Genome Biology

METHODOLOGY

Open Access

DeepISO: deep learning-powered prediction of protein–protein interaction rewiring generated by alternative splicing

Xiaokun Guo[1,2†], Linyang Jiang[1†], Jiajun Li[1], Mengdi Yuan[1], Dianke Li[1], Wenyu Shi[1*], Ziding Zhang[1*] and Stefan Wuchty[3,4,5,6*]

†Xiaokun Guo and Linyang Jiang contributed equally to this work.

*Correspondence: shiwy@cau.edu.cn; zidingzhang@cau.edu.cn; wuchtys@cs.miami.edu

[1] State Key Laboratory of Animal Biotech Breeding, College of Biological Sciences, China Agricultural University, Beijing 100193, China
[2] China Tobacco Hebei Industrial Co., Ltd., Shijiazhuang 050051, China
[3] Department of Computer Science, University of Miami, Miami, FL 33146, USA
[4] Department of Biology, University of Miami, Miami, FL 33146, USA
[5] Sylvester Comprehensive Cancer Center, University of Miami, Miami, FL 33136, USA
[6] Institute of Data Science and Computing, University of Miami, Miami, FL 33146, USA

Abstract

Isoforms from the same gene can significantly rewire protein interaction networks, but proteome-wide computational evaluation of these effects remains challenging. In this work, we present DeepISO, a deep learning framework for predicting isoform-specific interactions. DeepISO integrates two graph convolutional neural networks and a random forest model via a logistic regression model. To the best of our knowledge, this is the first approach to jointly leverage AlphaFold-predicted structures and ESM2 language model embeddings for this task. Compared with state-of-the-art PPI prediction tools, DeepISO demonstrates superior performance.

Keywords: Alternative splicing, Isoform, Protein–protein interaction, Deep learning, Structure

Background

Alternative Splicing (AS) is a crucial regulatory mechanism in many eukaryotes that significantly enhances the proteome diversity [1, 2], with over 95% of multi-exon genes undergoing AS in humans [3, 4]. During the splicing process, exons are selectively included or excluded, allowing a single gene to generate multiple isoforms [5] that provide functional diversity. For instance, isoforms can be involved in differing protein–protein interactions (PPIs), protein-DNA binding, and subcellular localization [6]. Despite their sequence similarity, isoforms may exhibit only minor differences in protein function, while many others show entirely distinct protein functions rather than slight variations [7, 8]. Currently, several studies have explored isoform function prediction, employing deep learning-based methods that predict isoform functions from sequence and expression information [9–13]. Additionally, databases such as scTEA-db [14] and IIIDB [15] provide isoform function information annotated through computational

methods. However, due to the lack of extensive experimental evidence, the large-scale functional regulation details through AS remain poorly understood.

Although AS diversifies the functional roles of specific genes, the extent of isoform contribution to functional complexity at the proteome and interactome scales remains unknown [8]. So far, extensive research has been conducted on protein interactions to delve into protein functions [16, 17]. However, most studies focused on the gene level, assuming that a single gene produces a reference isoform, significantly overlooking the impact of AS on the underlying isoform connectivity [18]. Notably, as one of the scant sources to date, a high-throughput interactomics study by Yang et al. found that isoforms extensively rewire protein interactions, indicating that most isoforms share fewer than 50% of interaction partners [8]. Although the prevalence of this rewiring effect remains unclear, its complexity is far greater than previously anticipated [19]. Although accurate, experimental identification of isoform-driven interactions at a proteome-wide scale is time-consuming and labor-intensive, considerably impeding our global understanding of isoform-driven PPI rewiring and highlighting the necessity for more rapid and cost-effective computational approaches.

Since PPI rewiring has been an important entry point for understanding AS-induced changes in gene function [20, 21], several computational methods have been elegantly developed to investigate isoform interactions. Ghadie et al. proposed a domain-based method (DIIP) to predict isoform interactions [20] that relies on known reference PPIs and isoform domain annotations to infer isoform interactions. Since this approach is based on isoform domain annotations, the coverage of protein interactions that DIIP can provide is significantly limited. Using an attention-based model, Yu et al. introduced IDMIL-III to predict isoform interactions on a genome-wide scale [22]. While Narykov et al. developed ALT-IN with a semi-supervised learning strategy [4], employing a triplet input approach to explore interactions between any two isoforms and a common protein, this method cannot directly determine the interaction relationships between isoforms. It is also worth noting that the source code of these existing isoform interaction methods is rarely available to the community, significantly limiting their applications.

Recent studies have demonstrated significant progress in predicting PPIs in general using deep learning techniques. In particular, convolutional neural networks (CNNs) were shown to effectively predict PPIs by embedding features from protein sequences [23]. Natural language processing (NLP) models have been applied to characterize protein features such as ESM2 [24] and ProtT5 [25], thereby greatly improving prediction accuracy. With the widespread application of AlphaFold [26], structural information about proteins is increasingly being used to predict PPIs [27–29]. For instance, Huang et al. demonstrated that incorporating structural information significantly enhances PPI prediction [29]. However, such methods focus on interactions between reference isoforms, ignoring interactions between isoforms. Notably, the availability of structural information is crucial for predicting interactions between isoforms, as they exhibit distinct structural conformations despite heightened sequence similarity. As a consequence, the prediction of isoform interactions is still in its infancy, and significant room for performance improvement remains in the deep learning era.

Here, we propose an integrated deep learning model, DeepISO, that leverages both sequence and structural information to predict interactions between protein isoforms.

To capture structural information of interacting isoforms, we represent each protein's structure as a residue contact map and train two graph convolutional neural networks (GCNs) to predict isoform interactions. As for sequence-based predictions, we also use a pre-trained protein language model to generate embedding vectors for each protein sequence, which are then used as input to a random forest (RF) model. By integrating the three individual models, the proposed DeepISO architecture can predict interactions between isoforms with maximum accuracy. To demonstrate the veracity of the predictions, we apply DeepISO to several case studies, allowing us to predict interactions between isoforms that explain the underlying biology.

Results

Architecture of DeepISO

DeepISO is an integrated prediction model that comprehensively integrates structural and sequence information for isoforms at the protein residue level, thereby enabling more accurate prediction of isoform interactions. In particular, we designed a three-pronged approach, in which we first represent each protein isoform using its contact map from AlphaFold (version 2.0). In the second step, such networked structures are subjected to two GCN models. Briefly, we used the pre-trained protein language model ESM2 to generate high-dimensional embedding vectors for every residue of interacting protein isoforms in the first GCN model (i.e., ESM2-based GCN model). In contrast, we relied heavily on structural features to encode each residue in the second GCN model (i.e., SF-based GCN model). In the third step, we used ESM2 solely to represent the underlying protein isoform sequence that is fed to an RF model. While such approaches can be used in isolation, we combined these methods in a final logistic regression (LR) step to predict interactions between protein isoforms (Fig. 1A).

Performance of DeepISO

As for training data, we collected isoform interactions from the large-scale study by Yang et al. [8], obtained via yeast two-hybrid (Y2H) screening. Specifically, the dataset captures 1,500 samples of isoform interactions between 498 proteins, with a balanced ratio of positive to negative samples. We randomly partitioned the dataset, using 80% of the samples for training and the remaining 20% as an independent test set. More details about the preparation and partitioning of the dataset are available in the Methods section.

To comprehensively evaluate DeepISO's performance, we obtained receiver operating characteristic (ROC) and Precision-Recall (PR) curves (Fig. 1B, C). Moreover, we quantified the predictive performance using the area under the ROC curve (AUROC) and the area under the PR curve (AUPRC). In particular, we observed that DeepISO performs extraordinarily well, with AUROC and AUPRC of 0.848 and 0.842, respectively (Table 1). At a false positive rate control of 20%, DeepISO achieved a sensitivity of roughly 80%. Furthermore, we conducted a thorough comparison of DeepISO with established PPI prediction methods, such as DIIP, which focuses on predicting isoform interactions, as well as D-Script [30], PPITrans [31], PIPR [32], and MAPE-PPI [28], which are utilized for general PPI prediction, and AlphaFold-Multimer [33].

