysis of all features used in our model. Among these features, the removal of the relative highly frustrated density (RHFD) or the interior contact area (ICA) has the most significant effect. The performance of our model after the exclusion of RHFD and ICA as well as only considering the RHFD and ICA is compared (Table II). The exclusion of either RHFD or ICA results in considerable declines in the model's AUROC, AUPRC, F1, and MCC. Moreover, if the feature set comprises only RHFD and ICA, the model can still realize an acceptable predictive performance. Taken together, RHFD and ICA are important for predicting IDP interaction. Additional discussions with regard to these two features are then conducted as follows.

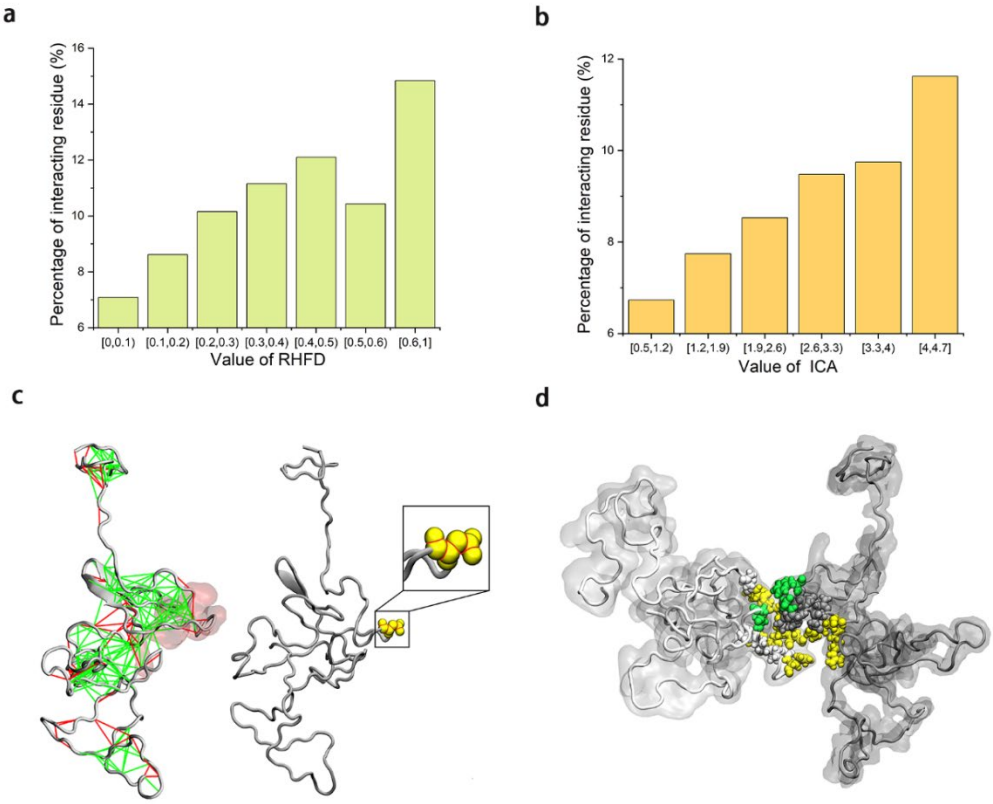

Fig. 4 | Analysis of ICA and RHFD. a-b, Percentage of the interaction sites in the residue groups with different RHFD or ICA **c,** Left, frustratogram of the representative IDP conformation highlighted the highly (red) and minimally (green) frustrated interactions, and PPI residues (red surface). Right, diagram of ICA of a residue. The red line represents the outline of the ICA. **d,** Rrepresentative conformation of IDP complex with highlighted interaction sites. The residues with high RHFD or high ICA are shown

in green and yellow, respectively.