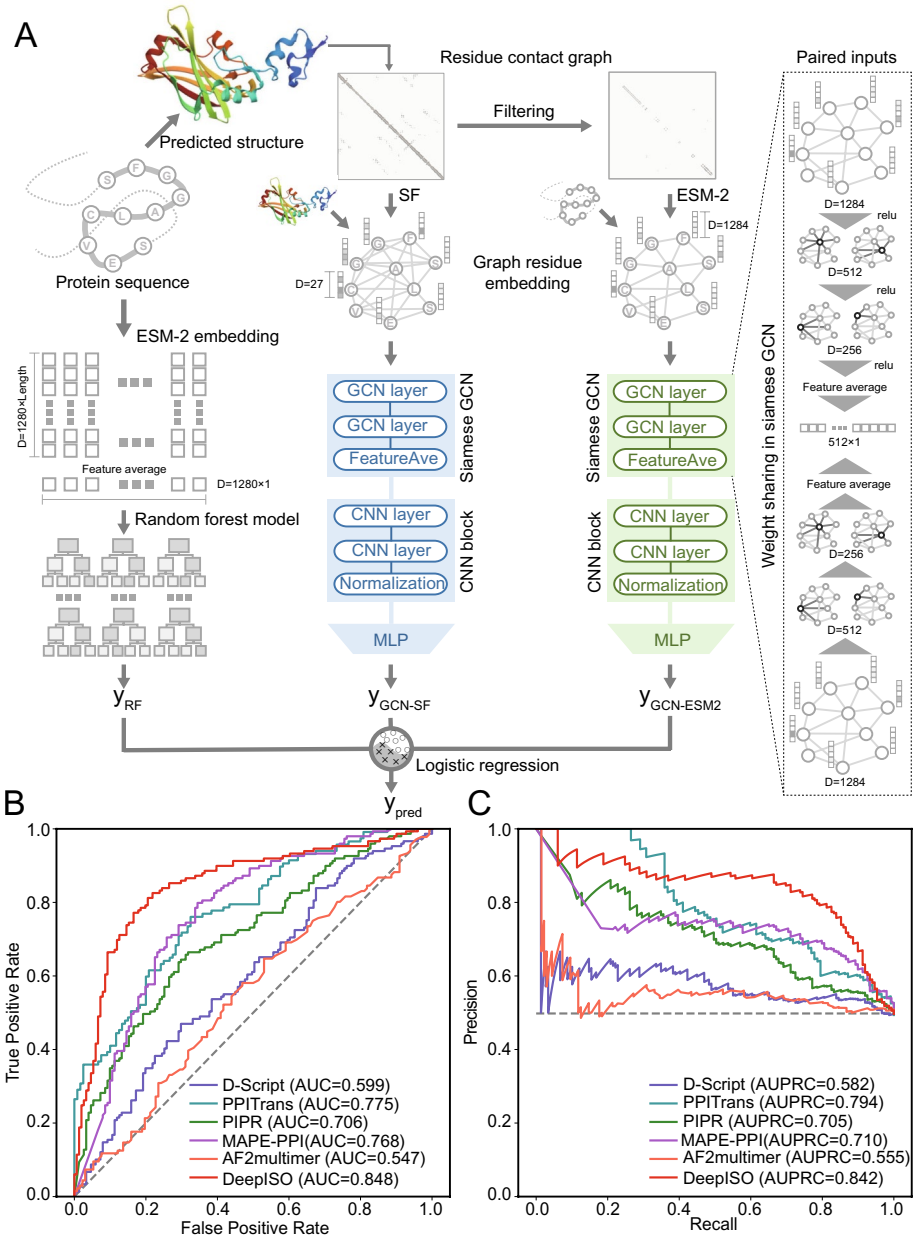

Fig. 1 Overall model architecture of DeepISO and performance comparison with other methods. **A** DeepISO consists of three individual predictive models: (i) a Siamese GCN model based on ESM2 features, (ii) a Siamese GCN model based on structural features (SF), and (iii) an RF model using ESM2 features. The final DeepISO model integrates the predictive scores from the three individual models through LR. Performance comparison of our DeepISO model with existing prediction methods was conducted using **B** ROC and **C** PR curves. Gray lines in B and C represent random performance

DIIP is designed to predict isoform interactions based on domain-domain interactions (DDI) by considering the different domains present in the isoforms. We implemented DIIP on our local machine based on its operational principles [20]. When tested on the independent test set, DIIP achieved a precision of 0.877, a recall of 0.087, and an F1 score of 0.159 (Table 1). In general, DeepISO considerably outperforms DIIP in terms of comparable precision and much higher recall and F1 scores (Precision = 0.763,

Table 1 Performance comparison of DeepISO and existing PPI predictors

Model	Precision[a]	Recall[a]	F1-score[a]	AUROC	AUPRC
DeepISO	0.783	0.826	0.804	0.848	0.842
DIIP	0.877	0.087	0.159	NA[b]	NA[b]
D-Script	0.613	0.255	0.360	0.599	0.582
PPITrans	0.703	0.709	0.706	0.775	0.794
PIPR	0.676	0.664	0.660	0.706	0.705
MAPE-PPI	0.693	0.772	0.730	0.768	0.710
AlphaFold-Multimer	0.667	0.015	0.029	0.547	0.555

[a] The value in each model was based on the default interaction threshold of 0.5

[b] NA indicates that the corresponding value is not available

Recall = 0.819, F1 score = 0.790). To allow a fairer comparison, DeepISO achieves a recall of 0.664 when we set the precision control to 0.877, which is much higher than the corresponding value of DIIP (0.087).

We further compared DeepISO to four state-of-the-art generic PPI predictors (D-Script [30], PPITrans [31], PIPR [32], and MAPE-PPI [28]). D-Script generates feature representations of protein structural information using a pre-trained model and estimates the interaction probability of the input protein pairs based on these features. PPITrans uses the pre-trained ProtT5 model to encode proteins and employs a transformer architecture to predict protein interactions. PIPR is a sequence-based PPI predictor that first generates protein embedding vectors and uses a recurrent convolutional neural network (RCNN) architecture to predict interactions between protein pairs. MAPE-PPI is a structure-aware model that predicts PPIs by learning microenvironment embeddings from 3D structural features. To ensure a fair comparison, we downloaded the source codes of these four methods, retrained the corresponding predictive models on our isoform interaction training dataset, and tested their performance on the isoform interaction test set. Judged by the ROC or PR curves, we found that DeepISO showed the best predictive performance, followed by PPITrans, MAPE-PPI, PIPR, and D-Script (Fig. 1B, C). In general, these four generic PPI methods still shared complementary results with DeepISO to some extent (Additional file 1: Fig. S1).

As a functional module provided by AlphaFold (version 2.0), AlphaFold-Multimer is widely used to predict the 3D complex structure of two interacting proteins, with parameters such as predicted Template Modeling (pTM) and interface pTM score (ipTM) used to evaluate the quality of the predicted structures [33]. Recently, a hybrid score that jointly considers pTM and ipTM (i.e., 0.8 ipTM + 0.2 pTM) has been proposed to distinguish protein interactions from non-interactions. A higher score indicates a higher likelihood of protein interaction. As shown in Fig. 1B, C, AlphaFold-Multimer's performance in predicting isoform interactions is inferior to DeepISO. Such a result implies that the simple metric inferred from AlphaFold-Multimer cannot accurately detect isoform interactions, even though AlphaFold-Multimer was specifically designed for 3D structure prediction of interacting proteins.

Furthermore, we explored the feasibility of predicting isoform interactions using models trained on the reference isoform-level PPI data. Specifically, we used the

HuRI dataset employed by Huang et al. [29] as the normal level PPIs, with a positive-to-negative sample ratio close to 1:1. For a fair comparison, we retrained PIPR and DeepISO on the reference-isoform PPI dataset and tested on the isoform test set. Additional file 1: Fig. S2 indicates that using normal PPIs to predict isoform interactions results in a dramatic performance drop, suggesting that isoform interactions may involve different interaction patterns compared to normal PPIs. However, even under these circumstances, DeepISO still outperforms other methods (Additional file 1: Fig. S2A, B).

Ablation and optimization experiments

Among the three individual predictors, the RF model is entirely based on protein sequence information, while the other two predictors within the Siamese GCN framework use protein structural information, prompting us to investigate the performance of the three individual predictors through ablation experiments. As shown in Fig. 2A, B, the ESM2-based GCN predictor achieved the best predictive performance among the three individual classifiers (AUROC = 0.833, AUPRC = 0.816), whereas the RF model yielded the weakest results. The integrative DeepISO achieved reasonable performance

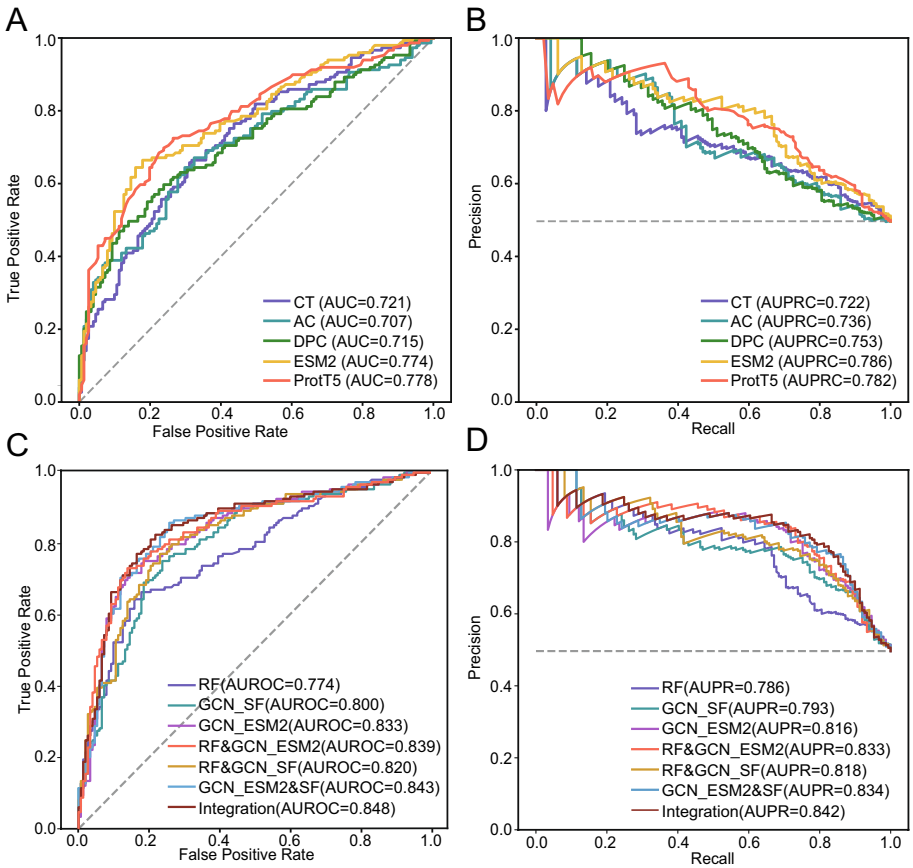

Fig. 2 Predictive performance of different models in ablation experiments. **A** ROC and **B** PR curves of the RF models with different encoding schemes. **C** ROC and **D** PR curves of DeepISO and its three individual models. Gray dotted lines indicate random prediction performance

improvement compared to the best individual predictor and the combinations of any two individual predictors (Fig. 2C, D).

We further examined some key factors affecting the performance of these three individual predictors. As for the RF model with ESM2, we also benchmarked it against another pre-trained protein language model called ProtT5 and three traditional sequence encoding schemes [Conjoint Triad (CT), Auto Covariance (AC), and Di-peptide Composition (DPC)]. More details about the calculations of these encoding schemes are available in the methods. As a result, the pre-trained models ESM2 and ProtT5 demonstrated significant advantages over traditional encoding methods (CT, AC, and DPC) (Fig. 2A, B). Compared with ProtT5, ESM2's performance is nearly the same (Fig. 2A, B). We also compared the RF model to other machine learning algorithms. In addition to yielding better performance than other traditional machine learning algorithms such as LightGBM and Adaboost [34], the RF model also outperformed deep learning models like MLP (Additional file 1: Table S1). Such an observation might result from the limited sample size, in which deep learning models cannot fully utilize their superior capabilities. When using protein structural information, we introduced two GCN models, converting protein structures into residue contact networks as a crucial step. Adjusting the threshold for edges in the residue contact map, we found that the model performed best when the residue contact threshold was set to 8 Å while edges formed by adjacent residues were removed from the contact matrix (Additional file 1: Fig. S3A). We also used the PeSTo [35] model to predict the probability of each residue in a protein participating in protein interactions. In particular, PeSTo is a geometric deep learning model that efficiently predicts protein binding interfaces using only atomic coordinates as input. Using a threshold of 0.1 (Additional file 1: Fig. S3B), we removed non-essential residues from the isoform structure, which optimized model performance. When we combined the three individual classifiers with an LR model, DeepISO achieved optimal performance.

(See figure on next page.)

Fig. 3 Examples of isoform interaction prediction with 3D complex structures. In panels **A**-**D**, red represents interacting partners, purple represents the reference isoform and cyan represents the differential regions between the AS isoform and the reference isoform. The solid black squares indicate interactions, while the white squares indicate no interactions. **A** 3D complex structure of the interaction between NDN_A (purple) and DTPBP1 (red). In the NDN_B isoform, this interaction interface is disrupted as a consequence of the loss of a sequence region (cyan), which includes some linear motifs. Despite such complex binding circumstances, DeepISO correctly predicted the presence/absence of interactions between DTPBP1 and the corresponding isoforms. **B** 3D complex structure formed by the interaction between HGS_1 (purple) and MAGED1 (red). Although a part of the domain sequence is lost (cyan) in the HGS_2 isoform, the interaction is not disrupted since the lost domain region is far from the interaction interface, interaction characteristics that DeepISO correctly predicted. **C** 3D complex structure of the interaction between ARNT2_F08 (purple) and ADAMTSL4 (red). The reference isoform ARNT2_F02 cannot interact with ADAMTSL4. However, in the ARNT2_F08 isoform part of a disordered region at the C-terminus is lost (cyan), enabling an interaction with ADAMTSL4. Notably, DeepISO correctly predicts both the absence and presence of the corresponding isoforms. **D** 3D structures of MVP_G/MVP_H (purple) in complex with RBPMS (red). The two figures at the bottom show the alignment of MVP_A with these two complexes, respectively, using the "Align" function in PYMOL. The longer isoforms of MVP (MVP_A, MVP_C) cannot interact with RBPMS. However, in the MVP_G and MVP_H isoforms, partial sequences are lost (cyan), causing changes in the protein's spatial structure, leading MVP_G and MVP_H to interact with RBPMS, complex circumstances that DeepISO correctly predicted

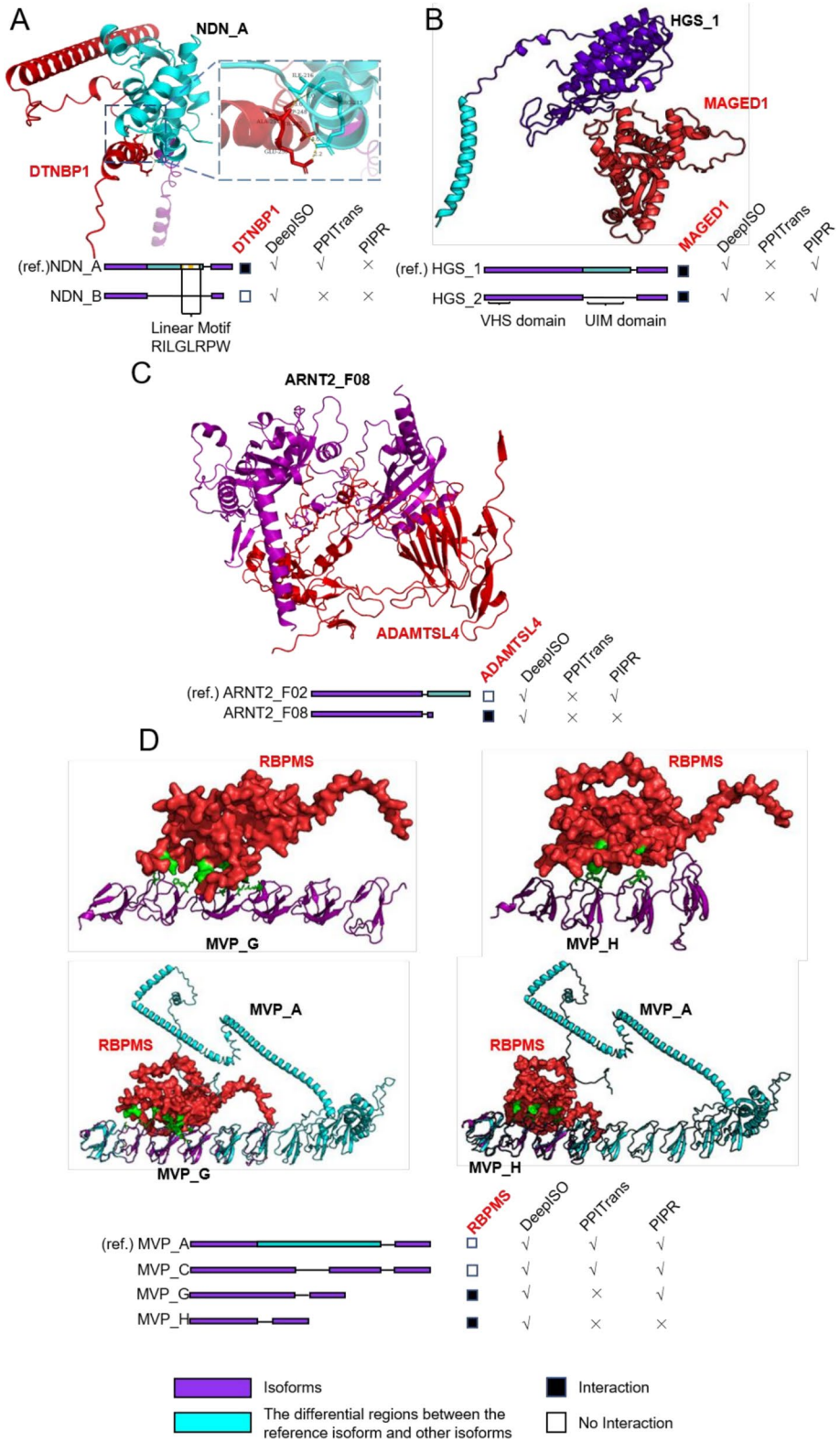

Fig. 3 (See legend on previous page.)