The relative highly frustrated density (RHFD) of a residue is defined as the ratio between the number of highly frustrated interactions and the number of all interactions that the residue forms[42]. Residues with higher RHFDs tend to encounter more severe local energy conflicts. Residues are categorized into groups according to their RHFD values, and the percentage of interaction sites in each group is calculated (Fig. 4a). It should be noted that the overall percentage of residues with high RHFD (i.e. between 0.6 and 1) is very low (Supplementary Fig. 2), these residues are combined into one group. With the increase of the RHFD, the percentage of interaction sites also increases, suggesting a clear correlation between the residue's RHFD and its tendency to participate in PPI. Such correlation with PPI involvement may be attributed to the fact that local energy conflicts within IDP monomer can be reduced by the interactions with other IDP. As revealed in a recent study, frustration in the complex is lower than in the free-state monomer[65]. The RHFD of residues in monomeric IDP is then compared with the cases of IDP complex. The average RHFD of the interaction sites in the monomeric IDP is 0.131. Interestingly, the average RHFD of interaction sites within the IDP complex significantly decreases to 0.076. Such observation corroborates our conjecture and illustrates the importance of the energy conflicts within IDP.

The interior contact area (ICA) is defined as the sum of the contact areas of all atoms within the residue and is calculated using the Qcontacts package[66]. We group the residues according to their ICA values and calculate the percentage of interaction sites in each group. As the ICA increases, the proportion of interaction sites also increases considerably (Fig. 4b). Residues with higher ICA values have more contracted side chains and their backbones are more exposed. Such conformations should favor the backbones to participate in the IDP interaction. Thus, the observed correlation may be attributed to the importance of backbone-involved interactions. According to our analysis of all complexes in the dataset, there are 62 % interaction residues whose backbone moieties participate in the interaction with the residues of the other IDP.

As demonstrated by these results, the extent of frustration and the backbone exposure of the residues are directly correlated to their tendencies to participating IDP interaction. Accordingly, for a given IDP conformation there are certain regions that have apparent propensities to mediate stable interaction with other IDP, i.e. the regions enriched of the residues with these features. Fig. 4d and Supplementary Fig. 3 illustrate representative conformations of IDP complex with highlighted interaction sites. The residues with high RHFD or high ICA (top one third of residues) are shown in green and yellow, respectively. There is an apparent enrichment of the residues with high RHFD or high ICA. We further analyze the complexes throughout the dataset, and there are more than 61 % of interaction sites with high RHFD or high ICA. The interaction tendency of the residues with these features and the existence of corresponding regions are essential to the predictability of IDP interaction.

Discussion

In this work, we develop a machine-learning method named GSALIDP to predict the behavior of IDP interactions. We propose a novel **G**raph**SA**GE-**L**STM Network. This framework models multiple conformations of IDP as a dynamic graph to describe the considerable fluctuation of IDP's flexible conformation. GSALIDP can effectively predict the interaction sites of IDP as well as the contact residue pairs between IDPs. Its performance is on par with or even better than the conventional models of structural proteins. To the best of our knowledge, this is the first method to predict the behavior of IDP interactions. Besides, the interpretation analysis of our model identifies two important features for the prediction of IDP interaction, i.e. RHFD and ICA. They reveal the importance of the energy conflicts and backbone moiety to IDP interaction. It should be noted that our dynamic graph-based architecture directly acknowledges the dynamic nature of IDP conformation, which plays essential roles in the IDP-mediated processes such as folding-upon-binding and fuzzy interaction[67–70]. Hence, our framework can be exploited to study these processes, advancing the understanding of IDP's biological function and the design of IDP-based therapeutics.

Methods

Features of the residue. The residues of IDP are characterized by a series of features. Here, statistics-based features, geometric-based features, and physicochemical-based features are all considered.

Statistics-based features. As demonstrated in our previous work, statistics-based indexes related to the frustrations of residue interaction are effective features for PPI prediction[42]. The frustration index at contact level is calculated for each residue-residue contact present in a protein structure, which is a statistical measure of how energetically favorable a native contact is with respect to a set of all possible contacts at that location. A recent study showed that local frustration is an evolutionary constraint that is related to protein stability and function[71]. As indicated in previous studies, residues involved in PPIs are enriched with highly frustrated contacts[72–75], suggesting the frustration index may be a suitable feature for PPI prediction. Thus, the features based on the frustration index, i.e. the highly frustrated densities (HFD) and the relative highly frustrated density (RHFD), are applied to encode the residues.