Understanding isoform interaction predictions in the context of 3D complex structures

To investigate structural changes in isoform interaction prediction, we collected some representative examples of AS-induced PPI rewiring from the literature and our independent test set. We analyzed the corresponding prediction results from DeepISO and two existing methods (PIPR and PPITrans) (Fig. 3) and used AlphaFold-Multimer to predict the 3D complex structures formed by the isoform interactions in these examples.

The first case study concerns the interaction prediction of NDN (necdin) isoforms and DTNBP1 (dysbindin). Specifically, NDN is a growth suppressor that plays an important role in nervous system development, while DTNBP1 is associated with synaptic function. The interaction between NDN and DTNBP1 may play a complex and significant role in the development, functional maintenance, and disease processes of the nervous system. As shown in Fig. 3A, NDN produces two isoforms, NDN_A and NDN_B. Compared to NDN_A, NDN_B lacks a segment of the linear motif, indicating that NDN_A can interact with DTNBP1, while NDN_B does not [36]. Using AlphaFold-Multimer, the complex structure formed by NDN_A and DTNBP1 indicates that the linear motif in NDN_A is located in the binding region with DTNBP1, suggesting that the missing linear motif in NDN_B may directly disrupt the interaction between NDN_B and DTNBP1. Among the three individual models, the ESM2-based GCN model produced prediction scores closest to the ground truth (Additional file 1: Table S2), indicating its strength in capturing sequence-level motif features. As a result, DeepISO accurately predicted the different interaction states of the NDN isoforms even in this motif-lacking scenario (Fig. 3A).

Protein domains are crucial in mediating PPIs and are responsible for recognizing and binding other proteins, indicating that a protein's change or loss of domains can significantly impact its interaction partners. The second example focuses on the interaction between HGS (Hepatocyte growth factor-regulated tyrosine kinase substrate) isoforms and MAGED1 (Melanoma-associated antigen D1). Specifically, HGS is involved in intracellular signal transduction, while MAGED1 participates in the apoptotic response in neuronal cells and inhibits cell cycle progression. The interaction between HGS and MAGED1 may connect apoptotic signaling and receptor transport, coordinating cell cycle regulation and protein degradation in neuronal cells. As shown in Fig. 3B, the HGS_1 isoform contains two domains (the VHS and UIM domains), while the HGS_2 isoform lacks the UIM domain. However, experimental results showed that both HGS_1 and HGS_2 can interact with MAGED1. Intriguingly, both DeepISO and PIPR made correct predictions in this example, while PPITrans failed. We further analyzed the reasons why DeepISO can achieve successful predictions. The 3D complex structure we predicted for HGS_1 and MAGED1 indicates that the missing UIM domain in HGS_2 is distant from the protein binding interface, which might explain why the loss of the UIM domain has little effect on the interaction. As shown in additional file 1: Table S2, the ESM2-based GCN predictor achieved the highest score in predicting the interaction between HGS_1 and MAGED1 among the three individual prediction models, while the SF-based GCN predictor obtained the highest score in predicting the interaction between HGS_2 and MAGED1. Thus, our two structure-based prediction models in

DeepISO can thoroughly characterize protein structure information, enabling accurate predictions of the isoform interactions.

In addition, disordered regions of proteins typically lack a fixed 3D structure, providing high flexibility and adaptability, allowing disordered proteins to adopt various conformations when binding to different partners. The third example focuses on the AS-induced interaction between ARNT2 (Aryl hydrocarbon receptor nuclear translocator 2) and ADAMTSL4 (ADAMTS-like protein 4) (Fig. 3C). Specifically, ARNT2 is a transcription factor (TF) that regulates the expression of various genes, while ADAMTSL4 is involved in cell apoptosis. Their interaction may help maintain the stability and functionality of the extracellular matrix, which is crucial for tissue structure and integrity. Compared to the ARNT2_F02 isoform, ARNT2_F08 lacks a disordered region at the C-terminus, yet still interacts with ADAMTSL4. However, the longer ARNT2_F02 cannot interact with ADAMTSL4, indicating that the disordered region at the C-terminus affects its ability to interact. Benefiting from the joint contributions of three individual predictors (Additional file 1: Table S2), DeepISO accurately predicts this PPI arising from the loss of the disordered region, while the PPITrans and PIPR results are incorrect.

Apart from the above structural differences, other conformational changes between different isoforms can also lead to the rewiring of PPIs, exemplified by the interaction between MVP (Major vault protein) and RBPMS (RNA-binding protein with multiple splicing). MVP is the largest known ribonucleoprotein, comprising approximately 70% of the vault protein mass and forming a tubular structure. As shown in Fig. 3D, both MVP_G and MVP_H can interact with RBPMS, whereas the longer MVP_A and MVP_C isoforms cannot. To investigate why the longer isoforms cannot interact with RBPMS, we structurally aligned MVP_A with the complex structures of MVP_G/MVP_H and RBPMS (Fig. 3D). The structure alignment revealed that MVP_A shares a similar N-terminal structure with MVP_G and MVP_H. However, MVP_A forms an extended structure region, which may create a severe spatial hindrance that prevents its interaction with RBPMS. A similar steric clash phenomenon can be observed in MVP_C. Notably, such complex structural considerations are best captured by DeepISO, enabling accurate predictions of interactions between MVP isoforms and RBPMS. In particular, the GCN_ESM2 model contributes dominantly to the correct predictions (Additional file 1: Table S2). In contrast, PPITrans and PIPR performed poorly and could not accurately distinguish the effects of interaction rewiring.

To help understand the potential limitations of the proposed DeepISO, we provided another case study on the interaction predictions between A2BP1 isoforms and RBPMS, where DeepISO yields incorrect predictions. A2BP1 is a human RNA-binding protein that generates three protein isoforms (A2BP1_A, A2BP1_B, and A2BP1_C) through AS, all of which are known to interact with RBPMS. Both PIPR and PPITrans successfully predicted all three interactions, whereas DeepISO only correctly identified the interaction between A2BP1_B and RBPMS. We further examined the prediction scores of the three individual models in DeepISO (Additional file 1: Table S2). Interestingly, the sequence model (RF_ESM2) alone correctly predicted interactions for all three isoforms, whereas the two GCN models produced lower prediction scores (Additional file 1: Table S2). After integrating the outputs via an LR model, only the interaction between A2BP1_B and RBPMS was predicted by DeepISO as positive. We further provide a

structural visualization of the predicted complex structure of the reference isoform A2BP1_C and RBPMS (Additional file 1: Fig. S4). The pTM value of the A2BP1_C structure from AlphaFold is only 0.27. Generally, a pTM score above 0.5 indicates that the predicted fold may be similar to the true structure. Therefore, the structural quality of A2BP1_C is low. The same phenomenon was observed in the predicted structures of A2BP1_A and A2BP1_B. The low quality of the predicted A2BP1 isoform structures may explain why our two GCN models failed to yield correct predictions, whereas RF_ESM2 and two existing sequence models (PPITrans and PIPR) are successful in this case.

AS-induced PPI rewiring in breast cancer

Breast cancer significantly impacts the health of hundreds of millions of people worldwide, and numerous studies have already demonstrated that AS plays a crucial role in the development and progression of breast cancer [37–39]. In this study, we conducted an in-depth analysis by integrating known interactomics and transcriptomics data related to breast cancer. Specifically, we employed DeepISO to quantitatively assess the PPI network rewiring induced by AS in breast cancer individuals. We collected RNA-Seq data on breast cancer, allowing us to identify 64,116 differentially expressed transcripts of 9,601 genes for subsequent analyses. More details about the RNA-Seq data processing are available in the methods. We employed the protein interaction network in breast cancer experimentally determined by Kim et al. [40], which consists of 997 gene-level PPIs among 697 proteins. Narrowing down the original PPI network by integrating the results of RNA-Seq analysis, we retained PPIs when at least one protein in a given PPI contains two or more differentially expressed transcripts. Such a filtering step allowed us to obtain a refined PPI network containing 498 pairs of gene-level PPIs from 420 proteins, which we used for further analyses.

We utilized DeepISO to predict isoform interactions within this PPI network. To provide an intuitive investigation of isoform interactions, we first compared the interaction profiles of two isoforms of each protein. As shown in Fig. 4A, approximately 26% of isoforms from the same protein share different interaction partners, indicating that AS significantly expanded the protein interaction capabilities. For network visualization, we plotted the refined breast cancer PPI network using a different color scheme to highlight the PPI rewiring effects caused by AS (Fig. 4A). In particular, if the isoforms of two proteins in a PPI have different isoform interaction modes, this PPI is defined as a rewired

(See figure on next page.)