The HFD and the RHFD are calculated using the Frustratometer tool[76,77]. This algorithm quantifies the frustration in the protein structure obtained by our simulation. The frustration index (FI) can be calculated in two modes at the contact level, i.e. mutational and configurational FI. In this work, the configurational FI is considered. It is calculated based on the decoys that change both the identity of the amino acids in contact and the contact distance accordingly. Thus, configuration FI can effectively characterize the frustrations of residue contact within the flexible structure. The FI value is calculated as a z-score of the distribution of the difference between the native energy of the protein and the energy of the decoys. The FI can be classified according to its value into minimally frustrated ($FI > 0.78$), neutral ($-1 < FI < 0.78$), and highly frustrated ($FI < -1$). After classifying the contacts, the HFD and the RHFD of the residues can be obtained. The residue's HFD is the number of involved highly frustrated interactions, while the residue's RHFD is the HFD divided by the number of all the involved interactions.

Geometric-based and physicochemical-based features. The commonly used geometric-based features like solvent accessible surface area (SASA) and relative solvent accessible surface area (RASA) as well as the lately developed features including interior contact area (ICA), exterior contact area (ECA), and exterior void area (EVA) are used to describe residues[78–81]. In addition, physicochemical-based features are also applied, including the hydropathy index (HI, two versions) and pKα (two versions)[82,83]. The calculation methods of these features are described in supplementary materials.

The GraphSAGE-LSTM Network. For each interaction event, multiple conformations of IDP are considered to describe the fluctuation of IDP's flexible conformation. These conformations are considered as a dynamic graph, where the features of the nodes (i.e., residues) and the connections between them (i.e., contacts) evolve dynamically over time. Similar dynamic graphs have been used to model a

variety of time-evolving networks, such as social networks, wireless networks, and others[50,51,54]. A series of machine-learning methods have been developed for embedding these dynamic graphs[49–51,54–56]. Especially, the framework that combines graph neural network (GNN) and recurrent neural network (RNN) is leveraged to capture both the spatial and temporal features of the dynamic graph[49,54,56].

Inspired by these studies, we develop the GraphSAGE-LSTM network. The main framework of the LSTM network is retained, but the computations of its input gate, forget gate, cell state, and output gate are replaced by GraphSAGE networks. The GraphSAGE-LSTM network sequentially processes the IDP conformations in the graph sequence $\{G_1, \ldots, G_N\}$ and outputs the hidden state h_N of the last graph, which serves as the embedding of the residues. For clarity, the processing of the n-th graph (where n=1,…, N) is described as follows:

First, the input gate i_n, forget gate f_n, and output gate o_n are determined using:

$$i_n = \sigma(W_i \otimes c_{n-1} + G_{ix}(X_n, A_n) + G_{ih}(h_{n-1}, A_n) + b_i),$$
$$f_n = \sigma\big(W_f \otimes c_{n-1} + G_{fx}(X_n, A_n) + G_{fh}(h_{n-1}, A_n) + b_f\big),$$
$$o_n = \sigma(W_o \otimes c_n + G_{ox}(X_n, A_n) + G_{oh}(h_{n-1}, A_n) + b_o).$$

c_{n-1} and h_{n-1} denote the cell state and hidden state of the preceding graph G_{n-1}, respectively. For the initial conformation (n=1), c_0 and h_0 are set as zero tensors. X_n represents the tensor of residue features and A_n is the adjacency matrix of the n-th graph. *G* represents the GraphSAGE network[52]. $W_{i,f,o}$, $b_{i,f,o}$ denote the trainable parameters of the LSTM network.

After the calculations of the i_n, f_n, and o_n, the cell state c_n and hidden state h_n of the n-th graph are calculated using:

$$c_n = f_n \otimes c_{n-1} + i_n \otimes \ tanh(G_{cx}(X_n, A_n) + G_{ch}(h_{n-1}, A_n) + b_c),$$

$$h_n = o_n \otimes tanh(c_n).$$

The c_n and h_n serve as the output of the n-th graph, and will be fed into the model for processing the subsequent graph. After scanning the entire sequence, the model outputs the hidden state h_N of the last graph. This h_N serves as the embedding of the residues. The dimension of h_N is $n_r \times d_h$, where n_r and d_h represent the number of residues and the hidden layer width, respectively.

A multilayer perceptron (MLP) consisting of four linear layers followed by a softmax layer is then utilized to output the prediction results. For the interaction site prediction, the MLP receives the embedding of the residues (h_N). The first three linear layers successively halve the dimension of the hidden layer width d_h, and the final linear layer reduces the dimension to 2 and outputs a tensor with a dimension of $n_r \times 2$. The softmax layer then normalizes this tensor into a probability distribution over the two predicted classes (non-interacting and interacting). For the contact residue pair prediction, the MLP receives the embedding of the residue pairs between IDPs, which

is obtained by concatenating the embeddings of the residues of IDPs. The linear layers successively reduce the dimension of the hidden layer width and output a tensor with a dimension of $n_r^2 \times 1$. The softmax layer then converts this tensor into the predicted contact probabilities of the residue pairs.

Evaluation metrics. The model's performance is evaluated by a series of metrics. The principle metric we considered is the area under the receiver operating characteristic curve (AUROC). AUROC is the area under the ROC curve that shows pairs of true positive rate and false positive rate values calculated at all possible thresholds[57]. AUROC reveals the overall performance of a classifier: 0.5 is for random classifiers and 1.0 is for perfect classifiers. Thus, model with an AUROC higher than 0.5 is considered to make effective predictions[57]. Similar to previous PPI prediction works of structural proteins, other metrics are also considered. Specifically, the area under the precision-recall curve (AUPRC), F1-score (F1), and Matthews correlation coefficient (MCC) are employed. These metrics are capable of effectively measuring the performance of PPI prediction models[30–32,38]. In addition, other metrics including accuracy (ACC), precision, and recall are also employed. The AUROC and AUPRC are independent of the thresholds. The other metrics are calculated using a threshold to convert predicted probabilities to binary predictions. And the threshold is determined by maximizing the F1-score of the model. The formulas for calculating these metrics are provided in the supplementary materials.

Implementation details. For the interaction site prediction, we perform the leave-one-out cross validation on the training set. Each complex takes turns to evaluate the model trained on the rest of the complexes in the training set. The resulting performances are then averaged as the overall validation performance, which is adopted to choose the optimal hyperparameters and training epoch. The chosen hyperparameters are as follows: hidden size = 128, learning rate = 1e-3, dropout rate = 0.3, weight decay of MLP = 1e-2, and weight decay of the GraphSAGE-LSTM network = 5e-4. The model is then retrained on the entire training set with the optimal hyperparameters and training epoch, and is evaluated on the independent test set. For the contact residue pair prediction, we perform 10-fold cross validation. The training set is randomly split into 10 folds. Each time, a model is trained on nine folds and evaluated on the remaining one fold. This process is repeated ten times and the performances are averaged to obtain the overall validation performance. And the chosen hyperparameters are as follows: hidden size = 128, learning rate = 1e-3, dropout rate = 0.3, weight decay of MLP = 1e-2, and weight decay of the GraphSAGE-LSTM network = 5e-4. The model is then retrained on the entire training set and evaluated on the test set.

We implemented the proposed model with Pytorch 1.12.1[84], PyTorch Geometric 2.3.0[85], and PyTorch Geometric Temporal 0.54.0[86]. And the Cross-Entropy loss with weights and Adam optimizer are employed for optimization[87].