Fig. 4 Breast cancer-centered PPI network perturbations induced by alternative splicing. **A** Distribution of interaction pattern differences between all possible pairs of AS isoforms in the breast cancer PPI network. A Jaccard distance of 0 indicates that the two isoforms share identical interaction partners, while a distance of 1 points to two isoforms that do not share any common interaction partners. A distance between 0 and 1 indicates that the two isoforms share partial common partners. **B** displays the breast cancer PPI network rewired by AS. **C** Network topological parameters (degree, clustering coefficient, and betweenness) between rewired nodes and other nodes through violin plots (** p-value < 0.01). Performing module clustering on the rewired breast cancer PPI network, we find 3 clusters that are significantly related to breast cancer: **D** Cluster 7 centers on TP53, **E** Cluster 10 revolves around ERα, and **F** Cluster 5 centers on BRCA1. Genes with mutations clinically associated with breast cancer are defined as "strongly correlated", while others are defined as "weakly correlated". In panels B and D-F, diamond-shaped nodes represent genes that are highly associated with breast cancer, while circular nodes represent genes that are weakly associated with breast cancer. Red edges indicate rewired interactions, while red nodes point to nodes that have undergone rewiring

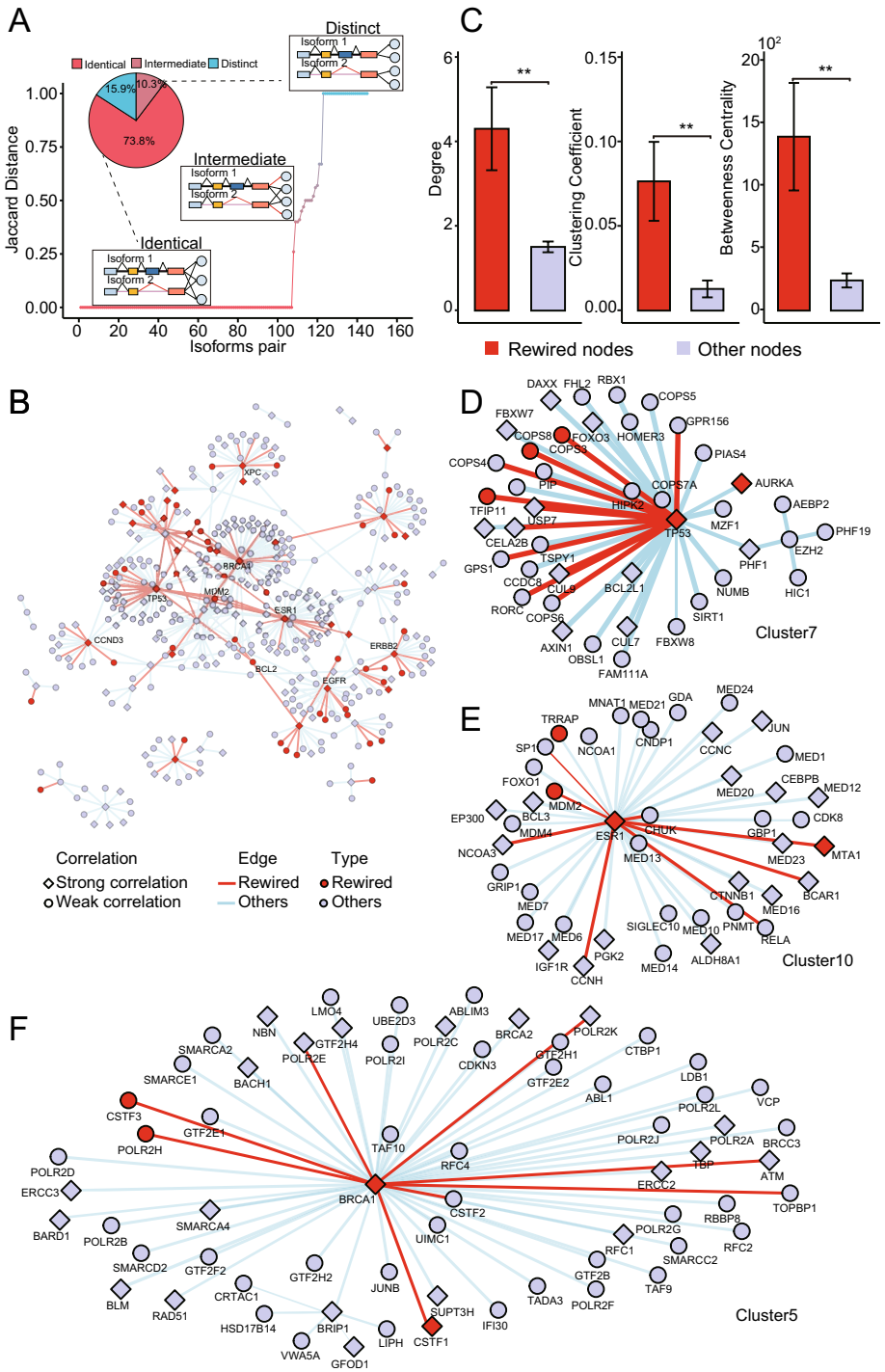

Fig. 4 (See legend on previous page.)

interaction (edge) in the network. Likewise, if the isoforms of a protein (node) in the network can have different interaction partners, the protein is defined as a rewired protein (node). As shown in Fig. 4B, 105 out of the 498 PPIs were rewired, and 69 out of the 420 proteins were considered rewired nodes. By comparing the network topological parameters (degree, clustering coefficient, and betweenness) between rewired nodes and other

nodes in the interaction network (Fig. 4C), we found that rewired nodes tend to occupy more significant positions within the interaction network.

To further quantitatively explore the rewiring effects of isoform interactions caused by AS in breast cancer, we also searched the ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar/) and obtained 4,462 genes strongly associated with breast cancer, each of which contains clinically proven mutation sites. 143 out of these 4,462 genes are involved in our newly compiled PPI network, participating in 457 PPIs. Interestingly, 103 out of the 105 rewired interactions contain SA genes (Fig. 4B and Additional file 1: Table S3). Statistically, SA genes are significantly involved in PPI rewiring (χ^2 test, p-value = 0.014), further indicating a substantial impact of AS on the breast cancer PPI network.

We also performed modular partitioning of the breast cancer PPI network using the igraph package (Fig. 4B) [41]. After removing modules with fewer than five nodes, 13 modules were obtained (Additional file 1: Table S4). We noted three clusters with a relatively higher number of nodes, including Cluster 7 containing cellular tumor antigen p53 (TP53) (Fig. 4D), Cluster 10 containing estrogen receptor (ERα) (Fig. 4E), and Cluster 5 harboring the breast cancer type 1 susceptibility protein (BRCA1) (Fig. 4F). We conducted functional enrichment analyses of these three clusters using MetaScape (Additional file 1: Fig. S5) [42]. All three clusters were significantly enriched for functions and metabolic pathways related to breast cancer and associated diseases. The enriched functions in Cluster 7 highlight TP53's central role in breast cancer. TP53 maintains genomic integrity by regulating DNA damage response, cell cycle, apoptosis, and protein stability. Mutations or loss of function in TP53 can lead to breast cancer cells evading normal proliferation control, thereby promoting tumorigenesis and progression [43, 44]. Cluster 10 highlights the critical role of ERα in breast cancer, particularly in regulating cell proliferation, apoptosis, and signaling pathways. By regulating the PI3K-Akt, MAPK, and Wnt signaling pathways, ERα66 promotes the growth and survival of breast cancer cells. Furthermore, the functional role of ERα66 in regulating cell adhesion and the senescence-associated phenotype also supports the development of breast cancer. Indeed, ERα is not only a biomarker for breast cancer but, more importantly, may constitute a potential therapeutic target [45]. In addition, Cluster 5 mainly involves DNA damage repair, particularly the repair of DNA double-strand breaks through homologous recombination (HR). BRCA1 interacts with the RNA polymerase II complex to regulate gene expression and maintain telomeres to ensure genomic stability [46] (Additional file 1: Fig. S5). It should be emphasized that the rewired nodes/edges detected in these three clusters by DeepISO (Fig. 4E, F) may provide rich information for further deciphering the biological functions of network modules.

Discussion

Regarding the model architecture, the proposed deep learning model DeepISO comprises three individual predictors (the RF, GCN_ESM2, and GCN_SF models), which are further integrated via an LR model. Compared with a specific isoform-interaction method and several popular PPI predictors, DeepISO achieves superior performance. The success of DeepISO in isoform interaction prediction can be attributed to its ability to effectively and comprehensively extract residue-level features from both sequence and structural information. It is worth mentioning that the contribution of the

sequence-based RF model to DeepISO is limited. In terms of AUROC or AUPRC values, the improvement introduced by the RF model is minor (less than 1%) (Fig. 2A, B). Given that protein structural information is not readily accessible in real-world applications and the predicted structures from AlphaFold may also be incorrect in some cases, we decided to keep the sequence-based RF model in DeepISO. As an option in DeepISO, the RF model can be easily used in isolation, allowing users to rapidly apply sequence information to infer protein isoform interactions in some large-scale prediction tasks. Since GCN_ESM2 and GCN_SF share the same model architecture, an alternative integration method is the concatenation of the ESM2 and structure features into a new feature vector to develop a single predictor. To examine this issue, we conducted a computational experiment to directly fuse the ESM2 and structure features within a single encoder. Since these two encoding schemes exhibit large variance, we conducted Z-score normalization for each feature before model training. The new predictor yielded intermediate performance between the two individual models (AUROC = 0.821, AUPRC = 0.802). In other words, the approach performed better than GCN_SF but worse than GCN_ESM2, implying that the current integrative strategy is suitable for maximizing the performance. Taken together, the model architecture is generally optimal for the current isoform prediction task.

It has been well established that a protein's structural properties, including motifs, domains, disordered regions, and overall conformation, play crucial roles in recognizing its interaction partners. Specifically, the AS isoforms may experience diverse structural changes compared to the reference isoform. To enhance model interpretability, we provided several case studies that explain why DeepISO can outperform two existing methods (PIPR and PPITrans) (Fig. 3). Through structural visualizations, these case studies clearly demonstrate that DeepISO can effectively handle diverse structural changes to yield proper PPI predictions. It should also be noted that, in some cases, the isoform structures from AlphaFold are incorrect, and these low-quality isoform structures can significantly affect the predictive performance, which is a potential limitation of the current DeepISO model.

Moreover, DeepISO can be applied to understand the rewiring of AS-driven PPI networks on a proteome-wide scale. To explore such an application, we applied DeepISO to the integrative analysis of known interactomics and transcriptomics data related to breast cancer. As expected, the introduction of DeepISO yielded informative results, as the isoform interactions between genes were predicted, allowing us to characterize PPI rewiring in breast cancer. Indeed, we observed extensive AS-induced perturbations in the PPI network, affecting disease-related metabolic pathways (Fig. 4B). We have shown the biological significance of the rewired nodes/edges detected by DeepISO through network topology analysis and module identification. In the meantime, the rewired nodes/edges may serve as potential new hints to further guide hypothesis-driven experimental efforts to interrogate the functional roles of candidate genes in breast cancer.

Conclusions

In this work, we propose DeepISO, a model that predicts isoform interactions by leveraging both sequence and structural information of isoforms. Although DeepISO can accurately predict isoform interactions, the current study still suffers from some

limitations. First, the limited number of isoform interaction samples may affect the generalizability of the proposed DeepISO model. Second, although we incorporated rich structural and sequence information, the direct use of atomic 3D coordinates remains limited; if appropriate methods are employed to leverage coordinate information [47], DeepISO's performance could be further improved. Third, the small training dataset also prevents us from taking full advantage of deep learning. For instance, the RF model used in our approach outperforms deep learning models for developing the sequence-based predictors. As experimental data increases in the future, we believe the generalizability of the updated DeepISO will be significantly enhanced, making it applicable in broader contexts.

Methods

Data collection and preprocessing

We collected the dataset of isoform interactions from Yang et al. [8], which was determined through high-throughput yeast two-hybrid (Y2H) screening. This dataset is currently the largest experimentally validated isoform interaction dataset, also recording negative samples (i.e., non-interactions with experimental evidence). In particular, this dataset contains 820 interactions and 635 non-interactions, capturing 332 isoforms from 140 genes. To enable the model to learn interactions between isoforms derived from the same gene, we retained only positive samples involving genes with multiple isoforms. As this training strategy differs from conventional PPI modeling, a total of 750 high-quality positive samples were selected for training and testing. Furthermore, a significant imbalance between positive and negative samples can severely impact the model's performance in pairwise input predictions. To ensure dataset balance, we added negative samples following strict criteria: (1) we randomly selected protein pairs from the dataset to construct negative samples and only retained those pairs whose proteins appeared in different subcellular localizations; (2) we downloaded DDI data from the 3did database and used HMMER to annotate all proteins with their corresponding domains (E-value < 1e-5) [48], excluding protein pairs with DDIs. After integration, we obtained a dataset of 1,500 isoform interaction samples, with a 1:1 ratio of positive to negative samples. We randomly split the dataset into 80% training and 20% testing sets.

The overall model architecture of DeepISO

The core idea of DeepISO is to predict isoform interactions by comprehensively extracting sequence and structural information of isoforms. DeepISO is an integrative model with three individual predictive models (Fig. 1A).

ESM2-based GCN predictor

Graph representation of protein structures Based on the prediction results of AlphaFold for isoform structures, we constructed an undirected residue contact graph for each protein structure [26]. Briefly, residues serve as nodes and residue-residue contacts serve as edges. If the distance between the Cα atoms of any two non-adjacent residues is less than a specific threshold (default 8 Å), we surmise that these two residues form an edge. Since not all residues contribute to isoform interactions, we predicted the probability

of each residue participating in protein interactions using PeSTo [35]. Specifically, we retained residues with a probability greater than 0.1 to serve as nodes in the graph. It should be noted that the removal of unimportant nodes (with probability scores less than 0.1) may further result in isolated nodes, which were also discarded. Considering a residue is generally in contact with multiple residues, the number of isolated nodes is limited.

Feature engineering In the ESM2-based GCN predictor, we employed ESM2 and the protein's secondary structure as node features. We used the esm2_t33_650M_UR50D pre-trained model to encode each residue in the protein, where each residue is encoded as a 1,280-dimensional feature vector. Then, DSSP [49] was used to assign the protein's secondary structure. Each residue's secondary structure state was categorized into four classes: helix, sheet, turn, and other, represented by a four-dimensional one-hot vector. Finally, each residue was converted into a 1,284-dimensional feature vector.

Model training of the ESM-based GCN predictor We utilized a Siamese GCN architecture to extract hidden features from protein structures. The protein is represented by a residue contact graph consisting of n nodes. The GCN input comprises two parts: an adjacency matrix $\mathrm{A} \in \mathrm{R}^{\mathrm{n}\times\mathrm{n}}$ and a node feature matrix $\mathrm{X} \in \mathrm{R}^{\mathrm{n}\times\mathrm{m}}$, where m is the length of the feature vector for each residue ($m = 1{,}284$ in this work). We employed two-layer GCN to update the hidden features of residues, with the update rules as follows:

$$H(l+1) = \sigma(\widetilde{D}^{-1/2}\widetilde{\mathrm{A}}\widetilde{D}^{-1/2}H^{(l)}W^{(l)}) \tag{1}$$

$$\widetilde{A} = A + I \tag{2}$$

In particular, H represents the hidden state during the l-th convolutional layer, $W(l)$ represents the weight matrix of the l-th layer, σ represents the nonlinear activation function (σ is LeakyReLU in this work), A denotes the adjacency matrix,$\widetilde{D}$ represents the diagonal degree matrix of $\widetilde{A}$ while I denotes the identity matrix. After two layers of GCN, the hidden features of all nodes are averaged. Then, the two input-paired proteins are concatenated, resulting in a final 512-dimensional feature vector that represents the hidden features of the underlying protein pair. Then, the concatenated feature vectors are fed into a CNN block. We employed 1D convolution with two layers of convolution kernels of size 2 to further extract feature information from an $l \times s$ dimensional array, where l represents the length, and s is the feature dimension (i.e., channels). After each convolution, feature transformation is performed using linear layers. Finally, the features extracted by the CNN block are input into an MLP with four fully connected layers, utilizing the sigmoid function to compute the probability of interaction between two proteins. In addition, to avoid potential overfitting, we applied appropriate weight initialization (Kaiming Normal), used a small learning rate (0.0001), and trained the model with a small batch size suited to the dataset. The model was trained for 30 epochs, considering that convergence was achieved based on the training loss (Additional file 1: Fig. S6). Both model training and evaluation were performed on a single NVIDIA RTX 3090 GPU with 24 GB of memory.

SF-based GCN predictor

The SF-based GCN predictor employs the same model architecture and graph-based protein structure representation as the ESM2-based GCN predictor. The major difference is that the SF-based GCN predictor more heavily utilizes structural features for residue representations.

The SF-based predictor primarily employs three types of features. First, it utilizes PeSTo to predict the probability of each residue participating in intermolecular interactions. In particular, PeSTo [35] operates directly on atomic coordinates without requiring additional physicochemical properties. It can predict molecular binding interfaces between proteins and other molecules such as proteins, nucleic acids, lipids, ions, and small molecules. Therefore, we obtain a 5-dimensional feature vector representation for each residue. Next, we utilized FoldSeek [50] to generate feature representations of the protein's 3D structure. FoldSeek initially aimed to rapidly and accurately search protein structures [50]. Specifically, FoldSeek employs a VQ-VAE model, encoding protein structural information into information tokens derived from 20 distinct 3D states, denoted as $P=(f_1, f_2,\ldots, f_n)$, where f_i represents the structural token at the i-th position, and n points to the sequence length. Consequently, we extracted intermediate layer information from FoldSeek, employing one-hot encoding to represent 3D structural features of each residue as a 21-dimensional feature vector. Finally, we also represented the secondary structure of each residue using a one-dimensional feature vector. Ultimately, we obtained a 27-dimensional structural feature vector for each residue. The GCN_SF model was trained similarly to the GCN_ESM2 model. The number of training epochs was determined to be 75 according to the convergence behavior of the training loss (Additional file 1: Fig. S6).