References

1. Rual, J. F. *et al.* Towards a proteome-scale map of the human protein–protein interaction network. *Nature 2005 437:7062* **437**, 1173–1178 (2005).
2. Stelzl, U. *et al.* Resource A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome. *Cell* **122**, 957–968 (2005).
3. Azzarito, V., Long, K., Murphy, N. S. & Wilson, A. J. Inhibition of α-helix-mediated protein–protein interactions using designed molecules. *Nature Chemistry 2013 5:3* **5**, 161–173 (2013).
4. Rickard, M. M., Zhang, Y., Gruebele, M. & Pogorelov, T. v. In-Cell Protein-Protein Contacts: Transient Interactions in the Crowd. *Journal of Physical Chemistry Letters* **10**, 5667–5673 (2019).
5. Scott, D. E., Bayly, A. R., Abell, C. & Skidmore, J. Small molecules, big targets: drug discovery faces the protein–protein interaction challenge. *Nature Reviews Drug Discovery 2016 15:8* **15**, 533–550 (2016).
6. Santos, R. *et al.* A comprehensive map of molecular drug targets. *Nature Reviews Drug Discovery 2016 16:1* **16**, 19–34 (2016).
7. Rodrigues, C. H. M., Pires, D. E. V. & Ascher, D. B. PdCSM-PPI: Using Graph-Based Signatures to Identify Protein-Protein Interaction Inhibitors. *J Chem Inf Model* **61**, 5438–5445 (2021).
8. Zhang, C., Freddolino, P. L. & Zhang, Y. COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information. *Nucleic Acids Res* **45**, W291–W299 (2017).
9. Brangwynne, C. P., Tompa, P. & Pappu, R. V. Polymer physics of intracellular phase transitions. *Nature Physics 2015 11:11* **11**, 899–904 (2015).
10. Uversky, V. N. Introduction to intrinsically disordered proteins (IDPs). *Chem Rev* **114**, 6557–6560 (2014).
11. Oldfield, C. J. & Dunker, A. K. Intrinsically disordered proteins and intrinsically disordered protein regions. *Annu Rev Biochem* **83**, 553–584 (2014).
12. Reichheld, S. E., Muiznieks, L. D., Keeley, F. W. & Sharpe, S. Direct observation of structure and dynamics during phase separation of an elastomeric protein. *Proc Natl Acad Sci U S A* **114**, E4408–E4415 (2017).
13. Bergeron-Sandoval, L. P., Safaee, N. & Michnick, S. W. Mechanisms and Consequences of Macromolecular Phase Separation. *Cell* **165**, 1067–1079 (2016).
14. Murthy, A. C. *et al.* Molecular interactions underlying liquid−liquid phase separation of the FUS low-complexity domain. *Nature Structural & Molecular Biology 2019 26:7* **26**, 637–648 (2019).
15. Zhang, M. *et al.* Sequence Tendency for the Interaction between Low-Complexity Intrinsically Disordered Proteins. *J Am Chem Soc* (2022) doi:10.1021/JACSAU.2C00414/ASSET/IMAGES/LARGE/AU2C00414_0006.JPEG.
16. Majumdar, A., Dogra, P., Maity, S. & Mukhopadhyay, S. Liquid-Liquid Phase Separation Is Driven by Large-Scale Conformational Unwinding and Fluctuations of Intrinsically Disordered Protein Molecules. *Journal of Physical Chemistry Letters* **10**, 3929–3936 (2019).
17. Altmeyer, M. *et al.* Liquid demixing of intrinsically disordered proteins is seeded by poly(ADP-ribose). *Nature Communications 2015 6:1* **6**, 1–12 (2015).

18. Molliex, A. *et al.* Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization. *Cell* **163**, 123–133 (2015).
19. Elbaum-Garfinkle, S. *et al.* The disordered P granule protein LAF-1 drives phase separation into droplets with tunable viscosity and dynamics. *Proc Natl Acad Sci U S A* **112**, 7189–7194 (2015).
20. Zosel, F., Soranno, A., Buholzer, K. J., Nettels, D. & Schuler, B. Depletion interactions modulate the binding between disordered proteins in crowded environments. doi:10.1073/pnas.1921617117/-/DCSupplemental.
21. Csizmok, V., Follis, A. V., Kriwacki, R. W. & Forman-Kay, J. D. Dynamic Protein Interaction Networks and New Structural Paradigms in Signaling. *Chem Rev* **116**, 6424–6462 (2016).
22. Wright, P. E. & Dyson, H. J. Intrinsically disordered proteins in cellular signalling and regulation. *Nature Reviews Molecular Cell Biology 2015 16:1* **16**, 18–29 (2014).
23. Schuster, B. S. *et al.* Identifying sequence perturbations to an intrinsically disordered protein that determine its phase-separation behavior. *Proc Natl Acad Sci U S A* **117**, 11421–11431 (2020).
24. Murthy, A. C. *et al.* Molecular interactions underlying liquid−liquid phase separation of the FUS low-complexity domain. *Nat Struct Mol Biol* **26**, 637–648 (2019).
25. Wu, S. *et al.* The Dynamic Multisite Interactions between Two Intrinsically Disordered Proteins. *Angewandte Chemie* **129**, 7623–7627 (2017).
26. Wang, W. & Wang, D. Extreme fuzziness: Direct interactions between two idps. *Biomolecules* vol. 9 Preprint at https://doi.org/10.3390/biom9030081 (2019).
27. Borgia, A. *et al.* Extreme disorder in an ultrahigh-affinity protein complex. *Nature* **555**, 61–66 (2018).
28. Sottini, A. *et al.* Polyelectrolyte interactions enable rapid association and dissociation in high-affinity disordered protein complexes. *Nat Commun* **11**, (2020).
29. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
30. Gainza, P. *et al.* Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat Methods* **17**, 184–192 (2020).
31. Zeng, M. *et al.* Protein-protein interaction site prediction through combining local and global features with deep neural networks. *Bioinformatics* **36**, 1114–1120 (2020).
32. Li, Y., Brian Golding, G. & Ilie, L. DELPHI: Accurate deep ensemble model for protein interaction sites prediction. *Bioinformatics* **37**, 896–904 (2021).
33. Zhang, B., Li, J., Quan, L., Chen, Y. & Lü, Q. Sequence-based prediction of protein-protein interaction sites by simplified long short-term memory network. *Neurocomputing* **357**, 86–100 (2019).
34. Qiu, J. *et al.* ProNA2020 predicts protein–DNA, protein–RNA, and protein–protein binding proteins and residues from sequence. *J Mol Biol* **432**, 2428–2443 (2020).
35. Zhang, J. & Kurgan, L. SCRIBER: Accurate and partner type-specific prediction of protein-binding residues from proteins sequences. in *Bioinformatics* vol. 35 i343–i353 (Oxford University Press, 2019).
36. Murakami, Y. & Mizuguchi, K. Applying the Naïve Bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites. *Bioinformatics* **26**, 1841–

1848 (2010).
37. Porollo, A. & Meller, J. Prediction-based fingerprints of protein-protein interactions. in *Proteins: Structure, Function and Genetics* vol. 66 630–645 (2007).
38. Yuan, Q., Chen, J., Zhao, H., Zhou, Y. & Yang, Y. Structure-aware protein–protein interaction site prediction using deep graph convolutional network. *Bioinformatics* **38**, 125–132 (2021).
39. Ahmad, S. & Mizuguchi, K. Partner-aware prediction of interacting residues in protein-protein complexes from sequence data. *PLoS One* **6**, (2011).
40. Afsar Minhas, F. ul A., Geiss, B. J. & Ben-Hur, A. PAIRpred: Partner-specific prediction of interacting residues from sequence and structure. *Proteins: Structure, Function and Bioinformatics* **82**, 1142–1155 (2014).
41. Sanchez-Garcia, R., Sorzano, C. O. S., Carazo, J. M. & Segura, J. BIPSPI: A method for the prediction of partner-specific protein-protein interfaces. *Bioinformatics* **35**, 470–477 (2019).
42. Zhou, X., Song, H. & Li, J. Residue-Frustration-Based Prediction of Protein-Protein Interactions Using Machine Learning. *Journal of Physical Chemistry B* **126**, 1719–1727 (2022).
43. Chen, R. *et al.* Prediction of protein-protein interaction sites in intrinsically disordered proteins. *Frontiers in Molecular Biosciences* vol. 9 Preprint at https://doi.org/10.3389/fmolb.2022.985022 (2022).
44. Lange, J., Wyrwicz, L. S. & Vriend, G. KMAD: knowledge-based multiple sequence alignment for intrinsically disordered proteins. *Bioinformatics* **32**, 932–936 (2016).
45. Elbaum-Garfinkle, S. *et al.* The disordered P granule protein LAF-1 drives phase separation into droplets with tunable viscosity and dynamics. *Proc Natl Acad Sci U S A* **112**, 7189–7194 (2015).
46. Brady, J. P. *et al.* Structural and hydrodynamic properties of an intrinsically disordered region of a germ cell-specific protein on phase separation. *Proc Natl Acad Sci U S A* **114**, E8194–E8203 (2017).
47. Majumdar, A., Dogra, P., Maity, S. & Mukhopadhyay, S. Liquid-Liquid Phase Separation Is Driven by Large-Scale Conformational Unwinding and Fluctuations of Intrinsically Disordered Protein Molecules. *Journal of Physical Chemistry Letters* **10**, 3929–3936 (2019).
48. Brangwynne, C. P. *et al.* Germline P granules are liquid droplets that localize by controlled dissolution/condensation. *Science (1979)* **324**, 1729–1732 (2009).
49. Cai Zhu, Y., Lyu, F., Hu, C., Chen, X. & Liu, X. Encoder-Decoder Architecture for Supervised Dynamic Graph Learning: A Survey. (2022).
50. Zaki, A., Attia, M., Hegazy, D. & Amin, S. Comprehensive Survey on Dynamic Graph Models. *International Journal of Advanced Computer Science and Applications* **7**, (2016).
51. Barros, C. D. T., Mendonça, M. R. F., Vieira, A. B. & Ziviani, A. A Survey on Embedding Dynamic Graphs. *ACM Computing Surveys* vol. 55 Preprint at https://doi.org/10.1145/3483595 (2021).
52. Hamilton, W. L., Ying, R. & Leskovec, J. Inductive Representation Learning on Large Graphs. *Adv Neural Inf Process Syst* **30**, (2017).
53. Hochreiter, S. & Schmidhuber, J. Long Short-Term Memory. *Neural Comput* **9**, 1735–1780 (1997).
54. Seo, Y., Defferrard, M., Vandergheynst, P. & Bresson, X. Structured sequence modeling with