ESM-based RF predictor

We employed an RF model to predict isoform interactions using protein sequence information. For each protein, the ESM2 model (esm2_t33_650M_UR50D) was used to generate a final representation of $L \times 1{,}280$, where L is the length of the protein. We averaged over the length-dimension of the representations to derive fixed-size vectors (i.e., 1,280 dimensionality) of each protein. We concatenated the feature vectors of the interacting proteins that were input into the RF model. Specifically, the RF model is implemented using the sklearn library (https://scikit-learn.org/) in Python, and parameter optimization was conducted using grid search. The specific parameters *n_estimators* and *max_depth* were optimally set to 50 and 15, respectively. Note that the hyperparameter optimization was based on the observed performance during the fivefold cross-validation on the training dataset.

Integrative predictor by logistic regression

To improve isoform prediction performance, we combined the scores from the three individual classification models into a vector. Subsequently, we trained an integrative predictor through LR. The implementation of the LR algorithm is based on the iLearn library in Python [34]. To prevent potential overfitting, a fivefold cross-validation strategy was employed on the training dataset to determine the optimal model parameters.

Other sequence-based encoding schemes

We also used three conventional sequence encoding schemes (CT, AC, and DPC) and another pre-trained protein language model called ProTrans to benchmark the ESM-based RF model.

CT

The CT encoding scheme first bins the 20 amino acids into seven groups (AGV, C, DE, FILP, HNQW, KR, and MSTY) based on the physicochemical properties of their side chains. Then, the composition of three consecutive amino acid groups (i.e., CT) in a protein sequence can be calculated. As a result, a protein is represented by a 343-dimensional ($7 \times 7 \times 7 = 343$) vector. Finally, a protein pair is represented by a 686-dimensional vector.

AC

AC encoding considers the effect of interactions between amino acids separated by a certain distance (the maximal sequence distance is set to 30 in this work) in the sequence. To measure this correlation, amino acid properties are characterized using seven indices: hydrophobicity, hydrophilicity, polarity, polarizability, side-chain volume, solvent-accessible surface area, and net charge index of side chains. More details about the AC encoding are available in our previous work [51]. Ultimately, each protein pair is transformed into a 420-dimensional ($30 \times 7 \times 2 = 420$) vector.

DPC

DPC represents the composition of two consecutive amino acids (i.e., dipeptides) within the entire protein sequence and is used to transform a protein into a 400-dimensional vector [52], pointing to the final dimension of DPC encoding for a protein pair of 800.

ProtTrans

The ProtTrans models are available at https://github.com/agemagician/ProtTrans. We used prot_t5_xl_uniref50 (ProtT5), which transforms each protein sequence into a 1,024-dimensional vector. Finally, a protein pair is represented by a 2,048-dimensional vector.

Performance assessment

In this work, we plotted ROC and PR curves and relied on the corresponding metrics (i.e., AUROC and AUPRC) to assess the overall predictive performance of the model. We also introduced four common performance metrics (Precision, Recall, ACC, and F1-score) for method evaluation and comparison. These four metrics are defined as:

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$\mathrm{ACC} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$F1 - score = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad (6)$$

where TP, FP, TN and FN represent the number of true positive, false positive, true negative and false negative samples, respectively.

Analysis of RNA-Seq data related to breast cancer

In this work, RNA-Seq data related to breast cancer were downloaded from the NCBI Sequence Read Archive (SRA) database with the accession numbers PRJNA1081147 and PRJNA1034678 [53], where the breast cancer samples were derived from primary breast tumor tissue. Subsequently, a standardized RNA-Seq data analysis pipeline was applied. After quality control of raw data through Trimmomatic (v0.39) [54], Hisat2 was used to align reads to the reference genome with default parameters [55], where the reference genome was obtained from Ensembl (version: Homo sapiens.GRCh38) [56]. Then, the SAM file from Hisat2 was converted to a BAM file using Samtools (v1.11). The transcripts were assembled according to reference genomes, while the corresponding expression levels (i.e., TPM values) were calculated using StringTie (v2.1.4) [57]. Finally, we identified differentially expressed transcripts in breast cancer samples using the R package "DESeq2" [58]. Transcripts with $|\log2FC| \geq 1$ and an adjusted p-value < 0.05 were considered differentially expressed transcripts.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-026-04057-3.

Additional file 1: This file contains Tables S1-S4 and Figures S1-S6. Table S1. The performance of ESM2 encoding on different machine learning methods. Table S2. Prediction scores of DeepISO and three individual models in the case studies. Table S3. The overlapping of SA genes and rewired nodes in the breast cancer PPI network. Table S4. The node information in each module of the breast cancer PPI network. Fig. S1. Venn diagram showing the overlapping predictions among DeepISO and four generic PPI prediction methods. Fig. S2. Performance comparison of DeepISO and PIPR trained on the normal PPI dataset. Fig. S3. Performance of DeepISO using different settings in residue contact graph construction. Fig. S4. Prediction results of DeepISO and other models on the interactions between the A2BP1 isoforms and RBPMS with 3D complex structure. Fig. S5. Gene expression patterns and functional enrichment analysis of the three clusters. Fig. S6. The training loss curves of the GCN_ESM2 model and the GCN_SF model.

Acknowledgements

We also appreciate the support of the High-performance Computing Platform of China Agricultural University for providing computational resources.

Peer review information

Andrew Cosgrove and Zhenrun Jerry Zhang were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Authors' contributions

Xiaokun Guo and Linyang Jiang constructed the model and performed the data analyses. Xiaokun Guo drafted the manuscript. Jiajun Li participated in the data analysis. Mengdi Yuan and Dianke Li contributed to the preparation of the figures and tables. Wenyu Shi, Ziding Zhang, and Stefan Wuchty supervised the study and significantly revised the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (31970645, 32270703) and Chinese Universities Scientific Fund (Grant No. 2025TC001 to W.S.).

Data availability

The source code of DeepISO, together with the datasets used in this study, is publicly available at Zenodo [59] (DOI: https://doi.org/10.5281/zenodo.17359893) and GitHub [60] (https://github.com/zzd-lab/DeepISO), both under the MIT License.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests except that Xiaokun Guo is currently employed by China Tobacco Hebei Industrial Co., Ltd.