graph convolutional recurrent networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **11301 LNCS**, 362–373 (2018).

55. Wu, T., Chen, F. & Wan, Y. Graph Attention LSTM Network: A New Model for Traffic Flow Forecasting. in *Proceedings - 2018 5th International Conference on Information Science and Control Engineering, ICISCE 2018* 241–245 (Institute of Electrical and Electronics Engineers Inc., 2019). doi:10.1109/ICISCE.2018.00058.
56. Chen, J., Wang, X. & Xu, X. GC-LSTM: graph convolution embedded LSTM for dynamic network link prediction. *Applied Intelligence* **52**, 7513–7528 (2022).
57. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit Lett* **27**, 861–874 (2006).
58. Caruana, R. *et al.* Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* vols 2015-August 1721–1730 (Association for Computing Machinery, 2015).
59. Du, M., Liu, N. & Hu, X. Techniques for interpretable machine learning. *Commun ACM* **63**, 68–77 (2020).
60. Azodi, C. B., Tang, J. & Shiu, S. H. Opening the Black Box: Interpretable Machine Learning for Geneticists. *Trends in Genetics* vol. 36 442–455 Preprint at https://doi.org/10.1016/j.tig.2020.03.005 (2020).
61. Katarya, R., Sharma, P., Soni, N. & Rath, P. A Review of Interpretable Deep Learning for Neurological Disease Classification. in 900–906 (Institute of Electrical and Electronics Engineers (IEEE), 2022). doi:10.1109/icaccs54159.2022.9785321.
62. Jing, F., Zhang, S.-W., Cao, Z. & Zhang, S. *An integrative framework for combining sequence and epigenomic data to predict transcription factor binding sites using deep learning*. *IEEE TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS CLASS FILES* vol. A (2018).
63. Che, D., Hockenbury, C., Marmelstein, R. & Rasheed, K. Classification of genomic islands using decision trees and their ensemble algorithms. *BMC Genomics* **11**, (2010).
64. Zeiler, M. D. & Fergus, R. *LNCS 8689 - Visualizing and Understanding Convolutional Networks*. (2014).
65. Freiberger, M. I., Wolynes, P. G., Ferreiro, D. U. & Fuxreiter, M. Frustration in fuzzy protein complexes leads to interaction versatility. *Journal of Physical Chemistry B* **125**, 2513–2520 (2021).
66. Fischer, T. B. *et al.* Assessing methods for identifying pair-wise atomic contacts across binding interfaces. *J Struct Biol* **153**, 103–112 (2006).
67. Holehouse, A. S. & Kragelund, B. B. The molecular basis for cellular function of intrinsically disordered protein regions. *Nature Reviews Molecular Cell Biology* Preprint at https://doi.org/10.1038/s41580-023-00673-0 (2023).
68. Borcherds, W. *et al.* Disorder and residual helicity alter p53-Mdm2 binding affinity and signaling in cells. *Nature Chemical Biology 2014 10:12* **10**, 1000–1002 (2014).
69. Elkjær, S. *et al.* Evolutionary fine-tuning of residual helix structure in disordered proteins manifests in complex structure and lifetime. *Communications Biology 2023 6:1* **6**, 1–11 (2023).
70. Berlow, R. B., Dyson, H. J. & Wright, P. E. Hypersensitive termination of the hypoxic

response by a disordered protein switch. *Nature 2017 543:7645* **543**, 447–451 (2017).