Received: 10 April 2025 Accepted: 20 March 2026
Published online: 27 March 2026

References

1. Markus MA, Yang YH, Morris BJ. Transcriptome-wide targets of alternative splicing by RBM4 and possible role in cancer. Genomics. 2016;107:138–44.
2. Bhasi A, Philip P, Sreedharan VT, Senapathy P. AspAlt: a tool for inter-database, inter-genomic and user-specific comparative analysis of alternative transcription and alternative splicing in 46 eukaryotes. Genomics. 2009;94:48–54.
3. Nilsen TW, Graveley BR. Expansion of the eukaryotic proteome by alternative splicing. Nature. 2010;463:457–63.
4. Narykov O, Johnson NT, Korkin D. Predicting protein interaction network perturbation by alternative splicing with semi-supervised learning. Cell Rep. 2021;37:110045.
5. Keren H, Lev-Maor G, Ast G. Alternative splicing and evolution: diversification, exon definition and function. Nat Rev Genet. 2010;11:345–55.
6. Kelemen O, Convertini P, Zhang Z, Wen Y, Shen M, Falaleeva M, et al. Function of alternative splicing. Gene. 2013;514:1–30.
7. Meininger I, Griesbach RA, Hu D, Gehring T, Seeholzer T, Bertossi A, et al. Alternative splicing of MALT1 controls signalling and activation of CD4+ T cells. Nat Commun. 2016;7:11292.
8. Yang X, Coulombe-Huntington J, Kang S, Sheynkman GM, Hao T, Richardson A, et al. Widespread expansion of protein interaction capabilities by alternative splicing. Cell. 2016;164:805–17.
9. Li W, Kang S, Liu CC, Zhang S, Shi Y, Liu Y, et al. High-resolution functional annotation of human transcriptome: predicting isoform functions by a novel multiple instance-based label propagation method. Nucleic Acids Res. 2014;42:e39.
10. Chen H, Shaw D, Zeng J, Bu D, Jiang T. DIFFUSE: predicting isoform functions from sequences and expression profiles via deep learning. Bioinformatics. 2019;35:i284–94.
11. Sommer MJ, Cha S, Varabyou A, Rincon N, Park S, Minkin I, Pertea M, Steinegger M, Salzberg SL. Structure-guided isoform identification for the human transcriptome. eLife. 2022;11:e82556.
12. Gao Y, Wang F, Wang R, Kutschera E, Xu Y, Xie S, et al. ESPRESSO: robust discovery and quantification of transcript isoforms from error-prone long-read RNA-seq data. Sci Adv. 2023;9:eabq5072.
13. Liu Y, Li H-D, Wang J. CrossIsoFun: predicting isoform functions using the integration of multi-omics data. Bioinformatics. 2025;41:btae742.
14. Barquin M, Kouzel IU, Ehrmann B, Basler M, Gruber AJ. scTEA-db: a comprehensive database of novel terminal exon isoforms identified from human single cell transcriptomes. Nucleic Acids Res. 2024;52:D1018–23.
15. Tseng Y-T, Li W, Chen C-H, Zhang S, Chen JJW, Zhou XJ, et al. IIIDB: a database for isoform-isoform interactions and isoform network modules. BMC Genomics. 2015;16:S10.
16. Zeng X, Meng F-F, Li X, Zhong K-Y, Jiang B, Li Y. GHGPR-PPIS: a graph convolutional network for identifying protein-protein interaction site using heat kernel with Generalized PageRank techniques and edge self-attention feature processing block. Comput Biol Med. 2024;168:107683.
17. Amjad H, Saleem F, Ahmad M, Nisar U, Arshad Dar H. Comprehensive bioinformatics-based annotation and functional characterization of bovine chymosin protein revealed novel biological insights. Food Chem. 2024;8:100191.
18. Yu G, Zeng J, Wang J, Zhang H, Zhang X, Guo M. Imbalance deep multi-instance learning for predicting isoform–isoform interactions. Int J Intell Syst. 2021;36:2797–824.
19. Tran JC, Zamdborg L, Ahlf DR, Lee JE, Catherman AD, Durbin KR, et al. Mapping intact protein isoforms in discovery mode using top-down proteomics. Nature. 2011;480:254–8.
20. Ghadie MA, Lambourne L, Vidal M, Xia Y. Domain-based prediction of the human isoform interactome provides insights into the functional impact of alternative splicing. PLoS Comput Biol. 2017;13:e1005717.
21. Zeng J, Yu G, Wang J, Guo M, Zhang X: DMIL-III: Isoform-isoform interaction prediction using deep multi-instance learning method. In: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2019 Nov 18–21; San Diego, CA, USA. p. 171–176. https://doi.org/10.1109/BIBM47256.2019.8982956.
22. Yu G, Zeng J, Wang J, Zhang H, Zhang X, Guo M. Imbalance deep multi-instance learning for predicting isoform–isoform interactions. Int J Intell Syst. 2021;36:2797–824.
23. Hu X, Feng C, Zhou Y, Harrison A, Chen M. DeepTrio: a ternary prediction system for protein–protein interaction using mask multiple parallel convolutional neural networks. Bioinformatics. 2022;38:694–702.

24. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science. 2023;379:1123–30.
25. Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, et al. ProtTrans: toward understanding the language of life through self-supervised learning. IEEE Trans Pattern Anal Mach Intell. 2022;44:7112–27.
26. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024;630:493–500.
27. Gao Z, Jiang C, Zhang J, Jiang X, Li L, Zhao P, et al. Hierarchical graph learning for protein-protein interaction. Nat Commun. 2023;14:1093.
28. Wu L, Tian Y, Huang Y, Li S, Lin H, Chawla NV, Li SZ. MAPE-PPI: Towards effective and efficient protein–protein interaction prediction via microenvironment-aware protein embedding. In: International Conference on Learning Representations (ICLR). 2024. Available from: https://openreview.net/forum?id=itGkF993gz.
29. Huang Y, Wuchty S, Zhou Y, Zhang Z. SGPPI: structure-aware prediction of protein–protein interactions in rigorous conditions with graph convolutional network. Brief Bioinform. 2023;24:bbad020.
30. 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;12:969-982 e966.
31. Yang S, Cheng P, Ju L, Liu Y, Feng D, Wang S. Exploring the knowledge of an outstanding protein to protein interaction transformer. IEEE/ACM Trans Comput Biol Bioinform. 2024;21(5):1287–98.
32. Chen M, Ju CJ, Zhou G, Chen X, Zhang T, Chang KW, et al. Multifaceted protein-protein interaction prediction based on siamese residual RCNN. Bioinformatics. 2019;35:i305–14.
33. Homma F, Huang J, van der Hoorn RAL. AlphaFold-Multimer predicts cross-kingdom interactions at the plant-pathogen interface. Nat Commun. 2023;14:6040.
34. Chen Z, Zhao P, Li C, Li F, Xiang D, Chen Y-Z, et al. iLearnPlus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization. Nucleic Acids Res. 2021;49:e60–e60.
35. Krapp LF, Abriata LA, Cortes Rodriguez F, Dal Peraro M. PeSTo: parameter-free geometric deep learning for accurate prediction of protein binding interfaces. Nat Commun. 2023;14:2175.
36. Corominas R, Yang X, Lin GN, Kang S, Shen Y, Ghamsari L, et al. Protein interaction network of alternatively spliced isoforms from brain links genetic risk factors for autism. Nat Commun. 2014;5:3650.
37. Yang Q, Zhao J, Zhang W, Chen D, Wang Y, Wang Z. Aberrant alternative splicing in breast cancer. J Mol Cell Biol. 2019;11:920–9.
38. Deguchi Y, Kikutake C, Suyama M. Subtype-specific alternative splicing events in breast cancer identified by large-scale data analysis. Sci Rep. 2024;14:14158.
39. Venables JP, Klinck R, Bramard A, Inkel L, Dufresne-Martin Gv, Koh C, et al. Identification of alternative splicing markers for breast cancer. Cancer Res. 2008;68:9525–31.
40. Kim M, Park J, Bouhaddou M, Kim K, Rojc A, Modak M, et al. A protein interaction landscape of breast cancer. Science. 2021;374:eabf3066.
41. Csárdi G, Nepusz T. The igraph software package for complex network research. InterJ Complex Syst. 2006;1695:1–9.
42. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun. 2019;10:1523.
43. Duijf PHG, Nanayakkara D, Nones K, Srihari S, Kalimutho M, Khanna KK. Mechanisms of genomic instability in breast cancer. Trends Mol Med. 2019;25:595–611.
44. Zhao M, Howard EW, Guo Z, Parris AB, Yang X. P53 pathway determines the cellular response to alcohol-induced DNA damage in MCF-7 breast cancer cells. PLoS ONE. 2017;12:e0175121.
45. Le Romancer M, Poulard C, Cohen P, Sentis S, Renoir JM, Corbo L. Cracking the estrogen receptor's posttranslational code in breast tumors. Endocr Rev. 2011;32:597–622.
46. Gorodetska I, Kozeretska I, Dubrovska A. BRCA genes: the role in genome stability, cancer stemness and therapy resistance. J Cancer. 2019;10:2109–27.
47. Meng L, Wei L, Wu R. MVGNN-PPIS: a novel multi-view graph neural network for protein-protein interaction sites prediction based on Alphafold3-predicted structures and transfer learning. Int J Biol Macromol. 2025;300:140096.
48. Mosca R, Ceol A, Stein A, Olivella R, Aloy P. 3did: a catalog of domain-based interactions of known three-dimensional structure. Nucleic Acids Res. 2014;42:D374-379.
49. Joosten RP, te Beek TA, Krieger E, Hekkelman ML, Hooft RW, Schneider R, et al. A series of PDB related databases for everyday needs. Nucleic Acids Res. 2011;39:D411-419.
50. Van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, et al. Fast and accurate protein structure search with Foldseek. Nat Biotechnol. 2024;42:243–6.
51. Yang S, Li H, He H, Zhou Y, Zhang Z. Critical assessment and performance improvement of plant–pathogen protein–protein interaction prediction methods. Brief Bioinform. 2019;20:274–87.
52. 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;57:1499–510.
53. Leinonen R, Sugawara H, Shumway M, International Nucleotide Sequence Database C. The sequence read archive. Nucleic Acids Res. 2011;39:D19-21.
54. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
55. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37:907–15.

56. Harrison PW, Amode MR, Austine-Orimoloye O, Azov Andrey G, Barba M, Barnes I, et al. Ensembl 2024. Nucleic Acids Res. 2023;52:D891–9.
57. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33:290–5.
58. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
59. Guo X, Jiang L, Li J, Yuan M, Li D, Shi W, Zhang Z, Wuchty S. DeepISO. Zenodo. 2025. https://doi.org/10.5281/zenodo.17359893.
60. Guo X, Jiang L, Li J, Yuan M, Li D, Shi W, Zhang Z, Wuchty S. DeepISO. Github. 2025. https://github.com/zzd-lab/DeepISO.

Publisher's Note