71. Freiberger, M. I. *et al.* The Evolution of Local Energetic Frustration in Protein Families. doi:10.1101/2023.01.25.525527.
72. Ferreiro, D. U., Hegler, J. A., Komives, E. A. & Wolynes, P. G. Localizing frustration in native proteins and protein assemblies. *Proc Natl Acad Sci U S A* **104**, 19819–19824 (2007).
73. Chen, M. *et al.* Surveying biomolecular frustration at atomic resolution. *Nat Commun* **11**, (2020).
74. Lindström, I. & Dogan, J. Dynamics, Conformational Entropy, and Frustration in Protein-Protein Interactions Involving an Intrinsically Disordered Protein Domain. *ACS Chem Biol* **13**, 1218–1227 (2018).
75. Freiberger, M. I., Wolynes, P. G., Ferreiro, D. U. & Fuxreiter, M. Frustration in fuzzy protein complexes leads to interaction versatility. *Journal of Physical Chemistry B* **125**, 2513–2520 (2021).
76. Parra, R. G. *et al.* Protein Frustratometer 2: a tool to localize energetic frustration in protein molecules, now with electrostatics. *Nucleic Acids Res* **44**, W356–W360 (2016).
77. Rausch, A. O. *et al.* FrustratometeR: an R-package to compute local frustration in protein structures, point mutants and MD simulations. *Bioinformatics* **37**, 3038–3040 (2021).
78. Sun, D. & Gong, X. Tetramer protein complex interface residue pairs prediction with LSTM combined with graph representations. *Biochim Biophys Acta Proteins Proteom* **1868**, (2020).
79. Liu, J. & Gong, X. Attention mechanism enhanced LSTM with residual architecture and its application for protein-protein interaction residue pairs prediction. *BMC Bioinformatics* **20**, (2019).
80. Yang, Y., Wang, W., Lou, Y., Yin, J. & Gong, X. Geometric and amino acid type determinants for protein-protein interaction interfaces. *Quantitative Biology* **6**, 163–174 (2018).
81. Zhao, Z. & Gong, X. Protein-Protein Interaction Interface Residue Pair Prediction Based on Deep Learning Architecture. *IEEE/ACM Trans Comput Biol Bioinform* **16**, 1753–1759 (2019).
82. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* **157**, 105–132 (1982).
83. Søndergaard, C. R., Olsson, M. H. M., Rostkowski, M. & Jensen, J. H. Improved treatment of ligands and coupling effects in empirical calculation and rationalization of p K a values. *J Chem Theory Comput* **7**, 2284–2295 (2011).
84. Paszke, A. *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Adv Neural Inf Process Syst* **32**, (2019).
85. Fey, M. & Lenssen, J. E. Fast Graph Representation Learning with PyTorch Geometric. *In ICLR Workshop on Representation Learning on Graphs and Manifolds.* (2019).
86. Rozemberczki, B. *et al.* PyTorch Geometric Temporal: Spatiotemporal Signal Processing with Neural Machine Learning Models. *International Conference on Information and Knowledge Management, Proceedings* 4564–4573 (2021) doi:10.1145/3459637.3482014.
87. Kingma, D. P. & Ba, J. L. Adam: A Method for Stochastic Optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (2014).

Acknowledgments
We thank Prof. Rui Shi and Prof. Qianyuan Tang for their comments on the manuscript. This work is supported by the National Natural Science Foundation of China (12175195 and 32371299).

Author contributions
J-Y. L. designed and guided the study. Y-C. Z. prepared the dataset, designed the model architecture, and wrote the first draft. Q-X. L. designed the model architecture and implemented the method. All authors discussed the results and contributed to the writing of the manuscript.